INTEGRATING HUMAN CONTEXT AND OCCLUSION REASONING TO IMPROVE HANDHELD OBJECT TRACKING

Daniel Parks

University of Southern California
Neuroscience Graduate Program
Los Angeles, CA, USA

Laurent Itti

University of Southern California
Department of Computer Science
Los Angeles, CA, USA

ABSTRACT

Tracking an unknown number of various objects involving occlusion and multiple entry and exit points automatically is a challenging problem. Here we integrate spatial knowledge of human-object interactions into a high performing tracker to show that human context can further improve both detection and tracking. We use the DARPA Mind's Eye Action Recognition Dataset, which is comprised of street level scenes with humans interacting with handheld objects, to show this improvement. We find that human context can greatly reduce the number of false positive detections at the expense of increasing false negatives over a large test set (>230k frames). To minimize this, we add occlusion reasoning, where object detections are hallucinated when a human detection overlaps an object detection. These components together result in an average F_1 improvement of 107% per object category and a 69% reduction in track latency.

***Index Terms*—** Human Context; Object Recognition; Occlusion Reasoning

1. INTRODUCTION

Semantic context has been shown to be an important part of object recognition in humans [1]. An example of using inter-object context was shown in [2] (Figure 1). Although the target object is hard to identify in isolation, once it is put in context of one or more contextual objects, it becomes much easier. Prior models have shown how knowledge of the local background category (e.g. road, building, trees, etc.) can improve recognition of foreground objects (car, person) [3]. Humans, however, are able to use more than just the contextual object's identity to infer the target object's identity. For discrete objects (cars, people, balls, etc), the relative position and scale of the contextual object can also be used to help infer identity. These kinds of inter-object influences, we argue, become crucial when one object is more reliably detected than the other. For example, some objects, like plastic containers, have simple and regular edge structure, resulting in many edge groupings in an image that might falsely activate the detector. Since humans are composed of more complex contours, a human detector can be more reliable than detectors for objects with more generic contours. We show the importance of human-based contextual cues in improving object tracking and detection in street-level scenes.

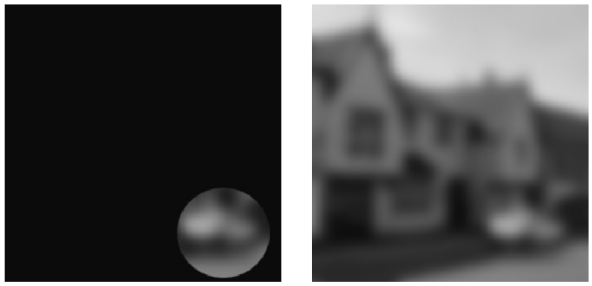


Fig. 1: [Taken from [2]] (a) The target object in isolation is hard to identify (b) With another more easily determined contextual object put in a stereotypical relative location with the target object, the identity of the target object becomes easier.

In DARPA's Mind's Eye Project the task was to identify and spatiotemporally localize actions occurring in video scenes (1-16 minutes each, mean 8 minutes) involving humans interacting with each other and with objects. This typically involves detecting and tracking the humans and objects in the videos. The scenes are challenging, because the humans and objects are frequently occluded by fixed pieces in the scene, and by other humans. The objects are also frequently carried and occluded by the human carrying the object. Further, it is important to track the object continuously over thousands of frames and across many instances of occlusion, to accurately determine the nature of the action. Finally, when determining what action is taking place, it is frequently important to know precisely when and where an object enters and leaves a scene, which is not typically measured when evaluating object detection and tracking performance.

We hand annotated humans and objects involved in actions across 14 videos (training=8 clips: 243,445 frames; test-

This work was supported in part by the DARPA Mind's Eye Program (W911NF-10-2-0063) and the National Science Foundation(CCF-1317433)

ing=6 clips: 231,567 frames), where a track was considered to extend until the object permanently exited the scene, or exited the scene for longer than at least 2 continuous seconds. If the object was occluded, the object was continuously tracked, unless the annotator could not estimate where it was, at which point the track was ended. The bounding box was set to be the smallest bounding rectangle that completely encloses the annotated object.

2. PRIOR WORK AND NOVEL CONTRIBUTIONS

Contextual information has been used to improve object detection in several ways. Scene context such as 3D scene information, scene category, and global image gist has been used to improve object detection on single images over the PASCAL dataset [4]. Models have improved the detection of objects (things) using nearby textural classification (stuff: grass, ocean, etc.) [3] as well as object co-occurence [5]. These previous models have not fully taken into account the spatiotemporal context that humans provide in a scene.

