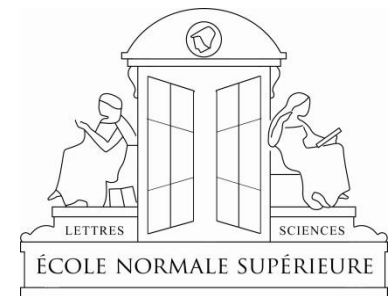


# People Watching: Human Actions as a Cue for Single-View Geometry

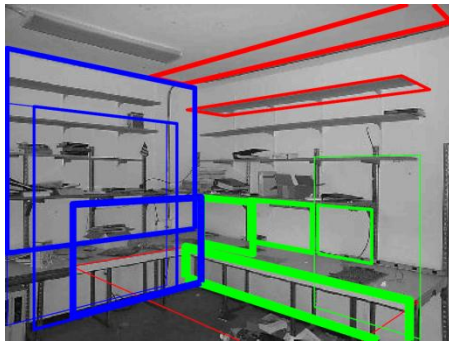
David Fouhey, Vincent Delaitre,  
Abhinav Gupta, Alexei Efros, Ivan Laptev, Josef Sivic

**Carnegie Mellon**  
THE ROBOTICS INSTITUTE

*inria*  
informatiques mathématiques



# Indoor Single-View 3D Geometry



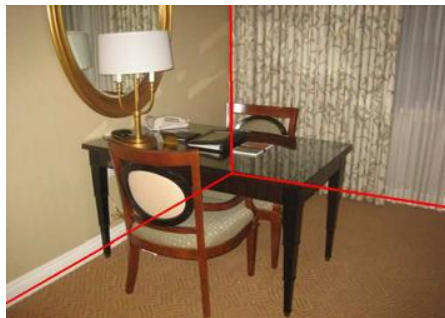
Yu et al., '08



Lee et al., '10



Del Pero et al., '12



Hedau et al., '09



Schwing et al., '12

## Man-Made Constraints



Where are the people?

# But People Are Interesting!



Gupta et al. '07



Yang et al. '11



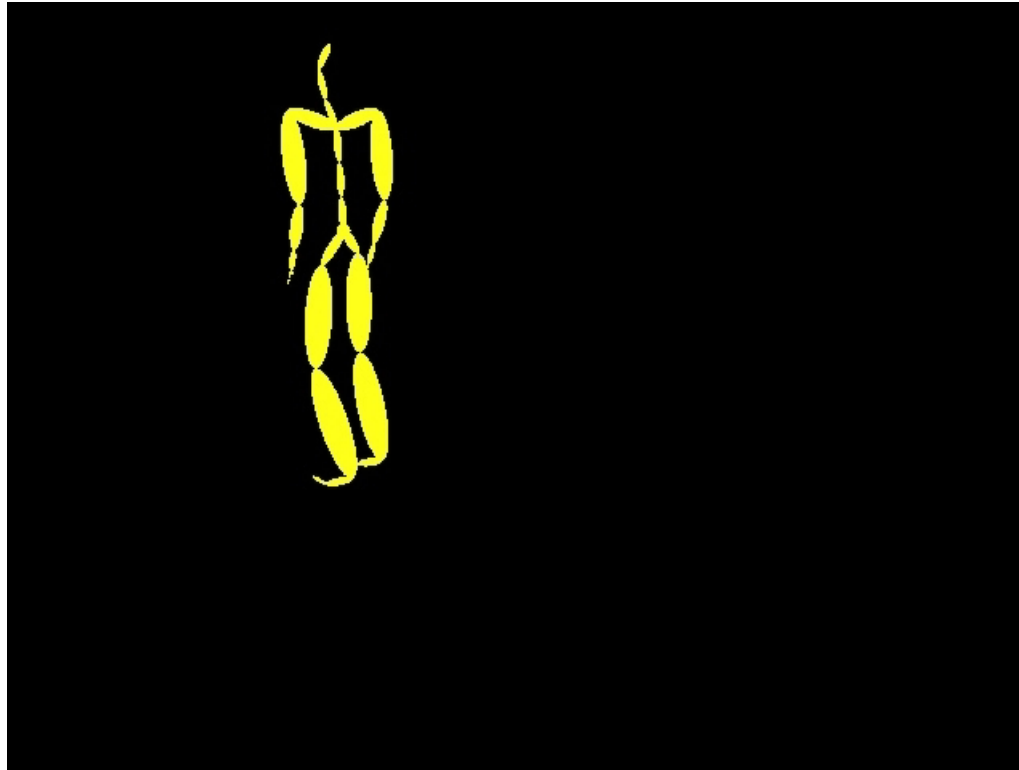
Yao et al. '10

# People as Clutter?

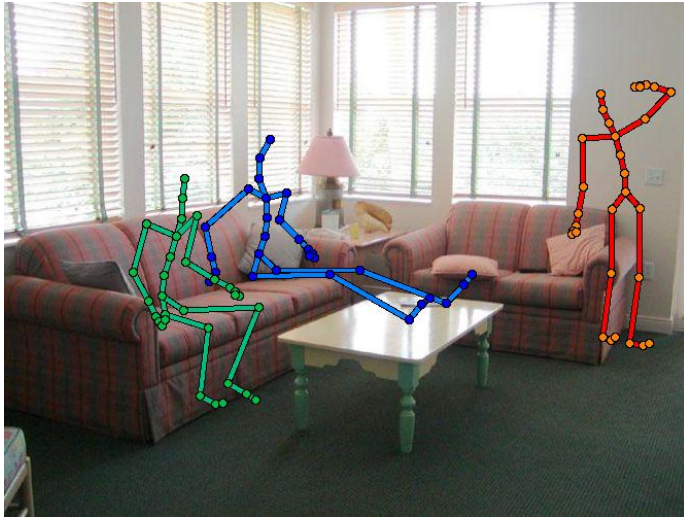


People Occlude the Scene!

