Abstract—Taking into account the dynamics of the scene is the most effective solution to obtain an accurate perception of unknown environments within the framework of a real autonomous robotic application. Many works have attempted to address the non-rigid scene assumption by taking advantage of deep learning advancements. Most new methods combine geometric and semantic approaches to determine dynamic elements that lack generalization and scene awareness. We propose a novel approach that overcomes the limitations of these methods by using scene depth information that refines the accuracy of estimates from geometric and semantic modules. In addition, the depth information is used to determine an area of influence of dynamic objects through our Objects Interaction module that estimates the state of both non-matched keypoints and out of segmented objects. The obtained results demonstrate the efficacy of the proposed method in providing accurate localization and mapping in dynamic environments.

Index Terms—Visual SLAM, Dynamic environments, Robust, Semantic segmentation, Depth influence

I. INTRODUCTION

Visual Simultaneous Localization And Mapping (vSLAM) has been widely investigated over the past decade for deploying many applications in robotics, Augmented Reality (AR), Virtual Reality (VR), Micro Aerial Vehicle (MAV), or Unmanned Autonomous Vehicles (UAV) [1]–[3]. This technology is based on extracting and interpreting visual information to locate the robot in an unknown environment [4]. Generally, the proposed methods rely on the rigid scene assumption to design high-performance vSLAM frameworks. However, their deployment in real applications has highlighted the impact of scene dynamics on their performance. Indeed, the vSLAM process attempts to locate the camera based on its environment using visual landmarks tracked during the camera’s movement. Thus, moving visual landmarks would imply an error in location estimation and mapping, affecting thus its performance. To solve this problem, many methods [5]–[8] deal with the dynamics of the scene through two main approaches used jointly or separately. These approaches classify keypoints into static and dynamic states in order to reject those in a dynamic state from the SLAM process. First, the semantic segmentation approach determines the dynamic objects according to their nature and the type of environment, where only humans are dynamic in indoor environments. Many indoor methods [6] perform semantic segmentation into person class in order to reject keypoints of likely dynamic objects. This approach is quite efficient but lacks flexibility due to the limited zone of the segmented region and the uncertainty around the mask edges. Subsequently, the geometric approach [9] was introduced to consider the whole image and deal with the other moving objects. It estimates the reprojection error of matched keypoints of two successive frames using the epipolar geometry constraint. A reprojection error above a certain threshold implies that the considered keypoint is dynamic and therefore should be rejected. Unfortunately, this approach has a limited impact since it only determines the state of the matched keypoints. To overcome these limitations we propose a novel approach based on the ORB-SLAM3 framework [10] by taking into account the influence of scene depth on the geometric and semantic segmentation reasoning. It is worth noticing that the geometric approach is based on 2D reasoning without considering the 3D information from the depth. This is one of its limitations. Indeed, the displacement of a distant point induces a smaller reprojection error than a closer moving point with the same motion amplitude. Therefore, we introduce an adaptive threshold related to the point depth to account for the relative reprojection error. Each reprojection error of a matched keypoint lower than this adaptive threshold is then fed into a normal probability distribution model to estimate the corresponding motion probability. In contrast, keypoints with an error higher than the threshold are considered dynamic and rejected. For semantic reasoning, some methods [7], [11] handle the state uncertainty of the points around the inner edges of the mask by using a probabilistic function based either on the minimum distance from these points to the edges or on the camera motion. This reasoning leads to ambiguity and imprecision in the way objects at different distances are taken into account in the estimation of the keypoints states. Indeed, a probabilistic function without depth consideration induces a constant size of the uncertainty zone, which would inevitably produce inconsistent results in the estimation of keypoints states. To solve this inconsistency, we introduce an adaptive impact factor in the proposed keypoints state probability. This allows to adjust the size of the area according to the depth of the segmented objects. Furthermore, to overcome the raised limitations of the existing methods, we propose a novel approach that combines information from
geometric and semantic approaches with depth information to determine the state of some keypoints that are neither matched nor from a segmented area. Our method is based on the assumption that in indoor environments dynamics result only from humans activity. Thereby, the probability that static objects become dynamic increases if they are within the zone of human influence, i.e. at a depth and an image position close to human. The neighborhood of the non-matched keypoints in this zone of influence is analyzed to determine if there are dynamic points resulting from the geometric approach, which would have the proximity of position and depth with this one. The spatial information about the elements within this neighborhood informs us about the probability of the state of this unpaired point. By doing so we could overcome the limitations of the geometric approach. The main contributions of our work are summarized as follows:

- An adaptive method to estimate the state of matched keypoints using geometric approach through an adaptive depth-related threshold.
- A refinement of the state estimation of keypoints near the inner and outer edges of segmented objects using an adaptive depth-related impact factor in a probabilistic function.
- A novel approach for keypoints state estimation based on the area of influence of dynamic objects with both non-matched keypoints and keypoints out of segmented areas.

