Tracking in three dimensions via multi-path branching

Alessandro Attanasio*,†, Andrea Cavagna*,†, Lorenzo Del Castello*,†, Irene Giardina*,†, Asja Jelić*,†, Stefania Melillo*,†, Leonardo Parisi*,§, Edward Shen*,†, Edmondo Silvestri*,†, Massimiliano Viale*,†

* Istituto Sistemi Complessi, Consiglio Nazionale delle Ricerche, UOS Sapienza, 00185 Rome, Italy
† Dipartimento di Fisica, Università Sapienza, 00185 Rome, Italy
§ Dipartimento di Informatica, Università Sapienza, 00198 Rome, Italy

Dipartimento di Matematica e Fisica, Università Roma Tre, 00146 Rome, Italy

(Dated: May 8, 2013)

Automated tracking of multiple moving targets is the key experimental gateway to quantitative dynamical information in fields as diverse as collective behavior in biology, turbulence, crowd and traffic control. In three dimensions tracking is hard because of optical occlusions: when two featureless targets overlap for several frames, loss of identity is likely. Occlusions are frequent in biological groups as flocks, schools and swarms, a fact that has severely limited collective animal behavior field studies in the past. Here we present a new tracking method in three dimensions able to cope with the most severe occlusions. The key strategy is to generate all possible paths running through an occlusion and to employ global optimization to select the physical trajectories. We apply the method to experimental field data of bird flocks and insect swarms, and show that it yields trajectories with negligible fragmentation. These data prove that birds collectively use equal radius turning on a topologically stable network. The excellent results for both birds and midges prove the potential applicability of our method to the most diverse experimental situations.

INTRODUCTION

In recent years there has been a growing interest in studying the motion of large groups of targets, both in two and in three dimensions: animals, humans, automotive vehicles, cells, microorganisms in field or laboratory experiments, and passive tracer particles used in complex fluid flows to study turbulence [1,2,3]. This kind of studies requires tracking, which is basically defined as the automated process of following in space and time individual objects using visual information from video sequences acquired via a synchronized and calibrated multi-camera system.

There are two main reasons for which tracking may be hard. First, when the average inter-object distance in the images is small compared to the average displacement of the objects between two time-consecutive frames, ambiguities arise when identifying individual objects in time between consecutive frames. This problem, which is present both in two and in three dimensions, is relatively mild, as it can be solved by increasing the temporal resolution of the camera system, thus reducing the objects’ displacement from frame to frame. Thanks to the rapidly developing technology of fast video cameras, this problem is no longer a serious issue in several tracking contexts.

The second problem is only present in three dimensions (3D) and it is far more serious, as it is largely hardware independent. When the average inter-objects distance in the images is small compared to their optical size, there is a high probability that two (or more) objects overlap in the image. When this happens, we have an optical occlusion, see Fig. 1(e). In the worse cases, occlusions can last several time frames. This happens if the motion of the tracked objects is highly coordinated, as in bird flocks and fish schools. Increasing the spatial resolution of the camera system, namely the pixel-count of the digital sensors and the optical quality of the lens, improves only marginally the problem: however sharp an image may be, when two objects optically overlap for several time frames it becomes very hard to keep track of their identities. While it is possible to control the density of the tracers used to seed a fluid flow, there is no control in biological groups where individuals aggregate at very high densities. In this last cases, optical occlusions make tracking very problematic. In some cases, experimental data may contain information on the object’s features, so that color- or pattern-matching strategies may be exploited to simplify the problem. This, however, is typically not the case when we must track objects belonging to large and extended groups, because to have the entire group in the camera’s field-of-view, individual objects only cover a couple of pixels, and no features are available. The hard featureless case is the one we focus on here.

Optical occlusions represent the main bottleneck in the majority of the existing tracking algorithms. As a consequence, in case of high optical density in the images, the reconstructed trajectories are highly fragmented: each time an ambiguity due to an occlusion occurs, there is a high probability that the trajectories of the objects involved in the occlusions are interrupted. Depending on what we need to measure, these interruptions may be more or less severe a problem. If we only need to sample the velocity field in a PTV – Particle Tracking Velocimetry – experiment in turbulence, having some interruptions is a minor problem, as it only decreases the effective number of sampling tracers. The situation is more problematic if we need the velocities to infer the inter-individual interaction within a group of animals:
several interruptions at any given time frame are equivalent to missing some of the individuals, hence potentially biasing the inferred interaction. But if we are measuring observables depending on the entire individual trajectories, interruptions become a real problem. For example, when we need to measure the long time diffusion properties [16], or the kinematic of turning [17] in collective animal behavior, it is of utmost importance to have the complete trajectories for all individuals in the group. Even in turbulence studies, the lack of complete trajectories can introduce serious statistical biases on some physical observables [18]. In fact, interruptions are the best case scenario when we have many optical occlusions. In the worse case scenario, when an occlusion presents, a wrong call is made on the link to the next time frame, so that the trajectory of object A crosses over into object B, creating an inexistente trajectory, see Fig. 1(e). Needless to say, the consequences of this situation on the physical and biological analysis of the data can be serious.

State-of-the-art tracking algorithms differ according to how they use the information available in the images. In the last thirty years, several tracking algorithms have been developed in the field of fluid dynamics [1–3]. In particular, the algorithm by Ouellette et al. [3], in case of ambiguities, optimizes the solution for all tracked particles locally in time. In the field of collective animal behavior, different algorithms have been developed to reconstruct the 3D-positions of individual animals in groups since the early experiments by Cullen [4]. The first time-resolved 3D-tracking of a single individual dates back to the work of Dahmen and Zeil [5], and only later Pomeroy and Heppner [6] succeeded tracking 16 birds in the field. The technical shortcomings and limited results of these initial studies have been the catalyst for the successive empirical investigations on collective phenomena such as flocking birds, swarming insects, and schooling fish. Between the most advanced tracking approaches published in the literature, Betke et al. [10] developed an algorithm based on a multi-dimensional assignment problem solved with a greedy approach. Zou et al. [11] implemented a tracking algorithm which uses a global correspondence selection scheme, and applies Gibbs sampling locally in time to reduce the problem complexity. Wu et al. [12, 13] implemented a different algorithm based on three linear assignment problems, making use of ghost objects to partially solve the problem of short-term optical occlusions. Despite these efforts, tracking in three dimensions large groups of targets for long times is still an open problem.

