Cataloging Birds in Their Natural Habitat

Teresa Ko, Stefano Soatto, Deborah Estrin
University of California, Los Angeles
{*tko, soatto, destrin*}*@cs.ucla.edu*

Angelo Cenedese
University of Padova
angelo.cenedese@unipd.it

Abstract

We devise a method to catalog novel objects from an image sequence even when the underlying scene exhibits sudden motion and appearance changes in consecutive frames, exemplified by the case of birds in their natural habitat. Cataloging birds in different ecosystems can provide important measures towards scientific models of global warming. However, images captured of the natural environment exhibit many visual "nuisances" that challenge standard detection and tracking methods that would allow for the cataloging of birds. We propose a method that specifically models the fine-scaled changes on the background due to motion, self-occlusion, and lighting changes. Regions that do not fit in this model are considered an instance of some bird. We then associate these regions with bird identities by allowing for either appearance similarity or location proximity as a guide. Birds are then clustered into visually similar groups that approximate species. Experiments show that we can maintain tracks for significantly longer periods of time as compared to classic mean shift tracking, and provide meaningful clusters for the end user.

1 Introduction

Fine-scale monitoring of natural environments is important for inferring meteorological and ecological trends. In particular, monitoring the behavior of small animals such as birds provides valuable insights and measures that feed into scientific models of global warming. While our community has produced a significant number of tools for generic object detection, localization, recognition and categorization, most are designed for man-made objects and environments, and fail in natural habitats. For instance, the visual appearance and aspect of an object can vary significantly with the vantage point, and data is collected in an automated manner that is different from the purposeful snapshot of a user that captures an image to upload to Flickr or other Internet databases. For obvious storage and computational limitations, data is captured at a reduced sampling rate, so while multiple views are generally available, they are not closely sampled in such a way that enables tracking from one image to the next. Finally, although birdwatching is a common pastime and there are plenty of people willing to spend hours labeling images, manual annotation can only be performed for a minuscule percentage of all available data, so one cannot rely on supervised datasets for training. For these reasons, the most common approaches to modeling object categories, via a distribution of "features," with or without spatial constraints, characterizing the object, trained from supervised datasets, yield performance far below that reported on benchmark datasets such as the Caltech 101 or Pascal Challenge.

Therefore, there remains the need to (a) devise a representation of objects that is less dependent on availability of full information at training, and (b) devise unsupervised learning approaches that can significantly cut on the labor cost of hand-labeling massive datasets being gathered from environmental monitoring stations.

In this paper, we present a 3-step process to this end. The first step extracts regions from each frame in the image sequence. We define as background the portions of the scene that, over relatively long observation times, remain within the field of view, even though they may move and even disappear temporarily due to partial occlusions. The second step associates these extracted regions with object identities. We define an object as a set of regions that occupies consecutive frames and are "sufficiently" similar in appearance or position in the image. We then devise a scheme for cataloging putative objects of interest into viewable clusters; our approach is based on a coarse descriptor, dubbed "object barcode", that represents the *occurrence* of certain visual words in multiple frames, rather than their *frequency* in a histogram. This proves to be more effective than bag-of-features for the task at hand. Because data capture is performed on pre-defined intervals, no "intelli-

gent sampling" is performed, so one can have multiple images of the same bird in the same pose, and fail to capture views with the birds in different poses. Thus, we propose a method to rank objects based on how "informative" each frame is. This ranking is then used to aid unsupervised learning (clustering).

2 Approach

Our goal is to associate foreground regions in an image sequence with an object and cluster identity. We are given a set of images $\mathcal{I} = \{I_t(x) : t \in \mathcal{T} = \{1, \dots, T\}; x \in \Omega\}$, where $I_t : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^+; x \mapsto I_t(x)$.

For each image I_t, we extract a set of sub-images each defined over a compact region in the image domain, $\mathcal{R}_t$, indexed w.r.t. *the image* I_t: $\mathcal{R}_t \doteq \{R_{t,n}(x) : n \in \{1, \dots, N_t\}; x \in \Theta_{t,n}\}$, where $R_{t,n} : \Theta_{t,n} \subset \Omega \to \mathbb{R}^+; x \mapsto R_{t,n}(x)$. Regions are not overlapping, that is for any two views $R_{t,k}$ and $R_{t,j}$, $R_{t,k} \cap R_{t,j} = \emptyset$.

An object O_i is defined as a collection of r_i views from consecutive frames $\{V_{t,i}\}$, where each view $V_{t,i}$ is a sub-region $R_{t,n}$ indexed w.r.t. *the associated object* O_i: $O_i \doteq \{V_{\bar{t},i}, \dots, V_{\bar{t}+r_i-1,i}\} = \{V_{t,i} : t \in \bar{\mathcal{T}} \overset{c.s.}{\subseteq} \mathcal{T}\}$, where $\bar{\mathcal{T}}$ is a compact set in $\mathcal{T}$ and views are associated into an object according to some similarity measure. The number of objects $|\{O_i\}|$ and the number of views for each object, $\{r_i\}$, are unknown a priori.

