Dynamic Visual Sensing based on MPC controlled UAVs

Christoforos Kanellakis, Sina Sharif Mansouri and George Nikolakopoulos

Abstract— **This article considers the establishment of a dynamic visual sensor from monocular cameras to enable a reconfigurable environmental perception. The cameras are mounted on Micro Aerial Vehicles (MAV) which are coordinated by a Model Predictive Control (MPC) scheme to retain overlapping field of views and form a global sensor with varying baseline. The specific merits of the proposed scheme are: a) the ability to form a configurable stereo rig, according to the application needs, and b) the simple design, the reduction of the payload and the corresponding cost. Moreover, the proposed configurable sensor provides a glpobal 3D reconstruction of the surrounding area, based on a modified Structure from Motion approach. The efficiency of the suggested flexible visual sensor is demonstrated in simulation results that highlight the novel concept of cooperative flying cameras and their 3D reconstruction capabilities.**

I. Introduction

Micro Aerial Vehicles and especially multi rotors are gaining more and more attention for accomplishing complex tasks, considering their simple mechanical design and their versatile movement. Their ability to hover over a target or fly close to an object make them suitable for a wide range of applications, such as environment inspection mission [1], as well as aerial manipulation tasks [2]. Advanced perception capabilities towards autonomous aerial agents could play a major role for a successful mission accomplishment. Conventionally, these tasks are handled by laser range finders, stereo/monocular and RGB-D cameras. In certain applications such as infrastructure inspection and monitoring as well as object tracking, visual feedback is essential, which can be provided by vision based sensors. However, depth perception in far ranges is challenging for camera sensors, specially in stereo cameras as the depth perception is bounded by the predefined baseline which degenerates them to monocular cameras [3]. For the monocular camera case sufficient parallax between camera frames is needed to provide accurate depth information. Varying baseline approaches could be addressed when scene-depth estimation is challenging. In this particular case, multiple cameras will form a global flexible sensor, where the distance between the UAVs will correspond to the new flexible baseline. To support these attributes control strategies and vision schemes are an important factor.

In the field of vision for aerial robotics, the concept of scene perception and tracking based on monocular stereo rigs has not been extensively investigated before. The main reason is due to the mathematical complexity in combining cooperative vision, with cooperative aerial agent navigation. Nonetheless, there exist a few approaches that study flexible stereo visual rigs formed by cooperative monocular cameras. In [4], a multi-camera feature based collaborative localization and mapping scheme in static and dynamic environments has been developed. In this study collaborative pose estimation, map construction as well as camera group management issues were addressed. In [5] the relative pose estimation of two MAVs in an absolute scale to form a flexible stereo rig, has been proposed. In this approach, the extracted information from the monocular cameras was fused with inertial sensors in an error state Extended Kalman Filter (EKF), to resolve the scaling ambiguity. However, in this article there has been no synchronous flying, one UAV per time, while there has been no contribution towards the controlling of the UAVs and the corresponding sparse reconstruction from the proposed scheme. In [6], a centralized system, where each agent performed pose estimation using monocular visual odometry was introduced. In this approach the information extracted from individual platforms was merged by a mapping ground station to create a global consistent map. Finally, in [7] a framework for multi-agent cooperative navigation and mapping has been introduced. The main idea in this approach was based on monocular SLAM using RGB-D sensors for solving the scaling problems and Inertial Measurement Unit (IMU) for data fusion. The UAVs were capable of high rate pose estimation as well as sparse and dense environment reconstruction. These works are mainly focused in the performance of odometry and mapping but were not considering posing specific vision constraints in the corresponding control cost functions.

Based on the presented state of the art, the main contribution of this article is the introduction and establishment of the flexible virtual stereo rig based on MPC controlled MAVs. The proposed approach delivers the framework for adjustable depth perception of the platforms by using the concept of varying baseline. The main sensory system for this approach are monocular cameras, thus retaining a low payload and low computational processing. The second contribution of the proposed scheme is a modified SfM approach for sparse reconstruction, which is adopted for the flexible stereo rig concept. Finally, the third contribution consists the evaluation of the proposed scheme in multiple close to reality simulations where the agents navigate using model based control while guaranteeing the overlapping field of view and formulate the visual sensor according to the application needs.

