Tracking by Planning

Haifeng Gong
*University of Pennsylvania*

Jiwoong Sim
*University of Pennsylvania*

Maxim Likhachev
*University of Pennsylvania*

Jianbo Shi
*University of Pennsylvania*, jshi@cis.upenn.edu

Follow this and additional works at: [http://repository.upenn.edu/cis_reports](http://repository.upenn.edu/cis_reports)

Recommended Citation
Haifeng Gong, Jiwoong Sim, Maxim Likhachev, and Jianbo Shi, "Tracking by Planning", January 2011.
Tracking by Planning

Abstract
We introduce a method for tracking multiple people in a cluttered street scene. We use global context to address the challenge of long occlusion by endowing each tracked object with a planning agent. This planner uses context of the street scene, people and other moving objects to reason about pedestrian intended behavior for tracking under occlusion and ambiguity.

We extract short but robust trajectories called tracklets by tracking people with a simple appearance model. We formulate the tracking problem as a batch mode optimization, linking tracklets into paths, each with supporting evidence by an agent’s goal directed behavior, and image partial matching along the trajectory gap. We propose a global criteria for consistent linking of the tracklet with planning that can correct local ambiguity in linking. We test our algorithm in a challenging real world setting, where we automatically estimate scene context and intended goals, then track multiple people from a moving camera.

Comments

This technical report is available at ScholarlyCommons: http://repository.upenn.edu/cis_reports/953
Tracking by Planning

Haifeng Gong, Jiwoong Sim, Maxim Likhachev, Jianbo Shi
hfgong@seas.upenn.edu, jshi@cis.upenn.edu
GRASP Laboratory, University Of Pennsylvania

Abstract. We introduce a method for tracking multiple people in a cluttered street scene. We use global context to address the challenge of long occlusion by endowing each tracked object with a planning agent. This planner uses context of the street scene, people and other moving objects to reason about pedestrian intended behavior for tracking under occlusion and ambiguity.

We extract short but robust trajectories called tracklets by tracking people with a simple appearance model. We formulate the tracking problem as a batch mode optimization, linking tracklets into paths, each with supporting evidence by an agent’s goal directed behavior, and image partial matching along the trajectory gap. We propose a global criteria for consistent linking of the tracklet with planning that can correct local ambiguity in linking. We test our algorithm in a challenging real world setting, where we automatically estimate scene context and intended goals, then track multiple people from a moving camera.

1 Introduction

Robust visual tracking of pedestrians, vehicles and other moving objects in cluttered urban scenes from a moving camera is an important problem in robotics, video surveillance and video analytics applications. For example, consider an autonomous vehicle with a stereo camera pair operating in a crowded urban environment. Tracking people reliably from this moving camera can enable safe autonomous operation by reasoning about pedestrian behavior, to determine if a pedestrian is about to cross the street, or about to run in front the car.

The fundamental challenge for tracking multiple people in complex environments is occlusion. Occlusions, particularly long occlusions, created among walking people, between people and cars, lead to a difficult data association problem. The key question is how to follow people as they appear from occlusion—to initialize a new track or to link to an existing one. Adding to this difficulty, people’s appearance change as they move, due to lighting, body pose and scale variations. Occlusion combined with appearance changes lead to ambiguous data association or ‘drifting’ in tracking, such that after an occlusion, measurements are associated with an occluder, who happens to have a similar appearance.

Our solution to resolve ambiguity in multiple-people tracking is to reason about pedestrian motion using context. For example, global context may include the positions of surrounding people and objects in the scene, reasoning about
Fig. 1. Tracking by planning. We use endow each tracked person with a planning agent, to reason about her behavior for tracking under occlusion and ambiguity. The person A, at \( t=12 \), crosses road with red-light. She zig-zags to avoid coming cars, and is occluded by the person with similar appearance at \( t=37 \). By estimating her intended goal, and obstacle avoiding path, we succeed in tracking her.

the intended goal location of the pedestrian, or reasoning about likely paths the pedestrian will take to get there. We propose to endow each tracked person with a planning agent with global context. This agent maintains a planned path for the pedestrian as it moves in the cluttered environment. We also associate an agent with the moving camera, who maintains its context in terms of its ego-motion, and a rough 3D scene geometry. In this approach, each tracked person is endowed with a planner, creating a 'virtual simulation' of intended pedestrian motion in a cluttered street. When a person is visible, we track him, and using his trajectory to narrow down a set of plausible goals/planned paths for him. When the person is occluded, we creates hypothesis explaining the occlusion, and predict his re-appearance based on the plausible set of goals/planned paths.

