Monitoring Giraffe Behavior in Thermal Video

Victor Gan[1,2]
[1]University of British Columbia
vhg@cs.ubc.ca

Peter Carr[2]
[2]Disney Research
carr@disneyresearch.com

Joseph Soltis[3]
[3]Disney Animal Kingdom
joseph.soltis@disney.com

Abstract

We present a solution for monitoring nocturnal giraffe behavior by reducing several hours of thermal camera surveillance footage into a short video summary which can be reviewed by experts. We formulate the video summarization task as a tracking problem: frames in which giraffes are successfully tracked are presumed to be typical poses/behaviors and not included in the summary; whereas frames containing track initializations or terminations are presumed to be atypical events and are therefore included in the summary. To implement our tracking-by-detection summarization approach, we explore various combinations of image features to determine the best combination for long infrared spectrum cameras, and devise a variant of the deformable parts model object detection technique using geodesic distances to handle the extreme variations of typical giraffe postures. Finally, we evaluate our summarization performance in terms of recall and compressibility, and show how a trade-off exists between these two measures using more fragile or robust tracking techniques.

1. Introduction

Observing animal behavior is critical for ensuring the long term health and safety of animals in the wild and under human care. Human experts can generate reliable data through direct observation, but such as approach is rarely sustainable for long periods of time. Instead, automated data collection and analysis are needed to construct a more comprehensive understanding of animal behavior. In some situations, intrusive data collection methods like on-body sensors and data loggers (or transmitters) may be feasible. However, that is not always the case. In this paper, we explore passive sensing of giraffe nocturnal behavior using thermal cameras.

Although recent computer vision work [4, 9] has performed well on object detection tasks, the majority of the datasets have focused on humans. Relative to pedestrians, most animals have a substantially larger pose variation, self occlusion, and strong visual similarity between individuals. As a result, directly applying established object detection techniques may not achieve sufficient reliability for long term animal surveillance.

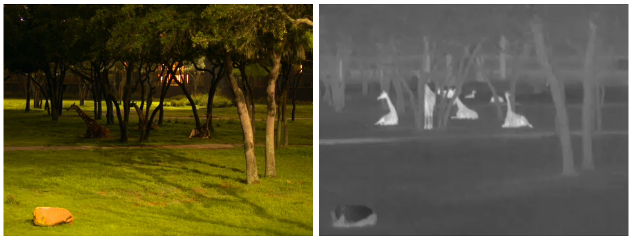

Figure 1. **Visual vs Thermal Imaging**. Sample frames from RGB (left) and thermal (right) video cameras. Although detection is significantly easier in thermal video, long term tracking remains a difficult challenge, as giraffes are easily occluded by trees and/or other animals.

Furthermore, we are interested in monitoring nocturnal behavior. Because visual spectrum cameras struggle in low-light conditions (see Figure 1), our study uses IR thermal cameras. The characteristics of thermal images are quite different from those captured in the visual spectrum. Therefore, many of the image-based features used in the visual spectrum (often designed to be invariant to illumination changes) may not be well suited to thermal images. For instance, when using thermal cameras to detect and track pedestrians, [5] employed *contour saliency maps* as the principle image-based feature.

In this paper, we describe a system for reducing several hours of surveillance data into a short summary video of observed rare behaviors which can be reviewed by human experts. We define rare events as atypical body poses. For example, when a giraffe eats, there is a significant deviation in the pose of the head and neck. The summary is generated by first detecting and tracking giraffes in video captured from stationary thermal cameras. The detectors are tuned to search for typical poses. The initialization or termination of a track constitutes a summary event. Because of occlusions with trees and other giraffes, the system also includes sequences where heads and necks are not tracked with high confidence.

Contributions We investigate a host of image-based features for building rigid part detectors for thermal video, and show how incorporating features specific to thermal cameras boosts performance over standard HOG+SVM implementations used in visual spectrum images. In order to reliably detect highly deformable objects (like giraffes), we reason about the connectivity of parts using geodesic distances, instead of the more standard Mahalnobis distance. Again, our experiments illustrate how this measure is able to correctly resolve ambiguous situations were the standard techniques fail. Finally, we formulate video summarization as a tracking problem: interesting events occur whenever tracks are initialized or terminated.