Here, we introduce a novel tracking algorithm which is able to resolve optical occlusions lasting up to tens of consecutive frames, and therefore to distinguish the identities of the tracked objects without creating interruptions, even when the optical density in the images is very large. The main idea of our method is the following: when an occlusion presents, we establish a redundant number of temporal links and we use these links to create all possible paths running through the occlusion, see Fig. 1(b) and (f). Many of these paths will be unphysical, but certainly the paths corresponding to the real objects will also be there. Then, the information from all cameras is assembled and global space-time optimization is used to select only the physically meaningful paths, namely those that optimize multi-camera coherence, see Fig. 1(g). The method produces negligible fragmentation even in the presence of severe optical density and frequent occlusions. We test the potential of our new method by applying it to original experimental field data of flocking birds and swarming insects. In the case of bird flocks, due to the reconstruction of the full trajectories of all birds, we are able to quantify the permanence of the structural topological network versus the variation of the orientation topological network.

RESULTS

Tracking

The first task of a tracking algorithm is the detection of the objects in the images, an operation known as image segmentation. Several approaches may be used to perform the segmentation, and the choice strongly depends on the type of objects. The second task is the stereoscopic linking of the detected objects, which consists of matching the individual objects across the images acquired by different cameras, see Fig. 2(c). Finally, the algorithm must time-match the individual objects from one frame to the next one, see Fig. 2(c), an operation known as temporal linking. Older approaches [3, 7, 8] establish the stereoscopic matches first, in order to reconstruct the 3D-positions of the objects using standard stereometric methods; temporal links are then assigned in the 3D-space. More recent algorithms [9, 10] assign the temporal links in the 2D-space of the images, and subsequently establish the stereoscopic correspondences between the reconstructed object paths in the 2D-space of the image of each camera. Our tracking methods do not depend on the particular procedure used to perform the three steps above, therefore the specific algorithms we use to perform segmentation, stereoscopic linking and temporal linking are described in online Methods. The only crucial point, as we shall discuss below, is that difficult calls must not be made in assigning the links. Rather, links must be included in a redundant way.

Standard algorithms are characterized by a one-to-one assignment of stereoscopic and temporal links, meaning that each object is linked with no more than one object in the other camera at the same time, and with no more than one object in the same camera at the following time frame. Such algorithms reconstruct the 3D-trajectories of the objects by simply percolating the assigned temporal
FIG. 1. Schematic drawings illustrating the advantage of the multi-path approach over the one-to-one linking approach, the idea behind the optimization algorithm, the exponential growth of the number of possible paths, and the recursion scheme. Panel (a) illustrates real examples of temporal links between consecutive frames (left images), and real examples of stereoscopic links across two camera views (right images). Panel (c) shows a generic partial temporal sequence of two objects A and B as seen by one camera, which overlap in the image for three frames. A standard algorithm assigns only one temporal link to each detected object, therefore it is able to reconstruct only one uninterrupted trajectory: a correct one (AA or BB) or a hybrid one (AB or BA). The same schematic example is instead solved by a multi-path approach, assigning multiple links to each object to produce all the possible paths connecting the detected objects. In this way, four uninterrupted trajectories are reconstructed, of which two are real (AA and BB) and two have hybrid object identities (AB and BA), as shown in panel (f). In panel (g), we schematically explain how the optimization “selects” the two correct paths over the four candidates proposed by the multiple link assignment in panel (f): the four possible paths in the left camera view (AA_L, BB_L, AB_L, BA_L) have to be matched with the two distinct possible paths in the right camera view (A_R, B_R). A simple score function is defined here as the number of stereoscopic links connecting the points of each 2D-path in different views, and a score matrix is built with the scores of each possible match between the paths in the left and right camera views. The optimization extracts the matches which maximize the global score: \{AA_L; A_R\} and \{BB_L; B_R\}. Panel (b) shows the possible paths of a few birds over a few consecutive frames in time extracted from a typical test event of flocking birds. In panel (c), we report the actual number of possible paths for a typical acquisition of roughly 179 birds as a function of the considered duration \(T\) (each dot refers to one interval of \(T\) frames). The dashed line represents the best exponential fit over the data. In panel (d), we show how the recursive approach breaks the acquisition lasting \(T\) frames into temporal intervals with length \(\tau_1 < T\), over which the complexity of the problem can be handled. By considering the resulting matched 2D-paths of objects as meta-objects, the procedure is iterated to match the 2D-paths of meta-objects across the different views and over time intervals of length \(\tau_2\) meta-objects. The procedure is applied recursively, until the product of the \(\tau\)'s of each iteration equals the duration of the entire acquisition, \(\tau_1 \tau_2 \ldots \tau_k = T\), so that the partial solutions retrieved at each iteration are combined into the solution of the full problem at the last \((k\text{-th})\) iteration.

links, and using the stereoscopic links to reconstruct the 3D-positions. Due to one-to-one linking, such algorithms
are not able to cope with optical occlusions, as they allow every object detected in the images to belong to only one trajectory. Let us consider the schematic example shown in Fig.1(e), which illustrates a generic partial temporal sequence of two objects \(A\) and \(B\) as seen by one camera. The two objects overlap in the image for three frames. A standard algorithm assigns only one temporal link to each detected object, therefore the points of occlusion will belong to only one uninterrupted trajectory, while the second recovered trajectory will be broken. This results in a potentially severe trajectory fragmentation. Furthermore, these algorithms assign temporal links using information which is purely local in time; as a consequence, they may lose identities of occluding objects (\(A\) and \(B\) in Fig.1(e)), so that the recovered uninterrupted trajectory may be a correct one (\(AA\) or \(BB\)) or a hybrid one (\(AB\) or \(BA\)).