The remainder of the paper is organized as follows. Section II summarizes the relevant related works. Section III is devoted to the description of our method. The obtained results are presented are discussed in section IV. Finally, concluding remarks and perspectives are provided in section V.

II. RELATED WORK

In this section, we present related works on vSLAM methods that attempt to deal with dynamic environments. These methods consider the non-rigid scene assumption and can be grouped into two categories. Most of them are based on the ORB-SLAM2 or ORB-SLAM 3 frameworks and include deep learning networks. The first category corresponds to dynamic vSLAM methods based on semantic information that combine geometric and semantic segmentation approaches. Among these methods, DynaSLAM [5] associates multi-view geometry with semantic information from the MASK R-CNN model to detect moving object to be rejected. DS-SLAM [8] extracts semantic and motion information through a segmentation model SegNet and a moving consistency check. These complementary tasks remove outliers and then improve the estimated pose. Dynamic-SLAM [6] enriches the SSD object detection method to a semantic level to determine the shape of dynamic objects. This novel approach is completed by a selective tracking algorithm applied to keypoints of dynamic objects that increases pose estimation robustness. DP SLAM [7] is based on the estimation of the moving probability propagation of the dynamic keypoints combining the epipolar geometry constraints and semantic segmentation into a Bayesian filter. This method processes the state probability ambiguity of the keypoints at the boundary of masks. DGS-SLAM [12] uses a multinomial residual network to detect dynamic objects combining the motion information from consecutive frames and potential motion information from the semantic segmentation. This approach is enhanced by a robust camera pose tracking strategy that uses a keypoints classification. These modules are then employed in a segmentation module which extracts potential moving points thanks to a semantic frame selection.
strategy. The second category consists of methods based on optical flow networks to estimate the scene dynamics without any prior semantic consideration. FlowFusion [13] performs simultaneously the camera pose estimation and the static background reconstruction. This method deals with dynamic environments by using an optical flow network that generates an indirect semantic segmentation of dynamic objects that makes the tracking more robust. RDMO-SLAM [11] is a multimodal dynamic vSLAM that combines geometric and semantic segmentation information approaches. The limitations of the semantic segmentation module due to its slow process are overcome by predicting missing semantic labels through a dense optical flow thread. Finally, the semantic information about the scene dynamic is improved by applying a moving probability estimation based on camera rotation. All these methods can only estimate the state of points contained in a segmented area or matched points. This limits their applicability in other zones of the scene.

III. Method

Our vSLAM system is illustrated in fig.1. RGB images undergo geometric and semantic segmentation modules. This first process allows to determine the keypoints state. The geometric module calculates the reprojection error of matched keypoints using the epipolar geometric constraint. This reprojection error is then compared to an adaptive threshold to decide on whether to reject a given keypoint or to fed it into the probability function of keypoint state. The semantic segmentation module is based on the YOLACT++ [14] method, which generates masks of people in the scene. Segmented areas are considered dynamic due to the nature of the segmented object category. In addition, the accuracy of the keypoint state is modeled by a probabilistic function based on the distance of the keypoints from the edges of the mask. These two modules and the depth images are connected to their respective adaptive modules, which will consider the influence of depth on these tasks. Once the keypoints are sorted, the remaining keypoints considered static are refined through the object interaction module. This module estimates the state of the keypoints in the zone of influence of people using depth information associated with the masks and the dynamic keypoints of the geometric module. (see fig.4).