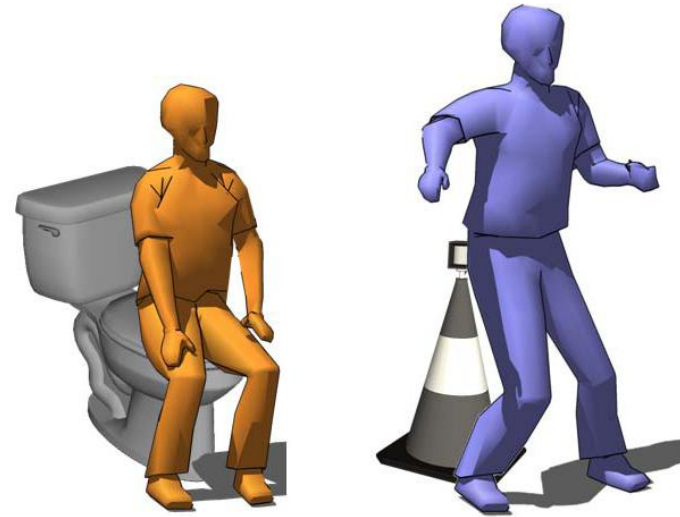
# People – Cues not Clutter



# Affordances – Where can I Sit?



Gupta et al. '11



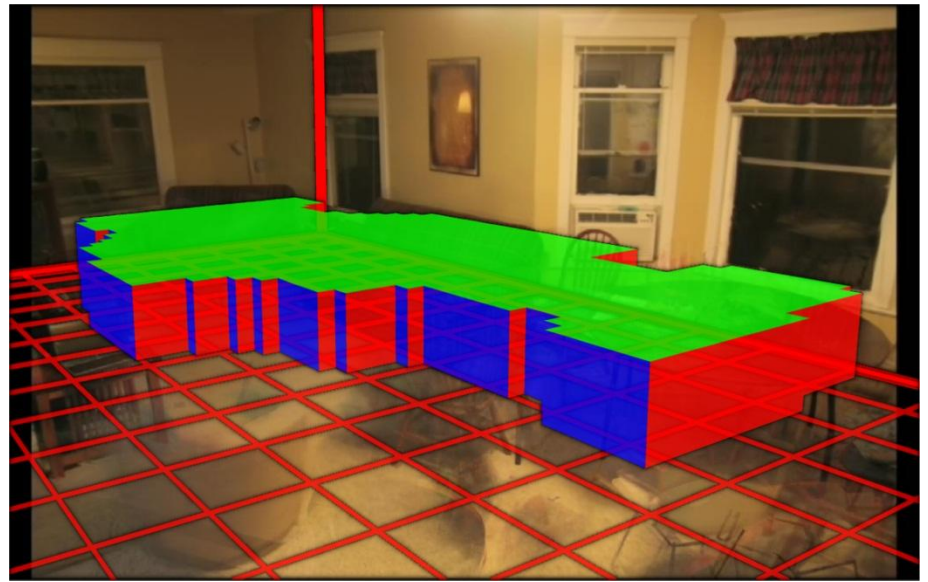
Grabner et al. '11

Affordances: Opportunities for interaction with the scene – J.J. Gibson

# Our Goal – Inverse Problem



Input:  
Timelapse



Output:  
3D Understanding

Humans as Active Sensors



# Our Approach

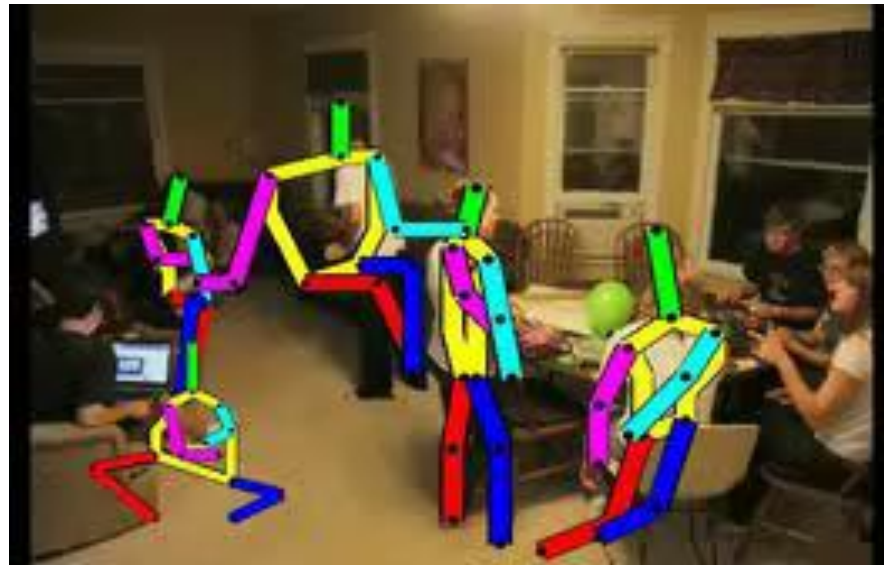


Timelapse

# Our Approach



Timelapse



Pose Detections

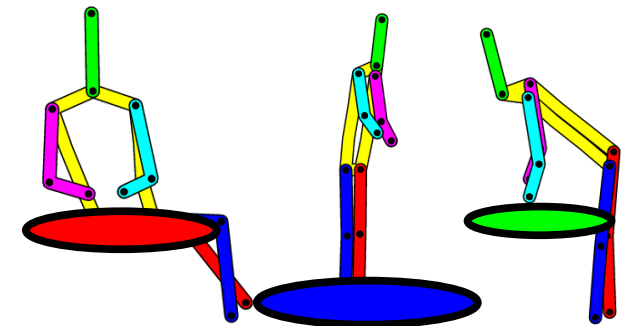
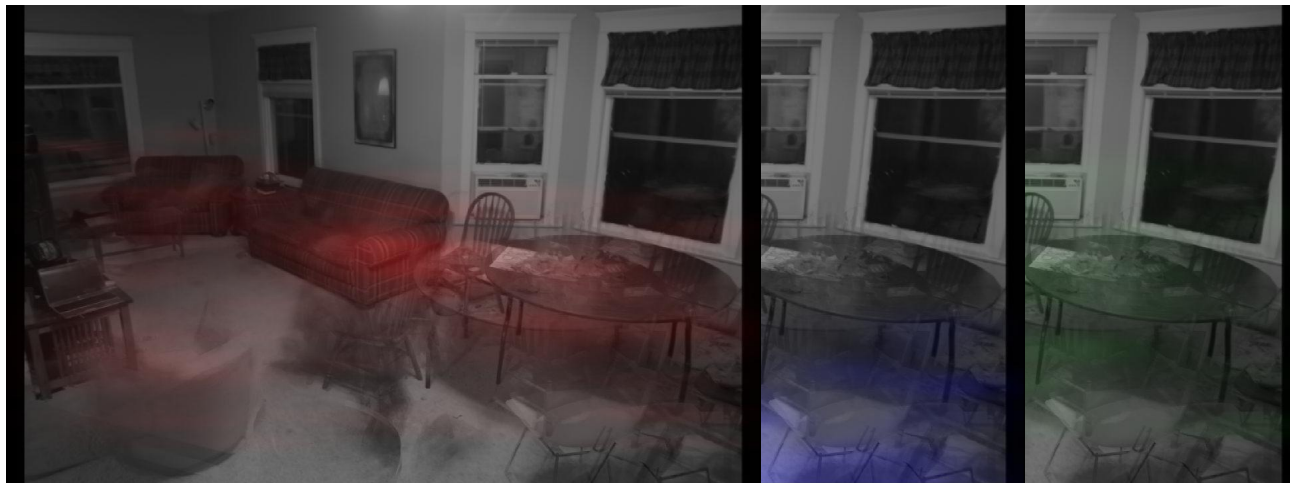
# Our Approach



