Colour and Feature Based Multiple Object Tracking Under Heavy Occlusions

Pabboju Sateesh Kumar, Prithwijit Guha and Amitabha Mukerjee

Computer Vision Group, IIT Kanpur
Kanpur - 208016, UP, India
E-mail: {psateesh,pguha,amit}@iitk.ac.in

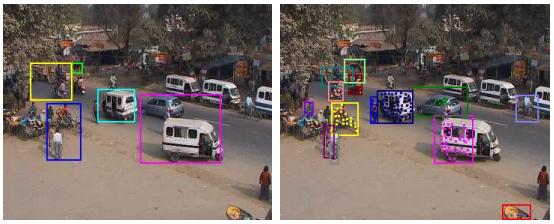
Tracking multiple objects in surveillance scenarios involves considerable difficulty because of occlusions. We report a composite tracker - based on feature tracking and colour based tracking - that demonstrates superior performance under high degrees of occlusion. Disjoint foreground blobs are extracted by using change masks obtained by combining an online-updated background model and flow information. The state of occlusion/isolation is identified by associating foreground blobs with object regions predicted using motion initialized mean-shift tracker (colour cue). The feature tracker is invoked in occluded situations to localize these with higher accuracy. We present results from dense traffic data with 5-15 objects in the scene at any instant. Overall tracking accuracy improves to 94.7% from 85.3% achieved by the colour only tracker.

Keywords: Tracking, Occlusions, Feature Correspondence

1. Introduction

An algorithm for tracking multiple agents in a monocular surveillance setup is reported. An early approach to this problem deals with tracking blobs obtained from the process of background subtraction.[1] However, such blobs do may form a group and get detected as a single blob or an agent can be detected as multiple blobs due to occlusions. The W4 system[2] differentiates people from other objects by shape and motion cues and tracks them under occlusions by constructing appearance models and detecting body parts. Several researchers[3] have employed particle filtering along with prior shape and motion models for multi-person tracking in cluttered scenes. Recently, Zhao *et al.*[4] proposed a Bayesian approach for tracking multiple persons under occlusions by computing MCMC based MAP estimates with prior information about camera model and human appearance along with a ground plane assumption. McKenna et al.,[5] on the other hand, presents a colour based tracking algorithm that performs in relatively unconstrained environments and works at three levels of abstraction, viz. regions, people and groups.

In the recent past, a number of approaches have distinguished the types of occlusion, treating these as a source of additional information to be used in the visual analysis.[6] This work builds on this approach, and uses the occlusion type to guide the tracking. We present a hybrid approach based on feature and colour based tracking, and contrast this with other approaches involving only colour.



(a) (b)

Fig. 1. Results of multi-object tracking using (a) only colour cue and (b) both colour match and feature correspondences. The scene contains a total of 16 objects, out of which 10 appear in crowds. (a) The colour based tracker properly localizes only 5 objects; (b) The composite colour-feature based tracker successfully localizes 14 objects in the scene.

The algorithm works by learning a background model as a pixel-wise mixture of Gaussians, change masks on which along with inter-frame motion information segments the objects as foreground blobs.[6] The object is characterized by its supporting region, weighted colour distribution, trajectory and a planar graph constructed by Delanuay triangulation of the feature point set extracted in its supporting region. The system maintains a set of objects to (from) which objects are added (removed) as they enter (exit) the scene. We identify the objects to be either *isolated* or *occluded* by associating motion initialized mean shift tracker predicted object regions with foreground blobs. More so, the dissociated object/foreground regions are detected to identify their *disappearances*/*reappearances*. The occluded objects are further tracked with higher accu-

racy by the feature-point graph structure constrained feature correspondences. The object features are selectively updated based on their occlusion states.

Some salient strengths of the proposed scheme are the following. First, the ability to identify occlusion states and using the same in selective feature updates. Second, the inherent ability of recognizing failure situations and automatic track restorations and finally, a relatively unconstrained approach that does not assume any priors on object shape, motion models and ground plane. Figure 1 shows the performance improvement achieved by the hybrid colour-feature based tracking algorithm as compared to a colour only tracker.

