Tracking Large Variable Numbers of Objects in Clutter

M. Betke[1], D.E. Hirsh[1], A. Bagchi[1], N.I. Hristov[2], N.C. Makris[3], and T.H. Kunz[2]

[1] Department of Computer Science and [2] Department of Biology, Boston University
[3] Department of Mechanical Engieering, Massachusetts Institute of Technology
{betke,dhirsh,bagchi}@cs.bu.edu, {hristov,kunz}@bu.edu, {makris}@mit.edu

Abstract

We propose statistical data association techniques for visual tracking of enormously large numbers of objects. We do not assume any prior knowledge about the numbers involved, and the objects may appear or disappear anywhere in the image frame and at any time in the sequence. Our approach combines the techniques of multitarget track initiation, recursive Bayesian tracking, clutter modeling, event analysis, and multiple hypothesis filtering. The original multiple hypothesis filter addresses an NP-hard problem and is thus not practical. We propose two cluster-based data association approaches that are linear in the number of detections and tracked objects. We applied the method to track wildlife in infrared video. We have successfully tracked hundreds of thousands of bats which were flying at high speeds and in dense formations.

1. Introduction

The object recognition and tracking scenarios that computer vision methods can interpret have become increasingly complex in the last decades due to significant research advances. Our work moves beyond the limit of what has been accomplished by formulating and solving a visual tracking task that involves an unprecedentedly large, unknown, and variable number of objects in clutter (Fig. 1). Our method solves a number of challenges:

- Initiation of the tracking process is performed automatically, without prior knowledge about the location or timing of the objects' first appearance.
- The objects of interest are automatically distinguished from objects that are not of interest and thus classified to be clutter.
- Both the number of objects to be tracked and the number of objects that are not of interest are unknown.
- Our method is able to track objects that have similar appearance.
- Previously tracked objects may not be observed in some frames due to occlusion or low signal-to-noise ratio. Tracking resumes as soon as the objects reappear.
- Our solution scales and works reliably for hundreds of thousands of objects.
- Our system tracks objects in near real time.

We have been motivated to tackle these challenges by a question that has puzzled conservation biologists for many decades – "how many bats do we have in North America?" The mid-summer population of one particularly gregarious species was "guesstimated" to consist of 150 million bats in the 1950s. Using our tracking system, we were able to make a new assessment, estimating that the current mid-summer population of this species consists of about 9 million bats [4]. Our results have been used to evaluate the ecological and economic impact of bats; for example, by preventing insect damage to cotton, each bat was shown to provide an economic service that saves farmers $0.02 per night in mid-June [7].

In addition to the computer-vision components, our system contains a rich visual interface that has attracted the attention of the United States National Park Service. The organization would like to build an installation of our system at Carlsbad Caverns, New Mexico, so that cave visitors can view the impressive flight tracks of hundreds of thousands of bats.

The core research contribution of our method is a solution strategy for the data association problem, which is the problem of determining the correspondence between previously tracked objects and current measurements. The main technical contributions of this paper can be summarized as follows:

- Two data association methods for the problem of tracking enormously large numbers of objects – a "greedy data association method" and a "cluster-based method." Worst-case complexity analysis and experiments showed that both methods are computationally efficient.

1-4244-1180-7/07/$25.00 ©2007 IEEE

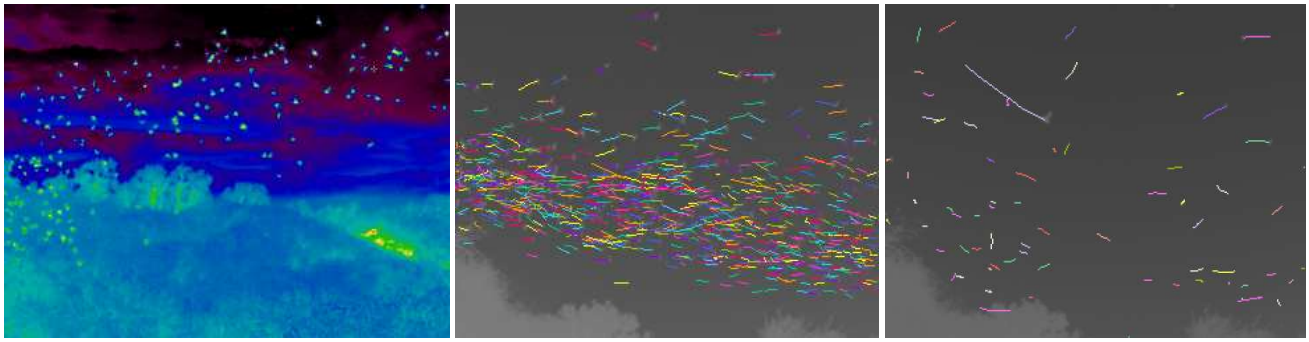

Figure 1. **Tracking wildlife.** An infrared input image shown in "false color" (left). The proposed approach was able to track bats whose flight paths were strongly correlated when they emerged from a cave (middle) and foraging bats, for which large variations in flight paths were observed (right). The image on the right was collected 1:53 hours after the image in the middle; video capturing was uninterrupted. The number of tracked bats in the two scenes were 616 and 66, respectively. For visibility, the colored tracks show the most recent track history only (2 frames).