2. Related Work

Sliding a window over an image pyramid is a very effective method for finding objects in arbitrary images. The combination [4] of histograms of oriented gradients (HOG) features and linear support vector machine (SVM) classifiers has been particularly popular due to its detection performance and computational efficiency. For non-rigid objects, a more complex approach is typically used: individual object parts are localized using HOG+SVM, and the existing of an object is inferred by the spatial arrangement of the parts [9]. For computational efficiency, the deformable parts model assumes parts have a spring-like potential relative to a reference location. In particular, [17, 18] found that giraffes were very difficult to detect (in visual spectrum images) because the neck could have significant deformation (and required to be modeled as several parts in order to detect successfully). Our thermal images lack the resolution and color texture information that was available in the images used by [17, 18].

The recent survey [10] by Gade and Moeslund gives a good overview of using thermal cameras for object detection. Because most objects of interest in thermal video are warmer than the background, exhaustive multi-scale sliding window techniques are often not necessary. Instead, candidate regions of interest can be found based on intensity, and object detectors can then be used to make a final decision as to whether an object is present or not. Almost all object detectors in thermal images use shape-based features, often derived from silhouettes identified by applying a threshold to intensity values. Gradient-based features, such as HOG, have also been used as well. Although the majority of investigations involving thermal cameras has focused on humans, some work has been done on animals (see the recent survey [3] by Cilulko *et al*. for a complete review). The most closely related previous work to ours is the roadside deer detection system [22]. A stationary thermal camera was used to monitor a road in a forested area. A computer vision system searched for deer in thermal images by finding candidate regions using an intensity threshold, and then making a final decision using HOG+SVM. The system was fast enough for realtime object detection, but no tracking was performed.

Many state-of-the-art tracking methods [2,11] are unable to track objects for long periods of time — especially in cluttered scenes. These trackers usually require manual initialization, and adapt to movement and appearance changes over time. However, once the tracker has lost the object, it is not trivial to find the target again. Tracking-by-detection, on the other hand, uses object detections to automate initialization and avoid drifting, and are equally competitive with online trackers [21]. As we require tracking multiple objects over long periods of time, we use a tracking-by-detection approach. However, we note that tracking is an intermediate goal. We use track initializations and terminations to identify important frames in the video.

Video summarization [15, 16] extracts representative sequences from a long video. Khosla *et al*. [13] use web images as priors to determine significant frames. Rodriguez *et al*. [19] compact multiple actions over time so actions happen at once in the video, compressing the amount of time needed to see all the actions in the video. However, this requires the actions to occur at different locations for multiple actions to be visually seen at once.

3. Method

Our goal is to develop a system which can automatically summarize several hours of surveillance video into short segments of observed atypical animal behavior for review by human experts. Because it may be impossible to define all types of atypical behavior in advance, we pose the summarization problem in terms of discarding frames which contain only well recognized animal behaviors — *i.e*. we want to find the most compact subset

$$\mathcal{F} \subset \{ \, t \mid 1 \leq t \leq T \, \} \tag{1}$$

of frames from a video of T frames that includes all exhibited atypical behaviors. As a result, a key requirement for our approach is the ability to detect common animal behaviors reliably.

Giraffes are particularly difficult to detect [17] because their composite shape is highly deformable. Furthermore, in thermal video, appearance cues (such as texture) are not present, and the only reliable visual characteristic of a giraffe is its long neck. Because the neck has extreme variations in pose, we detect giraffes in a two-stage process: we run head-only and body-only part detectors over each frame of the video (see Section 3.3.1), and then post-process the detection results to search for pairs of heads and bodies which appear to be connected by a giraffe-like neck (see Section 3.3.2). Finally, we link detected head-body pairs across multiple video frames (see Section 4) to determine the expected number of animals within the scene, as well as the on-set and off-set of atypical behavior.

3.1. Dataset

Our dataset consists of two hours of nighttime video recorded from a fixed vantage point using an Axis Q1910 video camera (see Figure 1). Four giraffes (*Giraffa camelopardalis*) and a variety of other African fauna were enclosed in a circular savanna approximately 180m in diameter. For convenience, the video was re-sampled at 1fps to generate 7410 frames of data (split into datasets of 361, 3060 and 3989 frames). Bounding boxes of heads and bodies were manually annotated in each frame, resulting in 24495 head-body pairs. Finally, the temporal sequences corresponding to infrequent behaviors (such as eating) were manually identified.