There are two key distinguishing factors for this 'tracking-with-planning' model. First, we explicitly model the interaction of the multiple people’s track as a goal directed obstacle avoiding path. Traditional HMM or Coupled HMM cannot model long range interactions between people’s path and objects easily. Second, context reasoning allow us to create more flexible and realistic pedestrian trajectories. Simpler models of dynamic social behavior [1], are limited in expressive power.

We recognize this 'tracking-with-planning’ model implemented natively would have many context variables to be estimated, including 1) people’s positions, 2) people’s goals, 3) other objects’ positions (such as cars); 4) 3D scene geometry, 5) camera ego motion. Instead of treating this as a full state estimation problem, we will ground those measurements which can be measured with high certainty. This includes, 1) visible people’s tracked trajectory, called tracklets; plus items 3 to 5 above. The ‘hidden’ states are partially occluded or occluded people’s
position, and people’s goal. Furthermore, to narrow our scope, we focus on a batch mode of tracking, that is we track a person at time $t$, given observation from $t-k$ to $t+k$. The system does not currently perform online tracking, it delays tracking decision until a period of observation.

We formulate the ‘tracking-with-planning’ problem as a) goal estimation, and b) tracklets association, extension with goal oriented motion plan as matching cost. This formulation was demonstrated using data collected from a car mounted with a stereo camera pair driven across an urban city. We show that this system can successfully track multiple people in a cluttered street environment, from a moving or stationary camera. In this paper, we present related work in section 2; formulate our approach for ‘tracking-with-planning’ in section 3; describe context measurement in section 4; and experiments in section 5.

## 2 Related Works

In many tracking works, the emphasis is on image feature selection and adaptation. [2] presents attributed relational graph (ARG) approach that adaptively updates the object model by adding new stable features as well as deleting inactive features. [3] learns a tracklet affinity by adaptively selecting features to maximize the discriminative power on training data by HybridBoost. [4] selects the features that best discriminate between object and background.

Better motion model can improve tracking. [5] uses a dynamical model to propagate the particle within a limited sub-space of target state. [6] introduces a mixture particle filter which is suited to multi-target tracking as it assigns a mixture component to each object. [7] describes a MCMC based particle filter which includes motion model to maintain the identity of targets throughout the interactions. Traditional dynamic models predict the location for each target based on its own history, without taking into account the remaining scene.

The most related approach to our method is that of [1] and [8] who introduces dynamic social behavior model, which is trained with videos recorded from birds-eye view, and applied as a motion model for multi-people tracking from a vehicle-mounted camera. Social force model [9] considers three terms for people tracking — 1) a term describing the acceleration towards the desired velocity of motion; 2) terms reflecting that a pedestrian keeps a certain distance from other pedestrians and borders; and 3), a term modeling attractive effects. The resulting equations of motion of nonlinearly coupled Langevin equations.

Multiple people and multiple hypothesis tracking is an active research area. [10] prunes an exponentially growing tree of Kalman filters by determining the k-best hypotheses in polynomial time. [11] relaxes the association in object tracking as a multi-path searching problem. It explicitly models track interaction, such as object spatial layout consistency or mutual occlusion.
3 Tracking by Planning

A person’s appearance can be categorized into two states: fully visible or fully/partially occluded. When a person is fully visible, we can detect and track him with high certainty. When a person is partially or fully occluded, we estimate his position. Tracking a person in the visible state leads to a short trajectory — we call a tracklet. We set a conservative threshold to terminate the trajectory when tracking score is too low. After termination, the same person may be picked up again by the detection algorithm, and tracked to produce associated tracklets.