The colour based multi-agent tracking algorithm and its extension to invoke feature correspondences is presented in section 2. Experimental results on a dense traffic video with ground-truth validation are reported in section 3.

2. Multiple Object Tracking

The object regions are segmented as a set of disjoint foreground blobs extracted by combining the cues derived from the change masks over the learned background models (pixel-wise mixture of Gaussians) and the inter-frame motion information. The object regions predicted by motion initialized mean shift trackers[7] are associated with the extracted foreground blobs[6] to detect the objects in either the state of *isolation* from other objects and background elements or the state of *occlusion* arising due to crowding and partial occlusions. Additionally, the cases of *entry/exit* and *disappearance/reappearances* are also identified. These occlusion cases guide the tracking algorithm in selective object feature updates and track restoration. The system maintains a set $\mathcal{S}(t)$ of objects, to (from) which objects are added (removed) as they enter (exit) the scene. The individual object features are updated as they are tracked across the frames. When an unmatched foreground blob detected, It is matched with disappeared objects based on color and position matching. Search region around each disappeared object is taken into consideration while matching with unmatched foreground blob. In the following sub-sections, we detail the limitations of colour based multi-object tracking algorithm[6] and the proposed extension to combine colour and feature based tracking.

2.1. *Tracking With Colour Cue*

The j^{th} object $\mathcal{A}_j(t)$ is characterized by the set of pixels $a_j(t)$ it occupies, the colour distribution $h_j(t)$ weighted by the Epanechnikov kernel[7] supported over the minimum bounding ellipse of $a_j(t)$ and the finite length position history of the centers $\{\mathbf{c}_j(t-t')\}_{t'=0}^{\tau-1}$ of the minimum bounding ellipse of $a_j(t)$. The object features are initially learned from the foreground blob extracted at its very first appearance and are updated throughout the sequence whenever it is in isolation.

An estimate of the center $\mathbf{c}_j^{(0)}(t)$ of the minimum bounding ellipse of $a_j(t)$ is obtained by extrapolating from the trajectory $\{\mathbf{c}_j(t-t')\}_{t'=1}^{\tau}$. The mean-shift iterations,[7] initialized at an elliptic region centered at $\mathbf{c}_j^{(0)}(t)$ further localize center of the minimum bounding ellipse of the object region at $\mathbf{c}_j(t)$.

The object region and foreground blob associations are computed to identify the various occlusion states. The supporting pixel set, weighted colour distribution and trajectory information are updated for isolated objects. For the occluded ones (same (foreground) object pixel in different (object regions) foreground blobs), we only update the trajectory. More so, we identify the dissociated objects (disappearance) and blobs (entry/reappearance) followed by object-blob association re-computation to restore tracks of the existing ones and log the new objects in $\mathcal{S}(t)$.

The colour only tracker employ's the mean-shift algorithm for object localization and is thus prone to erroneous drifts in the mean-shift iterations. The mean shift algorithm models the target as a weighted colour distribution learned over an elliptical domain. Thus, convex near-elliptic compact objects are successfully tracked with this algorithm. However, several real world objects have non-convex shapes with holes - e.g. *rickshaw*, *cycle*, *man on a motorbike etc.* In such cases, the mean-shift tracker learns the background colour distribution in the target model and hence drifts away in object localization iterations. More so, mean-shift trackers are also found to fail under severe occlusions, as it models the colour distribution of the whole target region and not by parts. To avoid these limitations, we extend the object characterization to include the feature points as well, which we describe in the following sub-section.