3.2. Features

Similar to [3, 7], we experiment with multiple features. We forgo an initial preprocessing step to calculate regions of interest. This is often done with thresholding, morphological operators and "blob-splitting" [10, 23] to create object proposals. However, occlusion with trees and other animals is common and makes this preprocessing prone to error without careful hand-tuning. Instead of using a segmentation preprocessing step, we incorporate segmentation features, such as thresholding and background subtraction, directly into a sliding window classifier as additional feature elements. In all, we found the following six features to be useful:

Intensity (INT) The raw grayscale pixel intensities values. These somewhat correlate with actual temperature, although the camera incorporates automatic gain and exposure controls.

Thresholding (THRESH) A threshold of $\frac{130}{255}$ is applied to the raw intensity values to coarsely segment warm-blooded animals.

Background subtraction (BG) A period of 11 minutes at 30fps (19981 frames) of the scene was captured without any animals to model the characteristics of an empty scene. A univariate Gaussian distribution was fit to the temporal history of each pixel. For all input frames, a binary mask was created by testing for pixels which differed by more than 5 standard deviations from the mean value.

Gradient magnitude (MAG) Similar to [7], we compute the gradient magnitude of each pixel.

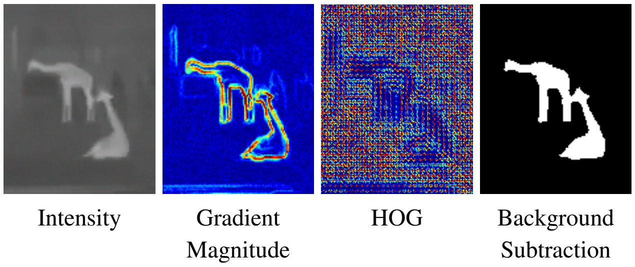

Figure 2. **Features**. A variety of channel features were used to detect rigid body parts.

Histogram of oriented gradients (HOG) We use Felzenszwalb *et al.* variant of HOG [9]. Due to the low resolution of the camera, we use a cell size of 4 instead of the default 8.

Unnormalized gradient histograms (UHOG) We bin oriented gradients without the post-processing contrast normalization normally set normally present in HOG. Illumination invariance is counterproductive for thermal imagery.

3.3. Giraffe Detector

To accurately detect the giraffes, we experiment with different feature sets and classifiers. We devise a head detector and a body detector and reason about how to best connect the parts to improve accuracy. We find a parts-based model with a suitable distance measure for connecting the parts superior to a single rigid whole body giraffe template.

3.3.1 Head and Body Classifiers

Both SVMs and random forests (RFs) were considered as classifiers. Combinations of features and classifiers were trained on one video of 361 frames and validated on another video of 3060 frames. The more effective feature-classifier combinations are shown in Figure 3. Generally, false head detections occurred on other animals and background structures like rocks, and false body detections occurred on other animals and animals clustered in groups. We found a combination of HOG, gradient magnitude and background subtraction features using linear SVMs to be a viable solution for detecting heads and bodies.

Implementation details We use the sliding window paradigm [4] to detect objects of varying sizes within the image. The detection windows are implemented using Piotr's Matlab Toolbox [6], with 8 scales per octave and a maximum scale factor of 1.6. The smallest bounding box sizes of the head and body are 24 × 24 pixels and 48 × 32 pixels respectively, determined by the smallest giraffe sizes in the training set. The HOG cell size was 4 × 4 pixels, and the stride between windows was also 4 pixels. Each

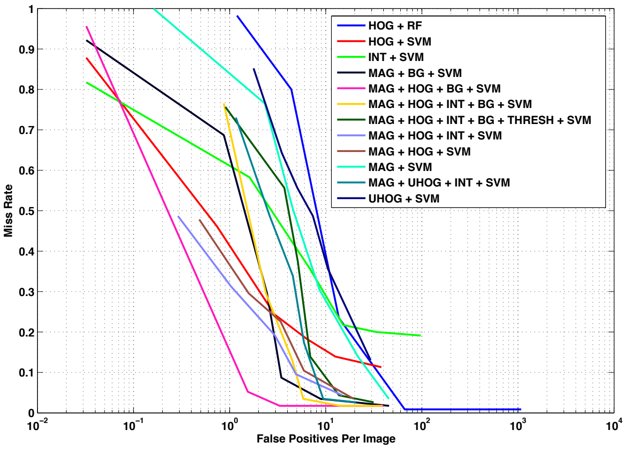

Figure 3. **Classifier Performance**. Various feature combinations were used in conjunction with SVMs and RFs. The performance of heads is shown here. The body detectors had similar performance curves.

classifier used a negative training set that is hard negative mined for the specific combination of features and classifier. Each classifier underwent eight iterations of mining, terminating early if perfect training precision was achieved. Results were greedily non-maximum suppressed with an overlap threshold of 0.3 as recommended in [14].