We tackle this problem by using a path-branching approach. Consider the example presented in Fig.1(e), in Fig.1(f). If, instead of one-to-one, we assign multiple temporal links, we can create all possible paths running through the occlusion. In this case, there are four paths, of which two are real trajectories (\(AA\) and \(BB\)) and two are hybrid objects (\(AB\) and \(BA\)). In order to build the set of all possible paths through the segmented objects, the temporal links are simply percolated for each camera view. Fig.1(b) shows, for a real flock, all possible paths of a few birds over seven consecutive frames in time. The great advantage of this approach is that through each trajectory can flow more than one object, which is exactly what happens during occlusions. The only problem then, is how to select the correct paths.

Paths selection can be done via global optimization, namely by maximizing a score function based on stereometric camera coherence. Fig.1(g) illustrates schematically how this works. The trajectories of the two real objects, \(A\) and \(B\), partly overlap in the left (\(L\)) camera view, generating four possible paths: the correct paths indicated as \(AA_L\) and \(BB_L\), and the hybrid paths \(AB_L\) and \(BA_L\). On the other hand, the two objects are distinct in the right (\(R\)) camera view, where they appear as two independent paths, \(A_R\) and \(B_R\). Given a pair of paths in the left-right cameras (a candidate match), we define its score as the number of stereoscopic links connecting the points of each 2D path in different cameras and a score matrix is then built (see Fig.1(g)). The optimization extracts the matches which maximize the global score, namely \(\{AA_L; A_R\}\) and \(\{BB_L; B_R\}\). Note that this optimization is global in both space and time, so that the final solution is selected on the basis of a score matrix built for all the objects, and considering the possible paths over the entire length of the acquisition. More details on the optimization algorithm are given in online Methods.

Due to the assignment of multiple dynamical links, the number of all possible paths obtained by percolation of the links through the entire acquisition (whose duration is typically hundreds or thousands of frames) grows exponentially with the duration of the acquisition. This is confirmed by Fig.1(c). The complexity of the optimization problem, therefore, also grows exponentially. The number of all possible paths \(\mathcal{N}\) can be estimated as

\[
\mathcal{N} = Ne^{\beta T},
\]

where \(N\) is the number of objects, \(T\) the duration of the acquisition and \(\beta\) is a bifurcation coefficient, which is zero in case of pure one-to-one temporal linking and greater than zero in case of multi-linking. Fig.1(c) shows the actual number of possible paths for a typical acquisition of roughly 179 birds as a function of the considered duration \(T\) (each dot refers to one interval of \(T\) frames). The value of \(\beta\) strongly depends on the optical density of the system. Typically, we have \(\beta \in [0.001; 0.2]\). To reduce the computational complexity, we use a recursive strategy. The acquisition is divided into temporal intervals with length \(\tau_1 < T\) – see Fig.1(d). The value of \(\tau_1\) must be small enough so that the complexity \(\mathcal{N}(\tau_1) = Ne^{\beta \tau_1}\) can be handled and the optimization described above can be performed in each time interval. Each selected path, which is a collection of objects, becomes then a meta-object, which can be linked dynamically and stereometrically to other meta-objects (paths). The procedure is then iterated as in the first step. A new time interval \(\tau_2\) is selected, where now \(\tau_2\) enumerates the number of blocks of size \(\tau_1\), i.e. of meta-frames. As shown in Fig.1(d), the procedure is applied recursively, until the product of the \(\tau\)'s of each iteration equals the duration of the entire acquisition, \(\tau_1 \tau_2 \ldots \tau_k = T\). Finally, the partial solutions retrieved at each iteration are combined into the solution of the full problem at the last (\(k\)-th) iteration.

**Test on flocks and swarms**

The implemented algorithm has been tested using newly acquired field data of flocking birds and swarming insects. A short description of the procedures we used to acquire data in the field is reported in online Methods. We analyzed 14 events of starling flocks [17], and 5 events of swarming midges, as summarized in Table I. Fig.3 shows the reconstructed trajectories for two events, the bird flock labeled 20111125a2 and the midge swarm labeled 20120702a3 – see Table I. The 2D-paths of a single bird and a single midge as captured by each camera are shown in Fig.3(a) and (e), respectively. Fig.3(b) and (f) show the corresponding reconstructed 3D-trajectories. The 3D-trajectories of the entire flock (482 birds, 365 frames) and the entire swarm (114 midges, 2200 frames) are displayed in Fig.3(c) and (g), respectively. The validity of all trajectories has been verified using stereometric coherence of the corresponding 2D-paths across camera
views, and the few trajectories lacking stereometric coherence have been discarded. In Fig. 3(d) and (h), we report the normalized length distributions of the reconstructed trajectories for eleven bird flocking events and seven midge swarming events, respectively. As shown, the distributions are strongly peaked on 1, indicating that the majority of the reconstructed trajectories are of full-length and that trajectory fragmentation is negligible. Such high-quality data have been used to perform the analysis presented in [17].

**Network permanence in flocks**

The 3D trajectories enable us to investigate quantitatively the permanence of the structural topological network of the flocks, i.e. how the angular positions between birds within a flock change over time. Let us consider the reconstructed flock at the initial time shown in Fig. 3(a). The red arrow indicates the barycenter velocity of the flock. A random bird is plotted in black, and the other birds are colored according to their angular position around this reference bird with respect to the barycenter velocity: green for the birds flying in front, light-blue for the ones on the sides, and orange for the birds flying behind the reference bird. In Fig. 3(b) we show the same flock 3.75 s later after performing a turn of 140°, with the same colors assigned previously to individual birds. The two figures reveal that the orientational topology changes with the change of the direction of motion, e.g., the birds which were in front of the reference bird (green) are now flying almost behind it. At the same time, the structural network of birds remains stable, i.e. their angular relative position does not change considerably. This is a clear signature of *equal radius* turning of the flock, namely a turn in which all birds have approximately the same radius of curvature and where trajectories cross, which was first observed in [6].