A. Semantic segmentation approach

The state of each keypoint of the semantic segmentation module is obtained by calculating their probability of displacement with respect to their distance from the the mask edges. First, this minimal distance $\Delta d_m$ between a keypoint $p_i$ and a point $m_i^n$ on the edge of the mask $n$ is defined as:

$$\Delta d_m = \min ||p_i - m_i^n||$$  \hspace{1cm} (1)

This distance is inserted into a binomial logistic regression model that will estimate the probability of the keypoints state according to the mask $n$ as follows:

$$P(S_{p_i^n}) = \frac{1}{exp(-\beta(z_{p_i^n}) \cdot \Delta d_m) + 1}$$ \hspace{1cm} (2)

Where $\beta(z_{p_i^n})$ is the adaptive impact factor proportional to the depth $z_{p_i^n}$ of the keypoint. The adaptive impact factor $\beta(z_{p_i^n}) \in [0.05; 0.25]$ will assign the size of the uncertainty area in and out of the mask as described in fig.3. Thus, the greater the distance, the more accurate the estimation of the state of this point, respectively static outside the mask and dynamic inside. We calculate the adaptive impact factor so as to obtain the limit of the uncertainty zone of a size corresponding to a state probability lower than 75%. In summary, the probability range of a keypoint to be dynamic can be defined as follows:

$$P(S_{p_i^n}) = \begin{cases} 
0.25; 0.5, & \text{if } p_i \in \text{Uncertainty area out}. \\
0.5; 0.75, & \text{if } p_i \in \text{Uncertainty area in}. \\
0.75; 1.0, & \text{if } p_i \in \text{Certainty area in}. 
\end{cases}$$ \hspace{1cm} (3)

Probabilities greater than 75% are considered reliable estimates of the state as shown in fig.3.

B. Geometric approach

The geometric module is based on the geometric epipolar constraint that links matched keypoints of two distinct images. Note the keypoint $q_i^{t-1}$ of a frame $t-1$ and its corresponding keypoint $p_i^t$ of the next frame $t$, $F_t$ denoting the fundamental matrix that describes the camera motion between these two
frames and the epipolar line \( l_{q_i} \), being the reprojection of the keypoint \( q_i \) in the frame \( f \) defined as follows:

\[
l_{p_i} = F \cdot (q_i)^T = \begin{bmatrix} X_{q_i} \\ Y_{q_i} \\ Z_{q_i} \end{bmatrix}
\]  

(4)

The reprojection error \( \Delta \varepsilon(p_i) \) describes the distance between the epipolar line \( l_{q_i} \) of the keypoint \( q_i \) and its matched keypoint \( p_i \) in the current frame. This error is denoted by the following equation:

\[
\Delta \varepsilon(p_i) = \frac{|p_i F(q_i)^T|}{\sqrt{||X_{q_i}||^2 + ||Y_{q_i}||^2}}
\]  

(5)

Keypoint \( p_i \) is considered dynamic if its reprojection error \( \Delta \varepsilon(p_i) \) is greater than an adaptive threshold \( \alpha(z_{p_i}) \) related to the keypoint depth \( Z_{p_i} \). This adaptive threshold based on the assumption that, the farther a point is, the more its reprojection error represents a large movement due to projective geometry constraints. Thus, this threshold \( \alpha(z_{p_i}) \in [0.5; 0.9] \) and is inversely proportional to the depth. If this error is smaller than the threshold, then it feeds the following normal probability density function to estimate the state probability of the keypoint:

\[
P(g_{p_i}) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\Delta \varepsilon(p_i)^2}{2\sigma^2}\right)
\]  

(6)

Where \( \sigma \) is set to 1 and represents the standard deviation of this distribution. The moving probability obtained is then normalized by the adaptive threshold.

C. Moving probability update

The moving probability \( P(p_i) \) represents the state probability of the keypoints where \( P(p_i) \in \text{State} \) with \( \text{State} = \{\text{static}, \text{dynamics}(d)\} \). The moving probability \( P(p_i) \) for a matched and segmented keypoint \( p_i \) combines the geometric probability model \( P(g_{p_i}) \) and the semantic segmentation probability model \( P(S_{p_i}) \) as follows:

\[
P(p_i) = \omega P(g_{p_i}) + (1 - \omega)P(S_{p_i})
\]  

(7)

