A Novel Consensus-based Formation Control Scheme in the Image Space

Edgar I. Chávez-Aparicio, Hector M. Becerra and J. B. Hayet[1]

***Abstract*— In this letter, we propose a novel distributed vision-based formation control operating in the image space, with free-flying cameras in a three dimensional space as agents. Two controllers are proposed, both formulated in terms of a formation image error, without using a global reference frame nor requiring the estimation of the 3D pose between agents. The proposed formation scheme allows flexibility in defining the desired formation, without constraining it to planar formations, for example. We give formal stability guarantees based on Lyapunov analysis and evaluate our approach in simulations under a variety of initial and desired conditions, numbers of agents and agents connectivity.**

I. Introduction

Formation control of multi-agent systems (MAS) aims at making a group of agents reach a predefined set of relative positions, distances or bearings between one another [1]. Most works in formation control assume that some of this geometrical information (e.g., relative positions) is available for each agent, through measurements, and very few leverage only information from the image space and communication with neighboring agents. We fill this research gap with an approach that allows formation control based on a single camera onboard each agent using information directly from the image space. This setup is appealing in the context of Unmanned Aerial Vehicles (UAVs), with applications ranging from monitoring to entertainment.

We tackle the formation control problem for free-flying agents equipped with a camera in the 3D space *without obstacles*, where they have to achieve a desired geometric pattern, as illustrated in Fig. 1. The proposed scheme combines two feedback-based control strategies; *image-based visual servoing (IBVS)*, which is a well known type of *visual servo-control (VS)* [2], and *consensus-based formation control* [3], which uses information from neighboring agents to reach a common goal. As opposed to the large majority of approaches that solve this problem in the Euclidean space, our distributed control scheme solves the formation control problem directly *in the image space*, which avoids the need for 3D measurements.

VS schemes are used to regulate the pose of a single camera, with a desired pose specified by a previously captured *reference image*. Within VS algorithms, image-based VS (IBVS) uses feedback information directly from the image space, while position-based VS (PBVS) uses feedback from data in the Euclidean space, estimated from image features. In the context of formation control, an option where 3D information is estimated from images is the use of unitary vectors known as bearings, which indicate the direction among agents to define a *geometric shape*. The bearing-based formation approach allows to reach formations up to a scale. In [4], bearings and a set of features related to the distance to the neighbors are used to reach a desired formation. Also, in [5], a controller enforces a desired relative pose to create a circular formation.

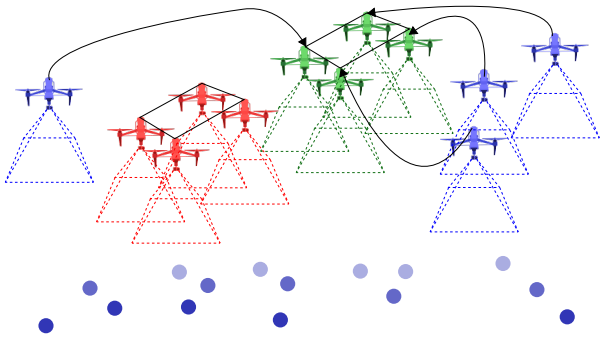

Fig. 1. A possible application of our approach: from a set of initial conditions (blue), UAVs have to reach a replicate (green) of the reference cameras (red), such that the set of replicates matches the set of references up to a geometric similarity. 3D image points (blue dots) are seen by the agents and their image projections are used to drive the agents to a formation.

A common image-based approach for formation control consists in placing a visual marker on a leader agent and in specifying a desired view of that marker for the followers. Formation arises when all the followers reach their corresponding reference poses. This approach has been used with UAVs [6], differential-drive robots [7] and with a group of satellites [8], where a leader-follower assignment is set to close a loop and reach a circular formation. These works are straightforward applications of IBVS but there is no communication between the agents. In contrast, our approach leverages the exchange of information between agents to improve the formation control.

Distributed vision-based formation control has been addressed using consensus-based approaches where communication between agents is essential. In [9], agents in formation collaboratively track a target; each agent computes individual steering actions from visual data and a global agreement for these actions is reached through consensus. In [10], one leader quadrotor uses IBVS to track a target and the other agents are driven by a consensus-based formation algorithm, assuming that the relative position between agents can be estimated. In [11], within a coordinated visual tracking scheme for a

[1]The authors are with Centro de Investigación en Matemáticas (CIMAT), Computer Science Department, Jalisco S/N, Col. Valenciana, Guanajuato, Gto., Mexico. edgar.chavez@cimat.mx, hector.becerra@cimat.mx, jbhayet@cimat.mx

moving target with UAVs, each agent measures the phase angles towards the target from image data to coordinate their separation. In [12], IBVS-driven consensus is applied with a fixed external camera for robot manipulators; an adaptive observer estimates the visual velocities and uncertainties.

Some works on distributed formation control propose to bypass the use of geometric 3D information by using visual information and without the need of a global reference frame. In [13], a consensus-based strategy for formation control of quadrotors uses specific elements of the homography matrix between views of two neighbor agents, each one carrying a camera pointing toward a planar surface. In [14], IBVS is used in a formation control scheme where followers track and synchronize with a leader, which must be in view to match each follower's image with its corresponding reference image. A consensus term formed from the image errors is included to improve the convergence.

We introduce a novel distributed vision-based formation control for agents equipped with a camera (which we will extend, in the future, to onboard cameras on quadrotors), expressed as a consensus of errors in the image space. We propose two controllers: the first one computes the cameras velocities proportionally to the consensus error; the second one includes an integral term to reduce steady state errors. In contrast to most works in the state of the art, our approach does not use a global reference frame or 3D relative poses between agents or geometric constraints. To the authors' knowledge, this is the first vision-based formation scheme operating directly in the image space, allowing flexibility in defining the desired formations, without limiting them to planar formations. Our controllers are evaluated in simulations for random initial and desired conditions, different numbers of agents and connectivities of the set of agents.

