Part-based Multiple-Person Tracking with Partial Occlusion Handling

Guang Shu[1], Afshin Dehghan[1], Omar Oreifej[1], Emily Hand[2], Mubarak Shah[1]
[1]Computer Vision Lab, University of Central Florida
[2]Computer Vision Lab, University of Nevada, Reno
{gshu, adehghan, oreifej}@eecs.ucf.edu, ehand@cse.unr.edu, shah@eecs.ucf.edu

Abstract

Single camera-based multiple-person tracking is often hindered by difficulties such as occlusion and changes in appearance. In this paper, we address such problems by proposing a robust part-based tracking-by-detection framework. Human detection using part models has become quite popular, yet its extension in tracking has not been fully explored. Our approach learns part-based person-specific SVM classifiers which capture the articulations of the human bodies in dynamically changing appearance and background. With the part-based model, our approach is able to handle partial occlusions in both the detection and the tracking stages. In the detection stage, we select the subset of parts which maximizes the probability of detection, which significantly improves the detection performance in crowded scenes. In the tracking stage, we dynamically handle occlusions by distributing the score of the learned person classifier among its corresponding parts, which allows us to detect and predict partial occlusions, and prevent the performance of the classifiers from being degraded. Extensive experiments using the proposed method on several challenging sequences demonstrate state-of-the-art performance in multiple-people tracking.

1. Introduction

The goal of our work is to automatically detect and track each individual target in a crowded sequence. Several challenges render this problem very difficult: First, the appearance of the target is often constantly changing in the field of view of the camera. Second, targets often exit the field of view and enter back later on; thus, a successful tracker needs to associate the two observations. Third, targets often become occluded by other targets or by objects in the scene. Therefore, traditional trackers [19, 7] suffer in such scenarios. On the other hand, discriminative tracking approaches with online learning have flourished recently [1, 2, 11, 6, 10, 8]. In such methods, an specific detector is trained in a semi-supervised fashion and then used to locate the target in consecutive frames. However, the online learned detector will often drift in long-term tracking. Additionally, such algorithms do not handle multiple targets. Therefore, several techniques [23, 13, 3, 5] were proposed to tackle multi-target tracking by optimizing detection assignments over a temporal window, given certain global constraints. Such methods employ offline trained detectors to find the targets and associate them with the tracks. Although they can handle several difficulties such as the uncertainty in the number of targets, occasional occlusions, and template drift in long term; they still suffer when faced with appearance changes and occlusion. In particular, when tracking a crowd of pedestrians, the data association often fail in the aforementioned approaches due to pose variations, partial occlusions and background changes.

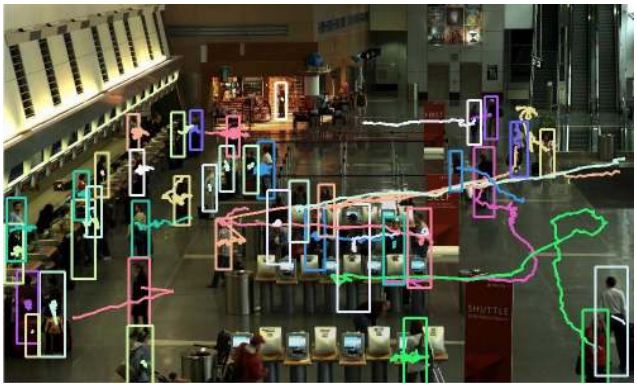

Figure 1. Multi-person tracking using our part-based tracker.

In this work, we address such difficulties by proposing a part-based representation in a tracking-by-detection framework. While the deformable part-based model [9] has shown excellent performance in static images, yet it has not been fully explored in tracking problems. Moreover, the availability of inexpensive high-definition sensors for surveillance yet provides the advantage to exploit such detailed appearance representations.