Where \( \omega \) is a weight that describes the relevance of the probabilistic model for different keypoint situations. In uncertain areas, \( \omega \) is set to 0.5, while in the reliable area, it is set to 0.1. In case of non-matched keypoint include in the mask area, the moving probability \( P(p_i) \) corresponds to the semantic segmentation probability \( P(S_{p_i}) \). Conversely, if it is a matched keypoint out of the mask area, then the moving probability \( P(p_i) \) corresponds to the geometric probability \( P(g_{p_i}) \). We update the moving probability function using Bayesian filter as expressed:

\[
\text{bel}(p_i) = \eta P(\Omega_1|p_i) \int P(p_i|q_i) \text{bel}(q_i) dq_i
\]  

(8)

Where \( \eta \) is an impact factor to normalize probabilities. Initials prior probability \( p_0 \) and observation likelihood \( P(\Omega_1|p_i) \) are set to 0.5. Keypoints with a moving probability greater than 0.5 are considered dynamic points and therefore rejected.

D. Objects interaction module

The Objects Interaction Module (OIM) assesses possible interactions between humans and inert objects. It is based on the assumption that, the more a point considered static has an image position and a depth close to a human, the more likely it is to interact with it. Thus, we define an interaction zone related to the human depth (see the colored areas in Fig. 4 that indicate the probability of objects interaction) in which we estimate the correlation between the position and depth distances of the point with the nearest human. Suppose a considered static keypoint denoted \( p_i \) is in this area and owns a position and depth distance lesser than an adaptive threshold. The dynamics keypoints \( p_{id} \) from the geometric module in its neighborhood are studied in that case. These dynamics keypoints are retained when they satisfy the following conditions:

\[
\begin{cases}
||p_i - p_{id}|| < \delta(z_{p_i}), \\
\Delta(i) = ||p_i(z) - p_{id}(z)|| < \rho & \rho = 0.7
\end{cases}
\]  

(9)

Where \( \delta(z_{p_i}) = 48 - 4 \times z_{p_i} \) is an adaptive position distance threshold in pixels, and \( \rho \) is a depth distance threshold in meters. We calculate the center of gravity \( G(p_i) \) of these dynamic points weighting it by its depth distance as follows:

\[
G(p_i) = \frac{\sum_{i=1}^{k} \frac{p_i - \Delta(i)}{\rho}}{\sum_{i=1}^{k} \frac{1 - \Delta(i)}{\rho}}
\]  

(10)

The distance \( \Delta(G_{p_i}) \) of this center from the point \( p_i \) provides information on the state of \( p_i \). Indeed, this distance informs us about the relevance of the link characterizing these dynamic keypoints and the point \( p_i \). A small distance implies for the static point either very close proximity to these dynamic points or that they encompass it. Thus, the OIM changes the keypoint state to dynamic when \( \Delta(G_{p_i}) \) is smaller than a threshold \( \gamma(z_{p_i}) \in [10; 28] \) inversely proportional to the depth.
IV. EXPERIMENTS AND RESULTS

We have evaluated our vSLAM method in the fr3 sequence of the public dataset TUM RGB-D dedicated to assessing vSLAM methods in dynamic environments. This dedicated sequence provides RGB and depth images, and ground truth trajectory. It is broken down into two types of sequences, namely sitting (s) and walking (w), which correspond to scenarios with low and high levels of dynamics. All experiments were achieved on an Intel XEON CPU and Nvidia GPU RTX2080 SUPER. To perform the segmentation, we used the official YOLACT++ model with the backbone Resnet50-FPN pre-trained on the COCO dataset with an image size of 550*550. Our method has been compared to other state-of-the-art dynamic vSLAM methods according to the Absolute Trajectory Error (ATE). First, we highlighted the contribution of the Objects Interaction Module (OIM) to our system by assessing the performance of our method with (w/-OIM) and without (w/o-OIM) this module (see tab.1). We observed a better improvement for high dynamic scene scenarios, especially for the ATE score. The Table II shows our benchmarking, where we highlighted in bold the best results of each assessment for better readability.