II. Preliminaries and Problem statement

Let $\mathcal{G}$ be a graph with a vertex set $\mathcal{V}(\mathcal{G})$ and an edge set $\mathcal{E}(\mathcal{G})$. In our case, $\mathcal{G}$ represents the communication relations (edges) between agents (vertices). The set of neighbors of a vertex i is denoted as $\mathcal{N}_i(\mathcal{G}) \triangleq \{j \in \mathcal{V} \mid (i,j) \in \mathcal{E}(\mathcal{G})\}$. The *Laplacian* matrix of $\mathcal{G}$ is $\boldsymbol{\mathcal{L}}(\mathcal{G}) \triangleq \mathbf{D} - \mathbf{A} \in \mathbb{R}^{N\times N}$ where $\mathbf{A} = [a_{ij}]$ is the adjacency matrix of $\mathcal{G}$ with $a_{ij} = 1$ if $i \neq j$ and $(i,j) \in \mathcal{E}(\mathcal{G})$ and $a_{ij} = 0$ otherwise, and $\mathbf{D} = \mathrm{diag}(d_1, \cdots, d_N)$ with $d_i \triangleq \sum_{j=1}^{N} a_{ij}$. If $\mathcal{G}$ is *undirected*, $\boldsymbol{\mathcal{L}}(\mathcal{G})$ is positive semi-definite and symmetric. If $\mathcal{G}$ is *connected*, $\boldsymbol{\mathcal{L}}(\mathcal{G})$ has a null eigenvalue with algebraic multiplicity one, paired with the eigenvector $\mathbf{1}_N = [1 \ \cdots \ 1]^T$.

In this work, the agents are supposed to be *cameras* with pose $\boldsymbol{x} = [\boldsymbol{p}^T, \boldsymbol{\theta}^T]^T \in \mathbb{R}^6$, where $\boldsymbol{p} = [x, y, z]^T$ and $\boldsymbol{\theta} = [\phi, \theta, \psi]^T$, with ϕ roll, θ pitch and ψ yaw Euler angles. We assume that the agent motion is described as a single integrator

$$\dot{\boldsymbol{x}}(t) = \mathbf{v}, \tag{1}$$

where $\mathbf{v} = [\boldsymbol{v}^T, \boldsymbol{\omega}^T]^T$, with $\boldsymbol{v} = [v_x, v_y, v_z]^T$ the linear velocity and $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$ the angular velocity of the origin of the camera reference frame.

We assume that the cameras are standard pinhole cameras and observe and track a set of M 3D points $\mathbf{X}_j$ for $j \in \{1, ..., M\}$ (blue dots in Fig. 1). Then each point is projected on the image plane and the coordinates of the projection are

$$\boldsymbol{s}_j = [\mathtt{x}_j, \mathtt{y}_j]^T = [X_j^c/Z_j^c, Y_j^c/Z_j^c]^T, \tag{2}$$

also known as *normalized image coordinates*, where $\mathbf{X}_j^c = [X_j^c, Y_j^c, Z_j^c]^T$ are the points coordinates in the camera reference frame. These projections are easy to obtain in simulation, and can be detected with fiducial markers or interest points in real images. The relationship between the time derivative of the image features coordinates $\boldsymbol{s}$ and the camera velocity is [2]:

$$\dot{\boldsymbol{s}} = \mathbf{L}_{\boldsymbol{s}}\mathbf{v}, \tag{3}$$

being $\mathbf{L}_{\boldsymbol{s}} \in \mathbb{R}^{2M\times 6}$ the *interaction matrix* associated to the points $\boldsymbol{s}$. This matrix can be assembled by stacking M 2×6 sub-matrices, corresponding to each point. They have the form:

$$\mathbf{L}_{\boldsymbol{s}_j} = \begin{bmatrix} -\frac{1}{Z_j^c} & 0 & \frac{\mathtt{x}_j}{Z_j^c} & \mathtt{x}_j\mathtt{y}_j & -(1+\mathtt{x}_j^2) & \mathtt{y}_j \\ 0 & -\frac{1}{Z_j^c} & \frac{\mathtt{y}_j}{Z_j^c} & 1+\mathtt{y}_j^2 & -\mathtt{x}_j\mathtt{y}_j & -\mathtt{x}_j \end{bmatrix}, \tag{4}$$

where Z_j^c is the *depth* of feature j, and is often unknown in practice. For desired image features $\boldsymbol{s}^*$ associated with a desired pose $\boldsymbol{x}^*$, IBVS typically uses control laws of the form $\mathbf{v} = \gamma\hat{\mathbf{L}}_{\boldsymbol{s}}^{+}(\boldsymbol{s} - \boldsymbol{s}^*)$ with a gain $\gamma > 0$ and an *approximate* interaction matrix $\hat{\mathbf{L}}_{\boldsymbol{s}}$ since the features depths have to be estimated. The $+$ operator indicates the pseudoinverse matrix; in this letter, the Moore-Penrose inverse $\mathbf{B}^+ = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ is used. This control law yields $\boldsymbol{s} \to \boldsymbol{s}^*$ [2], [15] and, under ideal conditions, $\boldsymbol{x} \to \boldsymbol{x}^*$. Even with an approximate interaction matrix, in practice this control shows local asymptotic stability in a surprisingly quite large [2] neighborhood at $\boldsymbol{x}^*$.