Furthermore, we conjecture that most people’s trajectories are purposeful, that is they have associated goals, and they anticipate to avoid obstacles (other objects, cars, and pedestrians) to reach the goals. This provides a strong constraint on evaluating which linked trajectory is more likely to be correct.

In total, we formulate the task of state estimation of partially/fully occluded person be equivalent to that of linking and extending tracklets into plausible goal-directed obstacle avoiding paths.

![Fig. 2. Tracking by linking and extending tracklets into complete goal directed trajectories. Blue line: a plausible planned path. Red solid: links with good value, evidence by consistence with planned path (1), and by image appearance matching (3). Red dash: links with bad value, due to impossible plan (2), and poor image similarity (4).](image)

3.1 Global context for linking tracklets

The main challenge in linking tracklets is that local greedy decisions are often bad, and one needs to delay the decision until longer period of observations, and consider its broader spatial context. We address the global context of correct tracklet linking:

1) **Appearance similarity context.** For example, a person appears as tracklet A, disappears and then reappears with tracklet B. When he reappears, he has changed his pose. Meanwhile, another person (tracklet C) with a similar appearance to tracklet A walks by. A local greedy decision would link tracklet A to the impostor C. If we observe the sequence longer, the person would change his
pose back, and return to his original appearance. To track this person correctly, a long period of observation is needed.

We address this problem in two ways. First, we adaptively update an appearance model of the person as he moves, see 3.4. Given a sufficiently long tracklets A, B, and C, the adaptive model would better capture their true appearances, and we can set up a better link between tracklets A-B, over trackets A-C.

Second, when tracklets are short—imagine tracklet B is fragmented into multiple short tracklets $B_1, \ldots, B_k$, we would link A to one tracklet $B_i$ that has a very close appearance, and re-estimate his position in between. To explain the gap between A-B, we seeks two pieces of evidence, 1) the feasibility of joining the tracklets and gaps to form a plausible path towards a target, and 2) local appearance similarity along the ‘hypothesized’ gap. The second piece amounts to HMM style estimation of the people’s position using partial image similarity matching.

2) Path planning plausibility context. A similar situation as before, when the person reappears with tracklet B, he happens to zig-zag to avoid an oncoming obstacle, while an impostor tracklet C happens to continue on the original path. Standard linking by trajectory continuity would link A-C. If we have observed the sequence longer, we would see tracklet C moves erratically, while tracklet B’s trajectory is well explained by its movement towards its goal while avoiding obstacles along the way.

As we will show in 3.3, we maintain multiple-plausible obstacle avoiding paths towards a set of hypothesized goals for each tracked person. We a) use visibility graph to compute a set of plausible shortest paths, and b) use these paths generate each a potential field, and 3) link tracklets based on energy defined by the potential field, see 3.3.

3) Path mutual exclusion. As shown in all previous multiple target tracking works, it is important to maintain mutual exclusion between the path formed by linking tracklets. We want to avoid joining two concurrent tracklets into one subsequence tracket.

3.2 Criteria for tracklets linking by planning

We defer the discussion of people detection and initial tracking until section 3.4. Assume we have a set of tracklets $\mathcal{T} = \{F_1, \ldots, F_{N_T}\}$,

where $F_i$ is the $i$-th tracklet, and $N_T$ is the total number of tracklets. Each tracklet

$$F_i = (t^0_i, t^1_i, x^i_0, \ldots, x^i_{t^1_i}),$$

(1)

where $t^0_i$ is the start time of $F_i$, $t^1_i$ is the end time of $F_i$ and $x^i_t$ is the object position at time $t$. Note that $x^i_t$ is defined in a fixed 3D world coordinate defined by the first camera. We first measure the people’s position in the stereo image frame, and map it to a fixed 3D world frame using the ego-motion estimation of the camera, as described in section(4).
Our solution is to link and extend these tracklets, $\mathcal{T}$, into complete trajectories, using the ‘estimated’ partial/full occlusion position to explain away the ‘gap’ form by the tracklets.