Timelapse



Pose Detections



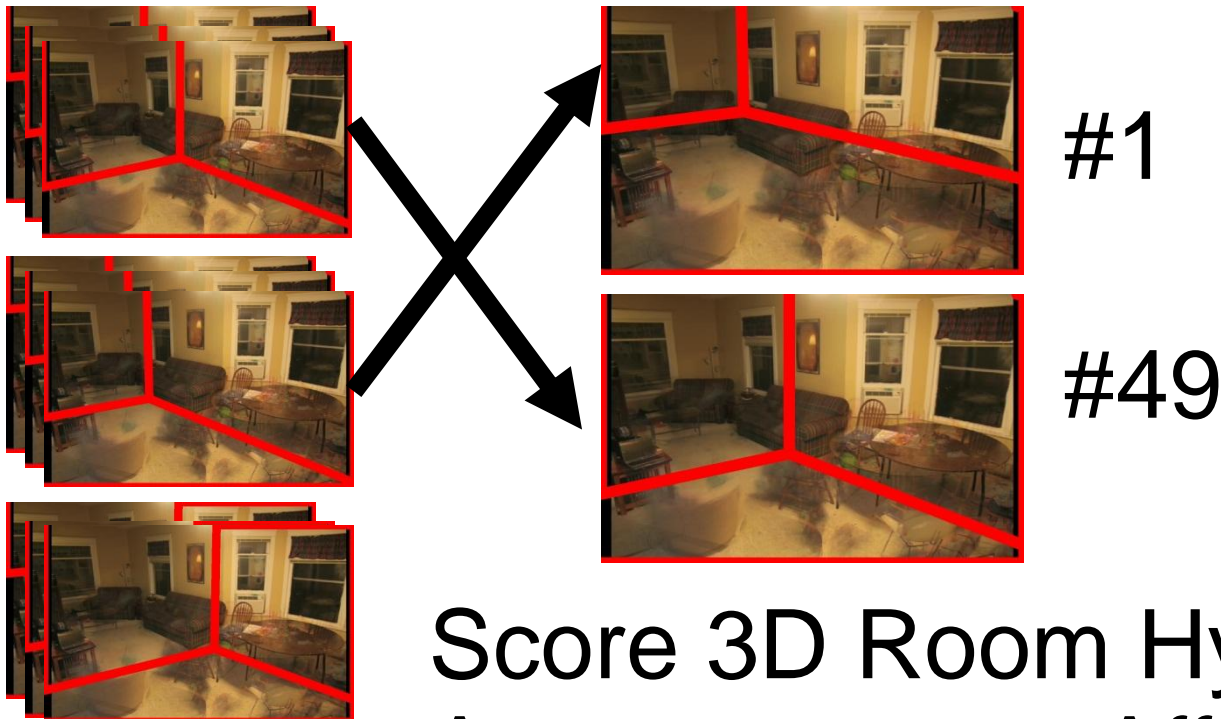
Estimate Functional Regions from Poses

# Our Approach



3D Room Hypotheses From Appearance

# Our Approach



Score 3D Room Hypotheses With Appearances + Affordances

# Our Approach



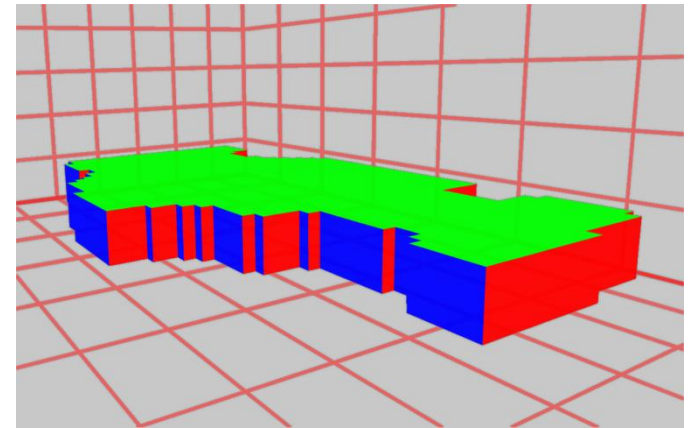
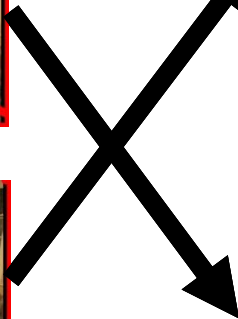
Timelapse



Pose Detections



Functional Regions



Estimate  
Free-Space

# Our Approach



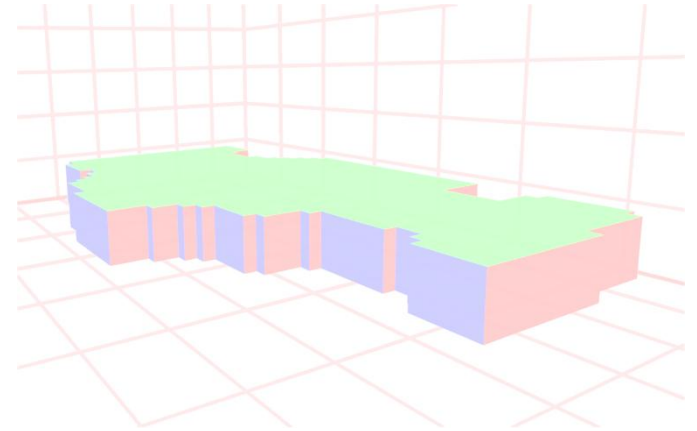
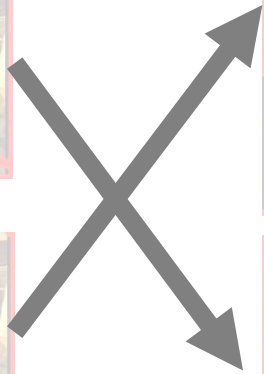
Timelapse