Recently, however, Zitnick and others built a spatial reasoning model between objects and humans in artificially generated scenes for activity recognition on a per activity basis [6]. We have developed their idea further by learning this over natural scenes, integrating this spatial reasoning with noisy detectors, and extending it with occlusion reasoning. This is difficult because, unlike artificially generated scenes, algorithmic detection of objects is inherently noisy and incomplete. Small, handheld objects can also be completely occluded or unrecognizable on a given frame, necessitating the use of temporal contiguity and contextual inference. Here, we take inspiration from [7], who used other human detections to reason about inter-human occlusion, and will do the same for human-object occlusions.

Our contributions are twofold: 1) learning where objects appear relative to humans and using this contextual knowledge to reduce false alarms; 2) reliably detecting object occlusions by humans and hallucinating objects at their predicted location during extended occlusion periods, to reduce track latency. On a large and very difficult video dataset, we show significant improvement by adding these two components to a state-of-the-art object detection and tracking framework.

3. BASE OBJECT TRACKING MODEL

Our system uses a state of the art pedestrian tracker [8] to track human locations. The Deformable Parts Model (DPM) [9] was used as the object detector for each small object (Bag, Briefcase, Gun, and Plastic-Container). The object detector, however, generates a large number of false positives across each image, some with high confidence, possibly because these objects lack distinctive textures (e.g., Figure 2). While DPM achieves high detection rates with low false alarms for more complex and textured objects like bicycles, cars,

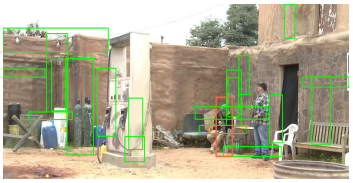


Fig. 2: Raw DPM gun detections. Red: First true positive; Green: all false alarms with DPM confidence higher than the true positive.

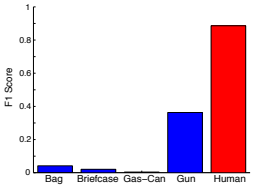


Fig. 3: F_1 score: Handheld objects had much lower scores compared to humans

etc., the objects studied here were difficult to identify with this detector alone. These false positives lead to low DPM F_1 (harmonic mean of precision and recall) performance (Figure 3).

To reduce false positives, especially in static background clutter, we applied background subtraction, color models, and frame to frame appearance tracking. The distance between the current detection and the background was calculated using a simple coarse binning histogram of the a* and b* components of the CIELAB color space. The background distance between the detection box histogram for the current frame and the first frame was then calculated using the L_2 norm. To reject false detections with clearly incorrect color, a clustered set of Gaussian color models are learned for each object category. The minimum L_2 distance between the detection color histogram and each Gaussian model is determined for the initial frame, and that model (e.g. "Bright Red Gas-Can") is set from that point on). The Track, Learn, Detect (TLD) appearance tracker [10] was initialized on the most promising resulting detections, and used to provide additional bounding box proposals.

Every frame, each DPM detection proposal is evaluated as a possible point to launch a tracker. A cost is calculated for each proposal based on the DPM detection confidence, the distance from the allowed Gaussian color models, the distance from the background, and the deviation from the expected relative size and position of the object given the human. Detections with costs below a threshold initiate a tracker from that point. Proposals that are too near the current best location of an existing tracker are ignored.

At a higher level, the Viterbi algorithm [11] is used to find the path with lowest cost through the set of box proposals for each frame. The cost function is a sum of exponentials:

$$cost(x) = \sum_i \alpha_i e^{\beta_i comp_i}$$

Each component of the model generates an unscaled cost, $comp_i$. Each component is then scaled by α_i and β_i which are discussed in the next paragraph. The cost function includes: the DPM detection confidence or TLD tracking confidence; the distance from the background; the distance from

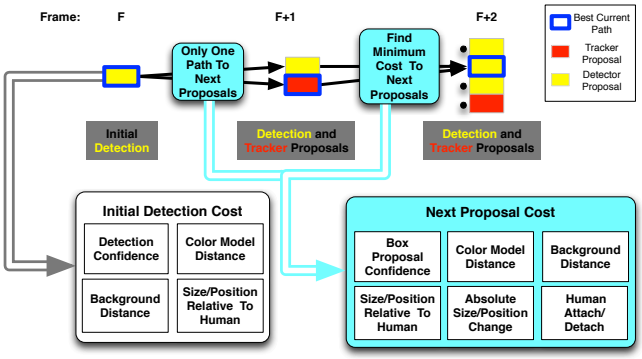


Fig. 4: The initial detection that spawned the tracker is shown at Frame F, which was calculated using the initial detection cost function. Moving away from the initial detection, a variable number of box proposals are evaluated for each frame. The lowest cost path from any prior box in the last frame is found for each current box using the next proposal cost.

