W. Choi and S. Savarese. **A unified framework for multi-target tracking and collective activity recognition.** ECCV 2012.

Presented by Aman Sikka
October 25th, 2012

Focuses on tracking more than 1 person, classifies the group behavior as well as individual behavior

Examples:
Individual behavior: Action + Pose, e.g. walking & facing-front, walking and facing-back
Pair behavior: walking together, moving-to-opposite (walking in opposite directions)
Collective behavior: crossing, gathering, standing in a queue

Tracklets - fragmented tracks for a person's motion, computed using optical flow

Goal is to jointly solve individual behavior, pair behavior, and collective behavior
Previous work has focused on only 1 of these tasks at a time

Contributions:

- Model that merges collective activity recognition and multiple target tracking
- Uses interactions for path selection / tracking
- Correlates activities and different levels of granularity
- Superior to state of the art

Bottom-up understanding:
Individual -> Pairwise -> Group
e.g. Walking -> Approaching -> Gathering

Discussion: Gathering and Crossing seem related, but gathering means to come together and stay there, but crossing means to continue afterwards and cross paths. Thus to complete this analysis we need to analyze the whole video

Features: HOG, Bag of Video

Graphical model:
Observations (using features) influence individual activities, which influence interactions, which influence collective behavior. Influence also flows downward from the collective behavior, and across all time points
O_c is the collective appearance - STL features on a larger window including multiple people

All variables are classified on a per time-step basis - on each time step we classify the individual and collective behavior. We also have links between successive time steps.

Factors:
Factor(Observation, Action) - low level observations influence the individual action and pose
Factor(Action, Interaction, Target associations) - where T is a function of f, individual actions combine to create joint interactions e.g. individual: standing, individual: standing -> together: standing in a line
Factor(Collective behavior, Interaction) - e.g. interaction: standing in a line, interaction: standing in a line -> collective behavior: queueing
Factor(Collective Observation, Collective Behavior) - use the average of STL features (from reference 1) - Note that these are not overall scene characteristics, because the scenes (background scene) for each sequence is the same; rather, these are overall motion features for the people
Factor: Smoothness factors across different timepoints
Factor: c transpose * f = cost of associating tracklets

Discussion: Each factor includes variables from multiple time slices within a temporal window

Discussion: Each collective behavior must hold for all people and all interactions in the video, e.g. if the collective behavior is gathering, then we assume that all people in the video are gathering

Ideal: complete tracks
Actual: Separated tracklets
The goal is to combine tracklets into complete tracks
We combine tracklets using color histogram and motion cues
c_ij is the validity of the match
f_ij is the correspondence (1 for a match, 0 for no match)

Discussion: How do they enforce temporal consistency in the C variable?
One idea is that they have separate time slices and allow each variable to change value across the video

Dataset:
Fixed camera, scripted actions

Results:
Baseline - dump all the features into an SVM
They are measuring collective activity classification
They also show the results of tracking - the quadratic model works the best

Discussion: It seems that the results drop significantly without O_c, the overall observation features (STL)
The results are well presented - they show the effect of removing every factor in their model

Computation time:

Given the original tracklets, it takes 1 minute to process the video

Discussion:
Tracklet is analogous to superpixels for segmentation - a small subpart that we are confident in, that we combine to make larger parts

Discussion:
The connection between O_c and C is important because the farther apart these variables are, the more the influence degrades. However, connecting them causes the influence to flow directly and makes a strong effect
This is very common in graphical models - to connect the observations directly to the decision variables
This is not how the brain works - the retina is not directly connected to the neo-cortex
But current graphical model inference algorithms are not advanced enough to allow for influence to flow well over many intermediary variables, so it is very common to shortcut this and have direct connections between observations and decision variables

Conclusions:
Contextual information helps target association and trajectory estimation

Discussion:
Tracklets are a big piece that requires its own discussion
Other concepts not fully discussed: branch and bound
How does pose variation affect their collective activity recognition? What if the people in the line are facing different directions?
The dataset is too contrived - fixed camera, fake actors, no major occlusions - it would be better to test it on a more natural dataset
The C variable should be flexible enough to consider only a subset of the people (if not all the people are performing the action) - we want to discover the group that is participating in the collective activity
How robust is this current method to noise, ie. if not all people are participating in the collective activity? This seems to happen to a small extent in their videos
Seems to be much more academic and "beautiful" than the paper **Online learned discriminative part-based appearance models for multi-human tracking**, possibly less hand-engineering

B. Yang and R. Nevatia. **Online learned discriminative part-based appearance models for multi-human tracking.** ECCV 2012.

Presented by Cameron Schaeffer
October 25, 2012

Online learned - no training set - learn on the fly
Discriminative - conditional probabilities

Tracklets - midlevel feature - a sequence of points in time

Human detection is easy (95% accuracy?)
What is hard is to associate different tracks (maintain identity) when people interact, occlude each other, and cross paths

Goal:
Person detectors are imperfect, so if we try to connect detection boxes, we are susceptible to missing some detections and losing the track
The goal is to overcome this
Given tracklets, goal -> combine tracklets to make longer tracks
We also lose frames at the beginning/end of the tracklet

Problems:
Identiity switches
False positives
Missed detections

Combines 2 previous approaches:
ABT (Association based tracking)
CFT (Category free tracking)

ABT (Association based tracking):
Links objects of similar appearances, using a pre-trained detector (usually pedestrians)
Computes a global solution
Kernel Density methods - assumes appearance doesn't change across change
Problem: Loses information if the detector fails

CFT (Category free tracking)
Requires an initial manual labeling, tracks each person that is labeled
Online learning - gains more information about the object as the object moves
Works for generic objects - fingers, motorcycles, etc

Algorithm:
Video frames -> pedestrian detector

Creates tracklets using ABT (basically nearest neighbors) from a pedestrian detector (only when detections are very close and when there are no interactions between people) - these are high confidence track pieces
Divides each person into 15 parts
Occlusion reasoning for human parts

Tracklets are created using a person detector
Then the CFT learns the appearance of each person to extend the length of each tracklet (ie. adds a few more frames to each tracklet)

Characterize each of 15 parts based on HOG and color histograms
For each tracklet, we create a positive set from these features, negative set from the background and "distractors" (nearby tracklets from other people)
Trained using RealBoost

Tracklets are extended using CFT using both appearance and motion cues
Motion cues use the velocity to predict the potential locations for the person in the next time step
This step is done conservatively - we only add these frames if we have high confidence. Thus we only add a few frames to the tracklet

Linking -
Uses the Hungarian algorithm to link tracks - e.g. this algorithm can assign people to tasks based on the cost of each person to minimize the total cost
Similarly, we match tracklets to other tracklets based on their probability of a match

Results
Real-life scenario using a surveillance camera, with **heavy occlusions**
Airport scene - Shows track identity being maintained even with heavy occlusions and overlaps
Outdoor scene - without CFT, creates track fragments, loses tracks, loses identity - with CFT, maintains identity
Sidewalk scene **with a moving camera!** - without CFT loses people behind poles, with CFT maintains identity

Summary discussion:
Seems to have many extra variables that might have been unnecessary - did not show the importance of each variable in their model
Did not explain tracklets or other concepts
Nice model
They don't model collective behavior
Heavy engineering (many parameters they had to set) but works very well