If we make the assumption that associations are formed independently for each pair of images, we can simplify the problem to matching views from 2 consecutive frames. That is, we would like to be able to define a decision function to assign $V_{t,j}$ with $V_{t+1,k}$ to the same object O_i.

Finally, each object O_i is also assigned with a single cluster $C_h \doteq \{O_i, i = 1, \dots, M\}$, based on some measure of appearance similarity. The number of categories $|\{C_h\}|$ is also unknown a priori.

2.1 Extracting Regions

Unlike many background approaches that represent each pixel independently [2], we model the entire background as a composition of several layers each warped independently. A sample image can be constructed by selecting the best warping from each canonical image, or layer, $\hat{I}_b$, such that,

$$\tilde{I}_t(x) = \sum_{b=1}^{B} \hat{I}_b(w_{t,b}(x))\chi(x), \quad (1)$$

where

$$\chi(x) = \begin{cases} 1 & \text{if } x \in \Omega \setminus \Theta, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 1. Bird shape and motion move quickly and in an unpredictable manner. They also change shape quickly and are not well approximated by an ellipsoid.

But for an observed image I_t, we do not know if it contains foreground objects, how the background is warped (unknown $w_{t,b}(x)$) or where the occlusions occur (unknown $\chi(x)$). The pixel-wise discrepancy is thus:

$$D_t(x) \doteq \|I_t(x) - \sum_{b=1}^{B} \hat{I}_b(\tilde{w}_{t,b}(x))\tilde{\chi}(x)\|_2, \quad (2)$$

where $\tilde{w}_{t,b}(x)$ is the estimated warp and $\tilde{\chi}(x)$ is the estimated occlusion map.

2.2 Inferring Objects

We propose using a weighted scoring approach to infer object identities from a set of views. A common assumption in tracking is that objects of similar appearance and in close proximity is the same object [1], or the motion can be predicted [4]. Because birds change appearance suddenly, this severely limits our ability to recognize the same bird across frames. For this reason, we propose a scoring function, s_V that is a linear combination of appearance and location similarity:

$$s_V(V_i, V_j) = w_l s_l(l(V_i), l(V_j)) + w_v s_v(f(V_i), f(V_j)), \quad (3)$$

where w_l and w_v are parameters used to weight the relative important of proximity versus appearance, and s_l and s_v measure the similarity between the position of the views and the values and appearance of the views, respectively. $l(V_i)$ is a function that converts the view into the relative positional information of the view V_i, such as position, size, velocity, etc. $f(V_i)$ is some function that extracts a feature vector from the view V_i.

2.3 Cataloging Objects

Given the set of views of an object O_i, we create a binary vector $\mathbf{b} \in \mathcal{B} = \{\mathbf{0}, \mathbf{1}\}^{\mathbf{D}}$ where a component $b_p = 1$ if a feature d_p is present in any of the views of the object. The features we use are the hue-saturation values of each pixel and the size of each view.

Figure 2. Mean shift fails when birds in consecutive frames move too suddenly (31) and when the appearance model contains much of the background (21 and 17), while our approach is able to recognize this two views as the same bird due to their similar appearance.

Figure 3. Mean shift loses the object track when the bird's appearances changes too quickly (81), while our approach considers the location of the bird and difference from the background to recognize this as the same bird.

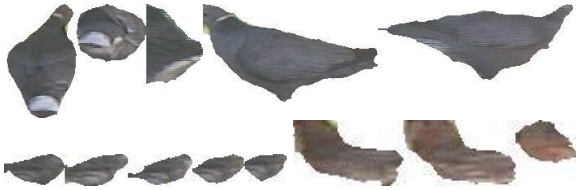

Figure 4. Sample regions extracted from the image sequences.

The similarity between two objects $\mathbf{b}_i$ and $\mathbf{b}_j$ (as column vectors) is given as

$$s_O(\mathbf{b}_i, \mathbf{b}_j) = \frac{\mathcal{N}(\mathbf{g}_i)^\top \mathcal{N}(\mathbf{g}_j)}{\left(\mathcal{N}(\mathbf{g}_i)^\top \mathcal{N}(\mathbf{g}_i) + \mathcal{N}(\mathbf{g}_j)^\top \mathcal{N}(\mathbf{g}_j)\right)/2}, \tag{4}$$

where $\mathcal{N}(\mathbf{g})$ is a Gaussian blur over the hue saturation space.

The capture mechanisms may result in multiple redundant views of the same object in the same pose. This generates a bias in the representation that is partially ameliorated by the barcode descriptor. Ideally, we would like the clustering algorithm to also place more emphasis on multiple aspects and appearance variation in multiple views. Thus, rather than quantifying the number of frames available per each object, we introduce a "completeness" measure as follows

$$\beta_i = \sum_{V_j} \left(1 - \max_{V_j^-} s_B\left(f(V_j), f(V_j^-)\right)\right), \tag{5}$$

where s_B is the Bhattacharyya similarity measure between normalized histograms of Ω in two views V_j and V_j^-, where V_j^- is any view before V_j.