Current tracking-by-detection methods use the final detection window as an observation model. In contrast, in our approach, we leverage the knowledge that all targets of

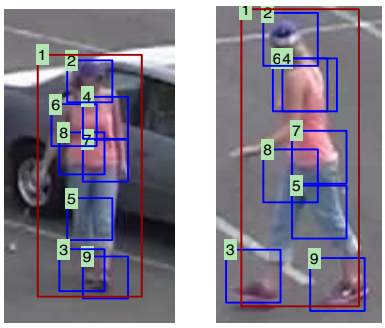

Figure 2. The part-based representation allows detailed correspondence between the articulated human bodies. The figure demonstrates how the parts of two instances of a human can be well-corresponded even with significant pose changes. Additionally, the background components in the detection boxes are automatically excluded from the correspondence.

one class (humans in this context) have similar part structure; thus, we employ the sets of detected parts as observation models. Therefore, our method provides several advantages: First, the combination of parts provides a rich description of the articulated body; thus, it represents the human better than a single detection window. In particular, since the spatial relations of the parts in an articulated body are often flexible, corresponding targets using a holistic model (one detection box) is error-prone and may compare dissimilar parts of the body. In contrast, a part-based representation allows parts to be strictly compared to their corresponding parts. An example is shown in figure 2, where the parts of two instances of a person are well-corresponded even with different poses and backgrounds. Second, the part-based model excludes most of the background within the detection window and thus avoids the confusion from background changes. Finally, since the part-based detector is offline trained by latent SVM using a large amount of training samples, it captures significant amount of discriminative information, which is essential for tracking.

Our tracking framework consists of the steps illustrated in figure 3. First, we use an extended part-based human detector on every frame and extract the part features from all detections. Person-specific SVM classifiers are trained using the detections, and consequently used to classify the new detections. We use a greedy bipartite algorithm to associate the detections with the trajectories where the association is evaluated using three affinity terms: position, size, and the score of the person-specific classifier. Additionally, during tracking, we reason about the partial occlusion of a person using a dynamic occlusion model. In particular, partial occlusions are learned by examining the contribution of each individual part through a linear SVM. This inferred occlusion information is used in two ways: First, the classifier is adaptively updated with only the non-occluded

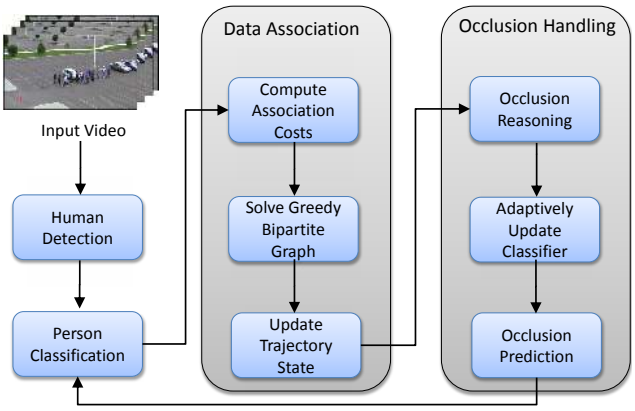

Figure 3. The various steps of our approach.

parts, which prevents from being degraded along the occlusion period. Second, the discovered occlusion information is passed to the next frame in order to penalize the contribution of the occluded parts when applying the person classifier.

In summary, this paper makes the following contributions: First, we adopt the part-based model in multi-target tracking to tackle occlusion and appearance change problems. Second, we extend the human detection approach in [9] which allows us to improve the detection in crowded scenes. Third, we propose a dynamic occlusion handling method to learn and predict partial occlusions, and thus improve the tracking performance.

2. Related Work

A significant amount of work has been reported for multi-target tracking-by-detection algorithms. In [23], Zhang et al. resolve the association between the detection and the tracking by optimizing a cost-flow network with a non-overlap constraint on trajectories. Brendel et al. [5] apply a maximum-weight independent set algorithm to merge small tracklets into long tracks. In that, information from future frames are employed to locate the targets in the current frame with a temporal delay. In contrast, we employ a greedy scheme in data association, which is more suitable for online tracking applications. On the other hand, several methods such as [15, 22, 14] employ social force models which consider the interactions between the individuals in order to improve the motion model. Such methods require prior knowledge of the 3D scene layout, which, however, is often unavailable in practical scenarios. Moreover, Benfold et al. [3] use MCMCDA to correspond the detections obtained by a HOG-based head detector in crowded scenes. Although they demonstrate promising results, using only the head is often not discriminative in various tasks.

