# MOTION BASED CORRESPONDENCE FOR 3D TRACKING OF MULTIPLE DIM OBJECTS

*Ashok Veeraraghavan*[1], *Mandyam Srinivasan*[2], *Rama Chellappa*[1], *Emily Baird*[2] *and Richard Lamont*[2]

[1]Department of Electrical and Computer Engg.
University of Maryland
College Park MD-20742, USA
{vashok,rama}@umiacs.umd.edu

[2]Research School of Biological Sciences
Australian National University
Canberra ACT 2601, Australia
{M.Srinivasan,emily.baird,richard.lamont}@anu.edu.au

## ABSTRACT

Tracking multiple objects in a video is a demanding task that is frequently encountered in several systems such as surveillance and motion analysis. Ability to track objects in 3D requires the use of multiple cameras. While tracking multiple objects using multiples video cameras, establishing correspondence between objects in the various cameras is a non-trivial task. Specifically, when the targets are dim or are very far away from the camera, appearance cannot be used in order to establish this correspondence. Here, we propose a technique to establish correspondence across cameras using the motion features extracted from the targets, even when the relative position of the cameras is unknown. Experimental results are provided for the problem of tracking multiple bees in natural flight using two cameras. The reconstructed 3D flight paths of the bees show some interesting flight patterns.

## 1. INTRODUCTION

Tracking objects using multiple cameras has the obvious advantages of 3D reconstruction of tracks and wider field of view. Moreover, when the cameras are sufficiently far apart objects that are occluded in one camera might still be visible in the other cameras. But the use of multiple cameras requires establishing correspondence across objects seen in the various views. When there is only one object in view then this correspondence is easily established [1]. But while handling multiple targets establishing this correspondence is a non-trivial task. Moreover, if the cameras are sufficiently separated then the appearance of the same target in the different cameras will be very different and therefore cannot be used as a cue for establishing correspondence. Also, when the targets are dim (very low signal to noise ratio) or are very far away from the camera (and therefore occupy very few pixels on the image), then appearance features cannot be used for establishing correspondence. Moreover, if the targets themselves resemble each other in appearance, as in the case of tracking several bees, then using appearance information could be ineffective. Therefore, one needs to develop alternate strategies for establishing this correspondence.

This work was partially supported by the NSF-ITR Grant 0325119.

Motion information that is implicit in the individual tracks obtained in the various views is an obvious candidate. But the tracks in the various camera views are perspective projections of true 3D tracks and therefore additional constraints are necessary to match tracks. There have been several attempts to use auxiliary information about motion to constrain the matching process. [2] uses the constraint that the motion of the feet of tracked people lies on the ground plane to recover extrinsic camera parameters and then to align and match tracks obtained in the two views. [3] computes the field of view of one camera on the field of view of the other cameras, again by assuming the presence of a ground plane on which subjects walk, to obtain correspondence across views. In our approach we use a theorem concerning the projection of 3-D trajectories of a moving object on to a 2-D image stated originally in [4] and then later again in [5], to establish correspondence between motion trajectories in the various cameras.

## 2. OVERVIEW OF THE APPROACH

Images from the different cameras are initially considered separately. The dynamic background is obtained for each video sequence during each frame by assuming that the background variations are much slower than the motion of the targets. The background subtracted frames are thresholded to obtain a binary foreground mask. Connected component analysis is performed on the binary foreground mask to obtain a set of blobs representing the hypothesised position of the several targets in each frame. A simple blob tracking model based on the constant velocity model is used to track the motion of the targets in the video. Thus there are several long tracks of targets available for each camera view. We establish correspondence between the various bee tracks in the different camera views by exploiting the properties of the spatio-temporal curvature of these tracks. We note that establishing this correspondence does not require one to know the exact relative position of the cameras. Once correspondence between tracks is established, we can infer the relative position of the cameras using these correspondences. We then reconstruct the 3-D trajectories of the targets using the standard triangulation algorithm. Therefore, the algorithm is distrib-

uted, in the sense that most processing is performed locally at each camera. The central processor only takes the tracked trajectories available from each camera and reconstructs the 3-D flight paths of the bees. The entire algorithm is completely automatic with no need for any manual inputs. We have successfully used this approach for tracking several hundreds of bees in several videos across two camera views and reconstructed 3D flight paths of the bees.