3 Preliminary Results

We examined 800 frame image sequence of birds at a feeder station. In this sequence, there are 81 bird visits that vary from 1 to 174 frames in length. The image sequences have been manually labeled with bounding boxes surrounding each bird.

To understand how well tracking could be used as a method for recognizing birds across frames, we ran an experiment using classic mean shift tracking to see how long a bird could be tracked through its visit. The longest track lasted for 11 frames.

In Fig. 2, mean shift is unable to track the bird when the bird moves too suddenly (bird 31). It also creates false positives when birds strongly deviate from an elliptical appearance model. Because the resulting representation contains much of the background, mean shift continues to track the birds even after they depart, as indicated by the labels 21 and 17. In Fig. 3, mean shift has trouble with bird 81 even though it is in the same location because the appearance changes too quickly.

3.1 Matching

We now illustrate the performance of our proposed approach. We use the same 800 images used in the motivating example, but now extract regions manually. Examples of the extracted regions are shown in Fig. 4. Using our proposed technique, we achieve an average precision of 56.22% as compared to 19.54% when using Elgammal's approach [2].

We examine the effect of s_l on our ability to track objects. In this experiment, we define $L(V_i) = [x\ y\ \Delta x\ \Delta y]'$, where $[x\ y]$ is the center of the bounding box of the object, and $[\Delta x\ \ \Delta y]$ is the size of a bounding box around the object. We assume that the object can move and change size from frame to frame, $L_{t+1} = L_t + \eta$ where $\eta \sim \mathcal{N}(0, \sigma^2)$, so that the most likely location of an object in the next frame is the current location. Using this approach, we are able to re-

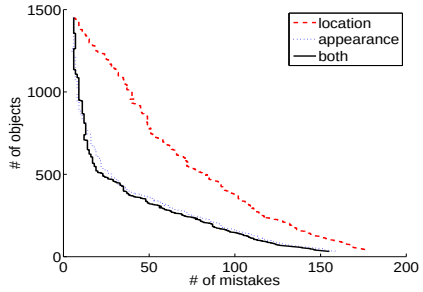

Figure 5. Using both appearance and location improves object cataloging.

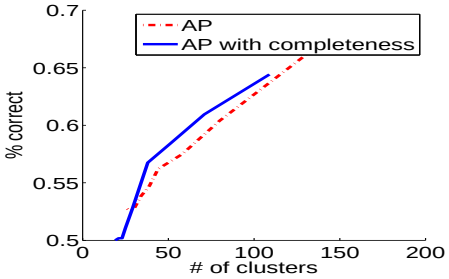

Figure 6. Affinity propagation clustering with the completeness measure improves the overall clustering of bird categories.

cover birds that change appearance rapidly, but stay in the same location, such as the one shown in Fig. 3.

We would also like to handle birds exhibiting sudden motion. We construct $f(V_i)$ as a color histogram of the view V_i and define s_B as the Bhattacharrya distance, $s_B(f(V_1), f(V_2)) = \sum \sqrt{f_{V_1,i} f_{V_2,i}}$. This method is able to interpret the views previously thought to represent two different birds in Fig. 2 as a single bird.

Using both s_l and s_v results in fewer mistakes (views of different objects mistakenly considered as the same object) while reducing the number of estimated objects more when compared to using s_l or s_v on their own.

3.2 Clustering

Objects are clustered using affinity propagation [3], with a fixed preference value for all objects. Affinity propagation is a good clustering choice for this problem because we can easily integrate the pairwise similarity between objects that we have explored above. The preference parameter in affinity propagation encodes the preference for an object to be a cluster head. We use this parameter to encode the completeness measure into the clustering in a flexible manner. We evaluate our clustering by looking at each proposed cluster, assign it to the ground truth label that is the most dominant, and accumulate the remaining objects in the cluster as a wrong label. The percent of wrong labels is plotted over the number of clusters proposed in Figure 6.

4 Discussion

These preliminary results indicate that this approach holds promise. The representation we use in these experiments are admittedly crude. The subtle variations exhibited by different birds are lost in the color histogram descriptor of the view and the barcode descriptor of the object. However, such subtle inter-category variabilities are swamped by the variability induced in the data by the complex illumination variability (different seasons, weather, times of the day), and by the imperfect background masks due to mimicry of the objects of interest with the background. Indeed, this is a challenging task for non-expert human observers.

Future exploration would include more sophisticated measures for proximity in space and appearance, and methods for automating the weighing process. More importantly, we would like explore an augmentation to $s(V_i, V_j)$ to include the similarity between objects and categories, such that, once a view is assigned to an object or category, we are able to incorporate that into our matching scheme.

We feel that this work represents a useful first-step, with significant work yet to be done before automated analysis of habitat monitoring data can be performed.

Acknowledgments. This material is based upon work supported by the CENS under the NSF Cooperative Agreement CCR-012-0778 and #CNS-0614853, by ONR 67F-1080868/N00014-08-1-0414, ARO 56765-CI and AFOSR FA9550-09-1-0427. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF, ONR, ARO, AFOSR.

References

[1] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[2] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *European Conference of Computer Vision*, pages 751–767, 2000.

[3] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[4] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.