The method proposed by Breitenstein et al. [4] is evidently the most similar to ours. In that, they propose a particle-based framework in which detections and interme-

diate detections' confidences are used to propagate the particles. Additionally, they employ a target-specific classifier to associate the detections with the trackers. Our method is different than [4] in that we employ a part-based model which is more robust and can handle partial occlusions. On the other hand, Wu et al. [21] train an edgelet-based part detector to track multiple persons by matching the parts using color features. Such method is, however, not as discriminative as our proposed online-learned classifiers which employ multiple features for reliable data association.

3. Human Detection with Occlusion Handling

We employ a deformable part-based model for human detection similar to [9]. However, such detector suffers when the human is occluded. In particular, the final score in [9] is computed from all the parts, without considering that some parts can often be occluded. Let H be the HOG feature of the image, and $p = (x, y)$ denotes a part specified by its position. The detection score at location (x_o, y_o) is computed in [9] as

$$score(x_o, y_o) = b + \sum_{i=1}^{i=n} s(p_i), \tag{1}$$

where b is a bias term, n is the number of parts, and $s(p_i)$ is the score of part i which is computed as

$$s(p_i) = F_{p_i} \cdot \phi(H, p_i) - d_{p_i} \cdot \phi_d(d_x, d_y), \tag{2}$$

where F_{p_i} is the part filter, and $\phi(H, p_i)$ denotes the vector obtained by concatenating the feature vectors from H at the subwindow of the part p_i. (d_x, d_y) is the displacement of the part with respect to its anchor position, $\phi_d(d_x, d_y) = (d_x, d_y, d_x^2, d_y^2)$ represents the deformation features, and d_{p_i} specifies the coefficients of the deformation features. Under this formulation, it is clear that even if the part was occluded, its corresponding score still contributes in the final detection score. This is a significant drawback especially when dealing with crowded sequences as shown in figure 4. In the figure, some humans appear fully in the image; however, several humans appear as only upper parts, or even only heads. Such impediment in [9] led previous works such as [3] and [18] to rely on only head detection and ignore the rest of the body. To address this problem, we propose to infer occlusion information from the scores of the parts and consequently utilize only the parts with high confidence in their emergence. Instead of aggregating the scores from the set of all the parts $P = \{p_0 \dots p_n\}$, we select the subset of parts $S = \{p_k \dots p_l\} \subseteq P$, which maximizes the detection score

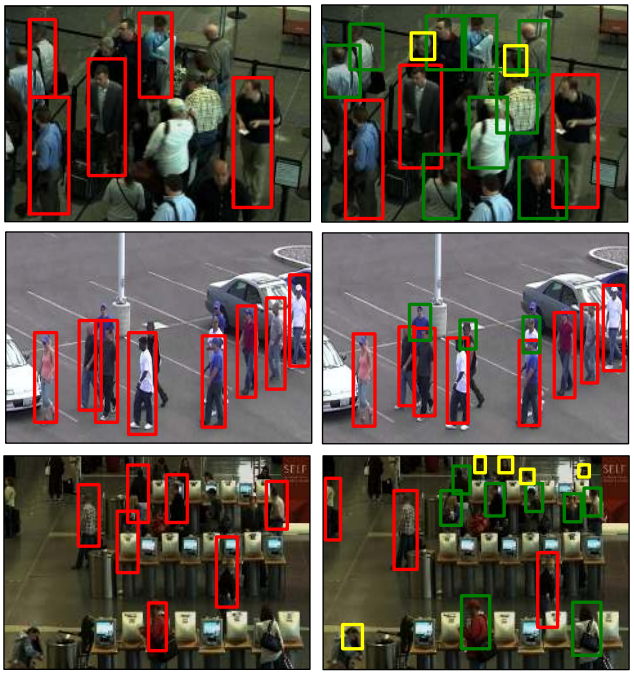

Figure 4. Left: Human detection results using [9]. Right: Human detection results using our approach where red boxes show the humans detected as full bodies, green boxes show the humans detected as upper bodies, and yellow boxes show the humans detected as heads only. It is clear that [9] failed to detect occluded humans since it does not have an explicit occlusion model, while our approach detects the occluded parts and excludes them from the total detection scores, thus achieving significant improvements especially in crowded scenes.

