Tracking Golden-Collared Manakins in the Wild

Anna Gostler, Nicole M. Artner and Walter G. Kropatsch
Pattern Recognition and Image Processing Group
TU Wien, Austria
E-mail: anna_g@prip.tuwien.ac.at

Leonida Fusani
Department of Cognitive Biology
University of Vienna, Austria

I. Introduction

The golden-collared manakin (*Manacus vitellinus*) is a small tropical bird, which lives in the Panama forest. The males perform elaborate, acrobatic displays to court mates [1]. During its courtship dance the male demonstrates its physical strength by jumping between saplings, producing loud wing snaps mid-flight. Mating success seems to be related to superior motor skills [2], which allow the male to execute its dance faster and more precisely. However, it is not fully clear yet how exactly the courtship dance has to be performed to impress a female. To gain more knowledge about their dance, biologists recorded the birds in the wild with high-speed cameras at 60 fps. One of the videos can be found at [1]. Manually annotating the male bird in every frame of the videos to enable analyzing their behavior is a tedious process. We propose a novel approach for automatic visual tracking of the male golden-collared manakin, combining a convolutional neural network, background subtraction and a Kalman filter.

The following properties of the videos make tracking the birds challenging:
Speed: While jumping, the bird moves very quickly (avg. 28 px per frame; avg. bounding box size: 113x95 px; frame size: 1928x1208 px).
Motion blur: Strong motion blur can make the bird hard to recognize as it loses most of its local features.
Size and shape change: The bird's bounding box changes in size and shape, e.g. when the bird opens its wings, turns, or moves away from the camera.
Occlusion: The bird can be partly or fully occluded by saplings and leaves.
Out of frame: The bird frequently leaves the camera's field of view (18 times in 78 videos).
Trajectory: The bird starts and stops abruptly and typically changes direction when starting a new jump. On average, the bird makes 3.7 (max. 10) jumps per video.
Background color: The forest is colored mostly green, yellow and brown – similar to the male bird, which has a green body, black head and a yellow neck.
Background motion: Leaves and branches move in the background. Saplings often move when the bird lands on them.

[1] *https://github.com/anna-gostler/ManakinTracker*

Tracking is made easier, however, by the static camera setup, i.e. **absence of camera motion**.

Figure 1: Close-ups of male (top row) and female (bottom row) golden-collared manakins

II. Related work

Most computer vision methods to analyze the behavior of birds worked on videos that were recorded inside custom-built arenas [3]–[7] or on videos of birds flying against the sky or distant landscapes [8], [9]. Therefore, background segmentation and tracking was not a particular concern for these studies. A method to get more precise information about the bird's position is to equip them with body markers [6]. This could potentially modify the bird's behavior during courtship, which the biologists wanted to avoid with the manakins.

In 2017, Oliva et al. [10] developed a visual tracker for golden-collared manakin males for videos that were recorded in 2016 by the same team of biologists with a different setup than the videos in this paper. Oliva's tracker detects foreground blobs using Mixture of Gaussians (MOG) [11] and finds the male manakins among these blobs based on the yellow color and saturation of their neck or, if it cannot find the bird this way,

predicts its location with a linear Kalman filter. This method relies on a distinct color difference between the background and the bird, which is not present in most of the current videos.

As we aim to track birds that strongly and abruptly change their appearance in videos recorded against a highly cluttered background, we have evaluated the top performing trackers of the VOT2016 challenge [12], which deal with visual tracking under similarly challenging conditions: TCNN [13] and C-COT [14].

Nam et al. [13] developed TCNN, a tracker that uses Convolutional Neural Networks (CNNs) arranged in a tree structure, where a CNN in a child node is a fine-tuned version of the CNN in its parent node. By keeping multiple models of the target object the tracker can handle appearance changes. The CNNs are based on a CNN pre-trained on ImageNet, but are adapted to output only two scores: a target and a background score.

Danelljan et al.'s tracker C-COT [14] is also based on a CNN pre-trained on ImageNet. C-COT extracts feature maps, which consist of the input image patch and convolutional layers, from the CNN and learns continuous convolution filters. The feature maps are convolved with the filters to obtain a continuous confidence score for every location in an area centered on the previous location of the target. This approach uses the CNN as a feature extractor. However, there are not many consistent local features (such as points and edges) in our target, mainly due to motion blur.

Both TCNN and C-COT do not model target movement, but instead search the target around its previous location, so they might not be able to follow a fast moving target such as the manakin.

III. ManakinTracker

In this paper, we propose the ManakinTracker, which

- detects moving objects with a Mixture Of Gaussians model (MOG) [11],
- decides if a candidate location visually resembles the male bird with a fine-tuned CNN,
- and estimates the location of the target using a Kalman filter [15] for frames without reliable visual cues (see Fig. 2).