- An event-based approach for multiple-hypotheses tracking that facilitates the correct interpretation of hundreds of newly appearing and disappearing objects in the presence of clutter and occlusion.
- A statistical framework for incorporating Bayesian recursive tracking methods, for example, the α-β filter [5], into a multi-object tracking system.
- Experiments to test the efficacy and speed of the proposed techniques. We present evidence of reliable tracking performance by (1) comparison of our results with manual "ground-truth" markings, (2) analysis of videos of the same scene taken from different view points.
- A comparison of our data association methods to a traditional data association method, called the "global nearest-neighbor algorithm" [5].
- Experiments on unprecedentedly large data sets.

2. Related Work

Our method builds on statistical data association approaches designed for tracking multiple targets in radar scans. These radar techniques were developed in the 70's and 80's [3, 17] and have created some interest in the computer vision community [8, 9, 16]. Cox [8] reviewed several data association methods, in particular, the nearest neighbor algorithm, the joint likelihood filter (JLF), the joint probabilistic data association (JPDA) algorithm, and the multiple hypothesis filter (MHF). Unfortunately, unless the visual tracking problem is relatively simple, e.g., the number of objects is small and track crossings are not common, these methods are often impractical to employ in practice. The JLF requires solving an NP-hard optimization problem and does not handle appearing or disappearing objects. The JPDA problem is NP-hard and assumes prior knowledge of the number of objects tracked. The MHF filter also attempts to solve an NP-hard problem. The exponential complexity of these data association methods is a serious disadvantage and makes them unacceptable as real-time approaches to our tracking problem.

Reid [17] provided a description of pruning techniques when he first proposed the multiple hypothesis filter. Cox and Miller [9] developed a technique to approximate the MHF and JPDA methods by applying Murty's algorithm and provided simulations to illustrate the resulting speedup for the MHF method. Rasmussen and Hager [16] extended the original JLM, JPDA, and MHF algorithms to track objects represented by complex feature combinations. Recently, Nillius et al. proposed a method to resolve multiple hypotheses via Bayesian networks [15]. This work follows the current trend to employ computer science techniques, e.g., belief propagation [10] and spatio-temporal reasoning [18], instead of, or in addition to, traditional radar techniques. This is an advance over previous work in computer vision that did not explicitly address the data association problem but rather focused on position estimation, or made the simplifying assumption that only one measurement would be near the predicted position and thus reasoning about correspondence would not be needed.

The reliability and efficiency of previously published multi-object tracking techniques were typically demonstrated for short video sequences that involved a limited number of objects, for example, a few walking people. Sometimes the application domain provided a lot of prior information about the movements to be estimated. Note that the correspondence problem is relatively easy if a low number of objects is tracked. A significant contribution of our method is that it can handle a very large number of objects, and this number does not need to be known in advance. Our method's reliable near real-time performance was shown for hundreds of objects per image frame and over long periods (hours), producing hundreds of thousands of tracks.

The analysis of wildlife video is a very challenging application area that has recently found considerable interest in the computer vision community (e.g., [2, 13, 19]).

Research in computer science, in particular, robotics and networking, has a tradition of using biological systems as sources of inspiration. The goal of wildlife image analysis, however, is to apply and extend computer science research to the study of biology [2] to better understand social insects, foraging and flocking behaviors, bird flight, colony censusing, and other movement patterns of wild life. There are also applications involving laboratory animals, for example, the study of the navigation behavior of drug-treated rats [14].

3. Methods

Our methods for solving the problem of tracking large variable numbers of objects in clutter include (1) an object detection method (Section 3.1), (2) the use of recursive Bayesian filters (Section 3.2), an event-based approach for multi-object tracking in clutter (Section 3.3), and two solutions for the data association problem (Section 3.5). A flowchart of our system is given in Fig. 2.

3.1. Detection Method

Because the motivation of our work was to census bats, which are active at night, we have worked with thermal cameras and developed a method that detects bats in infrared video. For each pixel in the field of view, our method builds a Gaussian model of intensity changes using a sliding temporal window. Each model is thus dynamic and representative of the significant changes that occur throughout the analysis of hours of uninterrupted video. The intensity changes are due to the reduction of ambient temperature during the night. The Gaussian models of pixel intensities are used to build foreground and background models. We classify foreground observations as either objects of interests (e.g., bats) or clutter (tree limbs moving with the wind, tourists wandering into the field of view, warm rocks, etc). Background observations are sky, clouds in the sky, and vegetation. Both foreground and background observations are generally quite noisy due to the thermal imaging process.