$$score(x_o, y_o) = b + \arg\max_{S_m} \frac{1}{|S_m|} \times \sum_{i \in S_m} \frac{1}{1 + \exp(A(p_i) \cdot s(p_i) + B(p_i))}, \tag{3}$$

where $|S_m|$ is the set cardinality, and the sigmoid function is introduced to normalize the scores of the parts. The parameters A and B are learned by the sigmoid fitting approach in [16]. Note that equation 3 corresponds to the average score of the parts in the subset. Since the average is sensitive to outliers, it is useful in capturing miss-detected parts. In other words, a subset S_m which contains occluded parts is likely to have less average score than a subset without occluded parts. Therefore, by maximizing equation 3 we obtain the most reliable set of parts and its corresponding probability of detection, which we use as the final detection score. We consider only three possible subsets of parts, namely, head only, upper body parts, and all body parts. We found such subsets representative enough for most realistic scenarios. Therefore, we do not need to search over all possible 2^n part combinations; instead, solving equation 3

involves only three evaluations which is a negligible overhead to the standard approach. Figure 4 demonstrates the advantage of our human detector over [9] in detecting occluded humans.

4. Tracking with Occlusion Handling

4.1. Person Classification

We train an online person-specific classifier for each individual target. In each frame, the human detections are classified by the person classifiers. Relevant previous work mostly used Adaboost classifier with Haar-like features; in contrast, our approach leverages the detected human parts and train a SVM classifier. We extract features from each individual part, and then concatenate them in a fixed order as a feature vector. We choose color histogram and Local Binary Pattern as features because they are highly discriminative for individuals and are complementary to the HOG feature which is used in the human detector.

The classifier is trained using the detections included in each trajectory. In particular, the positive examples are taken from all detections in the trajectory and the negative examples are taken from the detections of the other trajectories augmented with random patches collected from the surrounding background in order to improve the classifier's discrimination to the background.

4.2. Data Association

Many tracking applications require online forward tracking, i.e. the current trajectories should depend only on previous frames, not on future observations. To meet such requirements, we use a first-order Markov model in data association, in which trajectories are continuously growing as the tracking proceeds. In every frame, the detections are associated with existing trajectories by a greedy bipartite assignment algorithm [17] which has also been used in [4, 21]. In particular, for each frame, we construct an affinity matrix M for the trajectories and the detections. Consequently, the pair with the maximum affinity $M_{i,j}$ is selected as a match, and the i-th row and the j-th column are deleted from M. This procedure is repeated until no more pairs are available.

To evaluate the affinity of a trajectory i and a detection j, we use

$$M(i,j) = C(i,j) \cdot E(i,j) \cdot Z(i,j) \tag{4}$$

where three terms C is the output of the person-specific classifier, E and Z are affinities of position and size respectively. We calculate E and Z using similar methods to [21].

The detections which are not associated with any existing trajectories are used to initialize a new potential trajectory. Once the length of a potential trajectory becomes larger than a threshold, it gets formally initialized. On the other hand, when a new detection is associated to a trajectory, we update all its state variables, namely, the position, the size, the velocity, based on the new detection. However, when there is no detection associated due to occlusion or miss-detection, we use a correlation-based Kalman filter to track the head part of the target in the local area. This heuristic is particularly useful in crowded scenes where only humans' heads are observed.

