People watching: human actions as a cue for single view geometry
David Held

Motivation: Determine 3D geometry of indoor scene -> using people to help understanding!

Input: Time-lapse video, fixed-camera, disconnected shots of same scene

Goal: Output 3D understanding of the environment (determine walkable, sittable, reachable surfaces)

Start by generating hypotheses without people understanding:

1. Vanishing point estimation
2. Box (surfaces - walls/floor/ceiling) hypothesis generation consistent with vanishing points
3. Determine most likely hypothesis by training on labeled data (labeled surfaces)
4. Output ranked hypotheses

Using the appearance of the room alone, difficult to determine room geometry
Instead, leverage people for better understanding!

Start by detecting poses of humans in the time-lapse video

-Method of Yang and Ramanan (CVPR 2011)

-Felzenszwalb part-based model

Train separate detector for each pose (standing, sitting, reaching)

How do we use the detected poses to detect functional surfaces?
Sitting: Detect sittable regions when a sit pose is detected
Reaching: set reachable region where hands are at
Standing, Sitting, Reaching: Where feet are, set as walkable region

-Put a Gaussian distribution around these regions proportional to size of human

Re-estimating room layout:

-Re-rank original box hypotheses using functional surface information

-Prior to choose smaller rooms

-Helps to choose the correct room geometry from the ranked hypothesis!

Estimating Free-Space of the scene:

-Free-Space gives us information on which parts of the scene are occupied and which are not

-Mark voxels on the floor with occupancy

-Combine information from all sources to obtain Free-Space estimation

-How do they deal with functional surfaces like chairs that move?

Cool results showing inferred geometry and Free-Space, and how the system's beliefs change over time as the video is watched

-In one of the videos, floor seems to be slightly floating?

Appearance + People > Appearance Only or People Only

-Shown with quantitative results

Method also works for single images, not constrained to time-lapse videos:

-Pose detection on still images

-What about sittable regions on the wall?

-Functional surfaces for sitting are not used to help room geometry understanding

Conclusions:

-Humans are valuable cues for understanding scenes

-Although pose estimation is not perfect, there are good things that can be done with it

Video data comes from YouTube

-Quality of the videos don't seem to be great

-Seems difficult to detect people because of low quality features?

Activity Forecasting

Qifeng Chen

Activity Recognition vs Activity Forecasing

-Infer past observed actions vs Infer future UNOBSERVED actions

Given a starting point and ending point of a human in a scene, which path will he take?

-Different from tracking

-Find distribution of most likely path -> prefer to take sidewalk, avoid cars, etc.

Assume we know the destination of the human

Focus on predicting paths of humans, no prediction on other objects

Cast as a machine learning problem:

-Training data consists of demonstrated activities, extracted physical features (semantic pixel labeling)

-Make predictions on novel scenes!

Model human activity using a decision-theoretic framework (common in robotics)

-combine this with physical scene features

-motion model -> Dynamics of human

-policy model -> Decision of human

Activity sequence generated by a Markov Decision Process (MDP)

-activity sequence = (state, action, reward) ...

-the sequence of states is determined by the policy

-the policy is determined by the reward function and expected payoff

-infer reward function from video sequences

Intuitively, high reward areas are areas that can be walked through, whereas low reward ares are areas that cannot be walked through

Learning phase:

-Inputs: Trajectories and feature responses

-Output: Reward weights

Better forecasting performance compared to MaxEnt Markov Model and Markov Motion Model

Interesting idea: compare methods to human performance?

-it’s not clear that even a human could accurately forecast where a person would walk

Also shows results on novel scene types using knowledge transfer

-Can be done as long as policy is learned