| Sequences   | D2SLAM (w/o-OIU) | D2SLAM (w/-OIU) |
|-------------|------------------|------------------|
|             | ATE | RPE | RRPE | ATE | RPE | RRPE |
| fr3/w/xyz   | 0.0162 | 0.0236 | 0.648 | 0.0163 | 0.0237 | 0.642 |
| fr3/w/half  | 0.0291 | 0.0413 | 0.903 | 0.0290 | 0.0409 | 0.923 |
| fr3/w/static| 0.0160 | 0.0282 | 0.565 | 0.0101 | 0.0151 | 0.358 |
| fr3/w/rpy   | 0.0356 | 0.0466 | 1.047 | 0.0309 | 0.0444 | 1.003 |
| fr3/s/xyz   | 0.0141 | 0.0202 | 0.633 | 0.0138 | 0.0199 | 0.626 |
| fr3/s/half  | 0.0204 | 0.0293 | 0.842 | 0.0199 | 0.0287 | 0.826 |
| fr3/s/static| 0.0069 | 0.0101 | 0.332 | 0.0070 | 0.0102 | 0.328 |
| fr3/s/rpy   | 0.0231 | 0.0335 | 0.799 | 0.0224 | 0.0322 | 0.801 |

As shown in Table II the results obtained show that the performance of our method is globally better than that of the techniques considered. Indeed, we record relatively significant gains in terms of ATE for some sequences and slightly below the state of the art for others. This clearly demonstrates the interest of taking into account the parameters related to the proximity of humans and the depth in the estimation of the trajectory. The gain in accuracy in the estimation of the trajectory is even more significant in the case where the camera has high amplitudes of movement (sequences xyz, halfsphere and rpy).

V. CONCLUSION

Through this study we have shown the interest of taking into account the proximity of the points of interest by integrating the depth and distance information with respect to the dynamic persons to improve the estimation of ATE. This allowed us to take into account the interaction that exists between humans and inert objects by combining the information of the geometric module and the proximity information. The results obtained in terms of accuracy in the estimation of the trajectory are convincing and show the interest and the relevance of our approach. In this study we have limited ourselves to the analysis of the trajectory estimation. In a future work we will exploit this approach in the study of other aspects related to pose estimation.

REFERENCES

[1] Mur-Artal, R., Tardós, J. D. (2017). Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE transactions on robotics, 33(5), 1255-1262.
[2] Forster, C., Pizzoli, M., Scaramuzza, D. (2014, May). SVO: Fast semi-direct monocular visual odometry. In 2014 IEEE international conference on robotics and automation (ICRA) (pp. 15-22). IEEE.
[3] Qin, T., Li, P., Shen, S. (2018). Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 34(4), 1004-1020.
[4] Beghdadi, A., Mallem, M. (2022). A comprehensive overview of dynamic visual SLAM and deep learning: concepts, methods and challenges. Machine Vision and Applications, 33(4), 1-28.
[5] Bescos, B., Falcí, J. M., Civera, J., Neira, J. (2018). DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. IEEE Robotics and Automation Letters, 3(4), 4076-4083.
[6] Xiao, L., Wang, J., Qiu, X., Rong, Z., Zou, X. (2019). Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. Robotics and Autonomous Systems, 117, 1-16.
[7] Li, A., Wang, J., Xu, M., Chen, Z. (2021). DP-SLAM: A visual slam with moving probability towards dynamic environments. Information Sciences, 556, 128-142.
[8] Yu, C., Liu, Z., Liu, X. J., Xie, F., Yang, Y., Wei, Q., Fei, Q. (2018, October). DS-SLAM: A semantic visual SLAM towards dynamic environments. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1168-1174). IEEE.
[9] Hartley, R., Zisserman, A. (2003). Multiple view geometry in computer vision. Cambridge university press.
[10] Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M., Tardós, J. D. (2021). Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. IEEE Transactions on Robotics, 37(6), 1874-1890.
[11] Liu, Y., Miura, J. (2021). RDMO-SLAM: Real-time visual SLAM for dynamic environments using semantic label prediction with optical flow. IEEE Access, 9, 106981-106997.
[12] Yan, L., Hu, X., Zhao, L., Chen, Y., Wei, P., Xie, H. (2022). DGS-SLAM: A Fast and Robust RGBD SLAM in Dynamic Environments Combined by Geometric and Semantic Information. Remote Sensing, 14(3), 795.
[13] Zhang, T., Zhang, H., Li, Y., Nakamura, Y., Zhang, L. (2020, May). Flowfusion: Dynamic dense rgb-d slam based on optical flow. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 7322-7328). IEEE.
[14] Bolya, D., Zhou, C., Xiao, F., Lee, Y. J. (2020). Yolact++: Better real-time instance segmentation. IEEE transactions on pattern analysis and machine intelligence.