Let $L_{i,j}$ be the indicator of linking $i$-th and $j$-th tracklet:

$$L_{i,j} = \begin{cases} 1 & F_i \rightarrow F_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

To model the global context of linking tracklets into plausible goal-directed obstacle avoiding paths, we design the following criteria for tracking:

$$\max_L \epsilon(L) = \sum_{i,j: L_{i,j}=1} \left[ \text{AppScore}(i,j) + \text{PlanScore}(i,j) \right] - \beta |L| \tag{3}$$

where AppScore$(i,j)$ measures appearance similarity between tracklets $F_i$ and $F_j$, PlanScore$(i,j)$ measures a) how $F_i$ and $F_j$ are consistent with a plausible goal directed path; and b) how partial occlusion in the gap can be explained by appearance of $F_i$ and $F_j$. We introduce $\beta \neq 0$ to prevent aggressive linking.

The above criteria is subject to constraints:

1. The number of successor (predecessor) of a tracklet can not exceed 1: $L_{i,j} \in \{0,1\}$, and $\sum_i L_{i,j} \leq 1$, $\sum_j L_{i,j} \leq 1$.
2. Time ordering: if $t^i_1 < t^j_0$, $L_{i,j} = 0$.
3. Tracklets far apart in time can not be linked: if $t^j_0 - t^i_1 > t_{\text{Thr}}$, $L_{i,j} = 0$, where $t_{\text{Thr}}$ is a threshold for time duration.
4. Speed limit on bridging the gap between $F_i$ and $F_j$: $\|x^f_{t^0_i} - x^f_{t^1_i}\|^2 / (t^0_i - t^1_i) \geq v_{\text{Thr}}$, $L_{i,j} = 0$, where $v_{\text{Thr}}$ is a threshold for velocity.

This cost function is NP-hard, and we seek an approximate solution using Linear Programming. Let $A_{i,j} = \text{AppScore}(i,j) + \text{PlanScore}(i,j) - \beta$, and rewrite Eq (3) in form of

$$\max_L \sum_{i,j} L_{i,j} A_{i,j} \tag{4}$$

subject to the following constraints $\sum_i L_{i,j} \leq 1$, $\sum_j L_{i,j} \leq 1$, $L_{i,j} = 0$, $\forall (i,j) \in \text{InvalidSet}$, $L_{i,j} \in \{0,1\}$.

### 3.3 Planning score

We endow each tracklet with a path planning agent. The planning score is given by finding the best planned route to fill the gap between tracklet $i$ and $j$. The best route

1. is compatible with tracklet $i$, tracklet $j$ in space-time geometry,
2. allows possible partial matches by appearance during occlusions.
We define the following score

$$\text{PlanScore}(i,j) = \max_{r \in \text{routes}} -\text{Dist}(r, F_i) - \text{Dist}(r, F_j) + \text{OcclScore}(F_i, F_j, r)$$

where \(\text{Dist}(r, F_i)\) is the distance between route \(r\) and tracklet \(F_i\) and \(\text{OcclScore}(F_i, F_j, r)\) is the score for picking up the partial occlusions along the gap. Roughly, one can think each route \(r\) defines a potential field by distance transform on \(r\) and we measure closeness of \(r\) to \(F_i, F_j\) on this potential field.

We compute the distance between a tracklet and planned route as follows. First, we shorten the tracklet by keeping only the last \(K\) frames of \(F_i\) and first \(K\) frames of \(F_j\), to obtain a shortened tracklet \(F_i'\) or \(F_j'\). We denote \(x_1, \ldots, x_K\) the tracked positions in \(F_i'\) or \(F_j'\), and the length of the shortened tracklet by \(l\). Find the point \(p\) on the route \(r\) which is the nearest to the start point of \(F_i'\) or \(F_j'\). Using \(p\) as start point, we can obtain an arc on the route \(r\) with length \(l\), which results in a shortened route \(r'\). Finally, we divide \(r'\) into \(K - 1\) segments uniformly, to obtain \(K\) end points: \(r'_1, \ldots, r'_K\). We compute the distance:

$$\text{Dist}(r, F_i) = \sum_k \|x_k - r'_k\|^2.$$

**Partial occlusion Score** We want to find possible partial matches (by appearance) during occlusions. \(\text{OcclScore}(F_i, F_j, r)\) is defined by the score of two appearance models applied to the hallucinated trajectory bridging the gap. Given the route \(r\), which does not connect \(F_i\) and \(F_j\) perfectly, we first compute the hallucinated trajectory connecting \(F_i\) and \(F_j\) by a diffusion equation.

