Detecting and Tracking Communal Bird Roosts in Weather Radar Data Supplementary Materials

Zezhou Cheng
UMass Amherst
zezhoucheng@cs.umass.edu

Saadia Gabriel
University of Washington
skgabrie@cs.washington.edu

Pankaj Bhambhani
UMass Amherst
pankaj@cs.umass.edu

Daniel Sheldon
UMass Amherst
sheldon@cs.umass.edu

Subhransu Maji
UMass Amherst
smaji@cs.umass.edu

Andrew Laughlin
UNC Asheville
alaughli@unca.edu

David Winkler
Cornell University
dww4@cornell.edu

A Baseline detection models

Here we analyze the performance of the Faster R-CNN architecture variants on the roost detection dataset. These models were trained without the variational EM algorithm and correspond to the baseline detector presented in Table 1 in the main paper. The performance of the Faster R-CNN

- with VGG-M network:
  - with ImageNet pretraining MAP = 41.0%
  - without ImageNet pretraining MAP = 34.8%
- with shallow network, comprising the first three convolutional layers of VGG-M network:
  - with ImageNet pretraining MAP = 37.7%
  - without ImageNet pretraining MAP = 33.1%

The shallow network roughly corresponds to the computational pipeline of histogram of oriented gradients features (Dalal and Triggs, 2005), a classic image representation useful for detecting shape patterns. The performance with the shallow network is quite good at MAP=37.7%, but additional layers provide a significant improvement in performance.

B Roost tracking and rescoring details

From detections to tracks. We used a greedy heuristic to group detections to tracks across frames. We start a new track at a reliable roost detection (with score over 0.5) which has not yet been matched to existing tracks. Suppose the location and radius of the roost in a track at time instant t is (l_t, r_t). We match the detection $(l_{t+\delta t}, r_{t+\delta t})$ in the next frame to the track if $\|l_{t+\delta t} - l_t\| < \alpha\delta t$ and $\tau < r_{t+\delta t} - r_t < \beta\delta t$, where τ, α and β are fixed thresholds. If there is no detection matched to a track we simply add a "ghost" detection by interpolating the detection at a previous frame by assuming zero positional velocity and expansion rate of roughly 1000 meters/min based on an analysis of the ground-truth annotations of swallow roosts. There can potentially be multiple tracks competing for one roost detection. In this case we assign the detection to the track with the lowest percentage of previous ghost detections. A similar heuristic was used in (Ren, 2008) for tracking people in archive films.

Smoothing using a Kalman filter. Individual detections in a track can be noisy. Moreover the greedy grouping can introduce incorrect detections to a track. We use a Kalman filter to smooth the detections in a track by incorporating the temporal dynamics of the roosts. Kalman filters provides an optimal estimate of a constant velocity dynamic system and have been widely used for object tracking (Bishop, Welch, and others, 2001). To a rough approximation the bounding-box of a tree-swallow roost expands at a constant rate and the center slowly translates in the plane. We establish the following linear dynamics model to track the roost over time,

$$\begin{aligned} X_t &= \mathbf{\Phi} X_{t-1} + \xi, \\ Z_t &= \mathbf{H} X_t + \mu, \end{aligned} \tag{1}$$

where $X_t = [x_t, y_t, r_t, \dot{x_t}, \dot{y_t}, \dot{r_t}]$ represents the state at time t. The state contains the location $l_t = (x_t, y_t)$ of the center, its radius r_t, and their temporal derivatives. Z_t is the observation at time t that represents the roost detections from our single-frame detector. Since our observations are only the position and radius, the measurement matrix $\mathbf{H}$ simply selects the first three components of the state. The transition matrix $\mathbf{\Phi}$ captures the temporal dynamics, e.g., $x_t = x_{t-1} + \dot{x}_{t-1}\delta t$. ξ represents the uncertainty of the dynamics and μ represents the noise in the observation Z_t. These are modeled as zero-mean Gaussian vectors with a diagonal covariance.

Contextual rescoring. As a final step we improve the detections by incorporating features from the entire track. In particular, for a given detection we derive four features: (1) the detection score, (2) the average of detection scores within the track, (3) the sum of detection scores within the track, and (4) a bias term that indicates if the detection was assigned to a track. Using these features we train a linear SVM using bounding boxes with overlap of 0.5 or more with a ground-truth bounding box as positive examples, and those with overlap of less than 0.1 as negative examples.

C ELBO Derivation

Since x and u are always observed we temporarily drop these from the notation. We also derive the bound for a single data case and drop parameters from the notation on the right-hand side. The remaining variables are y, which is unobserved, and $\hat{y}$, which is observed. The derivation is then standard.

$$\begin{aligned}
\mathcal{L} &= \log p(\hat{y}) \\
&= \log \int p(y, \hat{y}) dy \\
&= \log \int q(y) \frac{p(y, \hat{y})}{q(y)} dy \\
&= \log \mathbb{E}_{y \sim q} \left[\frac{p(y, \hat{y})}{q(y)} \right] \\
&\geq \mathbb{E}_{y \sim q} \left[\log \frac{p(y, \hat{y})}{q(y)} \right] \\
&= \underbrace{\mathbb{E}_{y \sim q} \big[\log p(y, \hat{y}) \big] + H(q)}_{\text{ELBO}}
\end{aligned}$$

Bring back x, u and the parameters of each model, we have the following ELBO:

$$\begin{aligned}
\mathcal{L}(\theta, \beta) &\geq \text{ELBO}(\theta, \beta, \phi) \\
&= \sum_i \Big(H(q_\phi^i) + \mathbb{E}_{y_i \sim q_\phi^i} \big[\log p_{\theta,\beta}(y, \hat{y}_i \mid x_i, u_i) \big] \Big) \\
&= \sum_i H(q_\phi^i) + \underbrace{\mathbb{E}_{\{y_i \sim q_\phi^i\}} \left[\sum_i \log p_\theta(y_i \mid x_i) \right]}_{\text{Expected training loss}} \\
&\quad + \underbrace{\mathbb{E}_{\{y_i \sim q_\phi^i\}} \left[\sum_i \log p_\beta(\hat{y}_i \mid y_i, x_i, u_i) \right]}_{\text{Expected forward user model loss}}
\end{aligned}$$

Here $H(q_\phi^i) = \mathbb{E}_{y_i \sim q_\phi^i}[-\log q_\phi^i(y_i)]$ is the entropy of the variational distribution on data example i.

References

Bishop, G.; Welch, G.; et al. 2001. An introduction to the kalman filter. *Proc of SIGGRAPH, Course* 8(27599-23175):41.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*.

Ren, X. 2008. Finding people in archive films through tracking. In *CVPR*.