The rest of the article is organized as follows. Firstly, the

The authors are with the Robotic Team at the Control Engineering Group, Department of Computer, Electrical and Space Engineering, Luleå University of Technology, Luleå SE-97187, Sweden

This work has received funding from the European Unions Horizon 2020 Research and Innovation Programme under the Grant Agreement No.644128, AEROWORKS.

establishment of the novel cooperative virtual stereo rig is presented in Section II, while the proposed control scheme is studied in Section III. Section IV presents the extended simulation results that prove the efficiency of the proposed scheme and finally concluding remarks are drawn in Section V.

II. Dynamic Visual Sensor

This Section describes the establishment of the dynamic visual sensor formulated by monocular cameras. This multi-camera system is treated as a global sensor with independently moving components used for relative motion estimation and structure reconstruction. The key idea of this approach stems from conventional feature based Structure from Motion (SfM)[8], adjusted for a virtual stereo rig with a varying baseline. Within this research the reconstruction capabilities of the cooperative flexible rig are the main focus.

Before describing the methodology for the sensor some important assumptions are established. Firstly, the cameras are calibrated with known intrinsic (K) and distortion (d) parameters. Additionally, the sensor operates when the cameras share a common field of view. As mentioned in Section III, the controller regulates the quadrotors' attitude to guarantee an overlapping field of view for the cameras. Furthermore, a smooth camera trajectory is considered to avoid abrupt movements that will deteriorate the sensor performance. In this approach the left camera's coordinate frame is considered as the origin of the sensor. The camera frames are synchonised to maintain the epipolar constraint. Finally, the system does not consider any metric information about the environment, like 3D landmark positions or actual distances between objects. In the remaining part of this section two major steps are presented: a) the camera relative pose estimation, and b) the global structure and motion calculation.

A. Relative Pose Estimation

The first step towards building the 3D map of the environment, is to group spatially the cameras with overlapping field of views under a global stereo sensor. It is necessary to estimate the relative distance t and the orientation R between the cameras employing epipolar geometry [9] calculations. Figure 1 depicts the main idea for relative pose estimation among different viewpoints.

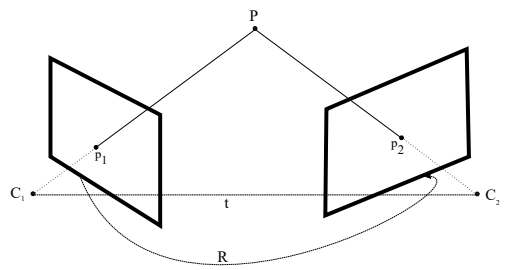

Fig. 1: Flexible Stereo Rig schematic.

The proposed approach is feature based, thus distinctive points need to be extracted from the background and matched across different viewpoints. Various methods have been proposed to track these salient features (points, shapes, lines etc.) in each view. Considering M input images $I_l\,(l\ \#images)$, the set W (where $W_l = \{(x_j, f_j)|\ j\ \#features\}$) should include feature locations $x_j \in \mathbb{R}^2$ described in descriptor f_j. In this study, the Oriented Fast Rotated Brief (ORB) feature extractor [10] is used, since it is a fast binary descriptor with rotation in-variance, which is important for the performance in wide baselines [11]. Furthermore, the detected features stored in sets W_l are matched with candidate sets within their local neighborhood from different viewpoints that define overlapping parts of the scene. More specifically, the matching process is based on the nearest neighborhood search technique described in [12], while the output of this process will be used to retrieve 3D positions of the detected features that have enough disparity for triangulation.

This appearance-based feature matching process is prone to outliers, without guaranteeing that the 2D points in the image plane correspond to the same 3D point in the world frame. Thus, to filter the induced outliers, an additional step is followed that verifies the geometric consistency satisfying the epipolar constraint [13].