Given end point of \(F_i = x_{i,t_i}^i\) and start point of \(F_j = x_{j,t_j}^j\), we find the nearest points on \(r\) for both \(x_{i,t_i}^i\) and \(x_{j,t_j}^j\), and use the arc bounded by these two nearest point as a reference for the gap trajectory. Let \(s\) denote the reference arc. We divide it into \(t_0^i - t_1^i\) segments uniformly and let \(s_t, t \in [t_1^i, t_0^j]\) be the nodes of the division. We solve for the occluded trajectory bridging the gap between
Let \( y = \{y_{t_1}, \cdots, y_{t_j}\} \) be the occluded trajectory. To obtain \( y \), we minimize the following energy function:

\[
\min_y \sum_t \left( \|y_t - y_{t-1} - (s_t - s_{t-1})\|^2 \right). \tag{7}
\]

The intuition is that \( y \) copies the shape and curvature of \( s \) but connects \( F_i \) and \( F_j \).

After an occluded trajectory is obtained, we project it to both cameras to pick up possible partial marches during occlusions. Let \( S_i(y, t) \) be the appearance score of appearance mode \( i \) apply to position \( y \) at frame \( t \),

\[
\text{OcclScore}(F_i, F_j, r) = \text{Sat}_{l_1, u_1} \left( \sum_t \text{Sat}_{l_2, u_2} \left[ \max_{y^\approx y_t} S_i(y, t) \right] \right)
\]

where \( y \approx y_t \) means \( y \) is in a small neighbourhood of \( y_t \), \( \text{Sat}_{l,u}(x) \) is a saturation function with lower bound \( l \) and upper bound \( u \).

**Multiple Hypotheses Planning** Each tracked person maintains a global context: the street scene layout, pedestrians, and other moving objects (cars). From the global context, we automatically generate a set of possible goals, as we will describe in next section. For each goal, we generate a set of plausible routes, that avoids the obstacles along the way.

We adopt traditional visibility graph based shortest path planning, and add to it a significant part for generating multiple plausible routes. In real life, people’s path are not always optimal in the shortest-path sense, due to the uncertainty in traversibility of the environment, ambiguity of people’s behavior nearby, as well as inherent bias/variations in one’s behavior. We show by reasoning on the visibility graph, we can arrive at a small set of topologically distinct paths towards a hypothesized goal, as following:

1) Visibility graph construction. Assume we have a set of obstacles, walkable area, both in polygon representation, and a pair of start point and goal. We collect a) all the vertices of all obstacles, b) vertices of the walkable region which are not in the convex hull, c) the start and goal, as nodes of the visibility graph, and connect nodes visible to each other as edges.

2) Reduced visibility graph construction. For each edge in the visibility graph, we check if it is

- a support line, on the same side of which the related obstacles lie, or
- a separate line, on the opposite side of which the related obstacles lie.

If it is neither support line nor separate line, it is removed.

3) Shortest paths. Find the shortest path from the start to the goal in the reduced visibility graph.

4) Plausible (sub-optimal) route candidates generation. A sub-optimal route must go through a vertex not on the shortest path. We enumerate all possible
vertices that are not on the optimal route as sub-goal, and we find the shortest path going through this sub-goal, from the start to goal.

5) Redundant sub-optimal routes pruning. For a pair of candidate routes, we can connect them into a polygon because they share the same start point and goal. If the polygon enclose no obstacle, the two routes are redundant.