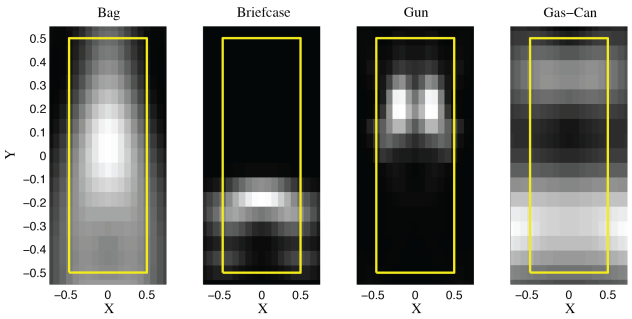


Fig. 5: Gray is neutral, white is high probability of object presence, and black is high probability of a false positive. The coordinates are normalized by the human width and length. The yellow outline is a normalized human having width and height equal to 1.

the expected color model; the change in position from the prior location in the path; and the change in size from the prior location in the path. After our Viterbi tracking system is run forward in the video, it is also run in reverse, and the tracks are continued backwards from their first frame. No new trackers are launched during the backward pass.

α_i and β_i parameters were learned using a direct maximization of the F_1 score. First, detection boxes box_{det} were checked with overlap of ground truth using Intersection over Union: $IOU = box_{det} \bigcap box_{gt} \div box_{det} \bigcup box_{gt}$, and if $IOU > 0.25$, it was considered a hit. Then the $comp_i$ for each detection was scored. Since some components rely on the prior box_{det}, the nearest detection hit (if present, otherwise just the nearest) from the prior frame was used. The optimal threshold $cost_{opt}$ to maximize the F_1 score for a given set of parameters was determined. The parameters were then optimized using: $\arg\max_{\alpha,\beta} = F_1(cost(box_{det}) < cost_{opt})$

4. HUMAN SUPPORT FOR OBJECTS

To enhance our tracking model, we added human context to improve the tracking results. The human detector performed much better than the handheld object detectors did, for a multitude of reasons, including that the objects were more often occluded, had less distinct shapes, and were trained using a standard object detection algorithm (DPM) instead of a custom built detector for human [8].

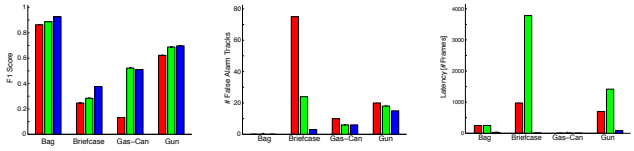
To harness these contextual cues, we learned the relative position and size of each small object relative to the human. Since we have the track of the human, we are also able to use the knowledge of which human the object is closest to (if any), and penalize paths that wander from human to human. Figure 4 depicts the complete system and the breakdown of the cost functions.

5. LEARNING HUMAN/OBJECT RELATIONSHIPS

The size and position of the object relative to the human were learned over the true positives and the false positives of the training data with left/right mirroring to generate more virtual examples. These densities were then used to provide a cost for a particular relative size and position of the object given a human detection. Both the distance from the object to the human and the scale of the object were normalized by the size of the human. Objects were frequently found in stereotyped locations of the human (Figure 5; plastic container carried at the end extent of the hand, rifle carried in the torso). Interestingly, the false positives were also occurring at stereotyped positions. For instance, the torso of the person was frequently detected as a plastic container, and the legs were frequently detected as a rifle. The inter-object occlusion was therefore useful in determining where an object was likely to be as well as what alternative explanations could exist for a particular detection to "explain away" that detection.

6. OCCLUSION REASONING AND MULTIPLE OBJECTS

In the Mind's Eye scenes, humans frequently occlude other humans and objects. During these times, there is little or no portion of the object visible, and the tracker and the object detector will be left with only false positive proposals. However, it is important to track the object across these occlusion events, to accurately generate the complete track of the object through the scene. Since many of these occlusion events involve other humans walking in front of the object, we can use the human tracks to estimate the likelihood that the object is occluded. If another human track overlaps the object, and the object detection is no longer present or much weaker, the probability of an occlusion is much higher. In these cases, a hallucinated detection is inserted, with the cost of the last



(a) F_1 Score (b) False Alarms (c) Track Latency

Fig. 6: Base System (red), with human context (green), with human context and occlusion reasoning (blue)

known good detection plus a hallucination penalty. This is allowed to continue until one of the following: the object is redetected at low cost, the occluding human is no longer occluding the position, or the maximum number of hallucinated frames has been exceeded. Once the hallucination is completed, if the object is not detected at low cost, the hallucination is rolled back and the tracker is ended. Otherwise, the tracker will continue as normal.