3.3.2 Neck Detection

Similar to the deformable parts model (DPM) [9], we combine parts to improve detection performance by explicitly considering pose variation. However the spatial relationship between giraffe body parts is complex. Figure 4 shows how the 2D Gaussian distribution used in DPMs is ill-fitted for estimating the location of a giraffe's head relative to its body (because of its highly deformable neck). Furthermore, giraffes frequently overlap within the video, making it necessary to include mutual exclusion constraints when matching parts (since each head can only be connected to one body, and vice versa). As a result, we experiment with alternative distance measures, and match discrete body and head detections through a linear assignment formulation. Four distance measures are explored:

1. **Euclidean**: The distance between the actual head location (relative to the body) and the learned average location of the head (relative to the body) in the training data.

2. **Mahalanobis**: The Euclidean distance between the centroids of the head and body detections is normalized by the mean and covariance observed in the training set.

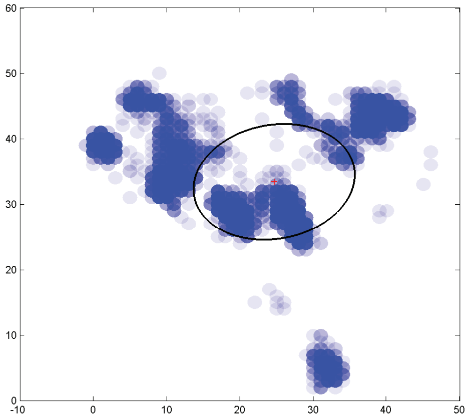

Figure 4. **Head-Body Spatial Relationship**. The positions of heads relative to their respective bodies (in pixels) for 9153 giraffe instances in a 51 minute video. The mean and one standard deviation contour is shown. The large space of body to neck positions prevents fitting a standard univariate distribution.

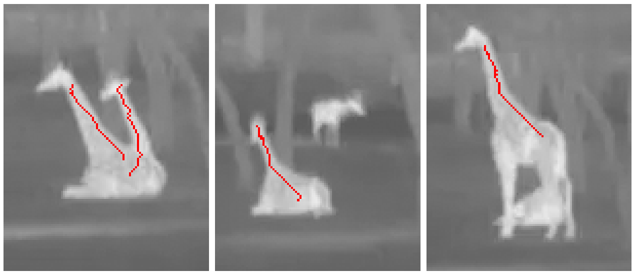

Figure 5. **Geodesic Distance**. Neck traces found by traversing the shortest path of the negative intensity image.

3. **Scale**: The size ratio of each body-head pair is subtracted from the ideal body to head ratio.

4. **Geodesic**: If a pair of head and body detections correspond to the same giraffe, then there should be a neck-like structure connecting them. Here, we compute the shortest path between the centroids of the head and body detections, where the cost of including a pixel is proportional to its negative intensity (see Figure 5). The resulting geodesic distance counts the number of pixels involved in the shortest path.

For each distance measure we experiment with thresholding at 2, 2.5 and 3 standard deviations from the mean of the training set. Any body-head pair with a distance greater than this threshold is considered not a match.

Because giraffes are frequently in close proximity to each other, a greedy association of the closest head to the closest body may fail (see Figure 6). Instead, we match heads to bodies in a non-greedy fashion by formulating a linear assignment problem. For n body detections and m

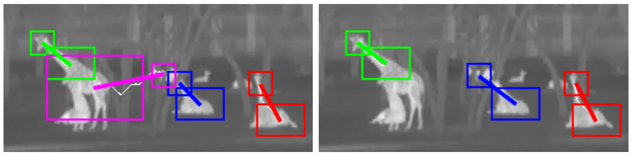

Figure 6. **Greedy vs Linear Assignment**. Head-body associations made in a greedy fashion using Euclidean distance can easily lead to incorrect solutions (left). The geodesic distance (right), in combination with a linear assignment formulation, significantly improves the correct association of heads to bodies.

