Tracking Down Under: Following the Satin Bowerbird

Aniruddha Kembhavi[†*], Ryan Farrell[*], Yuancheng Luo, David Jacobs, Ramani Duraiswami, and Larry S. Davis
†Department of Electrical Engineering, Department of Computer Science
University of Maryland, College Park, MD 20742
{anikem,yluo1}@umd.edu, {farrell,lsd}@cs.umd.edu, {djacobs,ramani}@umiacs.umd.edu

Abstract

Sociobiologists collect huge volumes of video to study animal behavior (our collaborators work with 30,000 hours of video). The scale of these datasets demands the development of automated video analysis tools. Detecting and tracking animals is a critical first step in this process. However, off-the-shelf methods prove incapable of handling videos characterized by poor quality, drastic illumination changes, non-stationary scenery and foreground objects that become motionless for long stretches of time. We improve on existing approaches by taking advantage of specific aspects of this problem: by using information from the entire video we are able to find animals that become motionless for long intervals of time; we make robust decisions based on regional features; for different parts of the image, we tailor the selection of model features, choosing the features most helpful in differentiating the target animal from the background in that part of the image. We evaluate our method, achieving almost 83% tracking accuracy on a more than 200,000 frame dataset of Satin Bowerbird courtship videos.

1. Introduction

Sociobiology seeks to understand the social behaviors of a given species by considering the evolutionary advantages these behaviors may have. To observe these social behaviors in their natural setting, biologists conduct a substantial portion of their research in the field, recording observations on videotapes. While fieldwork is very demanding, videotape analysis is truly exhausting. The corpus of video footage must be viewed in its entirety, during which time copious notes and qualitative observations are taken. Our collaborators add more than 2000 hours of video annually to a growing total of more than 30,000 hours. They desperately need computational video analysis tools.

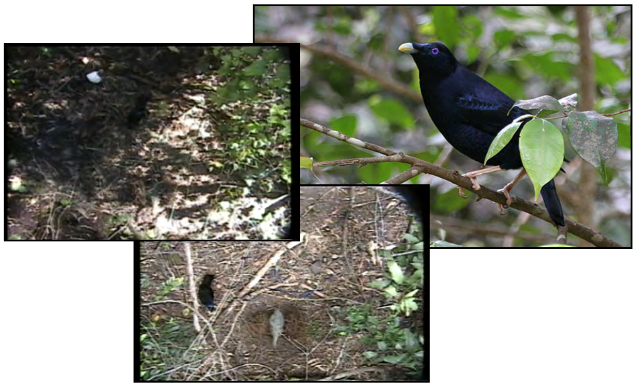


Figure 1. **Satin Bowerbird (*Ptilonorhynchus violaceus*).** A perched male Satin Bowerbird (above right [10]) and two frames taken from overhead courtship videos.

The approach we have developed addresses the challenges inherent in detecting and tracking animals in their native outdoor habitats. Characteristics of these field observation videos include: poor image quality; drastic illumination changes, some rapid due to varying cloud-cover overhead, others slow and spatial due to shadows cast by the rising sun; targets that are motionless for long stretches of time; and non-stationary background, such as vegetation swaying in the wind and also ground clutter kicked or shifted around by the animals being observed. Conventional computer vision techniques are not yet able to handle all of these challenges simultaneously.

Since our goal is to make the Biologist's video analysis much easier, there are several advantages in our favor. First, the video analysis will take place offline. This enables us to utilize all the information in the video's entire space-time volume. Second, we know a priori how many target objects need to be tracked. Third, domain-specific information about the target's appearance is available in the form of a coarse target model.

Our framework leverages these advantages and overcomes many of these problems. Our main contribution is a staged approach for target detection. We first use spatio-temporal volumes to isolate potential target regions. Our al-

[*] These authors contributed equally.

gorithm then combines target-specific information with local scene features to tailor individual models for different parts of the scene. Emphasis is thus given to those features which locally distinguish the target of interest.

We demonstrate our framework on an extensive data set of 24 videos comprising a total of more than 200,000 frames where we achieve 82.89% tracking accuracy. These videos contain courtships of the Satin Bowerbird (*Ptilonorhynchus violaceus*) and were collected by our collaborators, Jean-François Savard and Gerald Borgia.