A. Blob Detection

As the videos were recorded with stationary cameras, we can use a method based on background subtraction to segment the foreground. We chose Mixture Of Gaussians (MOG) because it can handle small movements in the background. For every frame, MOG generates a foreground mask from which we extract a set of moving objects, called blobs (Fig. 3). Out of these candidate blobs, we aim to select the ones that contain the target.

B. CNN architecture

To decide which candidate blobs contain the target, we use a CNN, as CNNs have shown top performance in object classification in images. Our CNN is based on the CNN AlexNet [16], which is pre-trained on ImageNet. We transfer the pre-trained layers of AlexNet to our CNN, except for the last 3 layers, which we replace with a new fully connected layer, a new softmax-layer and a new output layer to match our two classes: target (i.e. male golden-collared manakin) and background (i.e. Panama forest). The output of our CNN is a background score and a target score in [0, 1]. We fine-tune this new CNN with image patches of male golden-collared manakins and of background cropped from a set of videos in our dataset.

C. Kalman Filter

We use a linear Kalman Filter to predict the location of the bird if we could not obtain a reliable estimation of the location from III-A and III-B. In addition, the Kalman Filter's location estimation is used if we find more than one blob. In such cases, we select the blob that is closest to the Kalman Filter's location estimation.

D. Bird Tracking

The male golden-collared manakin's location is initialized in the first frame with the ground truth bounding box. For each following frame, moving foreground blobs are detected in the scene. To find the blobs that contain the male bird, the blobs are classified with the fine-tuned CNN. We keep the blobs that receive a high target score, and discard the others. If there is only one such blob, its position is selected as the current target location.

In case there is more than one blob, the one that is closest to the location predicted by the Kalman filter is selected as the main blob. Since the bird can be partly occluded (e.g. by the sapling it sits on) it can consist of more than one blob. Thus, we add blobs to the main blob that received a high target score by our CNN and that are close to the main blob. If we find a blob or combination of blobs, that fit these criteria, it becomes the current target location (Fig. 4).

If we find no blobs in a frame or none that fulfill the conditions described above we search the bird in the region around its previous location. This usually happens when the bird is sitting and thus not recognized as a foreground blob. We shift the bounding box from the previous position to its left, right, top, bottom and diagonal neighborhood, crop image patches at these candidate locations and classify them with the CNN. All candidate locations that receive a high target score are averaged, and selected as the current target location (Fig. 5).

In frames where the bird is not recognized by the CNN the Kalman filter is used if the bird is predicted to be flying. Otherwise, we use the bird's previous location as the current target location.

If the bird leaves the scene, candidate locations are placed along the edges of the frame to detect the bird when it re-enters the scene. To avoid false positive detections while the bird is outside the frame, only blobs are considered for detecting re-entering birds.