head detections, we define a $(n+m) \times (n+m)$ cost matrix $\mathtt{C}^{parts}$ (2) representing the negative log likelihood that a head-body pair corresponds to the same giraffe. The structure of the matrix is divided into four quadrants. The upper-left $n \times m$ block contains the distance d_{ij} between head detection i and body detection j, (where the specific distance measure changes between experiments). The upper-right and lower-left quadrants account for heads and bodies with no suitable match. The diagonal of these blocks corresponds to the maximum possible inlier distance $d_{\max} = k\sigma$, where σ models the standard deviation of inlier distances in training data, and $k > 0$ is a user defined parameter. The off-diagonal elements are inadmissible and have infinite cost. Finally, the lower-right block encourages null head detections to associate to null body detections by specifying a cost of zero in these cases

$$\mathtt{C}^{parts} = \left[\begin{array}{ccc|ccc} d_{11} & \cdots & d_{1m} & d_{\max} & \infty & \cdots \\ d_{21} & \cdots & d_{2m} & \infty & d_{\max} & \cdots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ d_{n1} & \cdots & d_{nm} & \infty & \infty & \cdots \\ \hline d_{\max} & \cdots & \infty & 0 & 0 & \cdots \\ \infty & \cdots & \infty & 0 & 0 & \cdots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \end{array}\right]. \quad (2)$$

The minimal cost assignment is found using the Hungarian algorithm, and generally includes head-body, head-null and null-body pairings. The linear assignment formulation prevents multiple heads associating to one body and vice versa. The maximum inlier distance $d_{\max}$ has the effect of discarding false positives generated by the head-only detector. Table 1 summarizes the performance of the different distance measures. The geodesic distance with a tolerance $< 3\sigma$ results in the best performance. Figure 6 qualitatively shows the results of using the Euclidean and geodesic distance measures. Both results reduce the number of false detections, and we see the body-head pairs are assigned differently if different distance measures are used.

Criteria	FPPI	Miss Rate	F1 Score
(2σ) Euclidean	0.8710	0.5114	0.5443
(2σ) Mahalanobis	**0.6452**	0.7763	0.3009
(2σ) Scale	2.7419	0.7614	0.2165
(2σ) Geodesic	1.0645	0.5059	0.5250
(2.5σ) Euclidean	1.0968	0.4886	0.5389
(2.5σ) Mahalanobis	0.7097	0.7089	0.3710
(2.5σ) Scale	2.7419	0.7614	0.2165
(2.5σ) Geodesic	1.1935	0.4886	0.5294
(3σ) Euclidean	1.7097	0.2614	0.6311
(3σ) Mahalanobis	1.6129	0.2386	0.6537
(3σ) Scale	3.3871	0.7273	0.2212
(3σ) Geodesic	1.5484	**0.2159**	**0.6732**

Table 1. **Distance Measures**. Each distance measure is normalized using the mean and variance of the training body-head pairs. In testing, any body-head pair that falls outside of k standard deviations (in brackets) from the training average is considered a impossible pair. The remaining body-head pair combinations are then optimally matched to minimize the distance costs from the training average, using the Hungarian algorithm. Note that detections are valid only if both the head and body have an intersection over union > 0.5 with their corresponding ground truth parts, a significantly harder task than the full giraffe object detection task seen in Table 2.

Criteria	FPPI	Miss Rate	F1 Score
Whole Giraffe	1.6452	0.4432	0.5213
Head Only	1.5484	**0.0522**	0.8015
Head + Neck/Body	**0.3871**	0.1023	**0.8827**

Table 2. **Detector Performance**. A single rigid template which looks for an entire giraffe achieved the lowest performance. A head only detector achieved the highest recall, but had a significant number of false detections. Filtering the head detections by searching for appropriate context of a body and neck drastically reduced the false positive rate, while still preserving good recall.

Implementation details Implementing the shortest path between n heads and m bodies takes $O(nmE)$, where E is the cost of the shortest path algorithm (for example, A-Star or Dijkstra's algorithm). Instead, we calculate geodesic images for each part [20], which allows the shortest path between a head and body to be found by adding the two part images and taking the minimum value. This results in $(n+m)a + (n+m)b = O((n+m)a)$, where $a(b)$ is the cost of the generalized geodesy method and b is the constant number of pixels in the image. In practice, the constant is non-negligible, but we use this method to conservatively hedge in the event of many detections.