2.2. *Combining Colour and Feature Cues*

Feature correspondence was proposed in the context of image registration,[8] later extended[9] for the selection of good feature points. Consider the consecutive images Ω_t and Ω_{t+1}, such that $\Omega_t(\mathbf{U}) = \Omega_{t+1}(\mathbf{U} + \mathbf{d_U})$. Tomasi *et al.*[9] have shown that the displacement vector $\mathbf{d_U}$ is sufficient during tracking feature points between successive frames approximating deformation value to zero. Feature points are tracked using the Kanade-Lucas-Tomasi (KLT) tracker.[10] The sum of squared difference between consecutive images is reduced to find the displacement vector. Tracking is based on symmetric definition for dissimilarity between two images unlike earlier approaches given as $\Omega_t(\mathbf{U} - \mathbf{d_U}/2) = \Omega_{t+1}(\mathbf{U} + \mathbf{d_U}/2)$. The displacement vector can be computed by solving the equation $\mathbf{G}(\mathbf{U})\mathbf{d_U} = \mathbf{e}(\mathbf{U})$, where the 2×2 symmetric matrix $\mathbf{G}$ and the residue vector $\mathbf{e}$ are obtained from,

$$\mathbf{G}(\mathbf{U}) = \int_{W(\mathbf{U})} \mathbf{g}_t(\mathbf{X})\mathbf{g}_t(\mathbf{X})^T w(\mathbf{X}) d\mathbf{X}$$

$$\mathbf{e}(\mathbf{U}) = 2 \times \int_{W(\mathbf{U})} (\Omega_t - \Omega_{t+1})(\mathbf{X})\mathbf{g}_t(\mathbf{X}) w(\mathbf{X}) d\mathbf{X}$$

Where, the integration is performed over a certain window $W(\mathbf{U})$ centered at $\mathbf{U}$, $\mathbf{g}_t(\mathbf{X}) = \nabla\Omega_t(\mathbf{X})$ is the image gradient and $w(\mathbf{X})$ is a weighing function defined over $W(\mathbf{U})$. The pixel position $\mathbf{U}$ is considered to host a *good* feature point, if both the eigen values (λ_1, λ_2) of $\mathbf{G}(\mathbf{U})$ are sufficiently high, i.e. $min(\lambda_1, \lambda_2) > \lambda$, where λ is a predefined threshold.

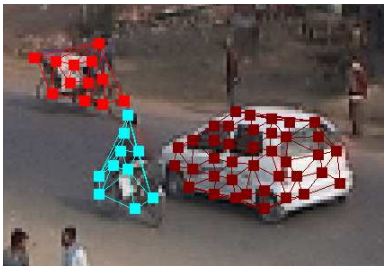


Fig. 2. Delaunay triangulation based graph model for (a *car*, a *cycle* and a *rickshaw*).

The feature tracker is invoked for objects under occlusions. We extend the object characterization of the colour only tracker to include the set of feature points in the object region. We perform a Delaunay triangulation (figure 2) over the feature point set forming a planar graph that represents a geometrical structure of the object. Isolated feature tracking can erroneously correspond to a) points in the background or b) points on other objects. In case (a), foreground segmentation can be used to eliminate the correspondence. In case (b), the faulty feature correspondence is detected by using the *Motion Consistency Hypothesis*: feature points on the same rigid body exhibit consistent motion. Even where the bodies exhibit large deformation (e.g. human motions), some branches of the graph exhibit relatively stable deformations. Object feature points are tracked in the consecutive images constrained by the feature-point planar graph structure to improve the tracking performance. However, in cases of (near) complete occlusions, where neither colour match nor feature correspondences can be established, we continue the tracking with motion predicted object position. Object is said to disappear, either it is found out as disappeared using predicates described in[6] or if it losses minimum threshold number of features. When object disappears it is tracked using only motion information, which is susceptible to error as number of frames increase. In the case of hybrid tracker the duration of disappearance for objects is less compared to color only tracker, as objects are tracked until the object completely disappears. Hence matching accuracy increased in hybrid tracker compared to color only tracker.

3. Results and Ground-truth Validation

We report an experiment on a traffic surveillance video involving a wide variety of vehicles like motorbikes, bicycles, rickshaws, cars, buses, trucks and tractors, as well as people and animals (cow). Next, we compare our results in multi-object tracking based on only colour and both colour and feature (figure 3).