4.3. Dynamic Occlusion Handling

If a partially occluded person is detected and associated to a trajectory, the classifier will be updated with noise and its performance will gradually degrade. Therefore, we employ an occlusion reasoning method to handle this problem. It was shown in [20] that in a detection window, occluded blocks respond to the linear SVM classifier with negative inner products. We adopt this approach to infer which parts are occluded for those detections with low classifier score. Assume that the detection's feature vector $\mathbf{x}$ consists of n sub-vectors corresponding to n parts, written as $\mathbf{x} = (\mathbf{s_1}, \ldots, \mathbf{s_n})$. The decision function for linear SVM classifier is

$$f(x) = \beta + \sum_{k=1}^{l} \alpha_k \langle \mathbf{x}, \mathbf{x_k} \rangle = \beta + \mathbf{x}^T \mathbf{w}, \tag{5}$$

where $\mathbf{x_k}$ is a support vector and $\mathbf{w}$ is the weighted sum of all support vectors. We can divide $\mathbf{w}$ to n sub-vectors $\mathbf{w} = (\mathbf{w_1}, \ldots, \mathbf{w_n})$, and find a set of $\{\beta_i\}$ such that $\beta = \sum \beta_i$, then the separate contribution of each part is represented by

$$f(\mathbf{s_i}) = \beta_i + \mathbf{s_i}^T \mathbf{w_i}. \tag{6}$$

Each time we re-train a person-classifier, we calculate β_i similar to [20] using the previously observed training samples. Consequently, we obtain the score for each individual part of $\mathbf{x}$. The part with a negative score is mostly likely to be occluded. Therefore, we adaptively update the classifier by only extracting features from the parts with high confidence which are likely to correspond to the non-occluded parts, while the features for the occluded parts are obtained from the feature vectors of the previous frames. Using this technique, the occluded parts will not be included in updating the classifier. Figure 5 demonstrates how occluded parts have negative responses to the SVM.

On the other hand, the occlusions are highly correlated in the adjacent frames of videos. Hence, when a partially occluded part is detected in one frame, it will have a high probability of being occluded in the consecutive frames. We harness such smoothness in occlusion to improve the classification performance by introducing an occlusion prediction method into the data association in order to improve the accuracy. First, we map the part SVM scores into positive values by

$$G(s_i) = \exp(\frac{f(s_i)}{\delta}), \tag{7}$$

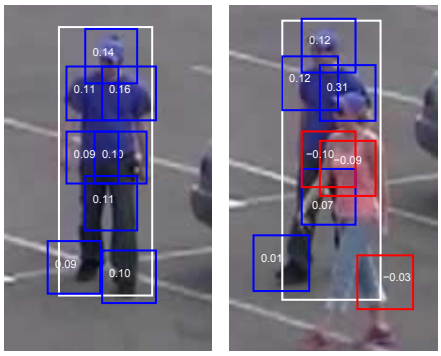

Figure 5. Left: a certain target human is detected, non of his parts are occluded (shown in blue); thus, all parts have positive responses. Right: the same target is observed again in another frame; however, three of his parts are occluded by another person (occluded parts are shown in red). The occluded parts have negative responses as shown in the figure. In our framework, the occluded parts are excluded when updating the person's model. Additionally, when performing person classification, the occluded parts have lower contribution since they are assigned weights relative to their score (refer to equation 8).

where δ is a constant. The occluded parts will likely have lower values as demonstrated in the example in figure 5. Therefore, when performing person classification, we weight each part by the non-occlusion confidence $G(s_i)$ and normalize the total score

$$C = \frac{n \sum_{i=1}^{n} G(s_i)(\beta_i + s_i^T w_i)}{\sum_{i=1}^{n} G(s_i)}. \quad (8)$$

In the following frames, we examine the occlusion again and update the occlusion confidence until the classifier score is larger than a threshold. This allows the occlusion information to be passed across continuous frames, and the person classifier to have higher weight corresponding to non-occluded parts.

5. Experimental Results

We extensively experimented on the proposed method using Oxford Town Center dataset [3], and two new datasets that we collected; the Parking Lot dataset, and the Airport dataset. The experimental datasets provide a wide range of significant challenges including occlusion, crowded scenes, and cluttered background. In all the sequences, we only use the visual information and do not use any scene knowledge such as the camera calibration or the static obstacles. It is important to notice that we selected the aforementioned datasets since they include high quality imagery which is more suitable to our approach since the part-based model requires detailed body information.