In order to acquire the relative camera pose, a transformation matrix that maps points from different views is required. In the proposed approach the cameras are calibrated and loosely coupled, therefore the transformation that describes the geometry relation on each image pair is expressed through the essential matrix E [9]. Given N homogeneous points $m_1,\ m_2 \in \mathbb{R}^3$ in 2 different views and the estimated intrinsic camera matrices K_1 and K_2 the geometric verification is showed in the following equation:

$$\hat{m}_2^T E \hat{m}_1 = 0 \tag{1}$$

where $\hat{m}_2 = {K_2}^{-1} m_2$ and $\hat{m}_1 = {K_1}^{-1} m_1$ are the normalized points respectively. The essential matrix E includes the relative rotation R and the relative translation D defined as:

$$E = R \times t = R[t]_x \tag{2}$$

with

$$[t]_x = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{3}$$

and E is estimated from point correspondences by utilizing a 5-point algorithm proposed by [14] with good performance in general purpose applications. In the sequel, Singular Value Decomposition (SVD) [15] is employed to recover the rotation matrix R and the translation vector t. The decomposition of the essential matrix provides a rotation estimation without any ambiguity but position is estimated up to a scale. The scaling problem is a well known issue for monocular cameras and it is inherited in this flexible configuration, since the cameras are freely moving with no knowledge about the environment. Thus, the scale is propagating among views and requires a solid estimation scheme. Within this article, and without a generality loss, the scaling issue is assumed

solved using information from external sources, e.g. a motion capture system or other on-board sensor fusion. Thus the main focus is to establish a controlled aerial stereo rig where Figure 1 depicts the general camera spatial configuration for the epipolar geometry. This algorithm continuously tracks the identified features in sequential overlapping frames. The pair of frames where the number of tracked features is below a threshold, are considered the new keyframes and the feature extraction as well as the relative pose estimation steps are repeated. In this manner, the scene is constantly updated when new points are identified. Algorithm 1 describes the aforementioned procedure.

Algorithm 1 Relative Pose

Require: I_1, I_2, K_1, K_2, d_1, d_2 ▷ d_1, d_2 are distortion parameters
function RELATIVEPOSE(I_1, I_2, K_1, K_2, d_1, d_2)
Detect features in frames I_1, $I_2 \rightarrow (x_{1,j}, f_{1,j})$, $(x_{2,j}, f_{2,j})$
Nearest neighborhood matching between f_1 and $f_2 \rightarrow (x_1, x_2)$ ▷ Matched features
Remove false matches $\rightarrow$ inlier matches $(\hat{x}_1, \hat{x}_2)$ ▷ Geometric Verification
5-point algorithm(I_1, I_2, $\hat{x}_1$, $\hat{x}_2$) $\rightarrow E$ ▷ Essential Matrix
SVD(E) $\rightarrow R$, $s \cdot t$ ▷ scaled relative camera pose
Retrieve scale $=s$ ▷ external sources
return R, t
end function

B. 3D Structure Calculation

This subsection describes the process where the structure is retrieved from a set of points $X^{3D} = \{X_j^{3D} \in \mathbb{R}^3 |\ j\ \#points\}$. The overall mapping framework consists of two layers. Firstly, an initial map is created and afterwards, a sequential map refinement is performed. Moreover, to estimate the relative pose pairwise, camera projection matrices $P = K[R|t]$ are defined. These matrices express the projection from the Euclidean $\mathbb{R}^3$ to the image $\mathbb{R}^2$ space. An initialization of the position of the identified points in the image pairs is performed using Direct Linear Triangulation [16]. As the visual system moves around, new parts of the scene are introduced and added to the global map. The absolute pose of the resulting virtual sensor in a global coordinate frame at each instance k can be calculated (Equation 4).

$$\begin{aligned} R_k &= R_{k-1} R_{k-1,k} \\ t_k &= t_{k-1} + R_{k-1} t_{k-1,k} \end{aligned} \tag{4}$$

The obtained map M is defined as:

$$M = (X^{3D}, I_k) \tag{5}$$

where I_k is the keyframe that contains the detected points.