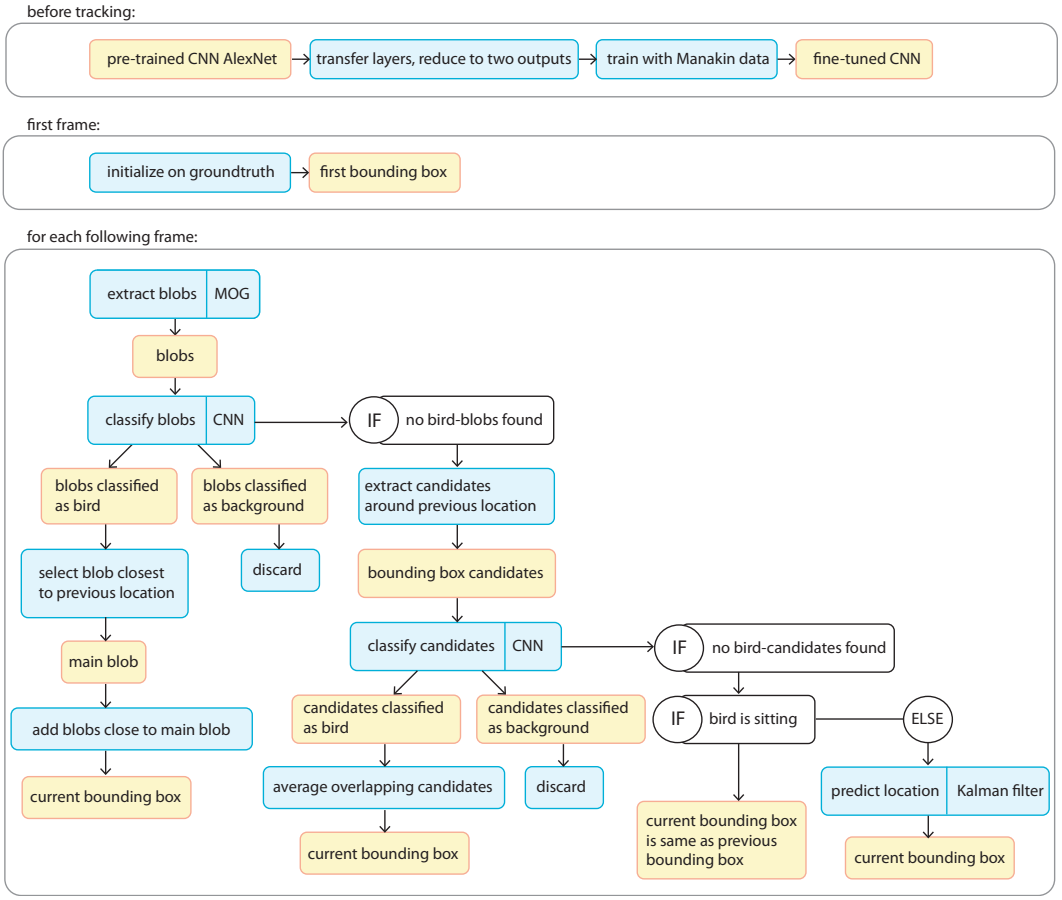

Figure 2: Flowchart of ManakinTracker.

Figure 3: Foreground mask (right) generated by MOG model of frame (left). Blue box: extracted blob.

The ManakinTracker uses a Kalman Filter and MOG, and thus does not rely only on the visual information in a single frame but detects motion based on multiple frames.

It uses a CNN, that is fine-tuned specifically to recognize male golden-collared manakins (including highly blurry and partly occluded) with high accuracy.

In most cases (particularly during jumps) blobs lead to very accurate bounding boxes and no further correction of the bounding boxes' dimensions is necessary. We can track the bird efficiently in most frames by classifying only a limited number of image patches extracted from blobs (usually 1-4 per frame).

In some cases the tracker recognizes the female bird as the target, even though the CNN was only trained on male birds. This suggests that the CNN does not rely only on the male bird's yellow neck for classification. The downside of this is, that if the male and female bird are both present in a frame the two have to be distinguished by the tracker. One solution would be to train a CNN on images of female birds also, which would require ground truth bounding boxes for the female birds. Currently, we handle this issue by choosing the blob that is closest to the location predicted by the Kalman filter if there are multiple blobs that get a high target score.

IV. Evaluation

To evaluate the performance of the ManakinTracker, we compared it to two of the trackers presented in Section II: TCNN and C-COT. A fair comparison with Oliva's tracker was not possible, because this tracker can neither be initialized nor re-started with a ground truth bounding box. Additionally, this tracker relies strongly on thresholds that were determined based on the dataset it was trained on, and outputs only bounding boxes for the male bird's neck, which keeps accuracy low even in case of successful tracking.

We assess the trackers' performance based on accuracy and number of re-starts. Accuracy is measured with the Jaccard index. A tracker is re-started at the next frame that has a ground truth annotation if the pre-

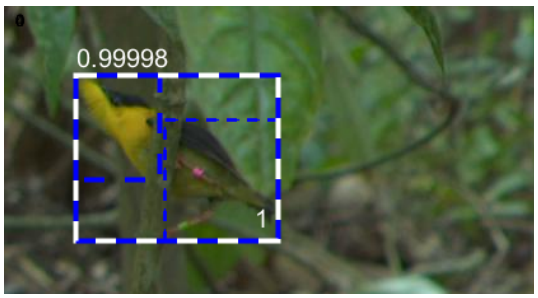

Figure 4: Two small blobs (small blue bounding boxes) are combined into a bigger blob (big blue bounding box). The white text indicates the blobs' target scores.

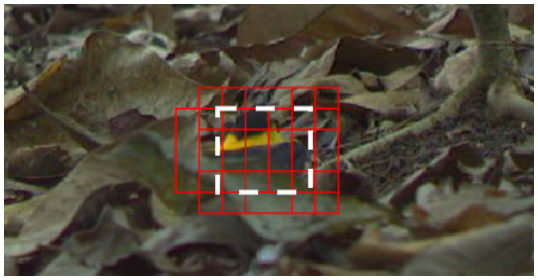

Figure 5: Bird is sitting and no blob was found (candidate locations with high target score (red boxes), final bounding box (white box))

dicted bounding box has zero overlap with the ground truth bounding box.

Our test dataset consists of 78 video sequences that show male golden-collared manakins performing their courtship dance. All videos were recorded with stationary high-speed cameras at 60 fps in the Panama forest. Every frame has a ground truth annotation (bounding box enclosing the male bird) provided by biologists. For testing, we split the dataset in half and train the CNN on one halve and run the tracker on the other halve.