Researchers in Prof. Borgia's lab study sexual selection (how various traits and behaviors influence mating success) in various species of the Bowerbird family [2, 13] , generally found in Australia, New Zealand and Southeast Asia. Male Bowerbirds attract mates by constructing a bower, a structure built from sticks and twigs, and decorating the surrounding area. Females visit several bowers before choosing a mating partner and returning to his bower. In part, because both courtship and mating occurs at this known location, Bowerbirds are a particularly good bird in which to study sexual selection. Of particular interest are the adjustments made by the male during courtship in response to the female. Their most recent study [16] evaluates how the male modulates his display (measured as distance from the female) based on the response cues given by a robotic female. An early prototype of our system was very valuable in facilitating the spatial tracking of the courting male, greatly reducing the days of work that would be required for manually tracking so many frames.

2. Related Work

The first step towards achieving the biologist's objectives is to accurately track the animals they are observing. While traditionally done by hand, our goal is to automate the tracking process. A typical method used in computer vision to find and track subjects moving within a scene is background subtraction. A sample of representative work includes algorithms based on Gaussian mixture models (Stauffer and Grimson [17]), non-parametric models (Elgammal *et al.* [4]), and local binary patterns (Heikkilä and Pietikäinen [6]). Typically, background subtraction algorithms are designed for online and sometimes even real-time analysis. These constraints are unnecessary for our purposes, hence affording the flexibility to use all available temporal information in a video, not just information from the recent past.

Recent work by Parag *et al.* [12] takes a similar approach to background modeling, selecting distinctive features on a pixel-by-pixel basis. A crucial advantage of our technique, however, is that we not only pick features that are distinctive for a given location in the scene, we choose the features which most effectively differentiate the target object of interest from that part of the scene.

While many effective background subtraction approaches have been and continue to be proposed, to our knowledge, they all encounter difficulties in handling all of the issues of natural outdoor environments such as those in our dataset. The general approach to dealing with background changes such as varying global illumination is to allow the model to evolve, discounting evidence from the more distant past in favor of that just observed. The primary difficulty with this method stems from its inability to simultaneously handle foreground objects that become stationary for some period of time (*e.g.* a sleeping person [18]), instead absorbing them into the background.

Efforts have been made to provide tools in support of field research. HCI researchers have recently built digital tools that allow biologists to integrate various observations and recordings while in the field [20]. In searching for the Ivory-billed Woodpecker, various teams have successfully employed semi-supervised sound analysis software to analyze the large volumes of recordings [5, 7] obtained in the field. However, there remains a need for automated tools capable of analyzing video recordings in natural outdoor environments.

We are aware of at least two projects that have previously focused on tracking animals. The Biotracking project at Georgia Tech's Borg Lab has conducted extensive research on multi-target tracking of ants [8] and bees [11] and also tracking larger animals such as rhesus monkey [9]. The SmartVivarium project at UCSD's Computer Vision Lab has investigated techniques for tracking and behavior analysis of rodents [1]. Their research also includes closely related work on supervised learning of object boundaries [3]. However, in these experiments the animals were observed in captivity, generally under laboratory conditions. While [9] used Stauffer and Grimson's background modeling technique, we have found this method to work very poorly in the Bowerbird courtship videos.

3. Our Approach

Our approach has three major phases: initial pixel classification, pixelwise background model selection and evaluation/final classification. In the first phase, the biologist provides a coarse initial model of the target (a male Bowerbird in our case) that he/she wishes to track throughout the video. This model is used to segment each frame of the video, extracting possible target pixels (in reality some target, some background), ideally leaving behind a set of only background pixels [1] . Here, we use information from all previous and all future frames of the video to take decisions (as opposed to just a few frames from the past). This helps us overcome the problem of the Bowerbird often being stationary for hundreds, even thousands of frames at a time.

[1] We define background pixels to be all those pixels that are not part of the target indicated by the biologist.

A key characteristic of unconstrained outdoor videos is the variation of the background scene, both from video to video as well as from one part of the image to another. Our second phase accounts for this. Here, we use the sets of background and target pixels and Principal Component Analysis (PCA) on a bag of features, to choose different features at different locations in the image, which can be used to build robust models. Our bag of features includes some that incorporate neighborhood information.

In the third phase, we use non-parametric Kernel Density Estimation (KDE) to build a background model for each individual image location (pixel). We then evaluate this pixel's value over all frames in the video, determining the probability in each frame that the pixel belongs to this model. We explain these three phases in greater detail in the following subsections.