In our implementation, we use the pre-trained pedestrian model with 8 parts from [9]. The feature vector for each part consists of 125-bin RGB color histogram using 5 bins for each channel and 59-bin LBP histogram. We apply normalization for each part and concatenate all 8 parts into one feature vector of 1472 dimensions. The training data for each person-specific classifier consists of up to 100 positive samples and 100 negative samples. When the number of collected samples exceeds this limit, we delete the oldest ones to ensure the model is up to date. Aside from the human detection, our tracker runs at 1 to 5 fps on a conventional desktop, depending on the number of humans in the sequence. Figure 6 shows example frames from the experiments' sequences with the tracking results overlaid. Additionally, figure 7 shows example results of our dynamic occlusion handling.

Table 1. Tracking results on The Town Center Dataset.

	TP	TA	DP	DA
Benfold et al. [3]	80.4	64.8	80.5	64.9
Zhang et al. [23]	71.5	65.7	71.5	66.1
Pellegrini et al. [15]	70.7	63.4	70.8	64.1
Yamaguchi et al. [22]	70.9	63.3	71.1	64.0
Leal-Taixe et al. [14]	71.5	67.3	71.6	67.6
Ours/detection from [9]	71.1	72.2	71.2	72.7
Ours/our detection	71.3	72.9	71.4	73.5

We evaluate our tracking results using the standard CLEAR MOT metrics [12], TP (tracking precision), TA (tracking accuracy), DP (detection precision) and DA (detection accuracy). Note that TP only measures the precision of tracked object positions, but TA measures false negatives, false positives, and ID-switches. Therefore TA has been widely accepted as the main gauge of performance of the tracking methods.

Town Center Dataset: The frame resolution in this dataset is 1920×1080, and the frame rate of 25 fps. This is a semi-crowded sequence with rare long-term occlusions. The motion of pedestrians is often linear and predictable. In table 1, we compare results with [3, 23, 15, 22, 14] using the results reported in [14]. With the same experimental settings, our method significantly outperforms all previous methods in TA. The improvement in our method is a result of two main factors: First, the part-based model could better represent the articulated body and thus improves the accuracy in data association. Second, our dynamic occlusion handling module allows us to robustly track partially occluded humans.

Parking Lot Dataset: The frame resolution in this dataset is 1920×1080, and the frame rate of 29 fps. This is a modestly crowded scene including groups of pedestrians walking in queues. The challenges in this dataset include long-term inter-objects occlusions, camera jittering, and similarity of appearance among the humans in the scene. The tracking results for the Parking Lot dataset are summarized in table 2.