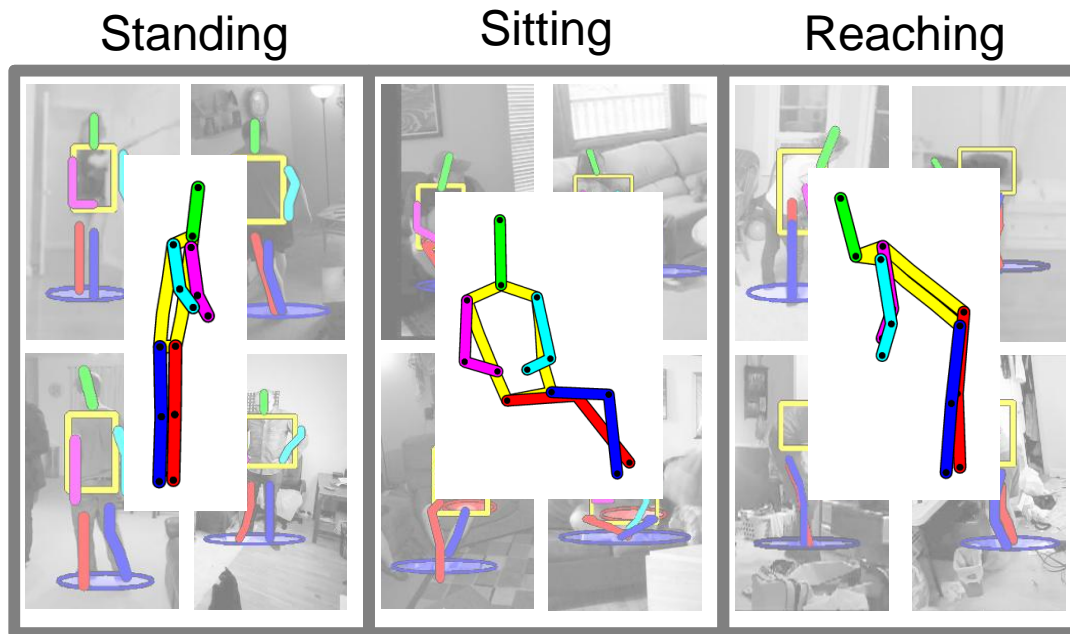
Pose Detections



Functional Regions



# Detecting Human Actions



Yang and Ramanan '11  
Train Separate Detectors for Each Pose

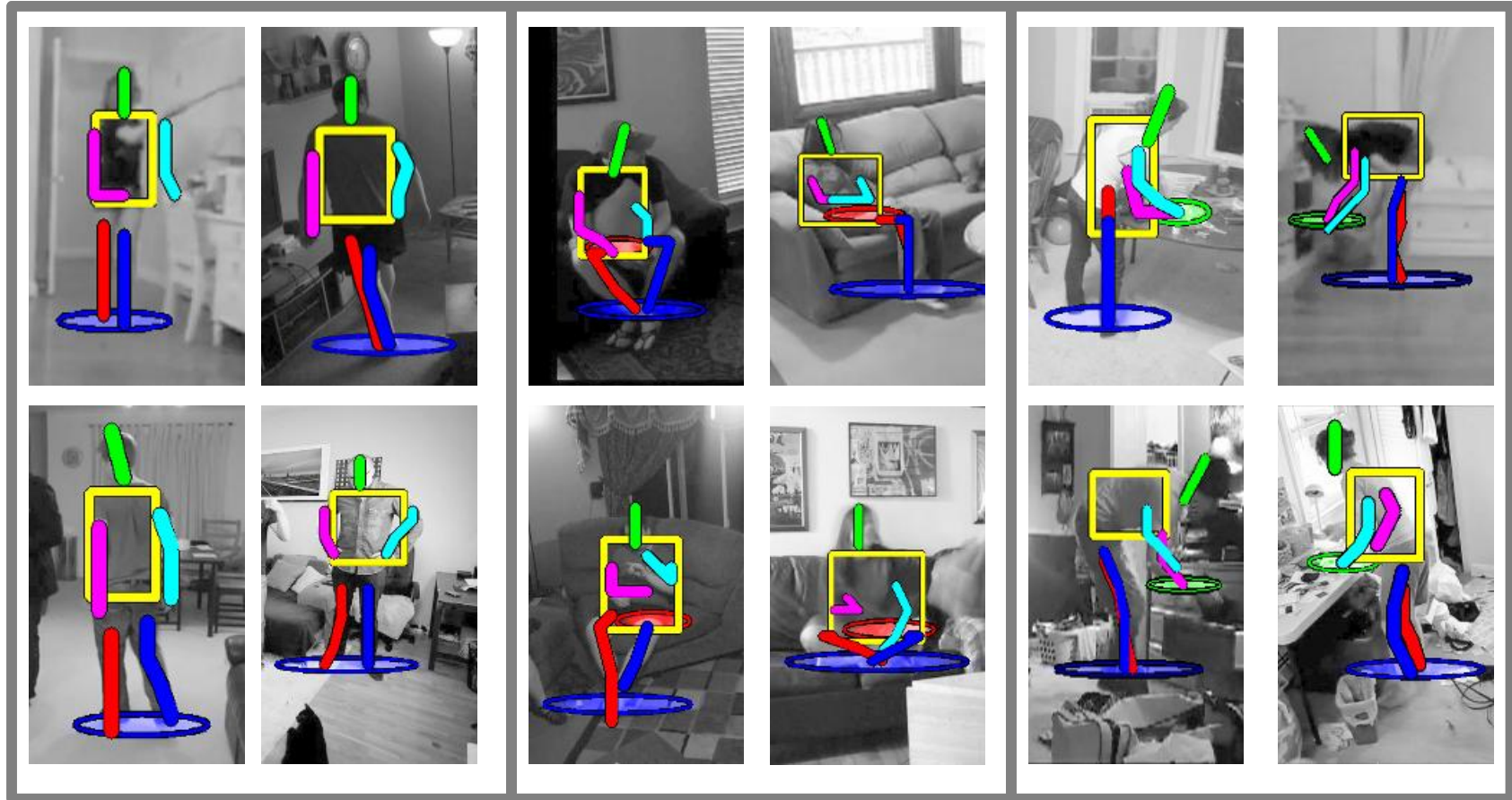


# Detecting Human Actions

Standing

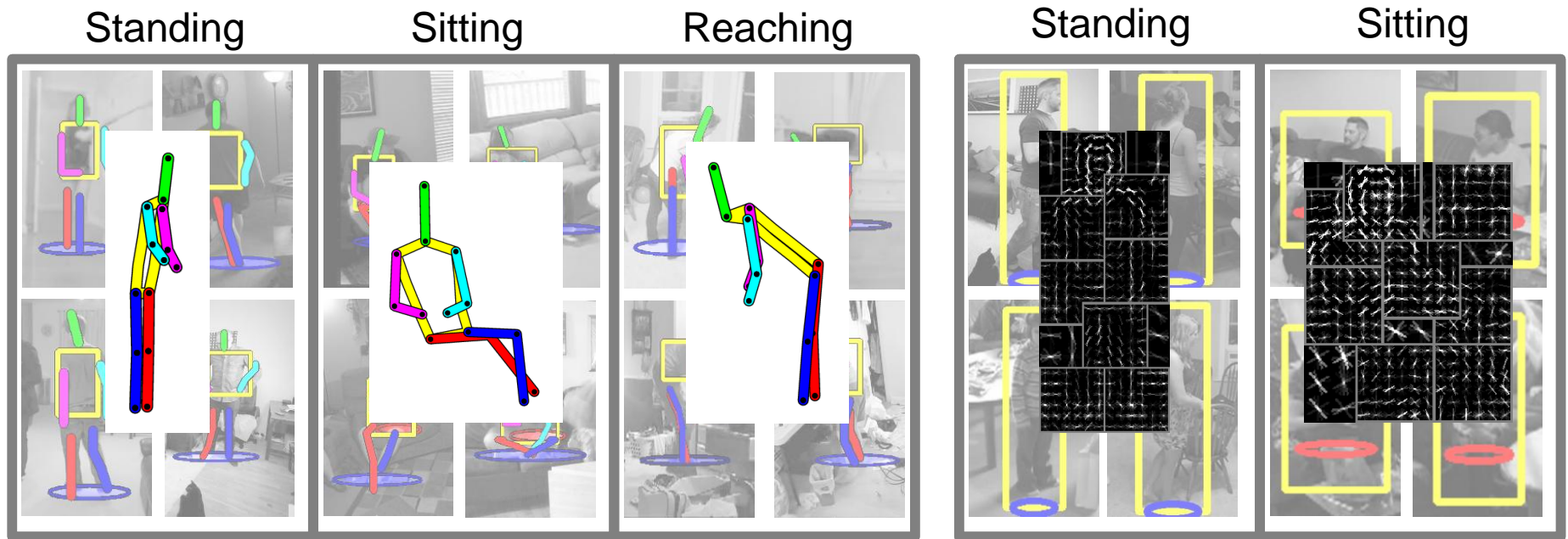
Sitting

Reaching



Yang and Ramanan '11  
Train Separate Detectors for Each Pose

# Additional Detectors



Felzenszwalb et al. '10  
Train Separate Detectors for Each Pose

# DPM Detections

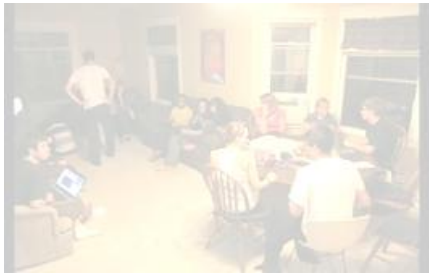
Standing

Sitting



Felzenszwalb et al. '10  
Train Separate Detectors for Each Pose

# Our Approach



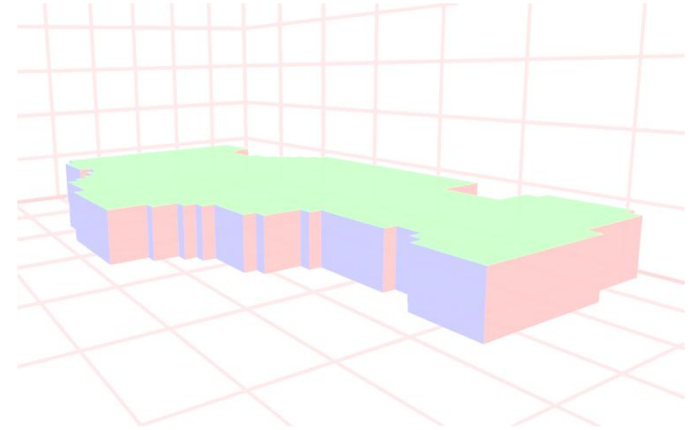
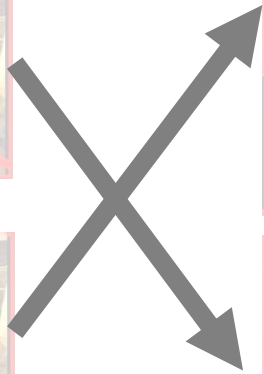
Timelapse