3.1. Initial Pixel Classification

Many of the videos in our dataset are affected by drastic changes in global illumination. These are caused by varying levels of cloud cover and by sunlight filtering through the canopy and foliage above. The automatic gain control setting on the camera also produces sudden global changes in the color and brightness of the video. To deal with such global illumination changes, we transform every image from RGB color space into a one dimensional rank-ordered space, equivalent to performing histogram equalization on the grayscale image. The rank feature space assumes that the feature distribution changes very little, instead just shifting due to a change in the overall illumination. It disregards the absolute brightness of a pixel in the scene, rather considering only its value relative to all the pixels in the image. It is invariant to multiplicative and additive global changes and thus is largely unaffected by these effects we have observed.

In order to tune our system to track the target, we require an initialization by the biologist. Before a video is processed, the biologist analyzes a small number of frames chosen randomly, and marks out the region enclosing the Bowerbird at every frame where it is present. These pixels are used to build a smoothed histogram to serve as a coarse initial model of the target. This model is used to classify every pixel in the video into one of two sets - "potential" target pixels and "high confidence" background pixels.

At each image location, the feature that is used for this initial pixel classification is a neighborhood histogram of rank intensity. While most traditional background subtraction approaches have relied on the information contained at a single pixel to build background and target models, we rely more on neighborhood information for the following reasons. First, it reduces the chance of noisy pixels being classified as target pixels. Second, while some background pixels might closely fit the target model, neighboring pixels around it are less likely to simultaneously fit the model as well. Third, our use of regional information allows us to "see through" occluding surfaces such as branches and foliage when the target is passing beneath them.

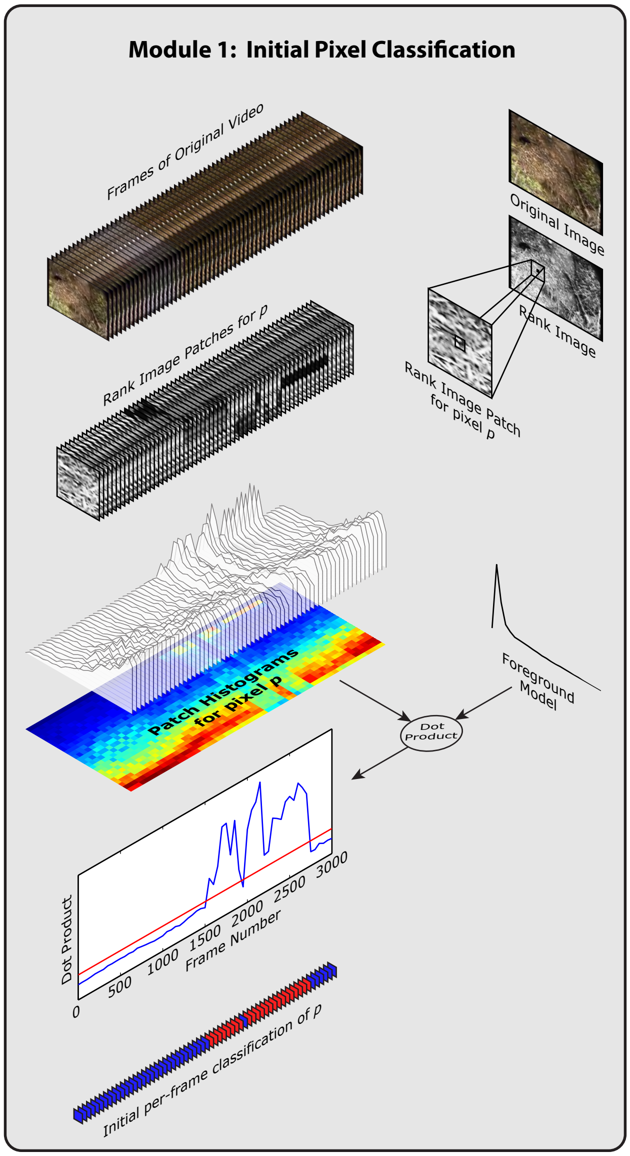


Figure 2. **Module 1: Initial Pixel Classification.**

Consider a tube of pixels $p_{ij} = \{p_{ij}^t\}$, where (i, j) denote the spatial location and $t \in \{1, 2, .., T\}$ denotes the frame number in the video sequence. We calculate a histogram of the neighboring patch at every time step to obtain a sequence of patch histograms as shown in Figure 2. Every histogram in this set is projected onto the target model to obtain a 1-D time series as shown in Figure 2. A high response at certain times indicates probable presence of the target at those times in the neighborhood of pixel (i, j). This process