We developed an adaptive filter that determines whether a pixel belongs to the background or foreground based on its current Gaussian intensity model [11]. If the intensity measurement $I(x, y, t)$ at the pixel deviates significantly from the intensity expected in the background at that time, e.g., the difference to the mean $\mu(x, y, t)$ is larger than $k = 5\%$ of the standard deviation $\sigma(x, y, t)$,

$$k\,\sigma(x, y, t) < |I(x, y, t) - \mu(x, y, t)|, \qquad (1)$$

the measurement is considered a foreground object. Classification of foreground objects of potential interest is performed by spatial analysis. For example, the warm thorax of a bat appears as a collection of pixels with a steep intensity peak and can be distinguished from the mostly flat intensity profile of a warm rock. Further classification of candidate objects is performed by temporal analysis (see below).

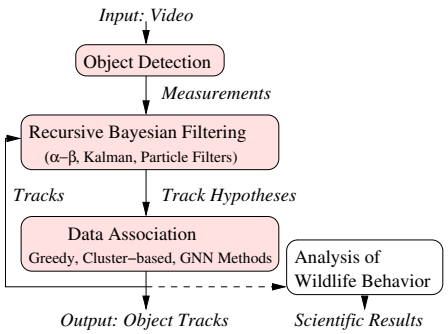

Figure 2. **System Overview.** The shaded system components are topic of this paper; the analysis of wildlife behavior has been published elsewhere [7, 4, 12].

3.2. Tracking with Recursive Bayesian Filters

Recursive Bayesian filters solve the problem of tracking a single object or feature in an image sequence recursively by estimating the state $s(t)$ of the object in the current frame t based on its state $s(t-1)$ in the previous frame and by "filtering" measurement $x(t)$ in the current frame. The Markov assumption is generally made that state $s(t)$ only depends on previous state $s(t-1)$ and a white noise sequence $u(t)$ of known distribution. We denote the estimator of $s(t)$ based on the measurements $X(t) = \{x(0), \ldots x(t)\}$ by $\hat{s}(t)$. More generally, the estimator of $s(t)$ based on the measurements $X(t') = \{x(0), \ldots x(t')\}$ is denoted by $\hat{s}(t|t')$.

In the case of a scalar state and measurement, the state sequence $s(t)$ is a Gaussian random process if the dynamics of the object can be described by a first-order Gauss-Markov process

$$s(t) = a\,s(t-1) + u(t) \qquad (2)$$

with a known parameter a, a white Gaussian noise sequence $u(t)$, and a Gaussian prior density that is independent of $u(t)$. If the observations $x(0), \ldots, x(t)$ can be described by the process

$$x(t) = s(t) + w(t), \qquad (3)$$

where $w(t)$ is a white Gaussian noise sequence, the estimator that minimizes the Bayesian mean square error $E[(s(t) - \hat{s}(t|t'))^2]$ is the Kalman filter

$$\hat{s}(t|t) = \hat{s}(t|t-1) + K(t)\,(x(t) - \hat{s}(t|t-1)), \qquad (4)$$

where $K(t) < 1$ is a gain factor that is independent of the measured data and can be computed offline. A simplified version of the Kalman filter is the α-β filter [5]. We implemented it with a four-dimensional state vector (i.e., 2D position, 2D velocity). The horizontal and vertical position components can each be estimated by

$$\hat{s}(t|t) = \hat{s}(t|t-1) + \alpha\,(x(t) - \hat{s}(t|t-1)), \qquad (5)$$

where α is a fixed gain factor. The velocity components can be updated similarly with a gain factor β. Typical

values used in our experiments were $\alpha = 0.8$ and $\beta = 0.5$ [1]. Since multiple objects must be tracked, each object is tracked by its own α-β filter. This means if n objects are tracked in each frame, $4n$ separate update equations must be evaluated (Eq. 5).

3.3. Event-based Approach for Tracking Appearing and Disappearing Objects in Clutter

The recursive Bayesian filters described above can be used if the number of objects is known and stays constant throughout the image sequence. Our problem, however, is much more complicated; the number of objects is unknown, objects may arrive anywhere in the image at anytime, clutter is present, and objects disappear. We use an $event$-oriented approach to generate hypotheses [17]. A measurement $x(t)$ may originate from an event, such as a new object, or clutter (1), which can appear anywhere in the V-pixel image, or a tracked object (2), which is assumed to appear near its predicted position $\hat{s}_i(t|t-1)$:

$$p(x(t)|event) = \begin{cases} 1/V & (1) \\ \mathcal{N}(x(t)|\hat{s}_i(t|t-1), C_i(t)) & (2), \end{cases} \quad (6)$$

where the likelihood function $p(x(t)|\hat{s}(t|t-1))$ is modeled as a Gaussian $\mathcal{N}$ with a filter-calculated covariance matrix C_i.