Triangulation is a process where uncertainties in the camera poses propagate to estimated points. This algorithm drifts fast to a non recoverable condition when there is lack of absolute measurements. Given a set of camera and world points local Bundle Adjustment [17] is employed in the current keyframes' neighborhood to optimize both the camera poses and 3D map points. The main idea behind this optimization step is the minimization of the re-projection error of the identified world points into the camera image plane under the assumption that are corrupted by Gaussian noise. The optimization is formulated in Equation 6.

$$X^{3D*} = \arg\min_{X^{3D}} \sum_{j=1}^{2N} (\|x_j - \pi(X_j^{3D}, P)\|^2) \tag{6}$$

where $\pi(\cdot)$ defines the camera projection model.

In the sequel, the calculated 3D landmark positions are assembled in a compact pointcloud form for further processing in order to be merged to a global map. More specifically, for each keyframe a pointcloud is generated and is compared with the previous one to minimize the relative pose, while the Iterative Closest Point method [18] is used to align the two clouds. In this way a global pointcloud is incrementally registered. The resulting map consists of the positions of the detected features, therefore it is sparse. The overall pipeline is briefly presented in the Algorithm 2.

Algorithm 2 3D reconstruction

function RECONSTRUCTION(I_1, I_2,K_1, K_2, d_1, d_2)
Import keyframes $I_1, I_2 \rightarrow$ RELATIVEPOSE(I_1, I_2, K_1, K_2, d_1, d_2)
Projection matrices P_1,$P_2 \rightarrow P_1 = [I|0]$, $P_2 = [R|t]$
Triangulate(I_1, I_2,$\hat{x}_1$, $\hat{x}_2$,P_1,P_2) $\rightarrow X^{3D}$ ▷ 3D points
Bundle Adjustment($\hat{x}_1$, $\hat{x}_2$,P_1,P_2,X^{3D}) $\rightarrow M$
end function
Initialize RECONSTRUCTION
for each image pair (I_1, I_2) **do**
track identified features in sequential frames $\rightarrow f_{tracked}$
if $f_{tracked} <$ threshold **then**
break
end if
update keyframes $I_{1,new}$, $I_{2,new}$
RECONSTRUCTION($I_{1,new}$, $I_{2,new}$,K_1, K_2, d_1, d_2)
if new map M_K generated **then**
ICP(M_{K-1}, M_K)
Update global map M
end if
end for

III. Model Predictive Control for Flexible Stereo Rig

The quaternion model of the UAV is extracted by representing the quadrotor as a solid body evolving in the 3D space [19]. In this case, the translation and the rotation are the two components of the motion of a rigid body and the following equations are obtained.

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -g \end{bmatrix} + \frac{T}{m} \begin{bmatrix} 2(q_0 q_2 + q_1 q_3) \\ 2(q_2 q_3 - q_0 q_1) \\ q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \tag{7}$$

where m is the quadrotor mass, T is the total thrust and q_i express rotation in quaternions. The rotational dynamics can

be summarized as:

$$\dot{w} = -I^{-1}[w \times Iw] + I^{-1} \begin{bmatrix} \tau_x \\ \tau_y \\ \tau_z \end{bmatrix} \tag{8}$$

where I is the inertia matrix, τ_x, τ_y and τ_z are related to the rotations of the quadrotor and $\dot{w}$ represent the angular velocities.