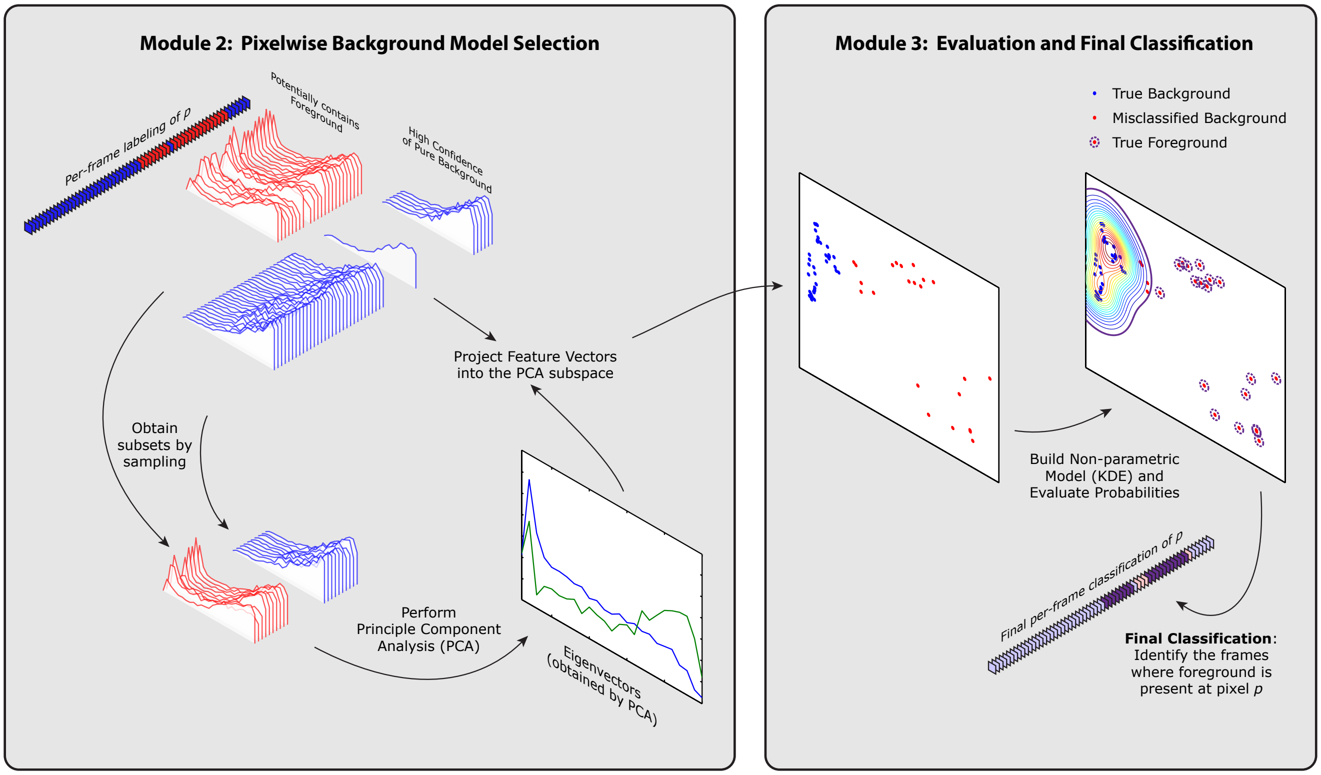

Module 2: Pixelwise Background Model Selection
Per-frame labeling of p
Potentially contains Foreground
High Confidence of Pure Background
Project Feature Vectors into the PCA subspace
Obtain subsets by sampling
Perform Principle Component Analysis (PCA)
Eigenvectors (obtained by PCA)
Module 3: Evaluation and Final Classification
True Background
Misclassified Background
True Foreground
Build Non-parametric Model (KDE) and Evaluate Probabilities
Final per-frame classification of p
Final Classification: Identify the frames where foreground is present at pixel p

Figure 3. **Module 2: Pixelwise Background Model Selection and Module 3: Evaluation and Final Classification.**

is repeated for every pixel location (i, j). In summary, we identify pixels whose neighborhoods, at times, change to more closely resemble the target model. Using all patches, both the past and the future frames, has its advantages. In the videos in our dataset, the Bowerbird jumps suddenly from one location to another, and then often waits at a single location for a lengthy period of time, sometimes even thousands of frames. Using a small quantile of the time series to model the response of background patches, we are able to easily identify frames when the bird might have visited the immediate neighborhood.