Computing the probability of a cumulative event conditioned on the full history of the measurements is an NP-hard problem [17] since the number of hypotheses generally grows exponentially with time. We propose a sliding-window approach to prune the number of hypotheses that must be considered at each frame. Our method discards unlikely hypotheses by removing the tracks from further consideration that are highly likely to be clutter. Initially, our method treats newly detected objects and clutter the same way. In both cases, the measurement is deemed to have originated from a new object j. A new tracker is initialized by assigning $x(t)$ to its initial state and using an initial co-variance matrix C_j.

Our method delays the decision whether $x(t)$ is clutter or a moving object for T frames. If at time $t+T$ an observation cannot be found that is likely to have originated from the object, $x(t)$ is determined to be clutter, the tracker is terminated, and the track s_j that it produced is removed. We call objects that have been tracked for at least T frames *persistent* and objects, whose states $\hat{s}(t-1|t-1)$ were estimated in the previous frame, *active*. We distinguish them from *inactive* objects which have been tracked successfully for at least T frames, but have not been observed in the scene recently.

Our assumption that the sliding-window length T is known is mild. It is defined by a lower bound on the number of frames that an object may be present in the video. For example, for the wildlife tracking application, it may be reasonable to assume that an animal is in the field of view for at least 0.5 s.

Our method can output the tracks of the persistent objects on-line during the video processing or in batch format once the last frame of the video has been processed. The number of trackers that our method maintains at any point in time is bounded from above by VT. In practice this number is significantly smaller, for example, at most a few hundred in our wildlife application.

3.4. The Data Association Problem and its Traditional Solution

The problem of data association is to determine the correspondence between measurements and tracked objects. In our system, the data association module obtains the measurements from the detection module (Sec. 3.1) and the estimated states of the tracked objects from the selected Bayesian filter (Sec. 3.2). The data association problem arises at time t when the task is to match the set of tracked objects $\{s(t)\}_{i=1}^{n(t)}$ with the set of measurements $\{x(t)\}_{j=1}^{m(t)}$ observed in the current frame. In the easiest case, the relationship is bijective, which means all objects present are also observed and each measurement was due to a previously tracked object. This is an unrealistic scenario in our application; typically $n(t) \neq m(t)$. More realistic are surjective or injective mappings. Surjective associations occur if all measurements can be matched to a previously tracked object. In this case, no new objects were detected, but tracked objects may have disappeared. In particular, tracked object s_i may not be visible in frame t because

- object s_i left the camera's field of view,
- object s_i is temporarily missing due to a false negative detection by the feature detector, possibly caused by a low signal-to-noise ratio,
- object s_i is occluded by an object s_j, which results in a single measurement.

Injective associations occur if all previously tracked objects can be matched to the observations $\{x(t)\}_{j=1}^{n(t)}$, $n(t) \leq m(t)$. In this case, additional objects or clutter may have been detected in the current image. The additional objects may be new objects entering the camera's field of view, previously tracked objects whose signal-to-noise ratios have improved to a level that made detection in the current frame possible, or previously tracked objects that emerged from occlusion. The most general and realistic scenario is that the mapping is neither surjective nor injective, which makes the problem of finding correspondences between tracked and detected objects extremely challenging.

The approach traditionally applied to the data association problem is the Hungarian algorithm, which can be used to find the measurement-to-track mappings in $O(m(t)^3)$ time [9]. The algorithm solves the weighted bipartite graph

matching problem. The nodes of the graph are on one side the measurements $\{x(t)\}$ and on the other side the predicted object states $\{\hat{s}(t|t-1)\}$. The weight of an arc between the nodes that represent the ith object and the kth measurement is the log likelihood $\log p(x_k(t)|\hat{s}_i(t|t-1))$ that $x_k(t)$ is the measurement of object $s_i(t)$. The Hungarian algorithm minimizes measurement-to-track assignment costs by maximizing the sum of the log likelihoods. If the likelihood function is assumed to be Gaussian, this means minimizing the sum of the Mahalanobis distances

$$d^2(x_k(t)) = (x_k(t) - \hat{s}_i(t|t-1))C_i^{-1}(t)(x_k(t) - \hat{s}_i(t|t-1)). \quad (7)$$

The Mahalanobis distance between predicted states and measurements generalizes the concept of the Euclidean distance between predicted object positions and measurements. Since all measurements are compared with all active tracks, the method is also called the global nearest neighbor (GNN) approach [5].