Consider a MAS composed of N cameras with joint state $\mathbf{x} \triangleq [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T]^T \in \mathbb{R}^{6N}$, whose motion is modeled as (1) and with an associated graph $\mathcal{G}$ modeling the network's connectivity. Let $\mathbf{x}^* \triangleq [\boldsymbol{x}_1^{*T}, \ldots, \boldsymbol{x}_N^{*T}]^T$ be an *arbitrary* reference joint state, then we address a formation control problem to design a distributed control law such that the set $\Xi_{\mathbf{x}^*}$ of equivalent formations to $\mathbf{x}^*$ is asymptotically stable with respect to the agent's dynamics (1). We define an *equivalent formation* to $\mathbf{x}^*$ as a state $\mathbf{x}$ resulting from applying a translation, rotation and scaling to $\mathbf{x}^*$, such that relative orientations are kept among agents.

III. Proposed vision-based formation approach

To solve this problem, reference images are taken from reference poses $\mathbf{x}^*$ to built the joint vector $\mathbf{s}^* = [\boldsymbol{s}_1^{*T}, ..., \boldsymbol{s}_N^{*T}]^T \in \mathbb{R}^{2NM}$, with $\boldsymbol{s}_i^* = [\boldsymbol{s}_{1i}^{*T}, ..., \boldsymbol{s}_{Mi}^{*T}]^T \in \mathbb{R}^{2M}$. Note that this sets a constraint on $\mathbf{x}^*$, as the objects (points) used as features should be visible. Similarly, from images taken at the current pose $\mathbf{x}(t)$, the vector $\mathbf{s}(t) \in \mathbb{R}^{2NM}$ is built from the corresponding visual features for each camera $\boldsymbol{s}_i(t) = [\boldsymbol{s}_{1i}^T, ..., \boldsymbol{s}_{Mi}^T]^T \in \mathbb{R}^{2M}$.

Our controller is inspired from displacement-based approaches, for instance [3], that use the camera positions $\mathbf{x}$ to calculate the displacement error for each agent i with respect to a reference $\mathbf{x}^*$. Instead, we define the *image formation error* as a function of the image projections, as follows

$$\mathbf{e}_i(t) = \sum_{j\in\mathcal{N}_i} \big((\boldsymbol{s}_j(t) - \boldsymbol{s}_j^*) - (\boldsymbol{s}_i(t) - \boldsymbol{s}_i^*)\big) \in \mathbb{R}^{2M}. \tag{5}$$

The joint image error vector $\mathbf{e} = [\mathbf{e}_1^T, ..., \mathbf{e}_N^T]^T \in \mathbb{R}^{2NM}$ is

$$\mathbf{e} = -(\boldsymbol{\mathcal{L}} \otimes \mathbf{I}_{2k})(\mathbf{s}(t) - \mathbf{s}^*), \tag{6}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{I}_q$ is the $q \times q$ identity matrix. Using (3), the time dynamics of $\mathbf{e}_i$ becomes

$$\dot{\mathbf{e}}_i = \sum_{j \in \mathcal{N}_i} (\dot{\boldsymbol{s}}_j(t) - \dot{\boldsymbol{s}}_i(t)) = \sum_{j \in \mathcal{N}_i} (\mathbf{L}_{\boldsymbol{s}_j} \mathbf{v}_j - \mathbf{L}_{\boldsymbol{s}_i} \mathbf{v}_i). \tag{7}$$

Hereafter, the goal is to achieve $\mathbf{e} = \mathbf{0}$, i.e., each camera must be driven to reach zero image formation error. Next, we describe a *distributed* control to drive the system to a formation *equivalent* to the one where the reference images were captured as seen in Fig. 1. Since this setup does *not* use the 3D geometry between cameras, only image features, we refer to our approach as *formation control in the image space*.

We make the following assumptions: 1) Every agent knows the references $\mathbf{s}_i^*$. 2) Cameras motions are done in an obstacle-free space and we neglect inter-agent collision. 3) The observed 3D points are *static*. 4) $\mathcal{G}$ is *undirected* and connected.

A. Proportional control

Taking inspiration on the IBVS approach, we propose the following control law that uses feedback from (5):

$$\mathbf{v}_i = \lambda \mathbf{L}_{\boldsymbol{s}_i}^+ \mathbf{e}_i, \tag{8}$$

with $\lambda > 0$ a gain factor and $\mathbf{L}_{\boldsymbol{s}_i}$ the *true* interaction matrix.

The closed loop dynamics (7) can be computed as:

$$\dot{\mathbf{e}}_i = \lambda \sum_{j \in \mathcal{N}_i} (\mathbf{L}_{\boldsymbol{s}_j} \mathbf{L}_{\boldsymbol{s}_j}^+ \mathbf{e}_j - \mathbf{L}_{\boldsymbol{s}_i} \mathbf{L}_{\boldsymbol{s}_i}^+ \mathbf{e}_i). \tag{9}$$

Let $\mathbf{M}_{\boldsymbol{s}_i} \triangleq \mathbf{L}_{\boldsymbol{s}_i} \mathbf{L}_{\boldsymbol{s}_i}^+$ and $\mathbf{M}_\mathbf{s} = \text{diag}(\mathbf{M}_{\boldsymbol{s}_1}, \ldots, \mathbf{M}_{\boldsymbol{s}_N})$. Then the joint system in closed loop is expressed as:

$$\dot{\mathbf{e}} = -\lambda (\boldsymbol{\mathcal{L}} \otimes \mathbf{I}_{2M}) \mathbf{M}_\mathbf{s} \mathbf{e}. \tag{10}$$

Theorem 1: For the system (10) with $\mathbf{e}$ as in (6), obtained by using the controller (8), the equilibrium point $\mathbf{e} = \mathbf{0}$ is asymptotically stable if the matrices $\mathbf{M}_{\boldsymbol{s}_i} > 0$, for $i \in \{1, ..., N\}$ and $\boldsymbol{\mathcal{L}}$ comes from a connected graph.