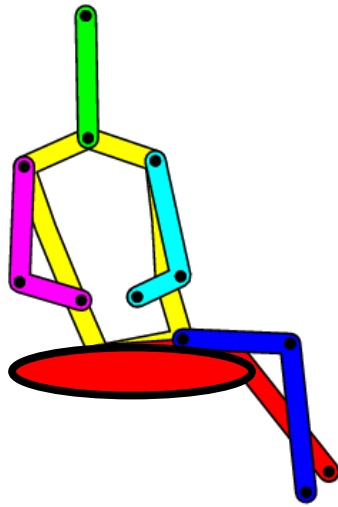
Pose Detections



Functional Regions

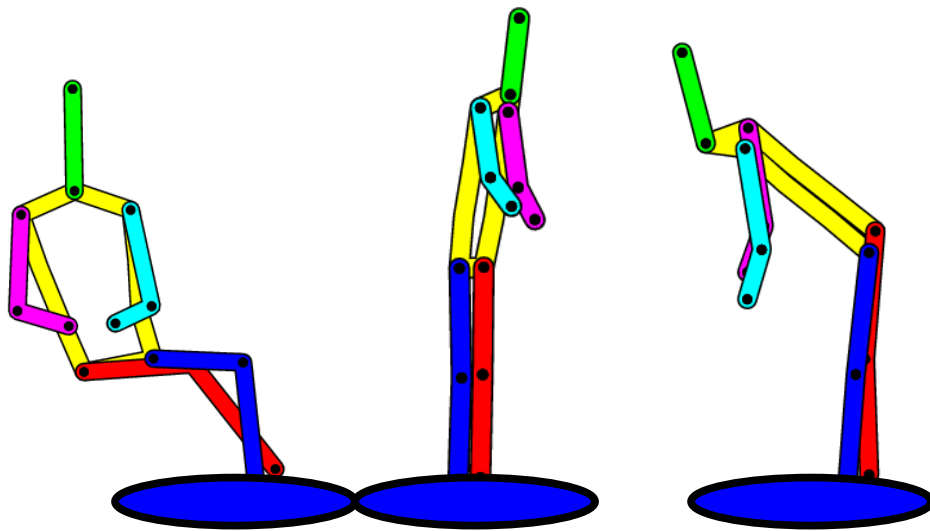


# From Poses to Functional Regions



Sittable Regions at Pelvic Joint

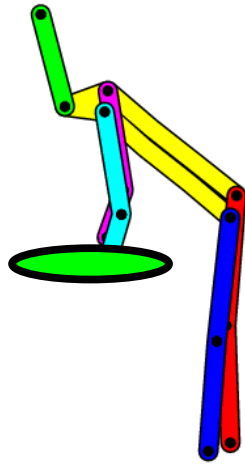
# From Poses to Functional Regions



Walkable Regions at Feet

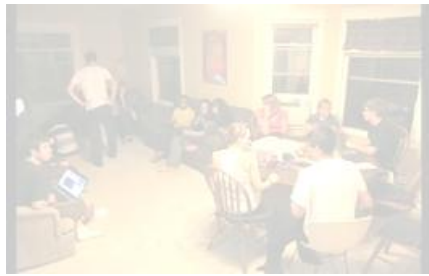


# Affordance Constraints

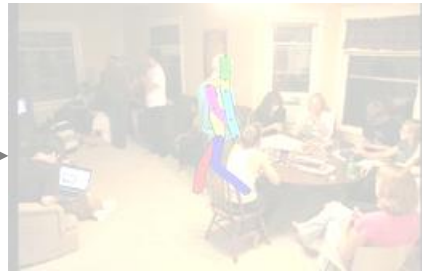


Reachable Regions at Hands

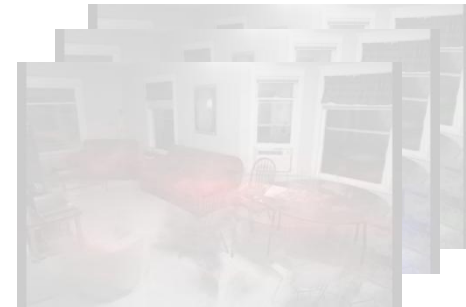
# Our Approach



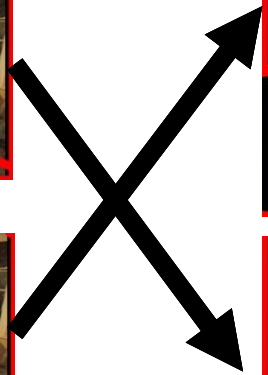
Timelapse



Pose Detections



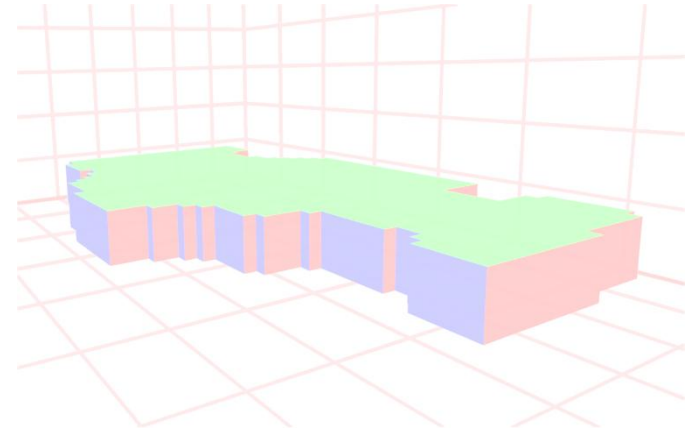
Functional Regions



#1

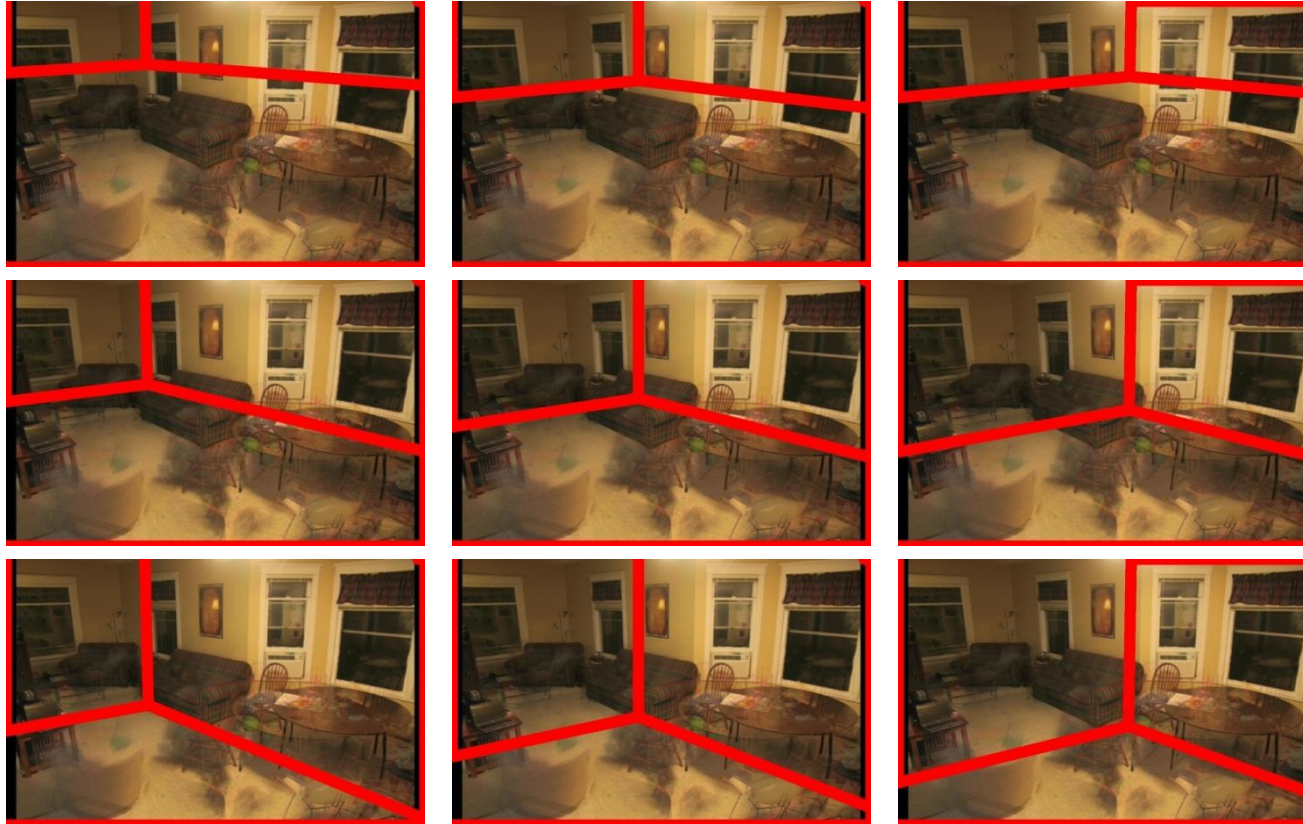


#49





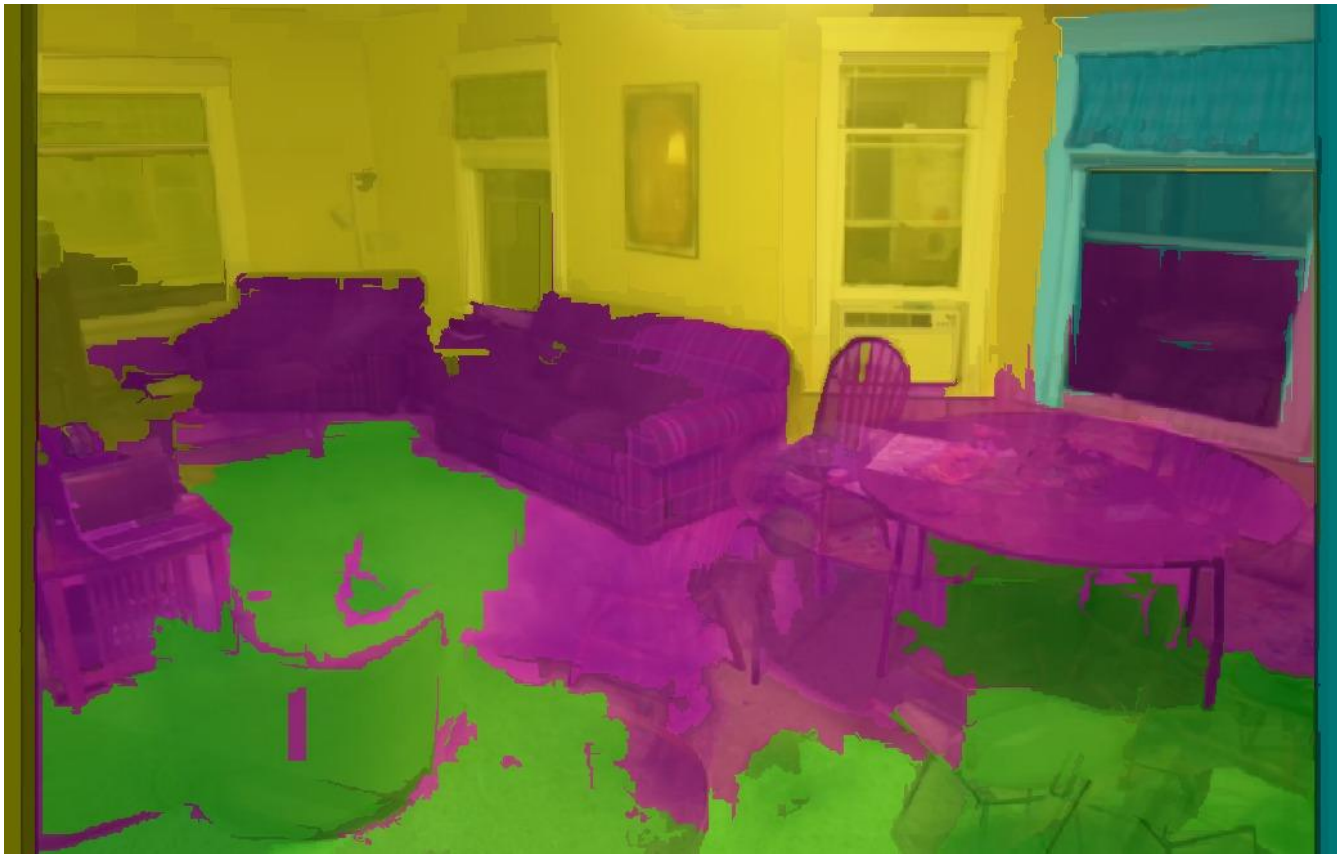
# 3D Room Hypotheses



Vanishing-point aligned hypotheses from  
Hedau et al., '09