We take great care not to allow target to be mixed with the background. This hypersensitivity in initial classification reduces the number of false negative target classifications at the cost of marginally increased false positive rates. At each pixel this gives us two sets, F_{ij} and B_{ij}, consisting of the frame numbers that are respectively classified as potential target and high confidence background pixels. In essence, we obtain an *over-background-subtracted* sequence of images. We can now use the reliable set B_{ij} to build more complex and robust background models.

3.2. Pixelwise Background Model Selection

Traditional background subtraction techniques rely on a fixed set of features to build their background and foreground models (R,G,B and gray values, gradients, edges and even texture measures). However in outdoor videos, such as the ones in our dataset, the background varies greatly in different parts of the scene as well as across different videos. The additional knowledge we have about the appearance of the target object should also play an important role in determining which features would be most effective at different places in the image. For example, sometimes the bird walks over grass-filled regions, where color might be an important cue. At other times, it walks over bright sunlit areas, where a histogram of neighborhood intensities might differentiate it. For highly textured targets, a bank of oriented filters might be appropriate. We utilize information about pixels from both sets, potential target and high confidence background, and choose the most appropriate features for every pixel location from a “bag of features”.

Consider pixel p_{ij}^t. At every time step t, we concatenate multiple features to form a joint feature vector f_{ij}^t. These could include any pixel-based or neighborhood-based features. We next determine which elements of the feature vectors are most important for distinguishing target and background pixels at location (i, j) for times $t = \{1, 2, .., T\}$. The set of potential target pixels has a large number of background pixels in it, because of the conservative thresholds we choose for the initial pixel classification. This prevents us from using a standard hard classifier to label the pixels as target and background. Instead, we use PCA to project our feature vectors onto a subspace that maximizes the variance, and KDE to classify them. This probabilistic framework allows us to remove many of the falsely classified pixels from the potential target set. We only use a small sample of feature vectors from the target set F_{ij} and from the background set B_{ij} to obtain a reduced subspace, as shown in Figure 3. Projecting the entire feature set f_{ij} onto this subspace gives us the set r_{ij}, in the reduced space. The reduced dimensionality of r_{ij} helps to drastically reduce the time required to build background models.

3.3. Evaluation and Final Classification

For every pixel we build a background model using Kernel Density Estimation on our reduced feature set and evaluate probabilities at all time frames that were initially classified as potential target F_{ij}. Suitably thresholding these probabilities allows us to break down the set F_{ij} into a set of target pixels and pixels that were misclassified as target by the first module of our system. For $t \in B_{ij}$ (background), $s \in F_{ij}$ (potential target) and kernel K, we obtain:

$$P(r_{ij}^s) = \frac{1}{N\sigma_1..\sigma_d} \sum_{t \in B_{ij}} \prod_{y=1}^{d} K\left(\frac{r_{ij,y}^s - r_{ij,y}^t}{\sigma_y}\right) \qquad (1)$$

This gives us a target silhouette at every frame of the video sequence, from which we are able to calculate the centroid of the detected region at every time step. We compare these centroid locations to ground truth provided to us by the biologists, and present our results in the following section.

4. Computational Considerations

Our implementation of the framework described in Section 3 incorporates highly optimized algorithms to facilitate the processing of these large videos. We utilize Integral Histograms [14] both to generate the patch histograms used in pixel classification and to generate features for background model selection. Further, to optimize the evaluation stage, we build KDEs and determine probabilities using the Improved Fast Gauss Transform (IFGT) [15, 19]. The framework is implemented in MATLAB, with computation- and memory-intensive algorithms such as Integral Histograms and IFGT implemented in C++ and compiled as mex routines. In addition to these algorithmic optimizations, we also employed many workstations[2] (a subset of the *vnode* cluster funded through NSF Infrastructure Grant CNS 04-3313) to process multiple videos in parallel.

A key strength of our background modeling approach is the use of a large spatio-temporal window. We consider image statistics, both in a large region around a given pixel and also over a large temporal interval (the entire video). Computing statistics for each image pixel over this large temporal window requires a tremendous amount of data storage. The amount of memory needed to store a single byte per pixel over 10,000 VGA sized frames is 2.86GB. We compute feature vectors per pixel that would require about 100 or more bytes of memory per pixel (25 or more floating-point features). If this entire structure were to be in memory at one time, it would require 100s of GB of memory, rendering this task impossible for even a modern PC. We are further-constrained by the memory limits of a 32-bit version of MATLAB (only about 1.2GB are available for variables).