3.5. Two Data Association Approaches

We propose two approaches [1] to solve the data association problem that are based on the idea of gating [5]. A gate is a surface of constant probability density; under the assumption of a Gaussian likelihood $p(x_k(t)|\hat{s}_i(t|t-1))$ it is an ellipsoid. A gate is defined with respect to the predicted state $\hat{s}_i(t|t-1)$. Gating means pruning the number of candidate measurements so that only measurements whose likelihood lies within the gate must be considered (Fig. 3). Gating reduces, but does not avoid, the possibility of conflict situations. Conflict situations arise when, for each object, the Mahalanobis distance between its predicted state and a particular measurement (red disk in Fig. 3) is lowest among all measurements that are likely to fall within the gate of the object. Our *cluster-based approach* creates a bipartite graph of the conflicting predicted states and all measurements that are likely to be in one of the gates (red and black disks in Fig. 3). It applies the Hungarian method to compute the assignments between the states and measurements in this cluster by minimizing the sum of their Mahalanobis distances.

Our second approach for data association, the "greedy approach," also computes clusters when conflict situations arise, but does not apply the Hungarian method. Instead, an assignment process is conducted that "greedily" favors objects with long observation histories. The process is started by creating a match between the longest-observed object and its nearest measurement. This object-measurement pair is then removed from the cluster. In the next step, the second-longest observed object in the cluster is matched with its nearest measurement, and so forth until all matches are made. One might interpret the greedy approach as a "sub-optimal" approximation algorithm for the "optimal" Hungarian method. However, it is not clear that the Hungarian method's criterion to minimize the sum of the Mahalanobis distances is the "best" criterion. The characteristic of the greedy approach to favor objects whose tracks were established earlier is a conservative attribute.

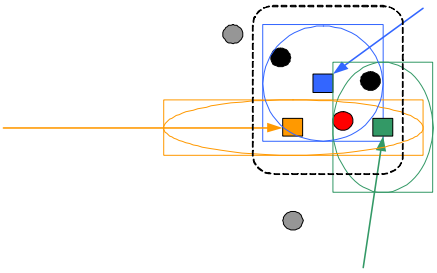

Figure 3. **Cluster-based and greedy data association approaches.** A cluster (black dotted rectangle) is created when the likelihood is high that a measurement (red disk) is due to any one of three objects (squares). The two observations (black disks) in the gates are part of the cluster, the two observations outside (gray disks) are not. Note that the gates are ellipsoidal probability density surfaces. In the greedy approach, which only maintains the object's position as its state, gates can be considered to delineate image regions (rectangular for efficiency).

We developed two versions of the cluster-based approach which differed in the definitions of the state vector. One version included and one excluded an appearance component, which was the intensity of the warmest pixel in object.

We implemented the greedy approach for the special case when the state simply represents the object's position. Then the Mahalanobis distance between the predicted state and the measurement corresponds to the Euclidean distance between the predicted position and the measurement. A gate can be defined as a region in the image and, for efficiency reasons, we used a rectangular subimage instead of an ellipse. These simplifications allowed us to predict positions and compute data associations in near real time.

3.6. Multi-Object Tracking with Data Association

We incorporated the detection, tracking, and data association methods described in the previous sections into the multi-object tracking system shown in Fig. 4. In the first step, a feature detector produces observations $x_1(t), \ldots x_{n(t)}(t)$ which may correspond to objects or clutter. Data association and Bayesian recursive filtering follow in steps 2 and 3. Step 4 concludes the processing of the current frame with an event analysis (Sec. 3.3).

Event analysis requires maintenance of the objects and their properties (states) in three datastructures, the "tracked, lost, and new object lists." State estimation (step 2) is performed only for active objects. This pruning technique ensures that the estimator does not associate current measurements with "old" objects that have long disappeared from the scene. The tracked object list contains persistent objects that are active. Objects whose measurements initiated new trackers are placed in the new object list. They stay

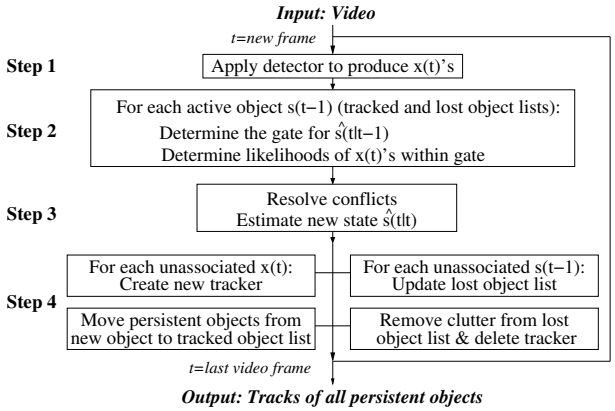

Figure 4. **Overview of the tracking system.** The first three steps must be performed sequentially; the tasks in step 4 (in four boxes) may be parallelized.

in this list as long as they remain non-persistent and active. Once such an object has been tracked for T frames, it is moved from the new object list to the tracked object list. If an object in the new object or tracked object lists could not be associated with measurements in the current frame, it is moved to the lost object list.