# Appearances

Geometric Context: Hoiem et al. '05



## LEGEND

Floor

Wall 1

Wall 2

Wall 3

Ceiling

Clutter

# Appearances Can Be Deceiving

#1



Score = -1.7754

...



#82



Score = -1.8859

# What If We Observe People?

#1



Score = -1.7754

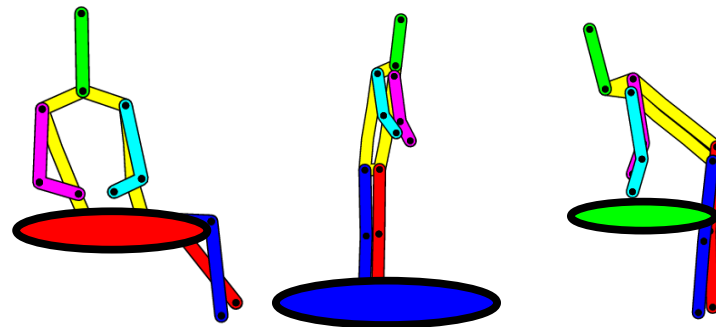
...



#82

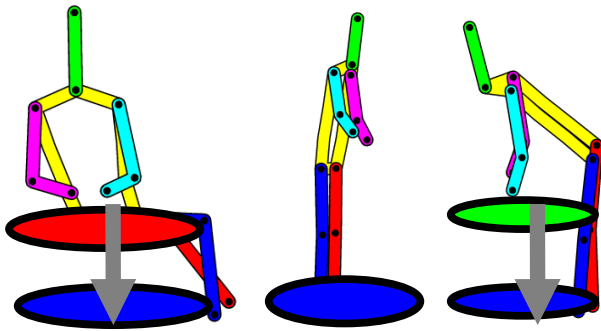


Score = -1.8859



# Penalties From Functional Regions

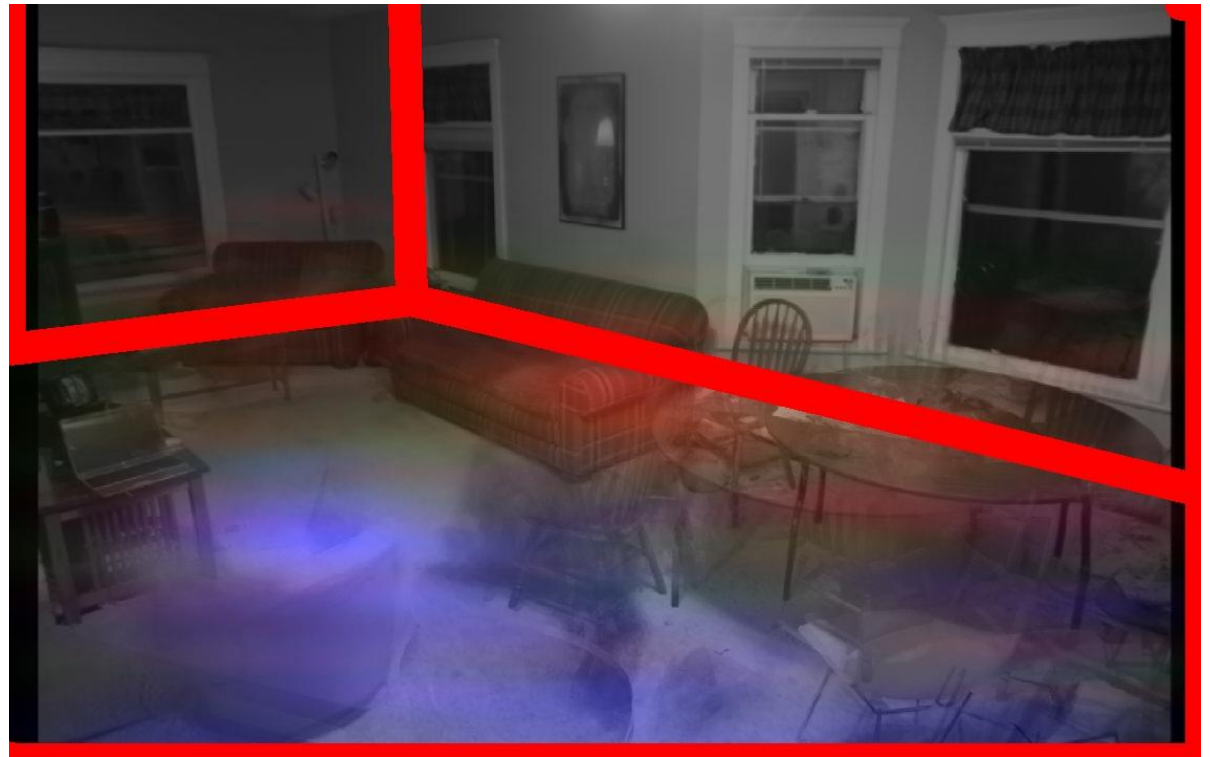
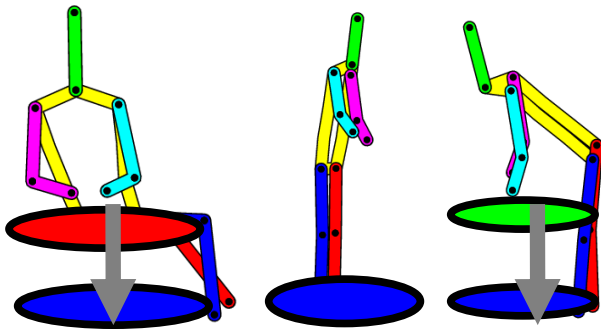
Not explained by room hypothesis



Non-overlap penalty: 0.256

Final Score:  $-1.7754 - 0.256 = -2.0319$

# Penalties From Functional Regions



Non-overlap penalty: 0.0006

Final Score:  $-1.8859 - 0.0006 = -1.8865$

# Reranking Results

Appearance  
Alone

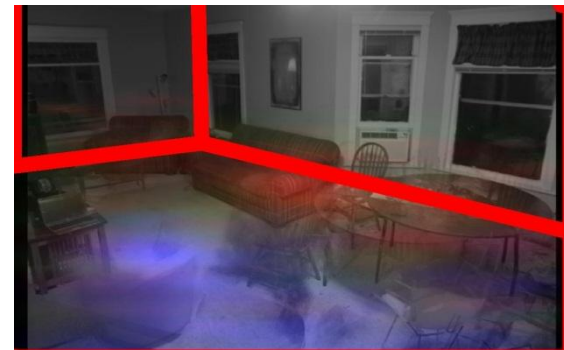
Appearance +  
People

#1



Score = -1.7754

...