In the case of single quadrotor the controller takes as input the position and orientaion references and generates T, τ_x, τ_y and τ_z. For simplicity, instead of $T, \tau_x, \tau_y, \tau_z$ the vector $U = \{u_1, u_2, u_3, u_4\}$ is used, while u_1, u_2, u_3, u_4 can be converted to rotor velocities. Equations 7 and 8 are utilized as the dynamic model to formulate the MPC, while the decision variables are u_1, u_2, u_3 and u_4. The corresponding problem can be mathematically formulated by Equation 9 for both quadrotors. The set of control inputs is called $U = \{u_1, u_2, u_3, u_4\}$ and the states are $X = [x, \dot{x}, y, \dot{y}, z, \dot{z}, w_x, w_y, w_z, q]$.

$$\begin{aligned} &\min_{U} J(k) \\ \text{subject to} \quad &g(X) \leq 0 \\ &\dot{X} = f(X, u) \\ &u \in U \\ &X \in X_{space} \end{aligned} \tag{9}$$

where $J(k)$ is the cost function for the motion control and it is defined as a quadratic cost:

$$\begin{aligned} &J(k) = \\ &\sum_{i=1}^{N_p} [x(k+i|k) - x^r(k+i|k)]^T Q_x [x(k+i|k) - x^r(k+i|k)] \\ &+ [y(k+i|k) - y^r(k+i|k)]^T Q_y [y_{k+i|k} - y^r(k+i|k)] \\ &+ [z(k+i|k) - z^r(k+i|k)]^T Q_z [z_{k+i|k} - z^r(k+i|k)] \\ &+ \Delta u(k+i|k)^T R \Delta u(k+i|k) \end{aligned} \tag{10}$$

where $x(k+i|k), y(k+i|k)$ and $z(k+i|k)$ are the predicted position of the quadrotor at the time $k+i$ in the time k and x^r, y^r, and z^r are the desired destination for the quadrotors. The first quadrotor act as a leader following the predefined path, while the second quadrotor follows the predicted position of the first quadrotor with constant offset in the y direction. Furthermore, N_p is the prediction horizon, Q_x, Q_y, Q_z are the state error weights and R is the weight for rate of change of control inputs. The term Δ corresponds to the manipulated variables that can effect the smoothness or aggressiveness of the obtained control action. Additionally, $X \in X_{space}$ stands for the feasibility of the obtained solution and $g(X)$ represents the constraints of the optimization. The following constraints are defined for this approach:

1) **Input Constraint:** The angular velocities of the quadrotor are bounded between minimum ($\Omega_{min} = 0$) and maximum values (Ω_{max}). Thus the input values are bounded and result in the following constraints:

$$\begin{aligned} 0 \leq u_1 &\leq 4b\Omega_{max}^2 \\ -b\Omega_{max}^2 \leq u_2 &\leq b\Omega_{max}^2 \\ -b\Omega_{max}^2 \leq u_3 &\leq b\Omega_{max}^2 \\ -2d\Omega_{max}^2 \leq u_4 &\leq 2d\Omega_{max}^2 \end{aligned} \tag{11}$$

where b, d are the thrust and drag coefficients.

2) **Velocity Constraints:** The longitudinal and angular velocities are bounded between v_{min}, v_{max} and w_{min}, w_{max} respectively and thus the following constraint for each quadrotor can be defined:

$$\begin{aligned} v_{min} \leq &\dot{x}, \dot{y}, \dot{z} \leq v_{max} \\ w_{min} \leq &w_x, w_y, w_z \leq w_{max} \end{aligned} \tag{12}$$

3) **Vision Constraints:** It is necessary that all quadrotors have a common field of view and thus the agents should have the same direction of monocular cameras, which can be established by the following constraints:

$$\begin{aligned} q_r \otimes q^{1^*} &= \epsilon \\ q^{1*} \otimes q^2 &= \epsilon \end{aligned} \tag{13}$$

where q_r is the reference orientation for the first quadrotor, q^{1^*} is the conjugate of the first quadrotor orientation and q^{1*} is the predicted quaternion value of the first quadrotor ($q^{1*} = [q_0^{1*}, q_1^{1*}, q_2^{1*}, q_3^{1*}]^T$).

In the MPC formulation, the prediction horizon N_P ($N_P \geq 1$) stands for the length of the time interval that the system behavior is predicted. For stability and good performance a long prediction horizon in required. Moreover, the length of the horizon should cover the slowest system's dynamic. The sampling time is recommended to be shorter than the fastest system's dynamic so that the MPC could react to external disturbances.