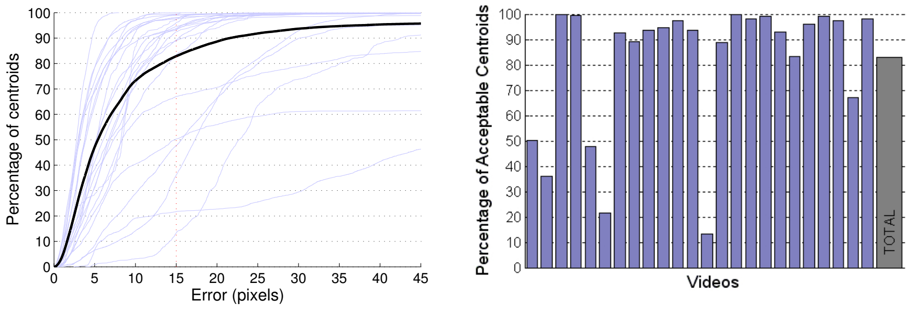
These considerations led us to implement our processing using data-decomposition as is frequently done in high performance scientific computing (though we process a given video serially on a single machine where a distributed system would run in parallel). We utilize two kinds of data-structures, *tubes* and *chunks*. Tubes refer to spatial subdivisions of the video (entire space-time volume), such that all frames for a particular subregion of the image fit simultaneously in memory. Chunks are temporal subdivisions, a contiguous set of frames in time that simultaneously fit into memory. These tubes and chunks must be created for not only the original image frames of the video but also for the large data structures that we accumulate during processing. At different stages, our algorithm requires reading in all the data, on a tube-by-tube or a chunk-by-chunk basis.

5. Experimental Evaluation

We evaluate our framework on a dataset of 24 videos comprising a total of over 200,000 frames captured at 29.97fps and a resolution of 720x480 pixels. The length of the bowerbird in the frames is roughly 90 pixels. While manually specifying the ground truth centroid for each of 200,000 frames is infeasible, we are fortunate to have what we consider a close second. In their study [16], our collaborators used an application implementing a very early prototype of our bowerbird tracking software. This version included a provision to manually correct erroneous tracking results. The biologists went through and refined the automatic tracking results for every frame in the dataset such that all centroids were then within the acceptable tolerance of 4.5cm in the real world (about 15 pixels in the image). We use these manually corrected results as our “ground truth” to quantitatively assess of our approach.

Given this, we seek to evaluate our approach using the following metrics: overall accuracy, per-video percentage within the biologist-specified tolerance, and false-positive and false-negative rates. In Figure 4(a), we present a cumulative distribution of overall accuracy. All videos are superimposed and the required tolerance is shown by the red dotted line. The overall cumulative distribution is shown by the solid line. Figure 4(b) shows the per-video percent-

[2] Workstations have dual 3.0Ghz Intel Xeon processors, 8GB RAM



(a) Overall Accuracy

(b) Percentage within Tolerance

Figure 4. **Evaluation** (a) Cumulative distribution of accuracy of every video is shown by faint lines. Overall accuracy is shown by the solid line. Red dotted line denotes the accuracy (in terms of centroid detection error) required by the biologists. (b) Per-video percentage of centroids within the biologists specified tolerance.

age of centroids within the specified tolerance. Overall, we are able to track the target within the biologists error tolerance in 82.89% of the frames in our dataset. For most of our videos this number goes beyond 90%. Having to hand label thousands of frames per video, biologists often spend days just tracking the object of interest. An accuracy of over 90% represents a very significant reduction in the time required for this process. We obtain overall false positive and false negative detection rates of 4.8% and 3.44% respectively. Our false positive detections are primarily caused by moving shadows cast by the overlying trees, and our false negative detections are primarily caused by severe occlusions by large branches and shrubs in the scene. It is often easier to manually correct false positives as compared to false negatives. The biologist can mark out a sequence of frames when the target is not present in the scene and all false positives within that range can be ignored. Figure 4(b) shows poor results for three of the videos in the dataset. These are caused by severe occlusions by large shrubs in the scene, making it very difficult to locate the target accurately.

Figure 5(a) shows a few frames from one of the videos in the database, sampled approximately every 300 frames. The stark illumination changes from one part of the video to another can be clearly seen. Figure 5(b) and Figure 5(c) show the results of the two modules in our staged approach to target detection. Some of the videos also had a very poor contrast between the target and background pixels, due to the dark shadows cast by the overlying trees, and the dark color of the male bowerbird. Figure 6 shows an example frame and detection results from one such video.