## 3. PRE-PROCESSING AND TRACKING

In this section we will discuss the nature of the pre-processing and tracking algorithm that we have used. The pre-processing and tracking algorithm initially runs independently on the video sequences obtained from the different cameras. We discuss the application of tracking multiple bees in free flight using two cameras. The bees are typically 25-50 metres away from the cameras and therefore are very small dim targets.

**Background Subtraction:** Since the cameras are static, the changes in the background are essentially due to changes in the environment and the illumination conditions. We assume that these changes are much slower (low frequency) when compared to the changes due to the foreground motion of the bees. Therefore by adequate and appropriate low-pass filtering, the slowly varying background can be reliably estimated for each frame. For each pixel location in the image we compute the median of a temporal window of about 10-20 frames in order to estimate the background. We have also noticed that this estimate of the background is fairly insensitive to the width of the temporal window. The background subtracted image is then thresholded to obtain a binary foreground mask. This is followed by connected component analysis to segregate the foreground mask into distinct separate blobs. So for each frame now, we have a set of blobs at various locations which are the hypothesised pixel locations of the bees. Figure 1 shows some of these binary background subtracted frames for one of the cameras. As can be seen, the targets are very small and therefore appearance models are ineffective for establishing correspondence across views.

**Tracking by Data Association:** Once the bees in each frame have been identified as blobs, tracking these blobs through the video sequence reduces to establishing correspondence between blobs in consecutive frames. For example, let us assume that we have $5$ bee tracks that are active at frame $i$, and at frame $i+1$ the background subtraction has determined that there are $6$ bees in this frame. Tracking is essentially determining which of these $6$ bees corresponds to which of those $5$ tracks present in the previous frame and also simultaneously identifying whether new bees have entered the frame. In order to do this we assume a simple constant velocity model for the motion of each bee. Using this constant velocity model, the location of each bee in the next frame can be predicted. We assume that the probability of the bee being a distance $r$ ($r \geq 0$) pixels from this predicted location is given by an exponential distribution, i.e., $P(d(\vec{p},\vec{a}) = r) = \frac{1}{\sigma} exp(-r/\sigma)$,

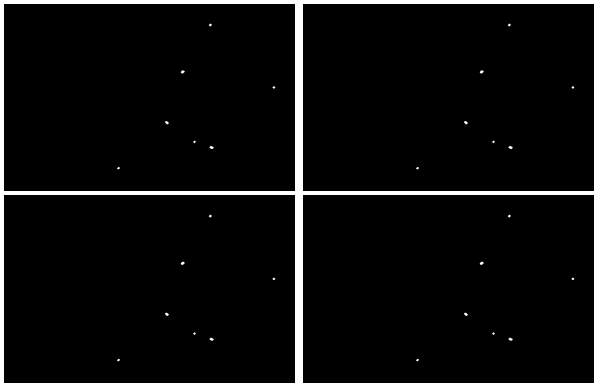

**Fig. 1**. Sample Background subtracted Frames from a tracked sequence of a several bees. Each blob represents a bee.

where $\vec{p}$ represents the predicted location of the bee and $\vec{a}$ represents the actual location of the bee and $d()$ represents the distance function. $\sigma$ is a scale parameter that represents how close to the constant velocity model the actual bee tracks are. The choice of an exponential distribution (as opposed to say gaussian) was motivated by the distribution of the velocity of the bees in several videos. Computing the maximum likelihood solution for this model is computationally expensive. If there are $N$ bee tracks and $N$ blobs in the next frame then $N!$ configurations have to be entertained.