The lost object list contains active and inactive objects, which may be clutter. Active objects that have been lost for more than T_{lost} frames are reassigned to be inactive. Maintenance of the lost object list requires removal of any object that has not been tracked for T frames and has been lost for more than T_{lost} frames. Such non-persistent, inactive objects are deemed to be clutter.

For all versions of the system, the computational complexity of each step, except data association, is linear in the number of observations. The version that uses the greedy method is linear in the number of tracks and measurements. The cluster-based method applies the Hungarian algorithm to each cluster, which has a complexity cubic in the number of measurements in the cluster. If this number is bounded by a small constant throughout the tracking, the overall complexity is linear. Such a bound cannot be assumed for the GNN method, which is cubic in the number of all measurements and tracks and therefore computationally expensive.

4. Experiments and Results

The proposed method has been applied to census Brazilian free-tailed bats (*Tadarida brasiliensis*), one of the most gregarious mammal species in the world. Numerous censuses have been conducted at this time and produced exciting new results in the field of wildlife ecology [7, 4, 12]. In the current paper, we focus on experiments that were performed to evaluate the efficacy of the proposed method, but it should be understood that the method has been applied to track and census millions of bats.

We collected videos of flying bats near caves known to be day-time roosts of bat colonies. Due to the low-light conditions during the bats' nightly emergence, infrared thermal cameras were used (Indigo Systems' Merlin Mid Imaging Camera), which recorded digital, non-interleaved video at a rate of $\sim$60 Hz. Each video frame contained $V = 320 \times 240$ pixels of 12-bit intensity values (Fig. 1 left). The cameras are sensitive to the infrared range of 1 to 5.4 μm.

We first experimented with video collected during an emergence of a small bat colony. The video contained 9,139 frames, i.e., 2:32 min of data. We established the "ground truth" on the number of emerging bats, 7,007, by visual inspection and manual marking. We consider this ground truth to be accurate because the apparent size of the bats in the images was sufficient for detection by visual inspection and because the density and range of size of the bats in the images were small enough so that bats occluded each other only briefly. We used the version of our system with the greedy method for data association. Our system detected and tracked a total of 834,979 objects. It pruned this number with a persistence threshold of $T = 32$ frames, which corresponded to 0.53 s of video. The threshold $T_{lost} = 5$ frames was used. On average, the system maintained tracking information of 132.7 active objects at any given point in time. Each persistent object appeared on average in 92.8 frames. The output of our system, 7,056 tracked persistent objects, compares favorably with the manual estimate of 7,007 bats; it is a difference of only 0.78%. The average processing rate was 10.8 Hz. The persistence threshold T was important for detecting clutter: with $T = 2$ frames, the number of tracked objects, 8,992, was considerably higher than 7,007.

We also established the ground truth for portions of the video collected during emergence and foraging periods at a second cave (Fig. 1 middle and right). In a four-minute video during the peak emergence, our system estimated 91,790 persistent objects. By manually tracking bats in a fraction of this video and then interpolating the numbers, we estimated 88,108 bats emerged from the cave during the four-minute period. The discrepancy of 4.2% may be due to mistakes by the tracking method or the manual estimate. Persistent objects appeared in 63.5 frames on average. The algorithm maintained 831 active trackers on average.

We also evaluated the accuracy of the tracking method by recording the same emergence from two different view points. The emergence lasted 86 minutes and resulted in two video data streams, each containing 309,600 frames. The number of bats tracked with the first camera was 431,205 and with the second camera 406,572, a difference of 5.7% (see Fig. 5). This difference was due to the relatively large distance between the bats and the second camera. It is important to note that our system can only work successfully when the apparent size of a bat in an image is at least a few pixels. On-the-spot adjustment of camera po-

sition once the bats started to emerge would have been dangerous due to the steep terrain and would have disrupted the recording of tens of thousands of bats due to their breathtaking speed (ca. 10 m/s).

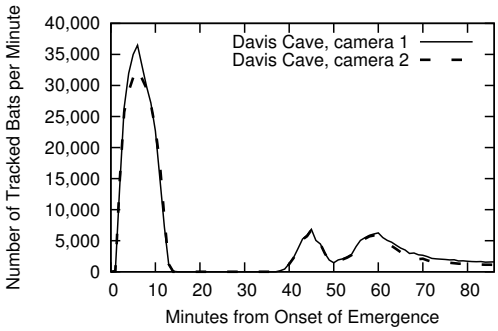

Figure 5. **Comparison of tracking results:** The number of bats emerging from a cave in Texas based on analyzing the data from two cameras with significantly different field of views.