3.3.3 Performance Comparison

A rigid whole giraffe detector was as a performance baseline. First, ground truth annotations of entire giraffes were found by taking the minimum size bounding box that encompassed both the head and the body bounding boxes. Each whole giraffe bounding box was then rescaled to 48×48 pixels (the average bounding box size of the training set is 53.8377×52.6849), and our HOG, gradient magnitude and background subtraction rigid features were calculated. During hard negative mining and testing, 48×48 square bounding boxes were used with an SVM classifier. This replicates the rigid detector used for the head and body. For the connected head-body detections, a similar bounding box is drawn encompassing both parts. This bounding box is compared to the ground truth. Table 2 shows the detection performance of the baseline whole body giraffe detector, only a head detector, and the body-head detections with the geodesic distance measure which incorporates both the individual head and body detectors. The detector with the geodesic distance measure drastically improves precision with minimal effect on recall.

4. Tracking

The giraffe head detector and subsequent neck filter were specifically trained to search for typical head poses. However, our primary interest is in *atypical* head poses. As a result, we need to know not when the detector is successful, but rather when the detector is unable to locate a giraffe head (presumably because the head had uncommon appearance or spatial context). Because it is impossible to directly test for missed detections, we instead identify detection failures by tracking detected objects: tracking initializations and terminations indicate missed detections before/after the current frame.

Two-Frame Hungarian Our baseline tracking-by-detection algorithm operates over pairs of sequential frames. Head detections $\mathcal{H}_t$ in the current frame t are matched to head detections $\mathcal{H}_{t-1}$ in the previous frame. In general, there are $n = |\mathcal{H}_{t-1}|$ detection in the previous frame, and $m = |\mathcal{H}_t|$ detections in the current frame. In order to handle missed and false detections, we formulate an $(n+m) \times (n+m)$ cost matrix $\mathtt{C}^{\text{two}}$, of the same form as (2). The upper-left $n \times m$ quadrant represents the negative log likelihood that detection h^i_{t-1} in the previous frame and detection h^j_t in the current frame correspond to the same giraffe. We assume the displacements are normally distributed and hence $\mathtt{C}^{\text{two}}_{ij}$ is the square Euclidean distance between the centroids of each detection. The head displacements between frames for the training video is shown in Figure 7. Though the majority of the displacements are within two pixels, during movement the head displacement is up to 18 pixels. Hence, we assume the maximum possible displacement between frames is 20 pixels, and the diagonals in the upper-right and lower-left quadrants

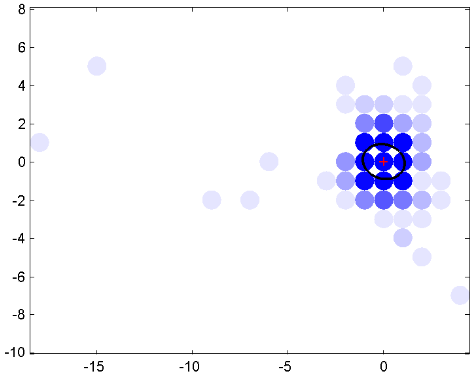

Figure 7. **Diraffe head frame-to-frame displacement**. The mean and one standard deviation contour of multiple giraffes in 361 are shown.

are 400. We use the Hungarian algorithm to determine the optimal association of head detections from the previous frame to the current frame. Track initializations occur when no detections in the previous frame are within 20 pixels of a detection, and the detection is matched to an element in the lower left quadrant. Similarly, terminations are identified by matches with a detection from the previous frame to an element in the upper-right quadrant, signifying there are no detections in the current frame within 20 pixels.

Multi-Frame Hungarian The two-frame tracking approach is very conservative: a single missed detection will justify including frame t in the final summary video. As a result, we also examine the multi-frame linear assignment method of [12]. In this alternative formulation, the entire set of detections $\mathcal{H}_{[1,T]}$ is considered. Here, an assignment $i \leftrightarrow j$ implies detections h^i and h^j are immediate predecessor/successor detections in a particular track. However, it is not strictly enforced that the two detections occur in frames $t-1$ and t respectively — tracks can continue while skipping frames with missed detections. In this formulation, there is now a $2n \times 2n$ cost matrix $\mathtt{C}^{multi}$ where $n = |\mathcal{H}_{[1,T]}|$. Similar to the two-frame formulation, values in $\mathtt{C}^{multi}$ encode the negative log probability $P(h^i \rightarrow h^j)$ of h^j being the next detection of the same object immediate followingly detection h^i. Like the two-frame tracking formulation, the spatial term is modeled using a 2D normal distribution of displacement between the centroids of the two detections. For the temporal term, we compute the elapsed time Δt between the two detections. The probability of h^j being the next immediate true detection after h^i needs to enforce temporal continuity and also take into account the detector recall r