We can actually be more quantitative. We calculate the velocity vector \( \mathbf{v}_i(t) \) of each individual \( i \) along its trajectory and the barycenter velocity \( \mathbf{V}(t) = \langle \mathbf{v}(t) \rangle_i \) as the velocity vector averaged over all birds. Let us name \( \mathbf{r}_{ij} \) the position of bird \( j \) relative to bird \( i \). We are interested in the orientational change of \( \mathbf{r}_{ij} \) between the initial time \( t = 0 \) and a later time \( t \). This is described by the angle \( \theta_{ij} \) between the two vectors \( \mathbf{r}_{ij}(0) \) and \( \mathbf{r}_{ij}(t) \), which is obtained from the scalar product between the two normalized vectors,

\[
\cos \theta_{ij}(t) = \frac{\mathbf{r}_{ij}(0)}{\|\mathbf{r}_{ij}(0)\|} \cdot \frac{\mathbf{r}_{ij}(t)}{\|\mathbf{r}_{ij}(t)\|}.
\]

(2)

Averaging over all pairs of birds \( i \) and \( j \), we obtain

| Event       | Object Type | Estimated # Objects | Duration (# frames) | Frame-rate (Hz) | Duration (s) | % Trajectories with Length > 90% |
|-------------|-------------|---------------------|---------------------|----------------|--------------|----------------------------------|
| 20110208a3  | birds       | 179                 | 440                 | 80             | 5.50         | 87.0%                             |
| 20110211a1  | birds       | 551                 | 360                 | 80             | 4.50         | 90.2%                             |
| 20110217a2  | birds       | 365                 | 128                 | 80             | 1.60         | 78.6%                             |
| 20111124a1  | birds       | 120                 | 310                 | 170            | 1.82         | 99.2%                             |
| 20111125a1  | birds       | 50                  | 1000                | 170            | 5.88         | 98.6%                             |
| 20111125a2  | birds       | 482                 | 365                 | 170            | 2.15         | 84.3%                             |
| 20111201a3f1| birds       | 117                 | 500                 | 170            | 2.94         | 88.7%                             |
| 20111201a3f4| birds       | 353                 | 400                 | 170            | 2.35         | 77.6%                             |
| 20111207a1  | birds       | 110                 | 661                 | 170            | 3.89         | 97.2%                             |
| 20111214a4f1| birds       | 153                 | 500                 | 170            | 2.94         | 82.5%                             |
| 20111214a4f2| birds       | 148                 | 708                 | 170            | 4.16         | 76.2%                             |
| 20111215a1  | birds       | 381                 | 960                 | 170            | 5.65         | 72.3%                             |
| 20111220a2  | birds       | 168                 | 300                 | 170            | 1.76         | 81.2%                             |
| 20111222a1  | birds       | 60                  | 609                 | 170            | 3.58         | 89.8%                             |
| 20110620a4  | midges      | 36                  | 900                 | 170            | 5.29         | 93.9%                             |
| 20110711a1  | midges      | 37                  | 2000                | 170            | 11.76        | 97.1%                             |
| 20110909a3  | midges      | 332                 | 465                 | 170            | 2.73         | 78.2%                             |
| 20110930a2  | midges      | 115                 | 1000                | 170            | 5.88         | 77.4%                             |
| 20120917a3  | midges      | 633                 | 250                 | 170            | 1.47         | 79.9%                             |

TABLE I. Summary of the field events analyzed with the new tracking software. For each event, we indicate the object type, the estimated number of objects, the duration expressed in frames, the acquisition frame-rate, the duration expressed in seconds, and the percentage of reconstructed trajectories whose length is greater than 90% of the acquisition duration.
FIG. 2. The reconstructed trajectories for two of the analyzed events, the bird flock labeled 20111125a2 (left column) and the midge swarm labeled 201120702a3 (right column) – see Table I. The 2D-paths of a single object as captured by each camera are shown in panels (a) and (e) (birds and midges, respectively). Panels (b) and (f) show the reconstructed 3D-trajectory of the same objects. The 3D-trajectories of the entire groups of objects (482 birds and 114 midges) are displayed in panels (c) and (g), respectively. In panels (d) and (h), we report the normalized length distributions of the reconstructed trajectories for eleven bird flock events and for seven midge swarm events, respectively. Note that the distributions are strongly peaked towards 1, indicating that trajectory fragmentation is negligible.
FIG. 3. Panels (a) and (b): reconstructed bird positions for the event labeled 20110208a3 in Table I at $t = 0$ and at $t = 3.75$ s, respectively; the red arrow indicates the barycenter velocity $\mathbf{V}$; a single reference bird is plotted in black, and the other birds are colored according to their angular position around the reference bird with respect to the barycenter velocity $\mathbf{V}'$ (green for the birds in front, light-blue for the ones on the sides, and orange for the birds behind). A cluster of three birds around the reference one is extracted and shown above the plots, revealing the permanence over time of the structural topological network.

Panel (c): temporal evolution of the terms $\theta_r(t)$ (dashed line) and $\theta_V(t)$ (solid line) for two of the analyzed flocking events. Panel (d): comparative plots of the terms $\theta_V(T)$ and $\theta_r(t)$ at times $t = 1$ s and $t = T$ for all the analyzed flocking events.