IV. Evaluation Results

The proposed method has been evaluated in the simulation environment Gazebo along with the Robot Operating System (ROS) framework. The Ascending Technologies Hummingbird quadrotor [20] has been selected for the simulations. The simulation environment provides multiple external sensory systems (e.g. cameras, laser scanners, odometry sensor etc.) that can be mounted on the agents, while each quadrotor is equipped with a monocular camera. The overall dynamic parameters of the Hummingbird considered in the controller part are depicted in Table I.

TABLE I: Quadrotor dynamic parameters.

Par.	Value
m_s	0.51 kg
I_{xx}	5×10^{-3} kgm^2
I_{yy}	5×10^{-3} kgm^2
I_{zz}	8.9×10^{-2} kgm^2

Par.	Value
b	2.9×10^{-5} Ns^2
d	1.1×10^{-6} Nms^2

The simulation studies contain two components: a) the global dynamic sensor (monocular cameras) and b) the

quadrotor control. The dynamic visual sensor is implemented using OpenCV, OpenGV and PCL libraries, while the controller is implemented within ROS framework. The proper interactions among the simulation environments have been implemented as indicated in Figure 2. Generally, the

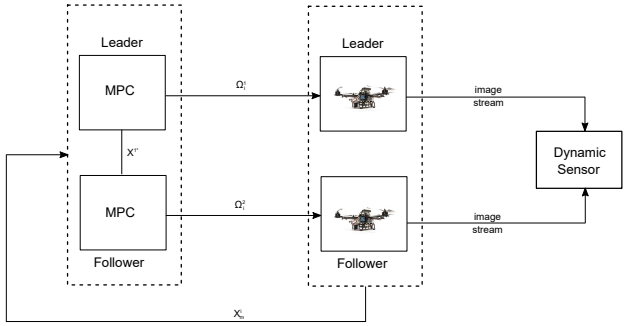

Fig. 2: Overall system architecture scheme.

proposed global sensor is applicable with any kind of robotic platform as long as field of view overlap is guaranteed. Within this paper, the simulation scenario considers the formulation of the global sensor from two MAVs that follow a sinusoidal movement keeping constant distance D_y among them. Additionally, the camera streams are synchronised through ROS framework. The designed virtual world consists of cubes with volume $1m^3$ placed in columns with different heights for visualization purposes during the 3D reconstruction. In Figure 3 the utilized Gazebo simulated virtual world is presented.

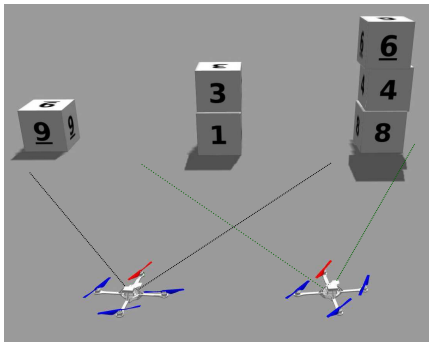

Fig. 3: The utilized flying virtual world in Gazebo.

The initial conditions for each quadrotor are $[\mathbf{0}_{1\times6}, 1, \mathbf{0}_{1\times3}]^T$ and $[-4, 0, -2, \mathbf{0}_{1\times3}, 0.7, 0, 0, 0.7]^T$ respectively. It is assumed that the second quadrotor starts with different position and different yaw angle $(\pi/2)$ and the vision constraints for the second quadrotor will be activated after obtaining the formation. The controller parameters are presented in Table II. The linear and angular velocity bounds are assumed to be same in all directions without losing the generality. The prediction horizon $N_p = 4$ is assigned for both inner and outer loop controllers, while the control horizon time is 0.1 s.

The established MPC scheme needs a full state feedback and thus in the presented scenario it is assumed that both

TABLE II: Controller parameters.