$$P_{\textit{temporal}}(h^i \rightarrow h^j) = \begin{cases} 0 & \text{if } \Delta t <= 0, \\ (1-r)^{(\Delta t - 1)} & \text{otherwise.} \end{cases} \tag{3}$$

As a result, the combined probability is $P(h^i \rightarrow h^j) = P_{spatial}(h^i \rightarrow h^j) P_{temporal}(h^i \rightarrow h^j)$. Similarly, the probability of h^i begin a false detection is directly proportional to the detector's false detection rate $(1-p)$, where p is the detector's precision. We refer the reader to [12] for details on how each $\mathtt{C}_{ij}^{multi}$ is computed from the various probabilities.

In practice, it is not feasible to process all detections simultaneously — especially if the video is long. We make the problem tractable by spitting the video into five minute chunks (300 frames), and perform multi-frame Hungarian algorithm on each chunk. We then perform two-frame tracking at the boundary of each 300 frame chunk to stitch the solutions together.

5. Summarization

We use the initialization and termination of automatically generated tracks to determine when the giraffe head detector fails. We assume these failure cases arise because the giraffe is currently exhibiting unusual behavior, and therefore the head and neck are in an unusual spatial configuration relative to the body. Of course, the tracks will also terminate whenever a giraffe enters or exits the camera view, as well as when it is significantly occluded behind a tree or other obstruction. However, because there is no set limit to the length of the summary video, our formulation automatically adapts to the complexity of the scene: typical behaviors lead to long, connected tracks and a shorter a summary video, whereas more difficult scenes with many occlusions result in a longer summary.

Of course, there are other ways to summarize a video. In this study, we compare against two baseline methods. Similar to Khosla *et al.* [13], our first baseline is a uniform temporal sampling: one frame every fifth minute is included in the summary

$$\mathcal{F}_{\mathsf{uniform}} = \{300, 600, 900, \dots\}. \tag{4}$$

For our second baseline, we use motion history images [1] to gauge the amount of movement in each frame. Frames with significant motion history energy $MH(t)$ are added to the summary

$$\mathcal{F}_{\mathsf{motion}} = \{t \,|MH(t) \geq \lambda\}. \tag{5}$$

A motion history image is computed as an exponentially decaying summation of previous inter-frame difference images. The exponential decay parameter $\gamma = 30$ was determined by visually evaluating typical giraffe movements. The energy threshold $\lambda = 0.2$ was tuned by finding the largest value of λ that still retained all rare frames in the training data.

Method	Compression	Recall
Uniform	0.8007	0.2857
Motion History	0.8444	0.8571
Two-Frame Tracking	0.5203	**1.0000**
Multi-Frame Tracking	**0.9124**	0.7143

Table 3. **Summarization Performance**. We evaluate the four methods on a 51 minute test video containing 7 rare events. The two-frame tracking algorithm is able to detect all rare events, but has the least amount of compression. The multi-frame tracking algorithm generates the shortest summary video while maintaining good recall.

5.1. Evaluation

Because a human expert benefits from context, we extract a short window of ± 5 frames (equivalent to 5 seconds) around each frame of interest $f \in \mathcal{F}$.

Each manually annotated ground truth event (standing up, sitting down, bobbing head) spanned multiple frames. Therefore, we use a one-dimensional equivalent to PASCAL VOC [8] and compute intersection over union scores between temporal ranges to determine whether an automatically detected summary event $[f-5, f+5]$ adequately coincides with a ground truth event. Because some ground truth events span only a few frames, we use a threshold of 0, which corresponds to at least one frame of overlap. However, a larger intersection over union requirement may be suitable for higher frame rates.