$\theta_r = \langle \theta_{ij} \rangle_{i,j}$, which quantifies the angular variation of the structural network of birds over time. In a similar way, we quantify with $\cos \theta_V(t)$ the deviation over time of the barycenter velocity, defining it as the scalar product between the normalized barycenter velocity vector at the initial time $t = 0$ and at time $t$:

$$\cos \theta_V(t) = \frac{\mathbf{V}(0)}{||\mathbf{V}(0)||} \cdot \frac{\mathbf{V}(t)}{||\mathbf{V}(t)||}.$$  \hspace{1cm} (3)

Some of the events in the data set refer to turning flocks, others to flocks flying almost straight. The temporal variation of $\theta_V(t)$ reflects this difference, as it reveals whether the flock is considerably changing direction or not. In Fig. (c) we show the evolution over time of $\theta_r(t)$ and $\theta_V(t)$ for two of the analyzed flocking events, comparing a straight-flying flock with a turning one. While $\theta_V(t)$ changes significantly in time in the turning flocks, marking a clear difference with the straight-flying one, the angle $\theta_r(t)$ does not increase considerably in time for any of the two flocking event. This indicates that the structural topology of the network of birds within the flock is persistent over the observed time interval, irrespective of the variation of the orientational topology due to turns of the flock.

The relative angle $\theta_r$, evaluated at $t = 1$ s and at $t = T$ (i.e. the last frame of the event), is plotted vs. the turning angle $\theta_V(T)$ in Fig. (d) for all the analyzed flocking events. In this representation, flocks cross over from non-turning to turning in moving from left to right. The plot reveals the absence of any correlation between the variation of the orientational network topology and the structural network topology, for events differing in
terms of number of birds, duration, and curvature of the barycenter trajectory, as qualitatively shown in Fig. 3(a) and (b). If flocks performed turns under the influence of some central force, they would turn on parallel paths, with a different turning radius for each bird. Hence, the angles $\theta_r$ and $\theta_V$ would be highly correlated, being only capped by the solid-body-rotation limit ($\theta_r = \theta_V$), which is represented by the straight line in Fig. 3(d). Instead, our results prove quantitatively that flocks perform equal radius turning. This is quintessential self-organized behavior: individual birds do not turn responding to any centralized, nor external stimulus, but by means of local democratic interactions.

**DISCUSSION**

We have presented a novel 3D-tracking algorithm able to reconstruct uninterrupted trajectories for large numbers of objects, even with frequent optical occlusions. The algorithm has three main features: the possibility for all detected objects to belong to more than one trajectory (branching paths); an optimization global in space and in time; a recursive approach to reduce the computational complexity of the global optimization. We implemented and tested our algorithm on new experimental field data of flocking birds and swarming insects, showing that it is capable of reconstructing 3D-trajectories with negligible fragmentation.

How does our method performs with respect to others? An estimate (when explicit information was not published) of the number of tracked objects $N_e$ and of the average trajectory length $\langle L_e \rangle$ is shown in Fig. 4 for a number of 3D-tracking results published in the literature. The points scattered on the plot have been classified according to the field of investigation for which the respective algorithms have been developed: fluid dynamics experiments (□), biological experiments (•), and the experimental data presented in this paper (△ and ▽ for birds and midges, respectively). Given that $N_e$ and $\langle L_e \rangle$ are both valid – but qualitatively different – criteria of evaluation, to compare different methods according to $N_e$ and $\langle L_e \rangle$ we can use a multi-objective optimization approach, the simplest of which is defining the Pareto frontier [19, 20] in the $N_e - \langle L_e \rangle$ plane. We sketched with a dashed line the Pareto frontier for the plotted data-points in the 2D-space of $N_e$ and $\langle L_e \rangle$. The best tracking performance is given by the points closest to the Pareto frontier. The plot clearly shows that fluid dynamics tracking algorithms have been optimized to track large number of tracer particles for short times; on the contrary, biological experiments focused on tracking relatively small numbers of animals, even though for longer times. The data we have presented in this paper reveal the potential of our algorithm to track large groups of animals for considerably long durations, without suffering from frequent optical occlusions of the tracked animals.

We obtained excellent results both on bird flock data, and on insect swarm data, despite these systems being very different from each other: insects in a swarm fly in a very jerky manner and occlude frequently in the images, but for very short times; on the contrary, the flight of birds in a flock is highly coordinated, so that occlusions are typically very long-lasting, and can involve several birds at the time. Because of this flexibility, we believe that our tracking approach can be successfully applied to process the most diverse experimental data.

**Acknowledgements:** This work was supported by grants IIT–Seed Artswarm, ERC–StG n.257126, and AFOSR–Z809101. We acknowledge the advice of Carlo Lucibello on the tracking algorithm.

![FIG. 4. Comparison of test-cases used for several tracking algorithms, quantified in terms of estimated average trajectory length $\langle L_e \rangle$ and estimated number of tracked objects $N_e$. Both estimates are based on the published data. Fluid dynamics turbulence experiments are marked as blue squares (□); biological experiments are marked as filled black circles (●); the experiments presented here and listed in Table II are marked as red triangles (△ and ▽ for birds and midges, respectively). The numbers next to the symbols correspond to the references to the paper from which $\langle L_e \rangle$ and $N_e$ are estimated.](image_url)
ONLINE METHODS

Data-taking procedures

Two distinct but similar experimental setups are used to acquire the video sequences of flocking starlings (Stur-
nus Vulgaris) and swarming midges (Chironomidae and Cera-
topogonidae) in the field. Despite the scale difference between the two setups, they both rely on the same system of industrial high-speed cameras IDT-M5 (www.idt-piv.com, monochromatic CMOS sensor 2288 × 1728 pixels @ 170 Hz). Each camera-head is connected via a double Camera-Link cable to a processing unit, which hosts a frame-grabber and a RAID hard-drive array for data storage. The three units are remotely controlled over a local network using a laptop. For midge experiments, Schneider Xenoplan 50 mm f/2.8 Compact lenses are used together with infrared filters to reduce the infrared component of the direct sunlight hitting the background. For bird video acquisitions, Schneider Xenoplan 28 mm f/2.0 Compact and Kowa LM16XC 16 mm f/2.0 lenses are used. Exposure parameters are set according to the illumination conditions: typically, the aperture was set between f/8 and f/2, and the exposure time between 700 and 3500 µs. The frame-rate is set to 80 or 170 frames/s, according to the event.