Proof: Let us choose $V = \frac{1}{2} \mathbf{e}^T \mathbf{M}_\mathbf{s} \mathbf{e}$ as a Lyapunov candidate function for this. Then, since $\mathbf{M}_\mathbf{s}$ is symmetric, its time-derivative is $\dot{V} = \mathbf{e}^T \mathbf{M}_\mathbf{s} \dot{\mathbf{e}}$, which is rewritten as

$$\dot{V} = -\lambda \mathbf{e}^T \mathbf{M}_\mathbf{s} (\boldsymbol{\mathcal{L}} \otimes \mathbf{I}_{2M}) \mathbf{M}_\mathbf{s} \mathbf{e}.$$

From the spectral properties of the Laplacian matrix in $(\boldsymbol{\mathcal{L}} \otimes \mathbf{I}_{2M})$, which is positive semi-definite for a connected graph, we deduce $\dot{V} \leq 0$. Let the invariant set making $\dot{V} = 0$ corresponds to $\mathbf{e} = \mathbf{0}$ and $\mathbf{M}_\mathbf{s} \mathbf{e} \in \ker(\boldsymbol{\mathcal{L}} \otimes \mathbf{I}_{2M})$. Both cases are covered with the constraint $\mathbf{M}_{\boldsymbol{s}_i} \mathbf{e}_i = \mathbf{M}_{\boldsymbol{s}_j} \mathbf{e}_j \; \forall i \neq j$.

Due to the Laplacian matrix property of zero row sums for a connected graph, we have $\sum \mathbf{e}_i = \mathbf{0}$. The sum in the invariant set can be transformed as $(\sum \mathbf{M}_{\boldsymbol{s}_i}^{-1}) \mathbf{M}_{\boldsymbol{s}_k} \mathbf{e}_k = \mathbf{0}$ for any $k \in [1, N]$ since we assume $\mathbf{M}_{\boldsymbol{s}_i}$ positive definite, as in most IBVS formulations [2]. Hence, both constraints require $\mathbf{e} = \mathbf{0}$ to be the largest invariant set yielding $\dot{V} = 0$, the desired equilibrium. By LaSalle's invariance principle [16, p. 115], $\mathbf{e} = \mathbf{0}$ is an asymptotically stable equilibrium point. ■

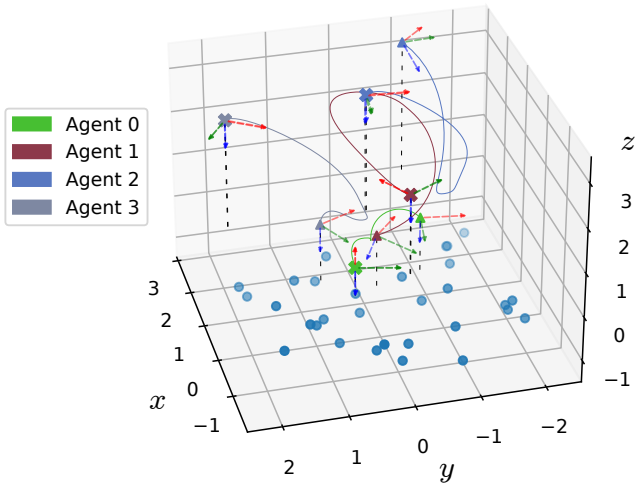

Fig. 2. Example scenario. 3D motion of each frame of reference (continuous lines) from the initial conditions (triangles) towards the formation reached at time 2000; the reference poses (crosses) are superposed with the final positions. Dashed lines show z-coordinate magnitude.

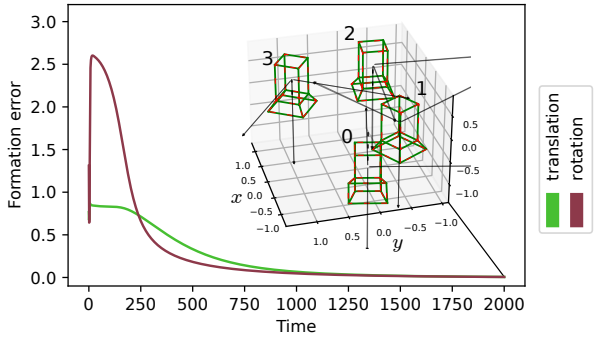

Fig. 3. Example scenario. Evolution in time of the 3D metrics and superposed scaled cameras at final poses (green) and reference poses (red).

B. Proportional-Integral (P-I) controller

As previously studied in [15], we should note that the assumption $\mathbf{M}_{\boldsymbol{s}_i} > 0$ might not always hold. As a consequence, the formulation is prone to local minima, i.e., steady state errors where $\mathbf{e}_i \in \ker(\mathbf{L}_{\boldsymbol{s}_i}^+)$. We propose to tackle this issue by using an integral control component: $\boldsymbol{\varepsilon}_i \triangleq \int_0^t \mathbf{e}_i(\tau) d\tau$. The error dynamics becomes of second order:

$$\begin{bmatrix} \dot{\mathbf{e}}_i \\ \dot{\boldsymbol{\varepsilon}}_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_i \end{bmatrix} + \begin{bmatrix} -|\mathcal{N}_i| \mathbf{L}_{\boldsymbol{s}_i} \\ \mathbf{0} \end{bmatrix} \mathbf{v}_i + \sum_{j \in \mathcal{N}_i} \begin{bmatrix} \mathbf{L}_{\boldsymbol{s}_j} \mathbf{v}_j \\ \mathbf{0} \end{bmatrix}, \tag{11}$$

for which, we propose the following feedback control:

$$\mathbf{v}_i = \mathbf{L}_{\boldsymbol{s}_i}^+ \left(\gamma_e \mathbf{e}_i + \gamma_\varepsilon \boldsymbol{\varepsilon}_i \right), \tag{12}$$

with $\gamma_e > 0$, $\gamma_\varepsilon > 0$ control gains. In closed loop, we get:

$$\begin{bmatrix} \dot{\mathbf{e}} \\ \dot{\boldsymbol{\varepsilon}} \end{bmatrix} = \begin{bmatrix} -\gamma_e (\boldsymbol{\mathcal{L}} \otimes \mathbf{I}_{2M}) \mathbf{M}_\mathbf{s} & -\gamma_\varepsilon (\boldsymbol{\mathcal{L}} \otimes \mathbf{I}_{2M}) \mathbf{M}_\mathbf{s} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{e} \\ \boldsymbol{\varepsilon} \end{bmatrix}. \tag{13}$$

Theorem 2: For the system (13) with $\mathbf{e}$ as in (6), obtained by using the controller (12), the equilibrium point $\mathbf{e} = \mathbf{0}$, $\mathbf{M}_{\boldsymbol{s}_i} \boldsymbol{\varepsilon}_i = \mathbf{M}_{\boldsymbol{s}_j} \boldsymbol{\varepsilon}_j, \forall i \neq j$, is asymptotically stable if the matrices $\mathbf{M}_{\boldsymbol{s}_i} > 0$, for $i \in \{1, ..., N\}$ and $\boldsymbol{\mathcal{L}}$ comes from a connected graph.

Proof: Let $V = \frac{1}{2}\gamma_\varepsilon(\boldsymbol{\varepsilon}+\mathbf{e})^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}(\boldsymbol{\varepsilon}+\mathbf{e}) + \frac{1}{2}\mathbf{e}^T\mathbf{M_s}\mathbf{e} + \frac{1}{2}(\gamma_\varepsilon^2+\gamma_e\gamma_\varepsilon)\boldsymbol{\varepsilon}^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}\boldsymbol{\varepsilon}$ be a Lyapunov candidate function with the following derivative:

$$\begin{aligned}\dot{V} =& \gamma_\varepsilon(\boldsymbol{\varepsilon}+\mathbf{e})^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}(\dot{\boldsymbol{\varepsilon}}+\dot{\mathbf{e}}) + \mathbf{e}^T\mathbf{M_s}\dot{\mathbf{e}} \\ &+ (\gamma_\varepsilon^2+\gamma_e\gamma_\varepsilon)\boldsymbol{\varepsilon}^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}\dot{\boldsymbol{\varepsilon}} \\ =& -\gamma_\varepsilon^2\boldsymbol{\varepsilon}^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}\boldsymbol{\varepsilon} \\ &- (\gamma_e-\gamma_\varepsilon)\mathbf{e}^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}\mathbf{e} \\ &- \gamma_e\gamma_\varepsilon\mathbf{e}^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}\mathbf{e}.\end{aligned}$$

Under the condition $\gamma_e > \gamma_\varepsilon$, we have $\dot{V} \leq 0$ due to $(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})$ being positive semi-definite and the fact that all the components are symmetric matrix products. In this case, the invariant set where $\dot{V}=0$ needs to fulfill the condition $\mathbf{e}^T\mathbf{M_s}(\boldsymbol{\mathcal{L}}\otimes\mathbf{I}_{2M})\mathbf{M_s}\mathbf{e}=0$. Similar to the proportional case, we have that the solution for that constraint is $\mathbf{e}=\mathbf{0}$. However, another condition $\mathbf{M}_{\boldsymbol{s}_i}\boldsymbol{\varepsilon}_i = \mathbf{M}_{\boldsymbol{s}_j}\boldsymbol{\varepsilon}_j \ \forall i \neq j$ also describes the invariant set for $\dot{V}=0$. Then, by LaSalle's invariance principle [16, p. 115], the set $\mathbf{e}=\mathbf{0}$; $\mathbf{M}_{\boldsymbol{s}_i}\boldsymbol{\varepsilon}_i = \mathbf{M}_{\boldsymbol{s}_j}\boldsymbol{\varepsilon}_j \ \forall i \neq j$ is an asymptotically stable equilibrium point. ■

Our approach only guarantees *local* asymptotic stability. However, convergence seems to be attained in a large neighborhood of the equilibrium point, as we will see in Sect. IV.

C. Formation Metrics

To assess how far a given state $\mathbf{x}$ is from a formation equivalent to $\mathbf{x}^*$, we use the following metrics:

$$\epsilon_t = \frac{1}{N}\min_{\alpha>0}\sum_{i=1}^{N}||\alpha(\bar{\boldsymbol{p}}_i) - \bar{\boldsymbol{p}}_i^*||^2, \tag{14}$$

$$\epsilon_\theta = \frac{1}{N}\sum_{i=1}^{N}\text{acos}\left\{\frac{1}{2}(\text{trace}(\mathbf{R}^T(\bar{\boldsymbol{\theta}}_i^*)\mathbf{R}(\bar{\boldsymbol{\theta}}_i)) - 1)\right\}^2, \tag{15}$$

where the bar operator indicates a normalization process in which: (1) the centroids are placed at the origin; (2) the average distance of the cameras to the origin is 1; (3) $||\bar{\boldsymbol{p}}_i - \bar{\boldsymbol{p}}_i^*||$ is minimized by rotation using the method in [17]. The notation $\mathbf{R}(\boldsymbol{\theta})$ represents the rotation matrix built from the camera pose parameters $\boldsymbol{\theta}$.

These metrics handle the difference in scaling between the desired and reached formations. The scaling is not an issue, as it is a typical feature of vision-based schemes [13]. In our approach, as in bearing-based formation schemes, the formation scale can be set by using a pair of leader agents acting as formation anchors. Moreover, the free scaling factor may be an advantage when a formation has to move in constrained environments with obstacles.