Since multiple humans and objects appeared in the scene, it was important to prevent trackers from oscillating between them as one becomes occluded and another is still visible. To prevent this, several steps are taken. First, a fixed human switch cost, ρ_{hs}, is applied every time an object switches to a different human track. Second, a Voronoi diagram is applied to the set of currently active trackers for a particular object type, based on their best current path, and a penalty, ρ_v is applied to all detections that are outside of a tracker's Voronoi region. Finally, hallucinated detections at the last known good location allows the best path for an occluded tracker to remain fixed and not migrate over to another tracker's location.

7. POST PROCESSING

We run a final pass through the results to find the best cut points based on the cost. During this process, spatiotemporally close tracks are merged, the bounding box path is smoothed with a simple fixed low pass filter, and weak tracks below a fixed threshold are removed. The best initial and average cost thresholds are found for each system when reporting the results.

8. SYSTEM RESULTS

We tried to improve both the detection and tracking performance of our system using human context and occlusion reasoning. To compare the different components of the system, we ran three versions of the model: a base system, one with human context, and one with both human context and occlusion reasoning. All three systems included a DPM object detector, a TLD tracker, a color model, and a background model. The per frame F_1 performance is shown in Figure 6(a), which improves on average by 94% with human context, but only 9% more with occlusion reasoning.

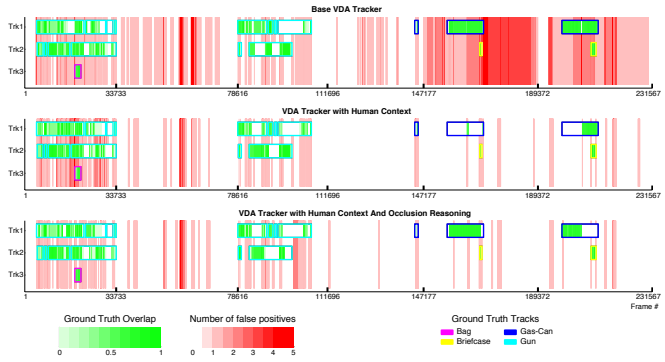


Fig. 7: Frame by frame performance of each system over whole test set. Up to 3 objects were present at any given time (Trk1 - Trk3). False positives from all detectors are shown.

To understand where the improvement is coming from, the change in false alarms for each system is shown in Figure 6(b). As more constraints are added, less false positives occur, at the expense of increased misses.

When trying to truly understand the interactions between a human and a set of objects, it becomes important to understand who brought the object into the scene and what happened to it (e.g., leaving a gun vs. leaving with a gun). These periods are frequently the most occluded points during the ground truth track, and don't count that much against a frame by frame detection score or tracking score, because they involve a relatively small number of frames at either end of the track. We measure these exit and entry point precision explicitly, and call it track latency. In Figure 6(c), we can see that although adding human context increased the average latency of the object detections by 97% (in many cases because the human wasn't detected yet), adding occlusion reasoning reduced the average latency 69% below the original system.

The results of the entire test set are shown in Figure 7, where the frames for each video are concatenated into a single set. Here we can see that adding human context lowers the false positives (the red heatmap), and the occlusion reasoning restores some of the detections lost when the human context was added.

9. CONCLUSION

Although context makes intuitive sense for use in improving object recognition, it has been difficult to show significant performance improvement using it. Here we show that in real-world cases where the object detector has low precision but humans are detected reliably, human context can play a sizable role. We further show that occlusion reasoning can significantly reduce track latency, which is important for determining the origin and fate of the object.

10. REFERENCES

[1] Aude Oliva and Antonio Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, Dec. 2007.

[2] Antonio Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, July 2003.

[3] Geremy Heitz and Daphne Koller, "Learning spatial context: Using stuff to find things," in *Computer Vision ECCV 2008*, David Forsyth, Philip Torr, and Andrew Zisserman, Eds., number 5302 in Lecture Notes in Computer Science, pp. 30–43. Springer Berlin Heidelberg, Jan. 2008.

[4] S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, and M. Hebert, "An empirical study of context in object detection," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, June 2009, pp. 1271–1278.

[5] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet, "Multi-class object localization by combining local contextual interactions," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 113–120.

[6] C. Lawrence Zitnick and Devi Parikh, "Bringing semantics into focus using visual abstraction," *Computer Vision and Pattern Recognition*, 2013.

[7] Bo Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, Oct. 2005, vol. 1, pp. 90–97 Vol. 1.

[8] Bo Wu and Ram Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, Nov. 2007.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, June 2008, pp. 1–8.

[10] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, July 2012.

[11] G. David Forney Jr, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268278, 1973.