We note that since the probability density function is exponential in the distance between the predicted and the actual location of the bee, the blobs that are far away from the predicted location of the bee are very unlikely to be associated with this bee track. This observation leads to a computationally efficient algorithm for tracking. We assume that the maximum distance between the predicted location of the bee and the actual location can only be $D_{max}$. This leads to two distinct advantages. Firstly, it reduces the computational burden. In practice we have noticed that this results in an order of magnitude decrease in computational complexity. Secondly, this leads to a very simple method for identifying new bees that enter the frame. If a certain blob is not associated with any of the bee tracks that were present in the previous frame then it is declared as a new bee that entered the field of view in the current frame. In effect this means that the probability of a bee being a distance $r$ (pixels) from its predicted location is given by,

$$P(d(\vec{p},\vec{a}) = r) = \begin{cases} \frac{1}{S} exp(-r/\sigma) & \text{if } 0 \leq r \leq D_{max} \\ 0 & \text{Otherwise} \end{cases}$$

where, $S$ is a scaling to normalize the density.

For each frame among all the various configurations (each configuration representing a set of correspondences between current tracks and blobs in the next frame),we pick the configuration with the maximum likelihood as the solution. Thus we have a simple maximum likelihood tracking algorithm based

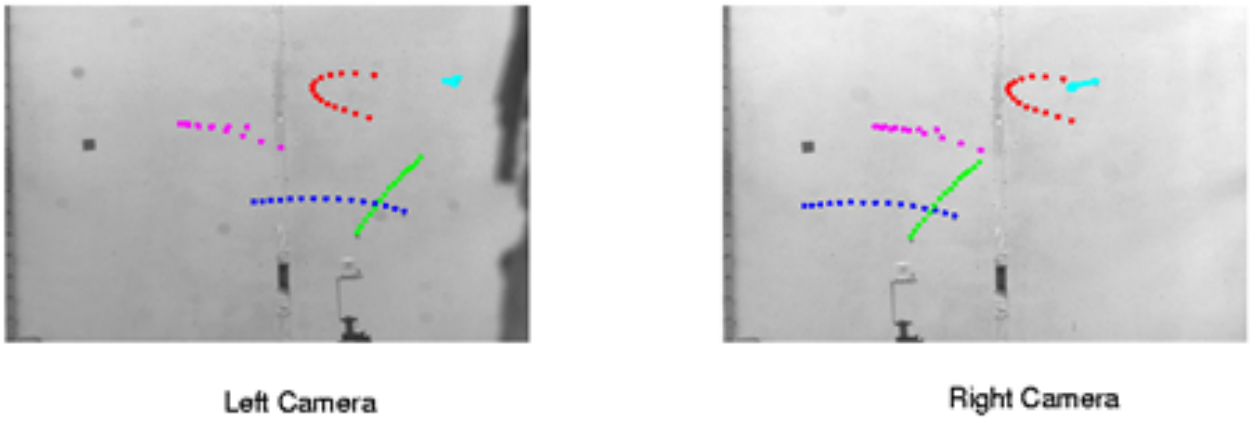


**Fig. 2**. The figures show 5 bees being tracked simultaneously in the two camera views. Each bee track is represented by a different color. Each dot represents the position of a bee in a particular frame. For simplicity, the images show the positions of the bees in the last 15 frames only.

on a constant velocity model that tracks the bees individually in each camera. Figure 2 shows multiple bees being simultaneously tracked in two cameras. The position of the bees in the last 15 frames have been marked in the image. As shown in this figure, the tracking algorithm produces a set of tracks one for each bee. Typically most of the bee tracks that we obtained were greater than 500 frames long.