The proposed algorithm is tested on different data sets. Figure 4 shows the results obtained on human data set. In this data set, we do not perform neglecting of feature points based on *Motion Consistency Hypothesis* described in section 2.2, because the tracked objects (humans) in this video are not rigid. Tracking accuracy in this case is less for both color only tracker and hybrid tracker compared to their respective Traffic video performances 3.

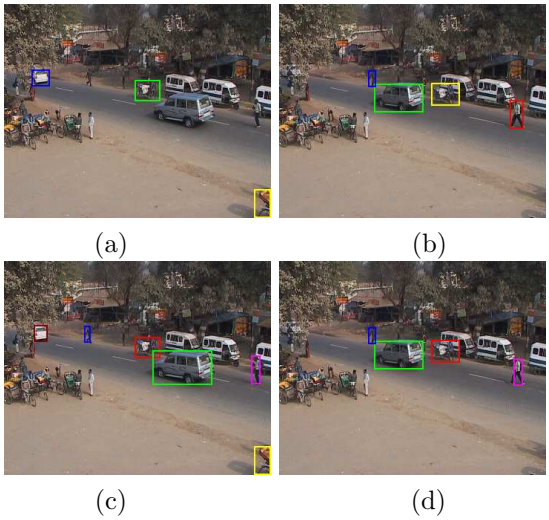


Fig. 3. Re-identification errors. In sequence (c,d) (Mean shift tracker), a silver SUV occludes the rickshaw at upper center; after the occlusion the rickshaw is misidentified as another object that disappeared earlier, and the SUV is wrongly matched as rickshaw. The man in the blue bounding box is matched with a car that left the scene. All these errors are overcome by the hybrid feature tracker (e,f), it is particularly good at re-identification.

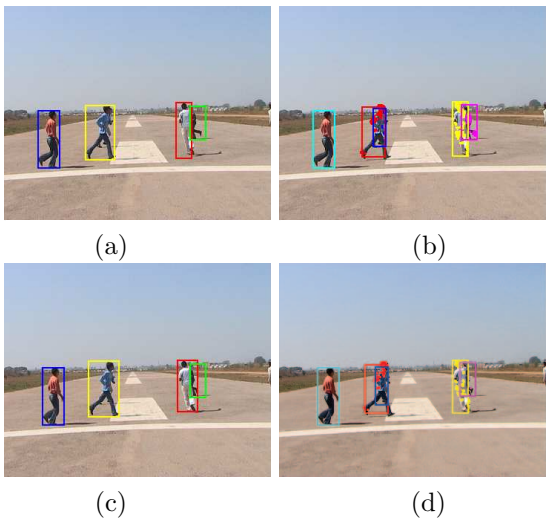


Fig. 4. Comparison of Mean shift and Modified trackers on Human data set

We observe, in the figures 4(a) and (c), that Mean shift tracker clearly fails to track the person under severe occlusion behind another person. In the figures 4 (b) and (d), we observe that Hybrid tracker able to track person under severe occlusion with **blue** bounding box.

Minimum distance ($MinDist$): Distance between point features can be changed. While selecting the point features, If the point feature that is going to be selected is near to already selected features, then it is not selected. It is based on the assumption that "neighboring pixels generally have similar goodness values" . The point feature is considered to be near other point feature if the distance between them is less than minimum distance. This can be used to speed up the process.

For results in figure 4 $MinDist$ is fixed as 0 while it is fixed as 5 in figure 3. We validate the results against a ground-truth data over 700 frames. We compute the following measures:

- **Total Tracking Accuracy** $accuracy_t = \frac{\sum_{i=1}^{t} b_i}{\sum_{i=1}^{t}(b_i+c_i)}$, where b_i denotes the number of well tracked objects and c_i is the number of track losses in the i^{th} frame.
- **Re-identification accuracy** $(re-ident_t)$: Let s_i be the number of disappeared objects in the i^{th} frame, e_i be the number of erroneously tagged reappearances[a], g_i the number of successfully registered object entries in the scene, f_i be the number of reappearances erroneously detected as scene entries, and h_i be the the number of scene entries erroneously detected as reappearances. Then we define the **Re-identification accuracy** as the ratio $re-ident_t = \frac{\sum_{i=1}^{t}(s_i-e_i-f_i+g_i)}{\sum_{i=1}^{t}(s_i+g_i+h_i)}$.
- **Tracking Accuracy in Crowds** is the ratio $crowd_t = \frac{\sum_{i=1}^{t} p_i}{\sum_{i=1}^{t}(p_i+q_i)}$, where p_i is the number of well tracked objects in the crowd and q_i is the number of track losses in crowd, as observed in the i^{th} frame.
- **Approximation of Object Localization Accuracy**, $localiz_t = \frac{\sum_{i=1}^{t} b_i-n_i}{\sum_{i=1}^{t} b_i}$, where n_i denotes the number of objects tracked with an ill-sized or misplaced bounding box in the i^{th} frame. As seen in table 1, results are significant improvements in all the categories, with strikingly improved re-identification.

4. Conclusion

We have proposed an algorithm for multi-object tracking under occlusion by combining multiple cues(Colour, Motion, Features) based on their importance in particular situation. The proposed scheme

[a]Let the agents a and b `disappear` in frames say `x, y` respectively. In frame `i (i > (x, y))`, the tracker identified agent b as re-appeared when actually it was agent a.

is not restricted by any prior object shape/motion models or ground plane assumptions and thus performs satisfactorily in relatively unconstrained environments; more importantly, since no camera calibration is needed, it can be placed anywhere and immediately put to work. The remaining limitations are significantly more difficult - e.g. when an object is nearly fully occluded (motion projection is the only option), or differentiating between multiple objects entering the scene together (before they split). A cue towards the latter may be to look for differences in deformations of the feature graph - resulting in several clusters of motion of point features on single blob, an approach that may work when the objects deform differently, or move with different speeds. Eventually, it would be important to extend these ideas to work in more general situations, e.g. cameras that move (initially with pan-tilt motions), and for dynamic backgrounds (trees, fountains).

References

1. C. R. Wren, A. Azarbayejani, T. Darell and A. Penaland, *Pattern Analysis and Machine Intelligence* **19**, 780(July 1997).
2. I. Haritaoglu, D. Harwood and L. Davis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 809(August 2000).
3. C. Needham and R. Boyle, Tracking multiple sports players through occlusion, congestion and scale, in *Proceedings of the 12th British Machine Vision Conference*, 2001.
4. T. Zhao and R. Nevatia, Tracking multiple humans in crowded environments, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, July 2004.
5. S. McKenna, S. Jabri, Z. Duric and A. Rosenfeld, *Computer Vision and Image Understanding* **80** (2000).
6. P. Guha, A. Biswas, A. Mukerjee and K. Venkatesh, Occlusion sequence mining for complex multi-agent activity discovery, in *The Sixth IEEE International Workshop on Visual Surveillance*, May 2006.
7. D. Comaniciu, V. Ramesh and P. Meer, Real-time tracking of non-rigid objects using mean shift, in *Computer Vision and Pattern Recognition*, 2000.
8. B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, in *Proceedings of 7th International Conference on Artificial Intelligence (IJCAI 1981)*, (Vancouver, British Columbia, 1981).
9. C. Tomasi and T. Kanade, *Detection and Tracking of Point Features*, Tech. Rep. Technical Report CMU-CS-91-132, Carnegie Mellon University (April 1991).
10. S. Birchfield, Klt an implementation of the kanade-lucas-tomasi feature tracker www.ces.clemson.edu/stb/klt/.

Table 1. Performance Comparision of colour only and colour-feature based hybrid tracker

	Colour tracker	Hybrid tracker
b_t	2904	3246
c_t	500	182
$accuracy_t$	85.31%	94.69%
s_t	29	23
e_t	17	1
f_t	2	2
g_t	5	20
h_t	30	0
$re-ident_t$	23.44%	93.02%
p_t	425	592
q_t	357	193
$crowd_t$	54.35%	75.41%
n_t	213	15
$localiz_t$	92.67%	99.54%