IV. Simulation Results

A simulation environment was developed in Python. Each agent is simulated as a pin-hole camera with intrinsic parameters encoded in the camera calibration matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ with entries $\mathbf{K}_{13} = \mathbf{K}_{23} = 512$ and $\mathbf{K}_{11} = \mathbf{K}_{22} = 200$.

Then, the normalized image coordinates $\boldsymbol{s}_j$ as expressed in (2) can be obtained from pixel coordinates $\boldsymbol{\zeta}_j$ as $[\boldsymbol{s}_j^T, 1]^T = \mathbf{K}^{-1}[\boldsymbol{\zeta}_j^T, 1]^T$. Note that the depth Z_j^c in (4) can not be obtained from pixel coordinates and in our simulations, we use the real Z_j^c. The use of approximate depths will be explored in future research. Approximate depths should not modify the positive definiteness of $\mathbf{L}_{\boldsymbol{s}_i}\hat{\mathbf{L}}_{\boldsymbol{s}_i}^+$ [2]. Each camera has 6 degrees of freedom (DOF) as described in Sect. II. We use two types of failure tests regarding the camera Field of View (FOV). The first test verifies that all M features are in front of the camera. In the following, a simulation is said to be a *failure* if it violates the condition $Z_j^C > 0$ for any $1 \leq j \leq M$. The second test is used to select the reference and initial configurations, and checks if all the features projections are inside the camera image sensor.

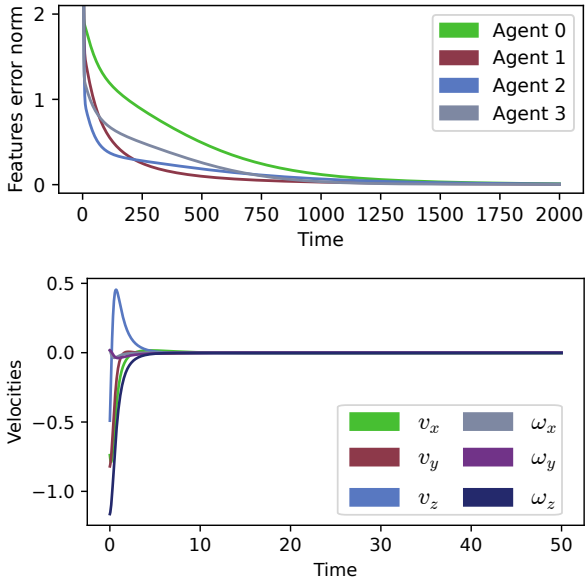

Fig. 4. Example scenario. Image formation errors $||\mathbf{e}_i||$ for each agent along time (top). Control inputs for one camera for the first 10 seconds (bottom-left) and for the whole simulation, using a wider y range (bottom-right).

The initial conditions are the poses of the cameras at $t=0$. For testing, 100 scenarios are defined by a set of reference cameras and a set of $M = 30$ 3D points. The reference cameras are sampled uniformly from $\Sigma_r \triangleq \{x \in [-2,-2], y \in [-1,2], z \in [0,3], \psi \in [-\pi,\pi]\}$, with $\phi = \theta = 0$ to emulate final poses of the cameras on a hovering quadrotor. The 3D points are selected randomly and uniformly in $\Sigma_p \triangleq \{[-2,-2],[-1,2],[-1,0]\}$.

In all the simulations, the roll and pitch control gains are diminished by a 1:2 gain ratio with respect to the other components, in order to reduce the occurrence of failures due to large rotations. The initial conditions are spawned randomly from the reference cameras as $\mathbf{x}(0) = \mathbf{x}^* + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}$ a random vector with a distribution described in the next sections.

A. Example scenario

Let us present a typical scenario where the desired performance is achieved using the proportional controller (8). Four cameras are simulated with a fully connected communication graph. The initial conditions are taken from a sample with entries of $\boldsymbol{\epsilon}$ bounded by ±2 (translations), $\pm\frac{\pi}{6}$ (roll) and

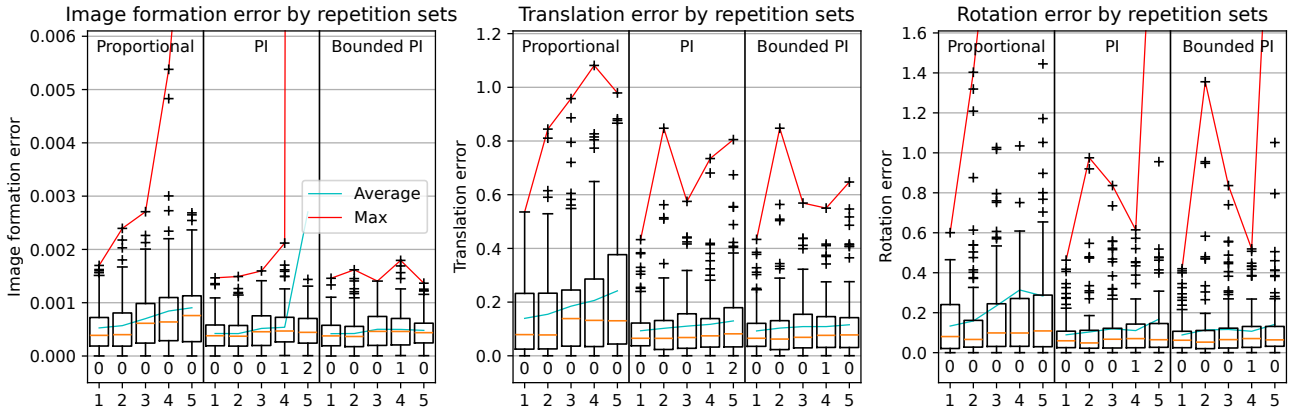

Fig. 5. Performance for initial conditions increasingly far from the reference camera pose. Each group of 5 box-plots represents the final error values for the three controllers (8), (12) and (16). The x-axis indicates the increment with respect to the initial conditions, i.e., the boundaries of $\boldsymbol{\epsilon}$. The maximum and average values for each box-plot are shown in red and cyan, respectively. The number of failure cases for each set is displayed below each box-plot.