The two setups also share the same geometry, which is schematically shown in Fig. 5(a): two cameras are separated by a baseline distance $D_{12}$, which is proportional to the working distance $Z_W$ (i.e. the distance of the flock or swarm from the cameras). We also use a third camera, placed at the shorter distance $D_{23}$ from the first one, which allow us to use the trifocal stereometric geometry – as described in the following section Stereoscopic temporal linking. In the case of birds, the mutual positions and orientations of the cameras are accurately measured on the field, as described by Cavagna et al. [21]. In this case, typically $D_{12} = 25.0$ m and $D_{23} = 2.5$ m. In the case of midges, the mutual positions and orientations are instead retrieved from a set of images of a known linear target [22]. For the midge setup, typically $D_{12} = 3.0$ m and $D_{23} = 0.25$ m. For both experiments, intrinsic camera parameters (i.e. optical distortions and aberrations) are also retrieved with a calibration routine which makes use of a set of images of a known planar target.

Image segmentation

We developed an image segmentation routine to detect the objects in the experimental images, specifically tailored to our data: irregularly shaped birds, dark on a bright background (see a sample image in Fig. 5(b)); in panel (a−I), a cropped image of the same flock reconstructed in Fig. 3(c); in panel (b−III), a crop of a few tens of birds; bright circular images of midges, back-

FIG. 5. In panel (a), the schematic drawing shows the geometry of the experimental setups used to acquire flocking starlings data and swarming midges data in the field. Two cameras are separated by a baseline distance $D_{12}$, which is proportional to the distance $Z_W$ of the target objects from the cameras. Their optical axes form the convergence angle $\alpha$. A third camera is positioned close to the first one, at a shorter distance $D_{23}$. In panel (b) we show the individual steps of the image segmentation routine applied to images of flocking birds: in panel (b−I), a crop of an original image of the same flock reconstructed in Fig. 3(c) is shown; in panel (b−III), a crop of a few tens of birds of the same original image is shown; panel (b−IV) shows the identified target objects, marked in blue, and their centroid coordinates, marked as red dots; clusters of adjacent or partly occluding object images are split using a watershed technique, as shown in panel (b−V) (flock crop) and panel (b−II) (full flock). Panel (c) illustrates an example of partly occluding objects in an image, which the watershed technique of our segmentation routine failed to split. The multi-path approach allows linking the four distinct objects detected at frame $t_0$ to the single object which represents the same four objects at the following frame $t_0 + 1$, and this one with the two distinct objects detected at the subsequent frame $t_0 + 2$. 


scattering direct sunlight at sunset, in front of a dark unsteady background.

The implemented routine computes the background of each individual image on the images within a time-window centered around the current frame. The background is subtracted, and a Gaussian denoising algorithm is used. An intensity threshold is applied to the images to extract binary pixel maps. A two-pass labeling algorithm \cite{23} finds interconnected regions on the binary maps, identifying them as target objects (marked in blue in Fig. 5(b – IV)). Using a weighted average \cite{24}, the centroid coordinates of each target object are extracted (marked as red dots in Fig. 5(b – IV)). A watershed technique permits splitting clusters of adjacent or partly occluded objects, as shown in Fig. 5(b – V) (flock crop) and Fig. 5(b – II) (full flock). Cases of unsolved occlusions are shown in Fig. 5(c). A detailed description of the entire segmentation routine is reported in Cavagna et al. \cite{21}.

**Stereoscopic and temporal linking**

The procedures we use to assign stereoscopic and temporal links share the same global optimization strategy used to match 2D-paths. The choice of the specific cost functions being optimized depends on the number of cameras used to acquire the scene, and on the individual and collective motion kinematics of the objects, for stereoscopic and temporal links, respectively.

We match the individual objects across the images acquired by three cameras optimizing globally in space (i.e., over all objects in the three images for each time instant) the function:

$$\begin{align*}
\min_{x_{ij}^T} & \sum_{ijk} f(d_{ijk}^S)x_{ijk}^S , \\
\text{with the constraints:} & \sum_{jk} x_{ijk}^S \geq 1 ,
\end{align*}$$

where $i$, $j$, $k$ are the object indices in the images of the first, second, and third camera, respectively; $d_{ijk}^S$ is the three-dimensional tensor of the stereometric distances between the $i$-th, $j$-th, and $k$-th objects; $x_{ijk}^S$ is a three-dimensional Boolean tensor of the variables of the optimization problem (true if the $i$-th, $j$-th, and $k$-th objects are matched, false if not); $f(d_{ijk}^S)$ is a cost function used to weight the stereometric distances. There are several possible definitions of the stereometric distance; we prefer to use the trifocal distance (see \cite{22}) to the sum of the epipolar distances, because the trifocal geometry presents less degeneracies. The cost function we choose is $f(d_{ijk}^S) = (d_{ijk}^S)^2$. Note that the possibility to assign multiple links per object is expressed in Eq. (5) by the non-strict equality requirement.