Par.	Value	Par.	Value
w_{min}	$-0.5\,\mathrm{rad/s}$	Ω_{max}	$250\,\mathrm{rad/s}$
w_{max}	$0.5\,\mathrm{rad/s}$	Q_x, Q_y, Q_z	10
v_{min}	$-3\,\mathrm{m/s}$	R	1
v_{max}	$3\,\mathrm{m/s}$	d_y	$1\,\mathrm{m}$
Ω_{min}	$0\,\mathrm{rad/s}$		

quadrotors provide the corresponding orientation, translation and their derivatives, without a loss of generality using a generic odometry sensor supported from the simulation environment.

The resulting sinusoidal path that the two quadrotors followed, based on the MPC is depicted in Figure 4. The reference path is assigned to the first quadrotor while the second quadrotor starting with different orientation (yaw) from a different position initially takes off and follows the optimal path until it reaches the constant distance with desired orientation relative to the leader quadrotor.

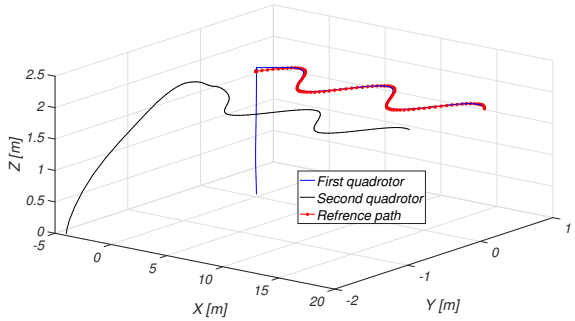

Fig. 4: 3D displacement of both quadrotors in the coordinated motion.

In this case, the second quadrotor tries to follow the leader considering the vision posed constraints as described in Section III. These constraints play a major role in the perception of the camera and the performance of the dynamic virtual rig. Therefore the dynamic sensing part is activated the moment both UAVs move in formation and guarantee overlapping field of views. In this manner it is feasible for the feature based algorithm to provide a sparse global map from distinctive landmarks of the scene. In the performed simulations, the virtual cameras were chosen to provide a resolution of 640×480 with a frame rate of 20Hz. Furthermore, the cameras were located in front of the aerial vehicles with 5° inclination to the Y-axis. Before multiple simulations were performed in various configurations for the cameras to identify problems and constraints of the approach presented.

In Figure 5 the resulting sparse 3D reconstruction during the cooperative flight of the two quadrotors and the utilization of the flexible cooperative rig based on the MPC is being depicted for the case of a constrained relative distance of 1m. From the obtained results, the overall established scheme is

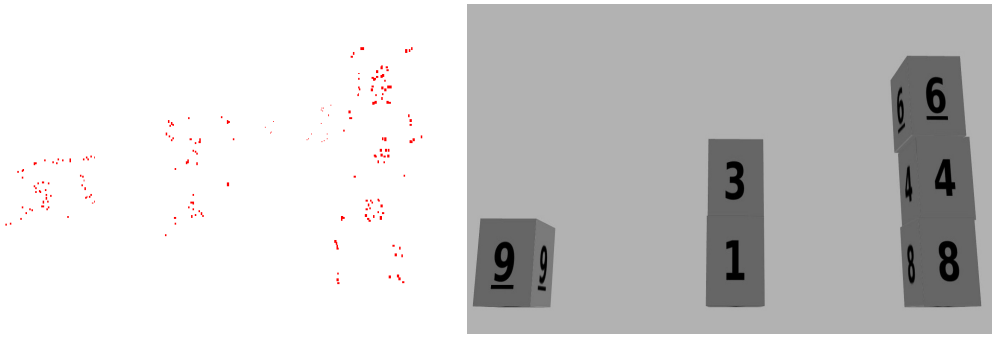

Fig. 5: 3D reconstruction during sinusoidal motion with the MPC constrained relative distance at 1m.