Our method allowed successful tracking in the presence of occlusion and track crossings (Fig. 6). An experiment showed that the cluster-based approach resolved conflict situations similarly to four independent volunteers, but the greedy approach did not (80% vs. 35% agreement in 20 situations). We timed the processing speed of the greedy, cluster-based, and GNN methods on several video sequences. The average respective frame rates were 7.51 Hz, 7.10 Hz, and 0.05 Hz when a 1.7 GHz processor was used. These results agree with our worst-case complexity analysis in Section 3.6. Analysis of the computed clusters showed that the average cluster size is small. When the number of active tracks is low, almost all clusters had a size less than 6. When the number of active tracks was greater than 300, only 10% of the clusters had a size between 6 and 12, 90% were smaller.

5. Discussion and Conclusions

The experiments showed that the proposed method scales extremely well and is accurate. A level of uncertainty remains ($< 6\%$), which is quite acceptable, especially when compared to the previous state-of-the-art. Methods used by biologists in the past produced census estimates that were one order of magnitude higher than our numbers and had confidence intervals involving millions of bats.

Our method has already made an impact in the field of conservation biology by helping to census millions of bats. Our techniques have the potential to make further important contributions, for example, provide the tools to analyze the interaction of wildlife (Fig. 7) and answer urgent economical and ethical questions about the mortality of birds and bats in wind energy parks. We hope this paper motivates others in the computer vision community to continue with the trend to address complex outdoor image understanding problems and thus make some contribution to help humankind understand and conserve its environment.

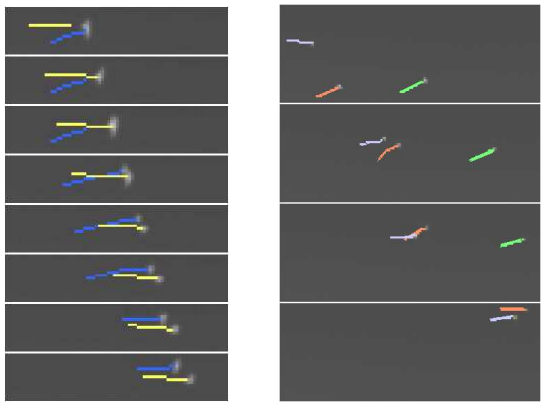

Figure 6. **Recognition of crossing tracks with and without occlusion. Left:** A horizontally flying bat (yellow track) occluded a bat (blue track) that was flying diagonally upwards and further away from the camera. Although the occlusion lasted for three frames, the occluded bat was not lost. For clarity, only small subimages of 8 consecutive frames are shown. **Right:** The path of a horizontally flying bat (blue track) was crossed by a bat (orange track) that was flying diagonally upwards. Conflict resolution was performed successfully. Only subimages of the most relevant frames (9390, 9399, 9403, and 9416) are shown.

Our current work can be extended in the following ways. First, the trade-off between reliability and efficiency should be explored further for the different filter and data-association versions of the proposed system. New cost functions may be developed. Second, additional object properties could be incorporated into the definition of a state, such as periodicity of movement, which would be useful for video surveillance of walking people as well as flying animals. Information about the appearance of an object could also be added. This is not straightforward in some wildlife scenarios, e.g., flying bats or birds look very much alike. Finally, although our method handles occlusion well, it would be interesting to explore other ideas of linking track fragments (e.g., [6, 19, 20]), which would be most important for scenes with high object density.

Acknowledgments

This study was supported by grants from the National Science Foundation (EIA-ITR 0326483 and DBI 9808396) and the Center for Ecology and Conservation Biology at Boston University. We thank Lisa Premerlani, Shuang Tang, and Marianne Procopio for data analysis and implementation support, and Edward Lee, Jonathan Reichard, and Luise Allen for data collection.

References

[1] A. Bagchi. A cluster-based data association technique for expedited multi target tracking. Master's thesis, Department of Computer Science, Boston University, Sept. 2006.

[2] T. Balch, Z. Khan, and M. Veloso. Automatically tracking and analyzing the behavior of live insect colonies. In *Proc. of the Fifth International Conference on Autonomous Agents*, pages 521–528, Montreal, Canada, 2001.

[3] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, Inc., 1988.

[4] M. Betke, D. Hirsh, N. Makris, G. McCracken, M. Procopio, N. Hristov, S. Teng, A. Bagchi, J. Reichard, J. Horn, S. Crampton, and T. Kunz. Thermal imaging reveals significantly smaller Brazilian free-tailed bat colonies than previously estimated. *Submitted to J. Mammalogy*, 2007.

[5] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House, 1999.