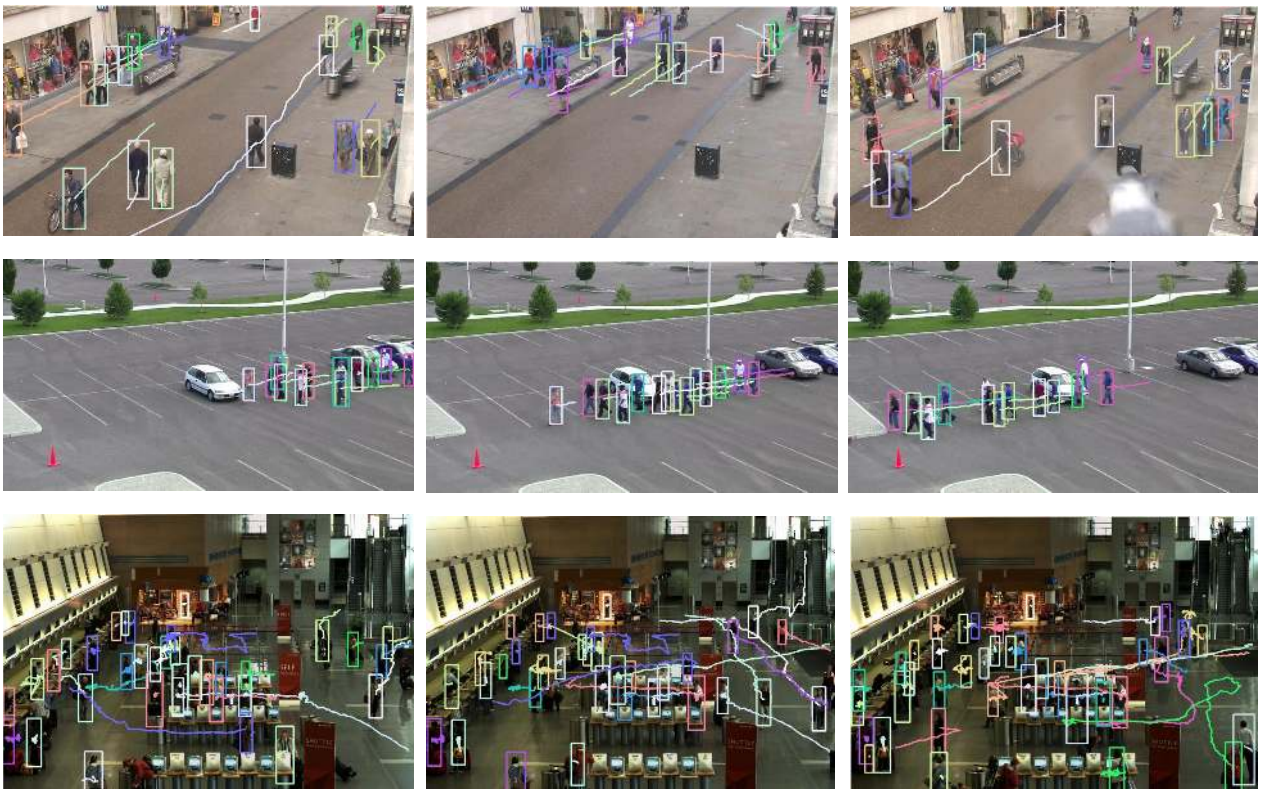

Figure 6. Example tracking results using our method. Top row shows the Town Center sequence, middle row shows the Parking Lot sequence, and the bottom row shows the Airport sequence.

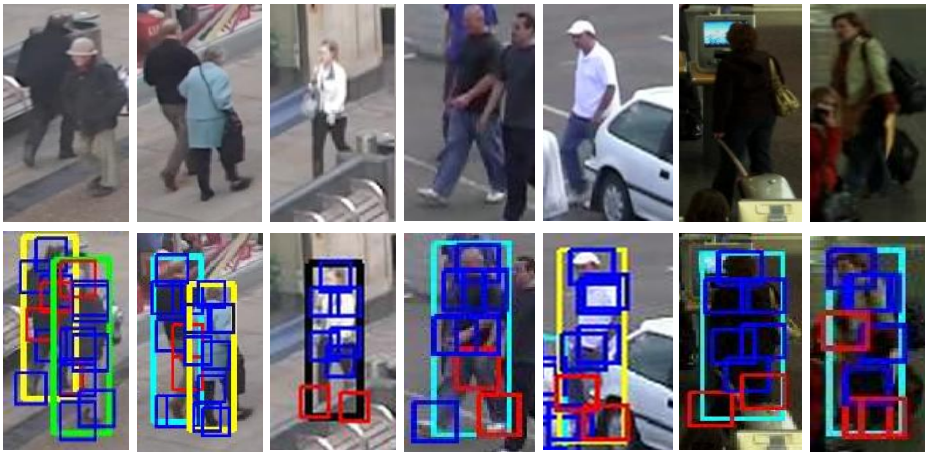

Figure 7. Examples results of our dynamic occlusion handling approach. Top row shows the original image, and bottom row shows the detected humans are their corresponding parts, where the occluded parts shown in red.

Table 2. Tracking results on the Parking Lot dataset.