#1

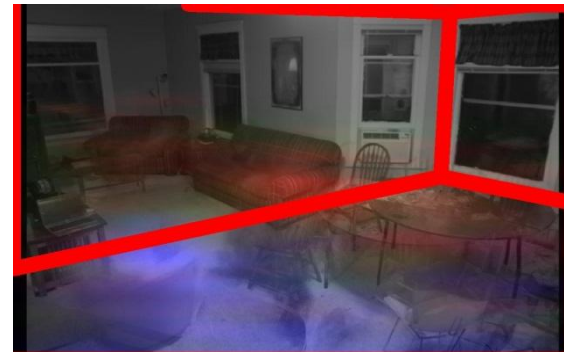
Score = -1.8865

...

#82

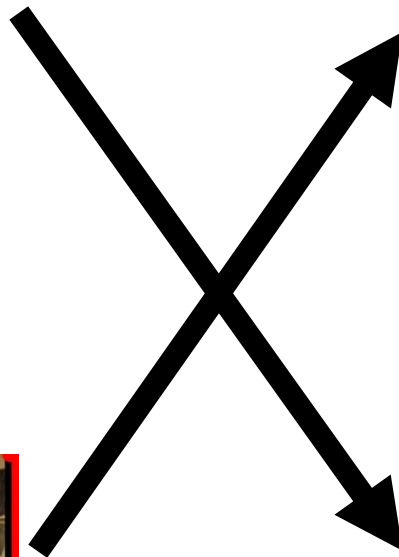


Score = -1.8859

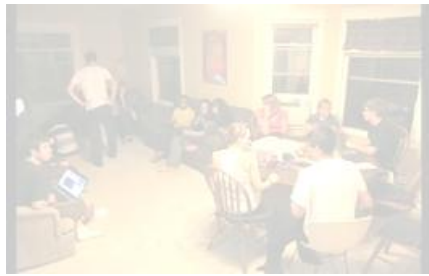


#49

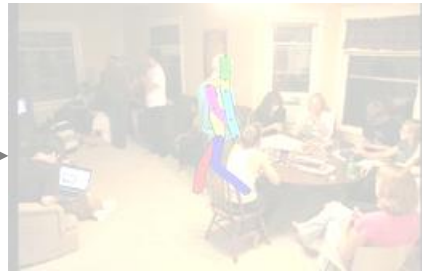
Score = -2.0319



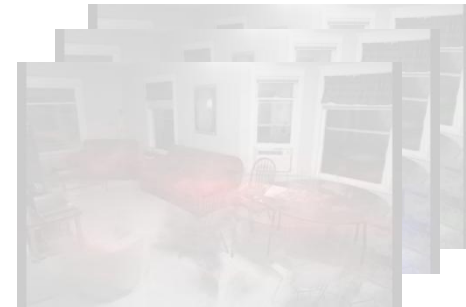
# Our Approach



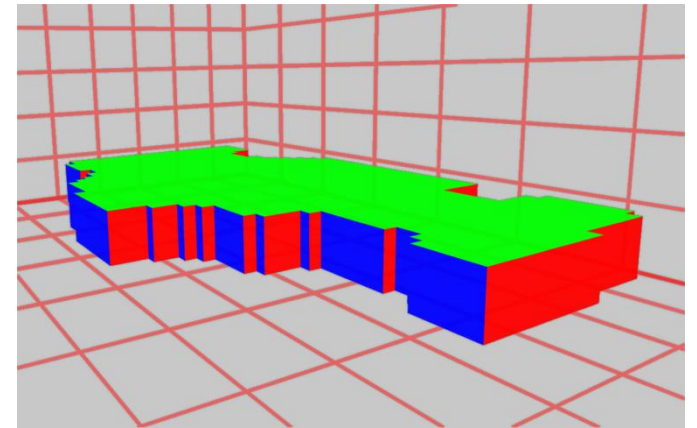
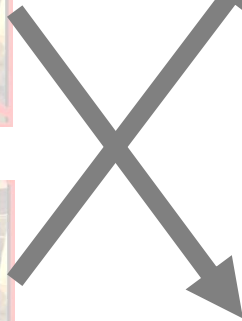
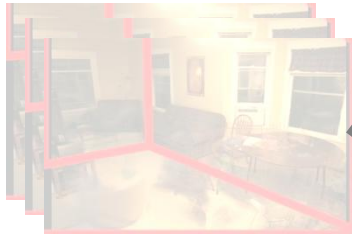
Timelapse



Pose Detections



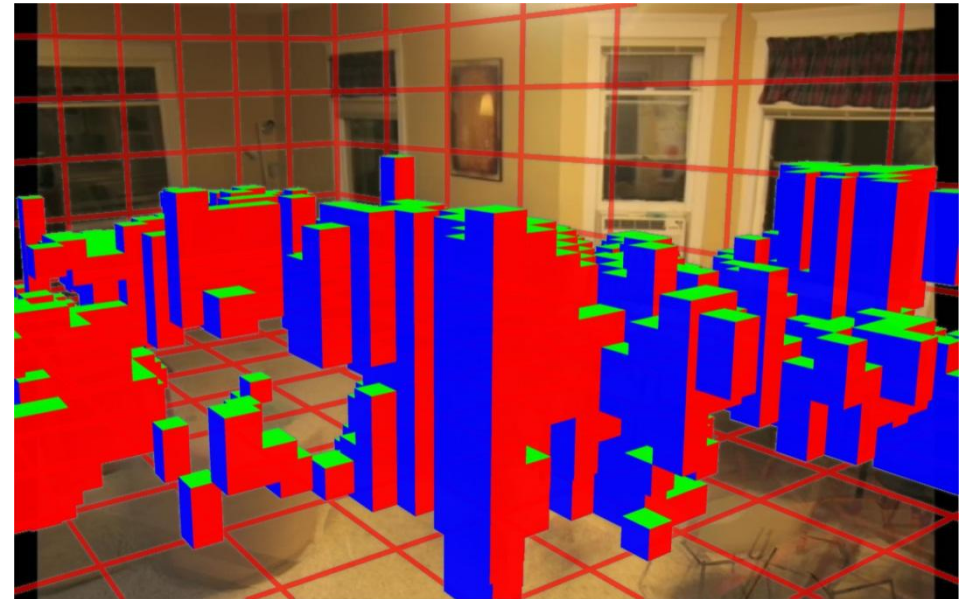
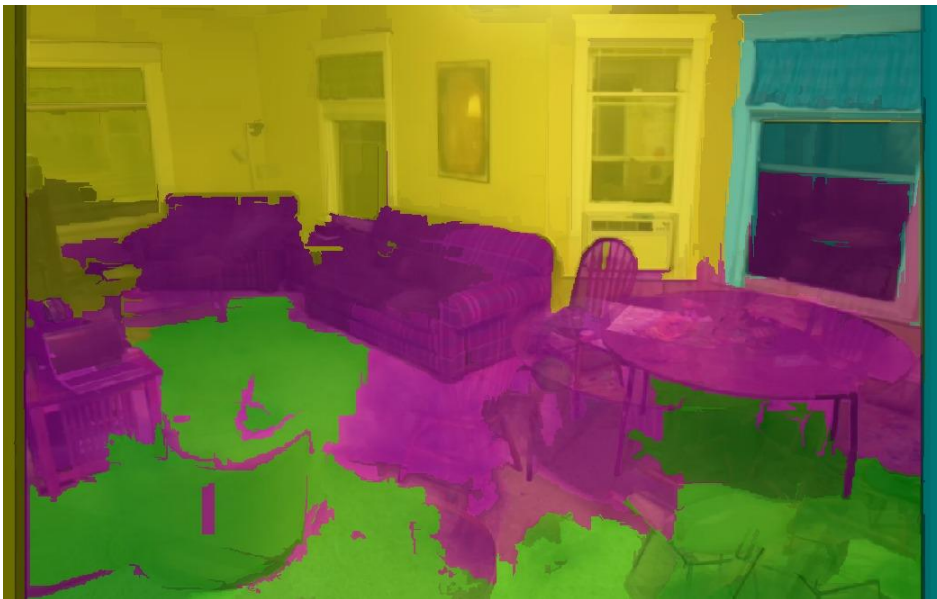
Functional Regions