![Fig. 4. Reduced Visibility Graph and Shorted Path — first, shortest path is obtained; then enumerate all possible midgoals, i.e. vertices of all polygon, and compute shorted path that goes through each midgoal; finally, prun the redundant paths.](image)

3.4 Appearance Feature

**Adaptive appearance model** For a pedestrian, we divide his image patch into three parts, head, torso and legs. Part based representation allows us to reason partial occlusion. For each part \( k \) at time \( t \), we collect the color histogram with \( 8 \times 8 \times 8 \) bins, denote this by \( p_t(k) \), and we also collect the histogram of surrounding background, denoted by \( q_t(k) \). We use simple color feature instead of more advanced shape features for computation efficiency reason. We maintain running means of the histograms as an object model: \( f_t = (1 - \alpha) \ast f_{t-1} + \alpha \ast p_t \), \( b_t = (1 - \alpha) \ast b_{t-1} + \alpha \ast q_t \). Denote Model\(_t\) = \((f_t, b_t)\) object appearance model.

**Tracklet creation.**

We use car detector\([12]\) to detect peoples and cars in the current frame. To track a person in frame \( t + 1 \) given the models of previous frame, Model\(_t\), we measure two scores:

- **Consistent Score** (\( S1 \)) to ensure that it is similar to foreground appearance model \( f_t \) and different from \( b_t \),
- **Contrast Score** (\( S2 \)) to ensure that the foreground is different from its surroundings in current frame.

We combine the following 2 scores for tracking people into initial tracklets:

\[
S1(\text{Model}_t, p_{t+1}) = \frac{1}{2} \left( \sum_k \sum_{\text{bin}} p_{t+1}(k) \log \frac{f_t(k)}{b_t(k)} + \sum_k f_t(k) \log \frac{p_{t+1}(k)}{b_{t+1}(k)} \right)
\]

\[
S2(\text{Model}_t, p_{t+1}) = \sum_k \text{KL}(p_{t+1}(k) || q_{t+1}(k)) = \sum_k \sum_{\text{bin}} p_{t+1}(k) \log \frac{p_{t+1}(k)}{q_{t+1}(k)}
\]

**Appearance score** The appearance score is obtained by testing the appearance model of tracklet $i$ on images patches of tracklet $j$ and vice-versa.

$$\text{AppScore}(i, j) = \text{Match(Model}_i, \text{Model}_j)$$

(11)

in which the model matching score is defined by the appearance model.

Let $\text{Model}_i(k) = (f_i(k), b_i(k))$ be the foreground and background appearance model of the $k$-th part of the $i$-th tracklet, we define

$$\text{Match(Model}_i, \text{Model}_j) = \sum_k \sum_{\text{bin}} \left[ f_i \log \frac{f_j}{b_j} + f_j \log \frac{f_i}{b_i} \right].$$

(12)

### 4 Context recognition

In our testbed, a car mounted with a stereo camera is driven through an urban city. To construct the map for the planner and predict goals of current human tracks, we construct the global context of the scene automatically, including ego-motion of camera, 3D scene geometry and obstacle positions.

![Fig. 5. [Context Recognition] From left to right : 3D Plane hypotheses generated from RANSAC plane detection based on stereo feature point matching. Over segmentation of the image. Detected Building surface, color of the surface denotes the direction of surface normal in horizontal plane. Linelets detected in individual frames. Linelets are accumulated to a unified coordinate system using ego-motion trasnformation. Final context recognition result, with goals (yellow star) and road area detected (green). Blue line indicates the car’s path computed by ego-motion.](image)

#### 4.1 Estimating Camera Motion

We estimate the rigid 3D transformation $R_{3 \times 3}, T_{3 \times 1}$ between consecutive frames using correspondences of reconstructed 3D feature points.
1) Find a set of $N$ feature point matches $m = (p^L_{i,t}, p^R_{i,t}, p^L_{i,t+1}, p^R_{i,t+1})$ that have stable correspondence over temporal as well as spatial domain.

2) Reconstruct 3D position of each points in frame $t$ and $t+1$ independently, which gives $m_{i}^{3D} = (p_{i,t}^{3D}, p_{i,t+1}^{3D})$.