We follow exactly the same approach to assign temporal links. First, we choose a kinematic prediction strategy with which we predict the position at time $(t+1)$ of the $i$-th object at time $t$. We then link the individual objects in time across the frames of the photo-sequences optimizing globally in space (i.e., over all objects for each time instant) the function:

$$\begin{align*}
\min_{x_{ij}^T} & \sum_{ijk} f(d_{ijk}^T)x_{ijk}^T , \\
\text{with the constraints:} & \forall i \sum_{jk} x_{ijk}^T \geq 1 , \\
& \forall j \sum_{ki} x_{ijk}^T \geq 1 ,
\end{align*}$$

where $i$ and $j$ are the indices of the objects at time $t$ and $(t+1)$, respectively; $d_{ijk}^T$ is the two-dimensional tensor of the two-dimensional distances between the predicted position of the $i$-th object and the $j$-th candidate; $x_{ijk}^T$ is a Boolean tensor of the variables of the optimization problem (true if the $i$-th and $j$-th objects are linked, false if not); $f(d_{ijk}^T)$ is a cost function used to weight the distances. We choose $f(d_{ijk}^T) = (d_{ijk}^T)^2$ again as the cost function. As before, the possibility to assign multiple links per object is expressed in Eq. 5 by the non-strict equality requirement. Different prediction strategies are used for birds and midges. We found that a roto-scale-translation predicts well the highly-polarized flight kinematics of birds flying within a flock, and we estimate its parameters at every time instant using the Kabsch algorithm \cite{25}. A two-times constant velocity prediction approach, or a three-times constant acceleration prediction approach \cite{3} are used in case of swarming midges, whose flight directions are highly uncorrelated. The difficulty of the temporal linking problem can be quantified in terms of the ratio $\xi = \langle v \rangle \Delta t / \Delta r$, where $\langle v \rangle$ is the average speed of the objects in the 2D-image-space, $\Delta t$ is the temporal delay between consecutive acquired images, and $\Delta r$ the average inter-object 2D-distance. Typically, our $\xi$ ranges between 0.02 and 0.3.

During the evaluation of the possible candidate objects for stereoscopic and temporal linking, we set a threshold on the distances $d^S$ and $d^T$ over which we discard physically impossible matches. Both optimization problems are solved using a linear programming algorithm, for which we use the C++ APIs of IBM ILOG CPLEX Optimization Studio v12.2 \cite{20}.

**Global optimization**

The procedure we use to match 2D-paths across camera views is applied to each temporal interval $\tau_1, \tau_2, \ldots, \tau_k$. The optimization is performed using a cost function which measures the coherence between each temporal 2D-path and its possible stereoscopic matches. As cost function, we use the sum of the trifocal distances \cite{22}.
between all the points of each 2D-path in the three camera views.

Let us consider the $p$-th object and the $i$-th 2D-path in the first camera. We define the variable $\pi_p^i$ as 1 if the $p$-th object belongs to the $i$-th path, and 0 otherwise:

$$\pi_p^i = \begin{cases} 
1 & \text{if } p \in i, \\
0 & \text{if } p \not\in i.
\end{cases}$$  \hfill (8)

In the same way, we define the variables $\pi_j^i$ and $\pi_k^i$ for the second and the third camera, respectively. The cost $c_{ijk}^G$ of the possible match between the $i$-th, $j$-th, and $k$-th paths is defined as the sum of the trifocal distances $d_{pqr}^S$ between all the points belonging to each path (as used for stereoscopic linking):

$$c_{ijk}^G = \sum_{pqr} d_{pqr}^S \pi_p^i \pi_q^j \pi_r^k.$$  \hfill (9)

In case of non-existing stereoscopic links, we assign to the distance $d_{pqr}^S$ the threshold value used for stereoscopic linking. We match the 2D-paths across the images acquired by three cameras optimizing the function:

$$\min_{x_{ijk}^G} \sum_{ijk} c_{ijk}^G x_{ijk}^G,$$  \hfill (10)

where $x_{ijk}^G$ is a three-dimensional Boolean tensor of the variables of the optimization problem (true if the $i$-th, $j$-th, and $k$-th paths are matched, false if not).

The optimization problem is solved using a linear programming algorithm, for which we use once more the C++ APIs of IBM ILOG CPLEX Optimization Studio v12.2 [26].

### Cleaner

The described procedure outputs a redundant set $\Gamma_{tot}$ of 3D-trajectories, due to the presence of partially and fully cloned trajectories. This is to be expected, since the set $\Gamma_{tot}$ has been produced with the constraint that each detected object has to belong to at least one trajectory. Because of this, unreal trajectories are created in order to include those objects which appear in the view of one camera only, or objects produced by over-splitting of the segmentation routine. It is therefore necessary to “clean” the set $\Gamma_{tot}$, removing its spurious elements.

Let us name $\Gamma \subseteq \Gamma_{tot}$ a reduced set of trajectories. Let $\gamma$ be a trajectory in the set $\Gamma$, and $\alpha$ an object (i.e., a single 3D-point) of one trajectory in the set $\Gamma$. Equivalent to Eq. 5 \[5\] $\pi_\alpha^\gamma$ is 1 if the object $\alpha$ belongs to the trajectory $\gamma$, and zero otherwise. We define a score $\Psi$ for the set $\Gamma$ as:

$$\Psi(\Gamma) = \sum_{\gamma \in \Gamma} e^{(-0.5c^G_\gamma)},$$  \hfill (12)

where $c^G_\gamma$ is equivalent to $c_{ijk}^G$ defined in Eq. 9. Let us also define $\Phi_k$ as:

$$\Phi_k(\Gamma) = \sum_{\alpha} \delta \left( \sum_{\gamma \in \Gamma} \pi_\alpha^\gamma - k \right),$$  \hfill (13)

where $\delta$ is the Dirac delta function. The function $\Phi_k(\Gamma)$ expresses the number of objects in the set $\Gamma$ which belong to $k$ trajectories: $\Phi_0$ represents the number of points not belonging to any trajectory, $\Phi_1$ the number of points belonging to only one trajectory, $\Phi_2$ the number of points belonging to two trajectories, and so on.