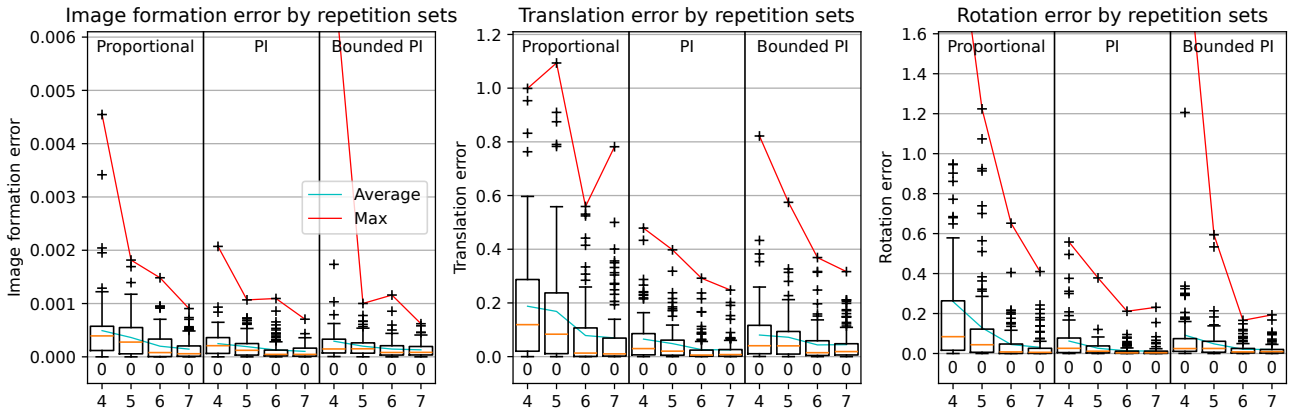

Fig. 6. Performance for different graph sizes and random connectivity. Each group of 4 box-plots represents the final errors values for the three controllers, similar than in Fig. 5. In this case, each step along the x-axis represents an increment in the graph size.

0 (pitch) to emulate initial poses of cameras on a hovering quadrotor, and $\pm\frac{\pi}{2}$ (yaw). The yaw limits are used to avoid undesired effects due to large rotations on the optical axis causing unbounded motions, a problem known as the Chaumette Conundrum [18]. Fig. 2 shows the 3D motion of the cameras reference frames; their final positions (crosses) are superposed with the reference cameras. Clearly, the set of reference cameras poses is an equilibrium point of the MAS that is expected; in practice, convergence to a broader class of equivalent formations in the sense defined in Sect. II has been observed. A deeper analysis of the set of possible equilibrium points in the 3D space is part of our future work and out of the scope of this paper. Fig. 3 presents the evolution of the 3D metrics along time, which reach values below 10^{-2}. Fig. 4, shows that each component $||\mathbf{e}_i||$ converges to zero, being consistent with the behavior of the control velocities.

In comparison with the closely related works [13], [14], our approach works under different conditions and can achieve a wider range of formations. The method in [13] relies on the estimation of a geometric constraint valid for planar scenes and can only solve planar formations, whereas our approach works for general scenes and can achieve planar and non-planar formations. In [14], each agent is constrained by IBVS to reach a fixed pose relative to a leader and the reached formation is indeed globally predefined, unlike our approach that can reach equivalent formations that are not limited to be tied to a common reference.

B. *Performance under different initial conditions*

To study the extent of the attraction region due to the local stability, 100 scenarios are defined by a set of reference cameras and image points. The entries of $\boldsymbol{\epsilon}$ are given for an incremental sequence of k steps, leading to boundaries of $\pm\frac{2k}{5}$ for translations, $\pm\frac{k\pi}{30}$ for roll, $\pm\frac{\pi}{2}$ for yaw and 0 for pitch. A fully connected graph is used. For each k, the performance is assessed with the final image formation error, the 3D metrics at time $8000s$ and the amount of failure cases.

The Proportional and P-I controllers are compared using $\lambda = 0.1$, $\gamma_e = 0.1$ and $\gamma_\varepsilon = 0.005$. Those values were determined empirically as follows: A series of tests were performed with different positive values while satisfying the condition $\gamma_e > \gamma_\varepsilon$. Then, the values leading to the best performance were selected. The results are depicted in Fig. 5. As seen in the middle plots, the P-I control reaches lower values for the image error formation and the 3D metrics, but

failures occur in 3 cases (among 500). Also, both the number of failures and the final errors increase as the initial conditions are farther from the references. We attribute these failures to the large velocities at the beginning of the simulations, a usual effect of consensus-based controllers which is an issue in practice due to maximum allowable control inputs. To handle this, a *bounded* version of the P-I controller is tested and compared to the other schemes (right plots):

$$\mathbf{v}_i = \tanh\left(\mathbf{L}_{\mathbf{s}_i}^{+}\left(\gamma_e \mathbf{e}_i + \gamma_\varepsilon \boldsymbol{\varepsilon}_i\right)\right). \tag{16}$$

The results in Fig. 5 show that the bounded control (16) performs better and reduces the number of failure cases.