3) Using RANSAC on the 3D points, we compute $R_{3 \times 3}, T_{3 \times 1}$ for the ego-motion between every consecutive frames. For every steps in our algorithm (road extraction, binocular detection, binocular tracking) where camera motion is involved, we use these transformation to unify our detection space into a single coordinate space. Figure (5) shows some result of our motion estimation.

![Figure 5. Object 1 example results — Top, no planning, middle, with planning, bottom top view with planning.]

### 4.2 Building detection

We extract building surface from each frames to provide information for road extraction. Our building surface detection algorithm consists two phases. First, we generate 3D plane hypothesis from the pair of stereo images using a specialized RANSAC, which requires only 3 corresponding stereo feature pairs for plane hypothesis.

Second, we generate a over-segmented regions of left image of current frame. We test our plane hypothesis to each segmented regions to classify these regions as building surface(we ignore planes with unlikely surface normals). We pick the best plane which transforms the given region into the other pair of stereo image, based on the chi-square distance of the color histogram over original segment region and the transformed region. After each segments are assigned with the best explaining plane, we classify those segment into building vs non-building surface using geometric characteristics such as surface size, shape and height. Figure 5 demonstrates this approach.
4.3 Goal detection, Road extraction

Detected building surfaces are then used to generate the boundary of sidewalk by intersecting the extension of surface with ground plane, which we call a linelet. For a given length (e.g., 60 frames) of clip of street scene, we extract the possible roads and goals on a common coordinate system using linelets and ego-motion. The task of this module is to detect the road lines to bound object trajectories, and find intersections of road lines as possible goals of objects.

The steps are as follows:

1. **Mosaic linelets**, using the ego-motion transformation between each pair of consecutive frames, we can mosaic a global ground plane and put all linelets from all frames together.
2. **Cluster linelets into lines**, given the set of linelets which are neither accurate enough nor completed enough, we cluster them into several dominant side lines for roads.
3. **Visibility analysis**, by finding intersections of those lines, and analyzing the visibility of each line segments from our camera, we can construct the plan of the street as well as possible goals.

5 Experiments

To test our algorithm we have collected a video from moving vehicle in the actual urban city. The stereo images are collected at 1024*768 resolution, with 6 FPS. We have a fully automated system for people detection, tracking, 3D scene layout/goal estimation, and camera ego-motion computation.

To compensate the lack of automated ground plane estimation, we calibrated a ground plane at the first frame of each sequences and propagated this estimation using our ego-motion transformation.

We have picked 6 sequences which contains multiple people, and have interesting interaction and occlusion. Details of each sequences are in table 5. There are total of 27 people in the all the sequences. None of these people can not be tracked in its entirety, as indicated by the number of tracklets (403) produced, and illustrated in figure 7. Tracking with planning significantly improves tracking: 24 out of 27 people are tracked to completion. We also did experiments to see the added value for using a full path planning algorithm. With simple linear plan, connecting the tracklets, but with the LP optimization, we succeeded in tracking 20/27 people. In hind-sight, it is expected, since most people walk normally without zig-zag or running red lights. However, it is in those outlier cases, we need a system that can track well, and the 4 additional pickups are significant.
Fig. 7. Top: tracking without planning. Each rows show the typical tracking result with template tracking—it drifts after occlusion. Bottom: tracking with planning. We are able to pick up the entire trajectory of a pedestrian, despite many long occlusions. The patches marked with red frames partially occluded frames which are recovered by the tracking-with-planning approach.


**6 Conclusion**

We present and test a system that can track multiple people in a clutter scene. For each tracked person, we endow it with an agent that plans with global context to predict its movement. We demonstrate the usefulness of tracking with high level planning on a challenging real world application.

![Fig. 8. Results of tracking with planning. Yellow stars are possible goals. Bold green lines are the planned routes that the objects follow. The dashed green lines are other planned routes that not follow by the person.](image-url)
Acknowledgement

This work is supported by ONR/MURI project (Rich Representations with Exposed Semantics for Deep Visual Reasonings). The authors thank Weiyu Zhang, Katerina Fragkiadaki Praveen Srinivasan and Jeff Byrne for proof-reading.

References