Tracker	Avg. accuracy	Avg. # re-starts
ManakinTracker	58.3093%	1.2051
C-COT	49.0720%	4.6795
TCNN	53.3405%	7.1154

Table I: Trackers' performance on test dataset.

V. Results

Table I shows that the ManakinTracker performed the best out of the three trackers, both in terms of accuracy (58.31% average overlap) and robustness (1.21 re-starts per sequence on average). TCNN achieves higher accuracy (53.34%) than C-COT (49.07%), but needs about 1.5 times more re-starts on average.

TCNN and C-COT both use CNNs pre-trained on ImageNet. The performance of networks trained on the ImageNet dataset, which consists of still images, decreases strongly if images are blurry [17]. In contrast, the ManakinTracker's CNN was trained also on blurry images, extracted from videos similar to the ones it was tested on. For a more detailed evaluation see [18].

VI. Conclusion

The ManakinTracker achieved better accuracy and needed less re-starts than two state-of-the-art trackers. Keeping the number of re-starts low was our main goal as we aim to minimize user input during tracking. Using a CNN trained on similar videos as the test set led to a high accuracy in detecting and tracking the male golden-collared manakin.

References

[1] M. J. Fuxjager, L. Fusani, F. Goller, L. Trost, A. T. Maat, M. Gahr, I. Chiver, R. M. Ligon, J. Chew, and B. A. Schlinger, "Neuromuscular mechanisms of an elaborate wing display in the golden-collared manakin (manacus vitellinus)," *Journal of Experimental Biology*, 2017.

[2] J. Barske, B. A. Schlinger, M. Wikelski, and L. Fusani, "Female choice for male motor skills," *Proceedings of the Royal Society of London B: Biological Sciences*, 2011.

[3] D. Kress, E. van Bokhorst, and D. Lentink, "How lovebirds maneuver rapidly using super-fast head saccades and image feature stabilization," *PLOS ONE*, vol. 10, no. 6, pp. 1–24, 06 2015.

[4] D. Eckmeier, B. R. Geurten, D. Kress, M. Mertes, R. Kern, M. Egelhaaf, and H.-J. Bischof, "Gaze strategy in the free flying zebra finch (taeniopygia guttata)," *PLoS One*, vol. 3, no. 12, p. e3956, 2008.

[5] D. L. Altshuler, E. M. Quicazán-Rubio, P. S. Segre, and K. M. Middleton, "Wingbeat kinematics and motor control of yaw turns in anna's hummingbirds (calypte anna)," *Journal of experimental biology*, vol. 215, no. 23, pp. 4070–4084, 2012.

[6] I. G. Ros, L. C. Bassman, M. A. Badger, A. N. Pierson, and A. A. Biewener, "Pigeons steer like helicopters and generate down- and upstroke lift during low speed turns," *Proceedings of the National Academy of Sciences*, vol. 108, no. 50, pp. 19 990–19 995, 2011.

[7] N. Ota, M. Gahr, and M. Soma, "Tap dancing birds: The multimodal mutual courtship display of males and females in a socially monogamous songbird," *Scientific Reports*, vol. 5, 2015.

[8] D. Song and Y. Xu, "Monocular vision-based detection of a flying bird," *Texas A & M University*, 2008.

[9] W. Li and D. Song, "Automatic video-based bird species filtering using periodicity of salient extremities," *Department of Computer Science and Engineering, Texas A&M University, Tech. Rep. TR2012-08-2*, 2012.

[10] L. Oliva, A. Saggese, N. M. Artner, W. G. Kropatsch, and M. Vento, "From trajectories to behaviors: an algorithm to track and describe dancing birds," *Proceedings of the 22nd Computer Vision Winter Workshop, Retz, A*, p. 1–9, 2017.

[11] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," p. 252 Vol. 2, 02 1999.

[12] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder *et al.*, *The Visual Object Tracking VOT2016 Challenge Results.* Cham: Springer International Publishing, 2016, pp. 777–823.

[13] H. Nam, M. Baek, and B. Han, "Modeling and propagating cnns in a tree structure for visual tracking," *CoRR*, vol. abs/1608.07242, 2016.

[14] M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *ECCV*, 2016.

[15] G. Welch and G. Bishop, "An introduction to the kalman filter," 1995.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[17] S. F. Dodge and L. J. Karam, "Understanding how image quality affects deep neural networks," *CoRR*, vol. abs/1604.04004, 2016.

[18] A. Gostler, "Tracking golden-collared manakins in the wild," PRIP, TU Wien, Tech. Rep. PRIP-TR-141, 2018.