C. Influence of the graph size and topology

We have also evaluated our approach under different graph topologies, with randomly placed reference cameras and increasing the numbers of agents (from 4 to 7). In total, 100 scenarios are used for each graph size with different values of $\boldsymbol{\epsilon}$, bounded as described in Sect. IV-A. The edges are set randomly as follows. Starting from an empty edge set $\mathcal{E}$, a random number of edges is chosen from a uniform variable bounded by the amount of possible edges that can form a connected graph. Then, an element of the complement of $\mathcal{E}$ is selected with a uniform distribution and added to $\mathcal{E}$. Last, we filter out the unconnected graphs. Note that isomorphic graphs may exist. The results in Fig. 6 are consistent when comparing the different proposed controllers between each other. We note a decrease in the final errors as $|\mathcal{V}|$ increases. This might be an effect of the average algebraic connectivity, which is related to the velocity of convergence in consensus systems [3].

D. No reference feature implies consensus

We experimented with the trivial scenario $\mathbf{s}^* = \mathbf{0}$ where the system should converge in the image space to $\boldsymbol{s}_i = \boldsymbol{s}_j$ for all i, j, meaning that all cameras reach the same pose in the Euclidean space. To evaluate this case, we use the same scenarios from Sect. IV-C. In *all* scenarios not ending in failure, the errors reach 0 values. Overall, only 4/400 experiments end up in failure for the proportional controller and 5/400 for the P-I controller. In those scenarios, at least one camera had opposing z-axis to the rest of the agents while maintaining the features in front of the cameras, leading to a problem similar to the Chaumette Conundrum [18] and the controller local solution sends the cameras to infinity.

V. Conclusions

We have presented a novel distributed vision-based formation approach for agents equipped with a camera, where the control is done in the image space. Two controllers are proposed, both relying on relative errors between image points: proportional and proportional-integral. Local asymptotic stability has been proven for both controllers. Our approach provides flexibility in the selection of the target formation, unlike existing methods that only work for planar formations or take a leader agent as a common reference. Both controllers, together with a bounded version of the P-I controller have been evaluated in simulations for a large set of random scenarios of initial and reference conditions, different number of agents and connectivities. The results show the validity of the proposed controllers to achieve formations only from visual information.

Our future work involves a deeper analysis on the reachable equivalent formations. Besides, we plan to extend our approach to implement it in real scenarios, which requires detection and matching of image points, an strategy to keep them in the FOV, as well as a collision avoidance strategy. With these features, experiments with quadrotors with onboard cameras could be realized.

References

[1] Y. Liu, J. Liu, Z. He, Z. Li, Q. Zhang, and Z. Ding. A Survey of Multi-Agent Systems on Distributed Formation Control. *Unmanned Systems*, 12(1), 2023.

[2] F. Chaumette and S. Hutchinson. Visual servo control. I. Basic approaches. *IEEE Robotics and Automation Magazine*, 13(4), 2006.

[3] W. Ren and R. W. Beard. *Distributed consensus in multi-vehicle cooperative control*. Springer, 2008.

[4] M. R. Rosa, A. Berkel, and B. Jayawardhana. Image-Based Visual Relative Information for Distributed Rigid Formation Control in 3D Space. *IEEE Control Systems Letters*, 8:658–663, 2024.

[5] K. Fathian, N. R. Gans, W. Z. Krawcewicz, and D. I. Rachinskii. Regular Polygon Formations with Fixed Size and Cyclic Sensing Constraint. *IEEE Transactions on Automatic Control*, 64(12), 2019.

[6] M. Bastourous, J. Al-Tuwayyij, F. Guerin, and F. Guinand. Image based visual servoing for multi aerial robots formation. *Proc. of the Mediterranean Conf. on Control and Automation, MED*, 2020.

[7] Z. Miao, H. Zhong, J. Lin, Y. Wang, Y. Chen, and R. Fierro. Vision-Based Formation Control of Mobile Robots With FOV Constraints and Unknown Feature Depth. *IEEE Transactions on Control Systems Technology*, 29(5), 2021.

[8] J. Pomares, L. Felicetti, G. J. García, and J. L. Ramón. Spacecraft Formation Keeping and Reconfiguration Using Optimal Visual Servoing. *Journal of the Astronautical Sciences*, 71(2), 2024.

[9] F. Poiesi and A. Cavallaro. A distributed vision-based consensus model for aerial-robotic teams. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018.

[10] H. Liu, Y. Lyu, and W. Zhao. Robust visual servoing formation tracking control for quadrotor UAV team. *Aerospace Science and Technology*, 106, 2020.

[11] V. Cichella and I. Kaminer. Coordinated Vision-Based Tracking by Multiple Unmanned Vehicles. *Drones*, 7(3), 2023.

[12] L. Wang and B. Meng. Distributed adaptive image-based consensus of networked robotic manipulators without visual velocity measurements. *IET Control Theory and Applications*, 8(18), 2014.

[13] E. Montijano, E. Cristofalo, D. Zhou, M. Schwager, and C. Sagues. Vision-Based Distributed Formation Control Without an External Positioning System. *IEEE Transactions on Robotics*, 32(2), 2016.

[14] D. Hu, X. Zhao, and S. Zhang. Robust image-based coordinated control for spacecraft formation flying. *Chinese Journal of Aeronautics*, 35(9), 2022.

[15] F. Chaumette. Potential problems of stability and convergence in image-based and position-based visual servoing. In *The confluence of vision and control*, pages 66–78, London, 1998. Springer London.

[16] H. K. Khalil. *Nonlinear systems; 3rd ed.* Prentice-Hall, 2002.

[17] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987.

[18] P. I. Corke and S. A. Hutchinson. A new partitioned approach to image-based visual servo control. *IEEE Transactions on Robotics and Automation*, 17(4):507–515, 2001.