able to provide a sparse reconstruction of the environment constituting the collaborative stereo rig functional. It should be also highlighted that the UAVs moved towards the virtual structure capturing mainly one part of it. Generally, the more passes the quadrotors are making over the object from different perspectives during the inspection task, the more points are registered to the global scene reconstruction and the more consistent the map will be. Overall, the requirement for the online execution of the sparse reconstruction and the corresponding induced time delay in the point cloud processing, creates a trade off in the mapping accuracy, while a sequential further refinement is needed to be performed offline for producing a dense map, from the presented sparse points. The mean absolute error between ϕ, θ, and ψ of the two quadrotors was calculated $0.02\,\mathrm{rad}$, $0.02\,\mathrm{rad}$ and $0.07\,\mathrm{rad}$ correspondingly, guaranteeing continuous overlapping field of view among the cameras making the global sensor functional throughout the whole exploration time.

V. Conclusions

In this article a collaborative visual sensor was constructed based on monocular flying cameras being coordinated to retain a flexible stereo rig by a MPC. The proposed collaborative control scheme had two aims: a) to retain the formation of two UAVs, while the distance between two quadrotors remains constant, and b) to guarantee a common field of view, which was a necessary factor for the existence of the flexible stereo rig. Based on the proposed scheme, during the flight, the cameras provided a sparse reconstruction of the environment to prove the efficiency and the overall concept of the proposed collaborative scheme. Future work on this topic consist the experimental evaluation of the scheme addressing challenges like scaling, measurement noise, processing time and disturbances.

References

[1] K. Alexis, G. Nikolakopoulos, A. Tzes, and L. Dritsas, "Coordination of helicopter UAVs for aerial Forest-Fire surveillance," in *Applications of Intelligent Control to Engineering Systems*. Springer Netherlands, June 2009, pp. 169–193.

[2] D. Wuthier, D. Kominiak, C. Kanellakis, G. Andrikopoulos, M. Fumagalli, G. Schipper, and G. Nikolakopoulos, "On the design, modeling and control of a novel compact aerial manipulator," in *Control and Automation (MED), 2016 24th Mediterranean Conference on*. IEEE, 2016, pp. 665–670.

[3] F. Kendoul, "Survey of advances in guidance, navigation, and control of unmanned rotorcraft systems," *Journal of Field Robotics*, vol. 29, no. 2, pp. 315–378, 2012.

[4] D. Zou and P. Tan, "Coslam: Collaborative visual slam in dynamic environments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 2, pp. 354–366, 2013.

[5] M. W. Achtelik, S. Weiss, M. Chli, F. Dellaert, and R. Siegwart, "Collaborative stereo," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 2242–2248.

[6] C. Forster, S. Lynen, L. Kneip, and D. Scaramuzza, "Collaborative monocular slam with multiple micro aerial vehicles," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 3962–3970.

[7] G. Loianno, J. Thomas, and V. Kumar, "Cooperative localization and mapping of mavs using rgb-d sensors," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4021–4028.

[8] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[9] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[11] R. Mur-Artal, J. Montiel, and J. D. Tardós, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[12] G. Shakhnarovich, P. Indyk, and T. Darrell, *Nearest-neighbor methods in learning and vision: theory and practice*, 2006.

[13] P. Lindstrom, "Triangulation made easy," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1554–1561.

[14] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[15] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[16] R. I. Hartley and P. Sturm, "Triangulation," *Computer vision and image understanding*, vol. 68, no. 2, pp. 146–157, 1997.

[17] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustmenta modern synthesis," in *International workshop on vision algorithms*. Springer, 1999, pp. 298–372.

[18] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek, "The trimmed iterative closest point algorithm," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3. IEEE, 2002, pp. 545–548.

[19] E. Fresk and G. Nikolakopoulos, "Full quaternion based attitude control for a quadrotor," in *2013 European Control Conference (ECC), July*, 2013, pp. 17–19.

[20] A. Technologies. Asctec hummingbird. [Online]. Available: http://www.asctec.de/en/uav-uas-drones-rpas-roav/asctec-hummingbird/