	TP	TA	DP	DA
Ours/detection from [9]	73.7	77.1	73.8	77.5
Ours/our detection	74.1	79.3	74.2	79.8

Airport Dataset: The frame resolution in this dataset is 4000×2672, and the frame rate of 5 fps. This is a very challenging real world scene with severe occlusions resulting from both static obstacles in the scene and inter-person occlusions. Additionally, the humans' appearance and pose significantly change along the sequence because of the wide field of view of the camera and the low frame rate. However, our approach still achieved promising results on this dataset. The tracking results for the airport dataset are shown in table 3. Note that TA is significantly higher using our detection than using [9] on this dataset since it is more crowded, and thus occlusions occur very frequently.

Finally, we analyzed the performance of our approach

Table 3. Tracking results on the Airport Sequence.

	TP	TA	DP	DA
Ours/detection from [9]	66.1	27.2	66.3	28.4
Ours/our detection	67.2	52.2	67.4	53.6

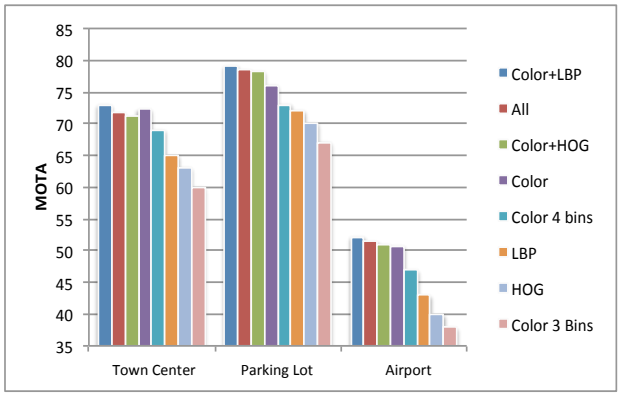

Figure 8. The performance of our tracking approach using different feature combinations.

using different feature combinations. Figure 8 demonstrates the obtained results. Color and LBP are evidently the most distinguishing features for all datasets.

6. Conclusion

We proposed an effective multiple-person tracking method using part-based model and occlusion handling. Our method captures rich information about individuals; thus, it is highly discriminative and robust against appearance changes and occlusions. We employ an extended part-based human detector to obtain human part detections. Consequently, distinguishing person-specific classifiers are trained using the parts' features and then employed to associate the detections with the tracking. We handle partial occlusions through dynamic occlusion reasoning and prediction across frames. We demonstrated by experiments that our tracking method outperforms state-of-the art approaches in crowded scenes.

Acknowledgment

This research is supported by the Pacific Northwest National Laboratory (PNNL) in Richland, Washington. We also thank Vladimir Reilly, Imran Saleemi and Hamid Izadinia for the exchange of ideas that helped improve this article.

References

[1] S. Avidan. Ensemble tracking. In *PAMI*, 2010.

[2] B. Babenko, M. Yang, and M. Hsuan. Visual tracking with online multiple instance learning. In *PAMI*, 2010.

[3] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011.

[4] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

[5] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *CVPR*, 2011.

[6] R. Collins and Y. Liu. Online selection of discriminative tracking features. In *ICCV*, 2003.

[7] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. In *PAMI*, 2003.

[8] T. B. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR*, 2011.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. In *PAMI*, 2010.

[10] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *BMVC*, 2006.

[11] Z. Kalal. P-n learning?: Bootstrapping binary classifiers by structural constraints. In *ICCV*, 2010.

[12] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation for face, text and vehicle detection and tracking in video: data, metrics, and protocol. In *PAMI*, 2009.

[13] C. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, 2010.

[14] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV Workshop on Modeling, Simulation and Visual Analysis of Large Crowds*, 2011.

[15] S. Pellegrini, A. Ess, , and L. van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, 2010.

[16] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

[17] K. Rangarajan and M. Shah. Establishing motion correspondence. In *Graphics and Image Processing: Image Understanding*, 1991.

[18] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, 2011.

[19] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994.

[20] X. Wang and T. Han. An hog-lbp human detector with partial occlusion handling. In *ICCV*, 2009.

[21] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. In *International Journal of Computer Vision*, 2007.

[22] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *CVPR*, 2011.

[23] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008.