A comparison of our two tracking methods to the two baselines are tabulated in Table 3 and visually summarized in Figure 8. The two-frame tracking algorithm is able to find all rare events, but generates the longest summary video ($\sim$25 minutes). The multi-frame tracking algorithm generates the shortest summary video ($\sim$5 minutes) and finds 5 of 7 events. Motion history does remarkably well for its computational efficiency. However, because the two-frame tracking algorithm is able to find all rare events, the multi-frame tracking algorithm should be able to obtain similar recall performance while keeping its high compression rate if more complex probabilistic models are incorporated, as well as correctly ignoring frames with significant non-giraffe related motion.

6. Summary

Although techniques for detecting objects in visual spectrum images can be applied to thermal images directly, we have found significant performance improvements when additional features are incorporated. For instance, absolute

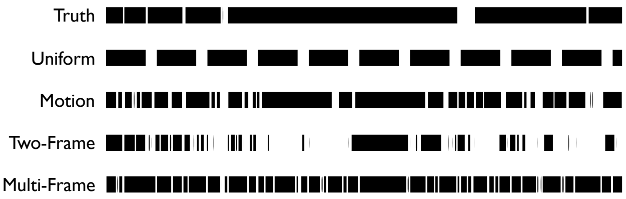

Figure 8. **Summarization Visualization**. The subset of frames $\mathcal{F}$ selected by each method to include in the summary are shown in white. The two-frame tracking method is able to recall all ground truth events of interest, but generates a very long summary video. The multi-frame tracking algorithm generates the most concise summary, while maintaining a fairly high recall.

gradient magnitude captures large intensity changes characteristic of the boundaries of animals in the scene, whereas HOG features provide fine-grain shape discrimination regardless of intensity values.

Giraffes are particularly difficult to detect because they have a large variation of permissible poses. Not surprisingly, a single rigid whole giraffe detector did not perform well in our experiments. Instead, we followed a similar approach to deformable part models, and trained rigid part detectors, and then reasoned about the existing of a complete object based on the local part evidence. In practice, we found a geodesic distance measure that searched for a suitable neck shape between a head and body drastically reduced the number of false head detections. Furthermore, enforcing one-to-one head/body matchings through a linear assignment problem avoided the pitfalls of successive greedy assignments.

Most of the giraffes in our training and testing videos appeared in sideways profile relative to the camera. However, the frontal view is noticeably different. Successfully dealing with this larger variation of deformations in an efficient manner is still an open problem.

Lastly, we have shown the viability of formulating video summarization as a track-by-detection problem. The approach naturally adapts the summary length based on the confidence of the detector and tracker, and the difficulty of the scene.

References

[1] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 2012. 7

[2] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *CVPR*, 2012. 2

[3] J. Cilulko, P. Janiszewski, M. Bogdaszewski, and E. Szczygielska. Infrared thermal imaging in studies of wild animals. *European Journal of Wildlife Research*, 2013. 2, 3

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 2, 3

[5] J. W. Davis and M. A. Keck. A two-stage template approach to person detection in thermal imagery. In *WACV*, 2005. 1

[6] P. Dollár. Piotr's Image and Video Matlab Toolbox (PMT). `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`. 3

[7] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 3

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *ICCV*, 2010. 7

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 1, 2, 3, 4

[10] R. Gade and T. B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision and Applications*, 2014. 2, 3

[11] S. Hare, A. Saffari, and P. H. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011. 2

[12] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*. 2008. 6, 7

[13] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 2, 7

[14] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. *ECCV 2014*, 2014. 4

[15] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 2008. 2

[16] S. P. Rajendra and N. Keshaveni. A survey of automatic video summarization techniques. *International Journal of Electronics, Electrical and Computational System*, 2014. 2

[17] D. Ramanan and D. Forsyth. Using temporal coherence to build models of animals. In *ICCV*, 2003. 2

[18] D. Ramanan, D. A. Forsyth, and K. Barnard. Detecting, localizing and recovering kinematics of textured animals. In *CVPR*, 2005. 2

[19] M. Rodriguez. Cram: Compact representation of actions in movies. In *CVPR*, 2010. 2

[20] P. Soille. Generalized geodesy via geodesic time. *Pattern Recognition Letters*, 1994. 5

[21] J. S. Supancic III and D. Ramanan. Self-paced learning for long-term tracking. In *CVPR*, 2013. 2

[22] D. Zhou. Thermal image-based deer detection to reduce accidents due to deer-vehicle collisions. 2013. 2

[23] T. T. Zin, R. Takahashi, and H. Hama. Robust person detection using far infrared camera for image fusion. In *Innovative Computing, Information and Control*, 2007. 3