## 4. CORRESPONDENCE ACROSS VIEWS

So far we have discussed how tracking can be accomplished on each of the video sequences separately for each camera. We are now left with multiple tracks for each camera. The actual correspondence across camera views (i.e., which track in view $i$ corresponds to which track in view $j$) is yet to be determined. In order to do this we exploit the following theorem from [4] which was recently restated in [5] in the context of view-invariant activity recognition.

**Theorem 1:** *The continuities and discontinuities in position, velocity and acceleration in the 3-D trajectory of a moving object are preserved in 2-D image trajectories under a continuous projection function.*

The proof of the theorem is given in [4]. Similar to [5] we consider the affine projection model for the projection of 3-D trajectories on to 2-D image trajectories. Each track of a bee is then a spatio-temporal curve given by, $r(\vec{t}) = [x(t)y(t)t]$, where $x, y$ represents the image coordinates in pixel units and $t$ represents the frame number. The velocity $v(\vec{t})$ and the acceleration $a(\vec{t})$ can be directly computed as,

$$v(\vec{t}) = r'(\vec{t}) = [x'(t)\ y'(t)\ 1] \quad (1)$$

$$a(\vec{t}) = r''(\vec{t}) = [x''(t)\ y''(t)\ 0] \quad (2)$$

The theorem states that the discontinuities in $r(\vec{t}), v(\vec{t}), a(\vec{t})$ are all conserved across the several camera views. Similar to the approach taken by [5], we identify *dynamic instants* as the maxima of the spatio-temporal curvature of these tracks. These dynamic instants are then conserved across the various camera views. Therefore we can compute the actual correspondences between tracks across views by matching the *dynamic instants* of the tracks across the views obtained by different cameras. The spatio-temporal curvature of the tracks $\kappa(t)$ is given by,

$$\kappa(t) = \frac{\|r'(t) \times r''(t)\|}{\|r'(t)\|^3} \quad (3)$$

where, '$\times$' represents the vector cross product and $\|\ \|$ represents the magnitude of a vector. Figure 3 shows the spatio-temporal curvature for corresponding tracks in two different camera views. We clearly see that the maxima in the spatio-temporal curvature (i.e., the dynamic instants ) match. We use the matching between the dynamic instants to establish the correspondences of tracks across camera views.

## 5. RECOVERING 3-D FLIGHT PATHS

Once we have established correspondence of tracks across views we now have for each frame the coordinates (in pixels) of each bee in all the cameras. For simplicity, let us consider the case of two cameras. For any given frame we have the position of each bee in both the cameras. Therefore, we can use simple triangulation to recover the 3D location of the bee for each frame. But in order to do triangulation, we need to know the internal and external camera calibration parameters. We assume that the internal calibration parameters, such as focal length are known. The relative orientation of the cameras can be recovered from the known correspondences as in [6]. We use a simple non-linear least-squares optimization to

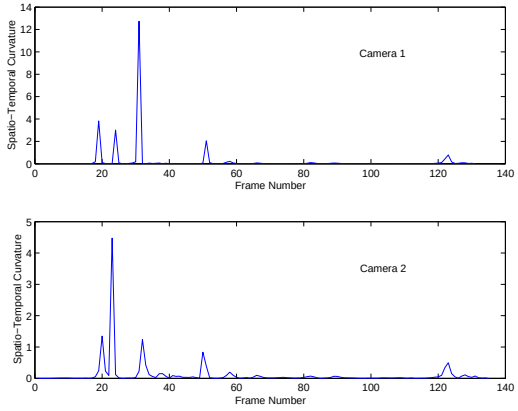


**Fig. 3**. The spatio-temporal curvature of corresponding bee tracks in two camera views. We see that the maxima of the spatio-temporal curvature match.

minimize the reprojection errors and compute the relative position and orientation of the cameras. In our set-up we also knew the distance between the two cameras approximately and used this to further constrain the optimization. Once we have recovered the external calibration (relative position and orientation of the cameras), the 3D reconstruction of flight trajectories is possible via triangulation. Precise camera alignment is difficult to achieve in a field study such as this, so the technique of auto-calibration used here is preferred.