6. Conclusion

Sociobiologists collect thousands of hours of video to study animal behavior. Detecting and tracking animals is a critical first step in this process. We improve on existing techniques with a two staged approach to target detection. We use spatiotemporal volumes to isolate potential target regions and then combine target-specific information with local scene features to tailor individual models for different parts of the scene. Our collaborators used an earlier prototype of our implementation for their study, and saved a considerable amount of time that they would have spent to manually track the target in every frame. We obtain accurate tracking within the biologists' error tolerance in over 90% of frames for many of the videos in our dataset.

Acknowledgements

We wish to express our appreciation to our collaborators, Jean-François Savard and Gerald Borgia. This research was funded in part by the U.S. Government VACE program.

References

[1] K. Branson and S. Belongie. Tracking multiple mouse contours (without too many samples). In *CVPR*, 2005.

[2] S. W. Coleman, G. L. Patricelli, and G. Borgia. Variable female preferences drive complex male displays. *Nature*, 428(6984), 2004.

[3] P. Dollar, Z. Tu, and S. Belongie. Supervised learning of edges and object boundaries. In *CVPR*, 2006.

[4] A. M. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction. In *ECCV (2)*, 2000.

[5] J. W. Fitzpatrick *et al.* Ivory-billed Woodpecker (Campephilus principalis) Persists in Continental North America. *Science*, 308(5727):1460–1462, 2005.

[6] M. Heikkilä and M. Pietikäinen. A texture-based method for modeling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):657–662, 2006.

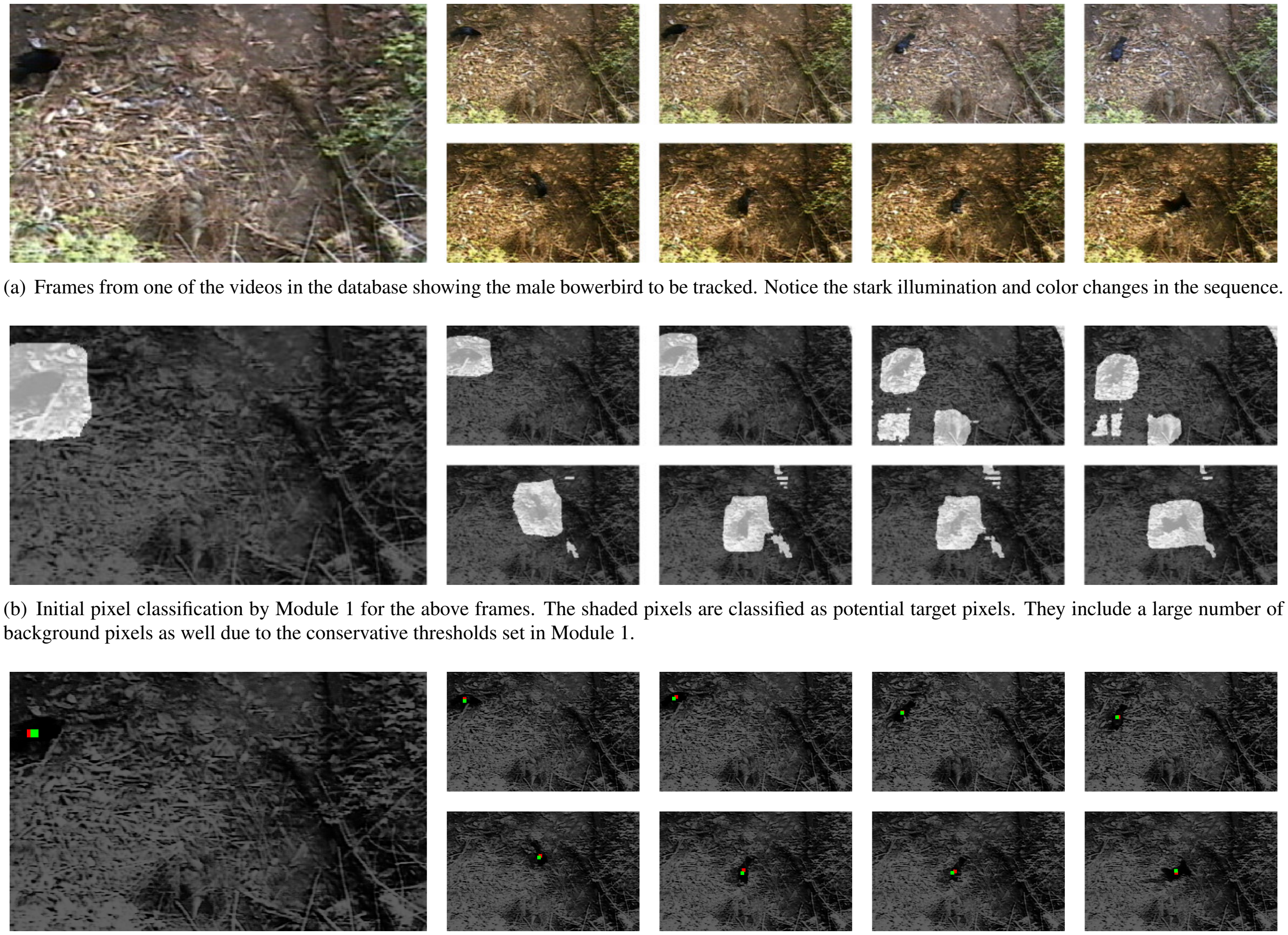
[7] G. E. Hill, D. J. Mennill, B. W. Rolek, T. L. Hicks, and K. A. Swiston. Evidence suggesting that ivory-billed woodpeckers (*Campephilus principalis*) exist in florida. *Avian Conservation and Ecology*, '06.