[6] M. Chan, A. Hoogs, R. Bhotika, A. Perera, J. Schmiederer, and G. Doretto. Joint recognition of complex events and track matching. In *CVPR*, pages 1615–1622, 2006.

[7] C. J. Cleveland, M. Betke, P. Federico, J. D. Frank, T. G. Hallam, J. Horn, J. D. L. Jr., G. F. McCracken, R. A. Medellín, A. Moreno-Valdez, C. G. Sansone, J. K. Westbrook, and T. H. Kunz. Economic value of the pest control service provided by Brazilian free-tailed bats in south-central Texas. *Frontiers in Ecology and the Environment*, 4(5):238–248, June 2006.

[8] I. J. Cox. A review of statistical data association techniques for motion correspondence. *Int J Comput Vis*, 10(1):53–66, 1993.

[9] I. J. Cox and M. L. Miller. On finding ranked assignments with application to multitarget tracking and motion correspondence. *IEEE Trans Aerosp Electron Syst*, 31:486–489, 1995.

[10] W. Du and J. Piater. Multi-view tracking using sequential belief propagation. In *ACCV*, pages 684–693, 2006.

[11] D. Hirsh. Evaluation of computer vision methods for analyzing infrared thermal video and censusing Brazilian free-tailed bats, May 2004. BA thesis, Department of Computer Science, Boston University.

[12] N. Hristov, M. Betke, D. Hirsh, A. Bagchi, and T. Kunz. Seasonal variation in colony size of Brazilian free-tailed bats at Carlsbad Caverns using thermal imaging. *Submitted*, 2007.

[13] Z. Khan, T. Balch, and F. Dellaert. A Rao-Blackwellized particle filter for eigentracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 980–986, 2004.

[14] T. V. Mukhina, S. O. Bachurin, N. N. Lermontova, and N. S. Zefirov. Versatile computerized system for tracking and analysis of water maze tests. *Behavior Research Methods, Instruments, & Computers*, 33(3):371–380, 2001.

[15] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking - linking identities using Bayesian network inference. In *CVPR*, pages 2187–2194, 2006.

[16] C. Rasmussen and G. D. Hager. Probabilistic data association methods for tracking complex visual objects. *IEEE Trans Pattern Anal Mach Intell*, 23(6):560–576, 2001.

[17] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Trans Automat Contr*, AC-24(6):843–854, 1979.

[18] D. Thirde, M. Borg, et al. Visual surveillance for aircraft activity monitoring. In *VSPETS*, pages 255–262, 2005.

[19] D. Tweed and A. Calway. Tracking many objects using subordinated condensation. In *Proc. of the British Machine Vision Conference*, pages 283–292, Cardiff, UK, 2002.

[20] T. Yang, S.Li, Q. Pan, and J. Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *CVPR*, pages 970–975, 2005.

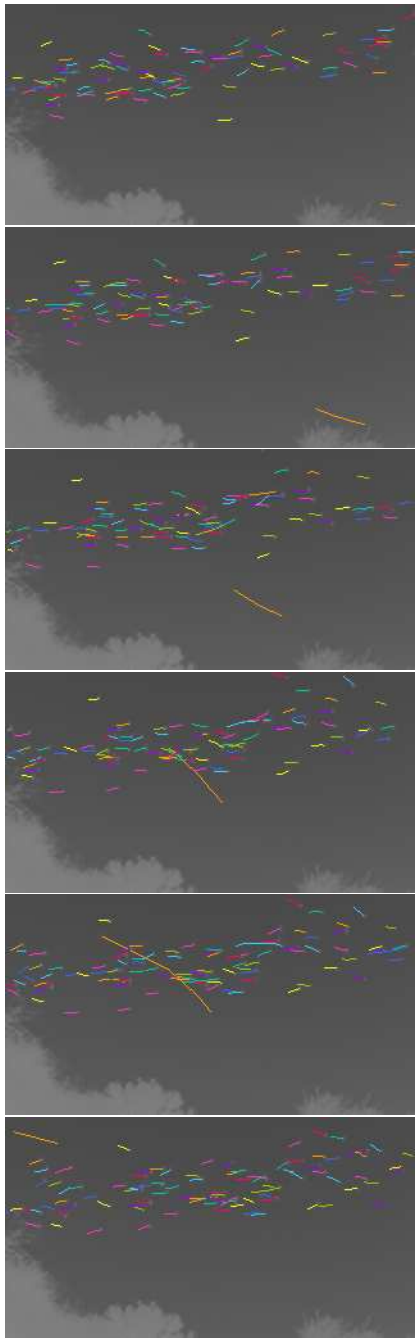

Figure 7. **Observing behavior of wildlife:** While hunting for bats, an owl (long orange track) is flying into the column of bats. Frames 9695, 9699, 9704, 9709, 9712, and 9719 are shown.