Let us assume that the imaged positions of the bee in camera 1 and 2 are given by $(x1, y1)$ and $(x2, y2)$ respectively. We know that the straight line passing through the camera center of camera 1 and the corresponding imaged point $(x1, y1)$ on its image plane, passes through the 3D coordinates of the bee. Similarly, the straight line passing through the camera center of camera 2 and the corresponding imaged point $(x2, y2)$ on its image plane, passes through the 3D coordinates of the bee. Therefore the 3D coordinates of the bee can be computed as the point of intersection between these two lines. In practice, these two lines might not actually intersect. In such cases an approximate solution is obtained by minimizing the reprojection error. Thus we automatically recover the 3-D coordinates of the bee in each frame.

Figure 4 shows the 3D volumetric reconstruction of the flight paths of 5 different bees in a video sequence. In practice the experimenters are interested only in bees that either go all the way from the feeder to the hive or from the hive to the feeder. Since the 3D coordinates of the feeder and the hive are available, we use these 3D coordinates to restrict our attention to bees that visit both the hive and the feeder. This enables us to study the nature of the flights of bees between the feeder and the hive during various conditions.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrated a method for automatically tracking several bees in 3D using multiple cameras. A novel

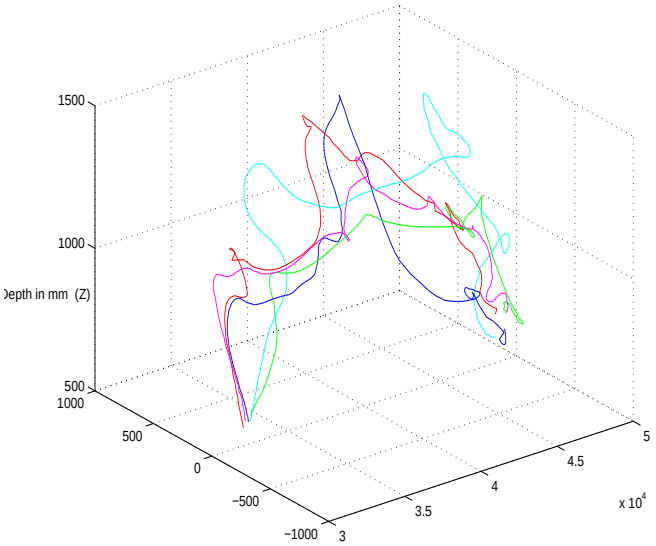


**Fig. 4**. 3-D Volumetric reconstruction of the flight path of 5 bees. This video sequence had more than 100 bee flights. We show only 5 of the reconstructions here for simplicity.

method for establishing correspondence across camera views by matching the maxima of the spatio-temporal curvature of the trajectories was presented. Experimental results were provided for several videos containing more than $10,000$ frames and consisting of a few hundred bee flights. The algorithm was used to recover the 3D trajectories of hundreds of freely flying bees.

## 7. REFERENCES

[1] Z. Yue, S. Zhou, and R. Chellappa, "Robust two-camera visual tracking with homography," *ICASSP*, 2004.

[2] C. Jaynes, "Multi-view calibration from planar motion for video surveillance," *Second IEEE Workshop on Visual Surveillance*, pp. 59–66, 1999.

[3] S. Khan, O. Javed, Z. Rasheed, and M. Shah, "Human tracking in multiple cameras," *International Conference on Computer Vision*, 2001.

[4] J.M. Rubin and W.A. Richards, "Boundaries of visual motion," *Tech. Rep. AIM-835, Massachussetes Institute of Technology, Artificial Intelligence Labaratory*, 1985.

[5] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *IJCV*, 2002.

[6] I. Ihrke, L. Ahrenberg, and M. Magnor, "External camera calibration for synchronized multi-video systems," *Journal of WSCG*, vol. 12, pp. 537–544, Jan. 2004.