[8] Z. Khan, T. R. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *ECCV*, 2004.

[9] Z. Khan, R. A. Herman, K. Wallen, and T. Balch. An outdoor 3-d visual tracking system for the study of spatial navigation and memory in rhesus monkeys. *Behavior Research Methods*, 37, August 2005.

[10] C. Moores. www.charliesbirdblog.com, *used with permission*.

[11] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and inference in parametric switching linear dynamical systems. In *ICCV'05*, Washington, DC, USA.



(a) Frames from one of the videos in the database showing the male bowerbird to be tracked. Notice the stark illumination and color changes in the sequence.

(b) Initial pixel classification by Module 1 for the above frames. The shaded pixels are classified as potential target pixels. They include a large number of background pixels as well due to the conservative thresholds set in Module 1.

(c) Final results for the above frames. The detected centroid of the target is marked with a green dot, and the ground truth is shown in red.

Figure 5. **Target detection for a sequence with stark illumination changes.**

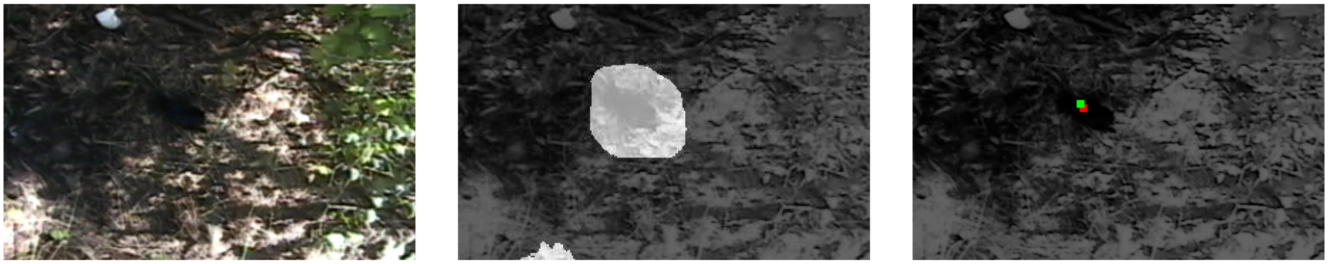


Figure 6. **Target detection for a sequence with poor contrast between target and background.**

[12] T. Parag, A. M. Elgammal, and A. Mittal. A framework for feature selection for background subtraction. In *CVPR*, 2006.

[13] G. L. Patricelli, J. A. C. Uy, G. Walsh, and G. Borgia. Sexual selection: Male displays adjusted to female's response. *Nature*, 415(6869):279–280, 2002.

[14] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. *CVPR*, 1:829–836, 2005.

[15] V. C. Raykar and R. Duraiswami. Fast optimal bandwidth selection for kernel density estimation. In *SDM*, 2006.

[16] J.-F. Savard, J. Keagy, and G. Borgia. Spatial dynamics and modulation of courtship in satin bowerbirds, *Ptilonorhynchus violaceus*. 44th annual meeting of the Animal Behavior Society, 2007.

[17] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1999.

[18] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *ICCV*, 1999.

[19] C. Yang, R. Duraiswami, and L. S. Davis. Efficient kernel machines using the improved fast gauss transform. In *NIPS*, 2004.

[20] R. Yeh, C. Liao, S. Klemmer, F. Guimbretière, B. Lee, B. Kakaradov, J. Stamberger, and A. Paepcke. Butterflynet: a mobile capture and access system for field biology research. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006.