In order to extract the best set $\Gamma$ of trajectories out of the total set $\Gamma_{tot}$, we need to maximize the score $\Psi$, to minimize $\Phi_0$, and to minimize $\Phi_k$ for $k > 1$. We define an energy $E$ for the set $\Gamma$ as:

$$E(\Gamma \mid V_0, V_\Psi) = V_\Psi \Phi_0(\Gamma) + \sum_{k=1} \left( (k-1)^2 \Phi_k(\Gamma) \right) - V_\Psi \Psi(\Gamma),$$  \hfill (14)

where $V_0$ and $V_\Psi$ are two parameters used to weight the contributions of the three terms. The space of the trajectory set $\Gamma$ is explored using a simulated annealing (low-temperature Monte Carlo simulation) in order to build the Pareto frontier [19, 20] of the three-dimensional space of the three terms of Eq. 14. The parameters $V_0$ and $V_\Psi$ are varied to further explore the Pareto frontier. Between all the possible configurations $\Gamma$ on the frontier of $E(\Gamma)$, we choose the configuration for which the number of trajectories is closer to the expected number of objects:

$$\Gamma = \arg \min \left( |\Gamma | - N_e \right),$$  \hfill (15)

with $N_e$ the average number of detected objects over all frames and all cameras, representing the best estimate for the number of expected objects.

[1] Adrian, R.J. Particle-Imaging Techniques for Experimental Fluid Mechanics. Annu. Rev. Fluid Mech. 23, 261–304 (1991).

[2] Lüthi, B., Tsinober, A., Kinzelbach, W. Lagrangian measurement of vorticity dynamics in turbulent flow. J. Fluid Mech. 528, 87 (2005).

[3] Ouellette, N.T., Xu, H., Bodenschatz, E. A quantitative study of three-dimensional Lagrangian particle tracking algorithms. Exp. Fluids 40, 301313 (2006).

[4] Cullen, J.M., Shaw, E., Baldwin, H.A. Methods for measuring the three-dimensional structure of fish schools. Animal Behaviour 13, 4, 534–536 (1965).
[5] Dahmen, H.J., Zeil, J. Recording and reconstructing three-dimensional trajectories: a versatile method for the field biologist. *Proc. R. Soc. B* **222**, 107–113 (1984).

[6] Pomeroy, H., Heppner, F. Structure of turning in airborne rock dove (Columba livia) flocks. *AuK* **109**, 256–267 (1992).

[7] Malik, N.A., Dracos, T., Papantoniou, D.A. Particle tracking velocimetry in three-dimensional flows. *Exp. Fluids* **15**, 279–294 (1993).

[8] Doh, D.H. *et al*. Single-frame (two-field image) 3-D PTV for high speed flows. *Exp. Fluids* **29**, 1 Suppl., S085–S098 (2000).

[9] Willneff, J. 3D particle tracking velocimetry based on image and object space information. *Int. Arch. Photogrammetry and Remote Sensing and Spatial Inform. Sci.* **34**, 601 (2002).

[10] Wu, Z., Hristov, N.I., Hedrick, T.L., Kunz, T.H., Betke, M. Tracking a Large Number of Objects from Multiple Views. *Proc. 12th Int. Conf. Computer Vision (IEEE)*, 1546–1553 (2009).

[11] Zou, D., Zhao, Q., Wu, H.S., Chen, Y.Q. Reconstructing 3D motion trajectories of particle swarms by global correspondence selection. *Proc. 12th Int. Conf. Computer Vision (IEEE)*, 1578–1585 (2009).

[12] Wu, H.S., Zhao, Q., Zou, D., Chen, Y.Q. Acquiring 3D Motion Trajectories of Large Numbers of Swarming Animals. *Proc. Workshop 12th Int. Conf. Computer Vision (IEEE)*, 593–600 (2009).

[13] Wu, H.S., Zhao, Q., Zou, D., Chen, Y.Q. Automated 3D trajectory measuring of large numbers of moving particles. *Optics Express* **19**, 8 (2011).

[14] Liu, Y., Li, H., Chen, Y.Q. Automatic Tracking of a Large Number of Moving Targets in 3D. *Proc. 12th European Conf. Computer Vision, Part IV*, 730–742 (2012).

[15] Ardekani *et al*. Three-dimensional tracking and behaviour monitoring of multiple fruit flies. *J. R. Soc. Interface* **10**, 78 (2012).

[16] Cavagna, A., Duarte Queirós, S.M., Giardina, I., Stefanini, F., Viale, M. Diffusion of individual birds in starling flocks. *Proc. R. Soc. B* **280**, 20122484 (2013).

[17] Attanasi *et al*. Superfluid transport of information in turning flocks of starlings. arXiv:1303.7097 [cond-mat.stat-mech] (2013).

[18] Biferale, L. *et al*. Lagrangian structure functions in turbulence: A quantitative comparison between experiment and direct numerical simulation. *Phys. Fluids* **20**, 065103 (2008).

[19] Pareto, V. *Cour d’economie politique*, Librarie Dronz, Geneve, 1964 (first edition in 1896). See also http://en.wikipedia.org/wiki/Pareto_efficiency.

[20] Messac, A., Ismail-Yahaya, A., Mattson, C.A. The normalized normal constraint method for generating the Pareto frontier. *Struct. Multidisciplinary Opt.* **25**, 2, 86–98 (2003).

[21] Cavagna, A. *et al*. The STARFLAG handbook on collective animal behaviour: Part I, empirical methods. *Animal Behaviour* **76**, 217–236 (2008).

[22] Hartley, R., Zisserman, A. *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.

[23] Shapiro, L., Stockman, G. *Computer Vision*, Prentice Hall, 2002.

[24] Parthasarathy, R. Rapid, accurate particle tracking by calculation of radial symmetry centers. *Nat. Methods* **9**, 724–726 (2012).

[25] Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica* **32**, 922 (1976).

[26] CPLEX Optimization Incorporated. *Using the CPLEX Callable Library*, Incline Village, Nevada, 1994.