Estimate  
Free-Space



# Estimating free space



## LEGEND

Floor

Wall 1

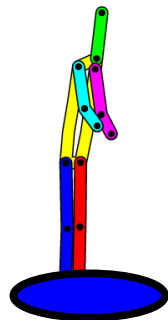
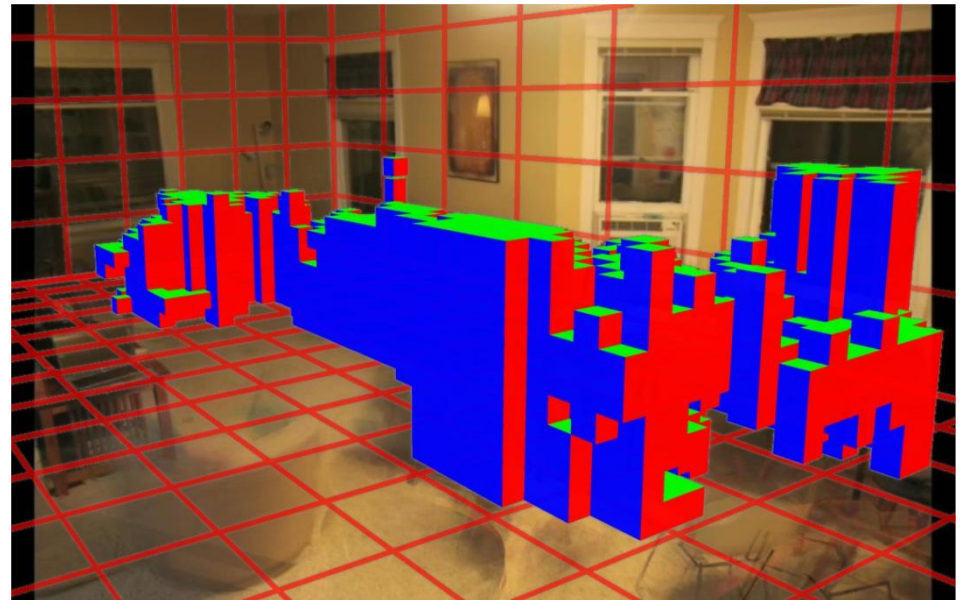
Wall 2

Wall 3

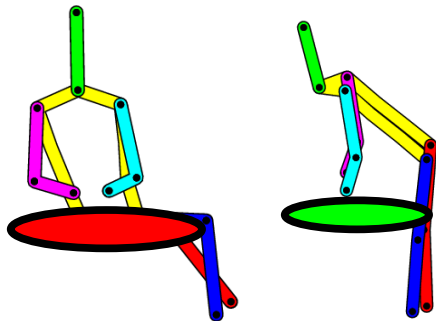
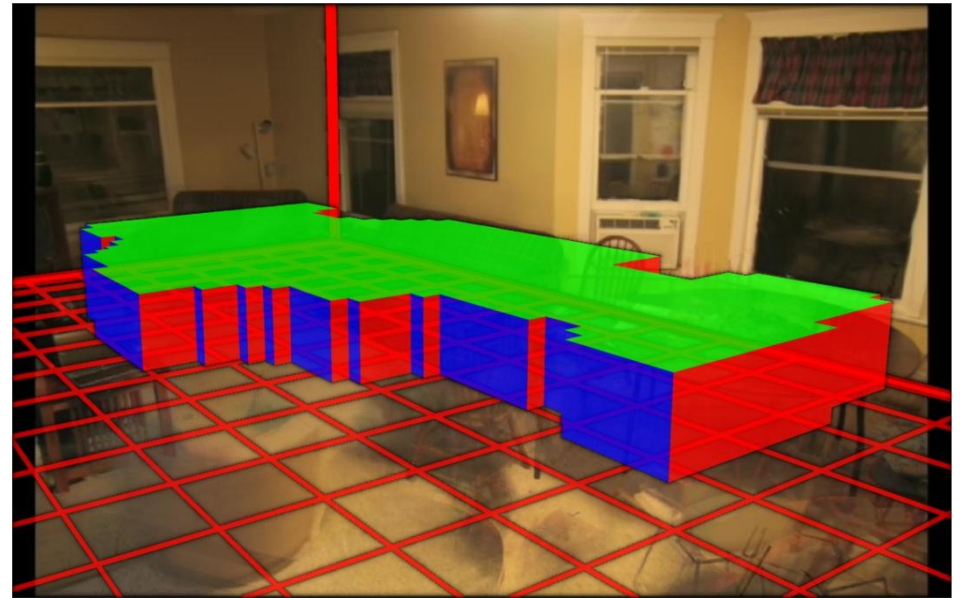
Ceiling

Clutter

# Estimating Free Space



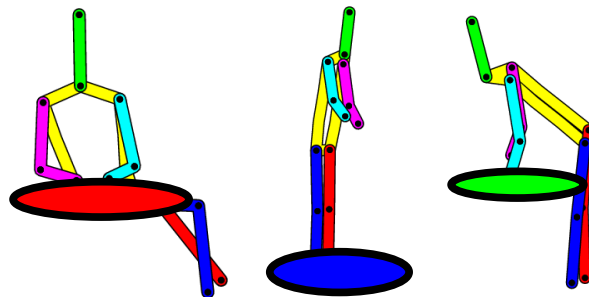
# Estimating Free Space



# Results

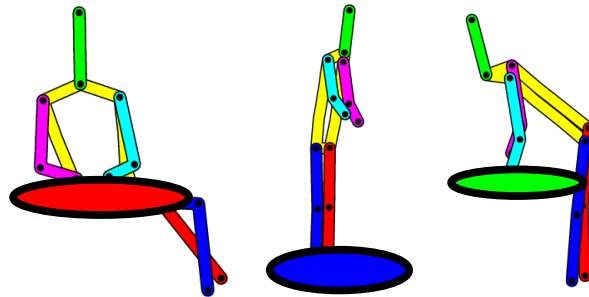
# Qualitative Example

Original video



# Qualitative Example

Original video



# Quantitative Results - Timelapses

40 Timelapse videos from Youtube  
Evaluated on room layout estimation.

Location	Appearance Only		People Only	Appearance + People
	Lee et al. '09	Hedau et al. '09		
64.1%	70.4%	74.9%	70.8%	<b>82.5%</b> <b>(+7.6%)</b>

Does equivalently or better 93% of the time

# Single Images with People



## Appearance Alone

All results at:  
<http://graphics.cs.cmu.edu/projects/peopleWatching/>



# Single Images with People



## Appearance + People

All results at:  
<http://graphics.cs.cmu.edu/projects/peopleWatching/>


# Quantitative Results – Single Image

100 images from Internet  
Evaluated on room-layout estimation.

Location	Appearance Only		Appearance + People
	Lee et al. '09	Hedau et al. '09	
66.4%	71.3%	77.0%	<b>79.6% (+2.6%)</b>

Does equivalently or better 88% of the time

# Limitations



Original video

- Can we learn relationships?
- Can we recover semantics?

# Our Work on Semantics

Input image



Scene Semantics  
from Long-term  
Observation of  
People

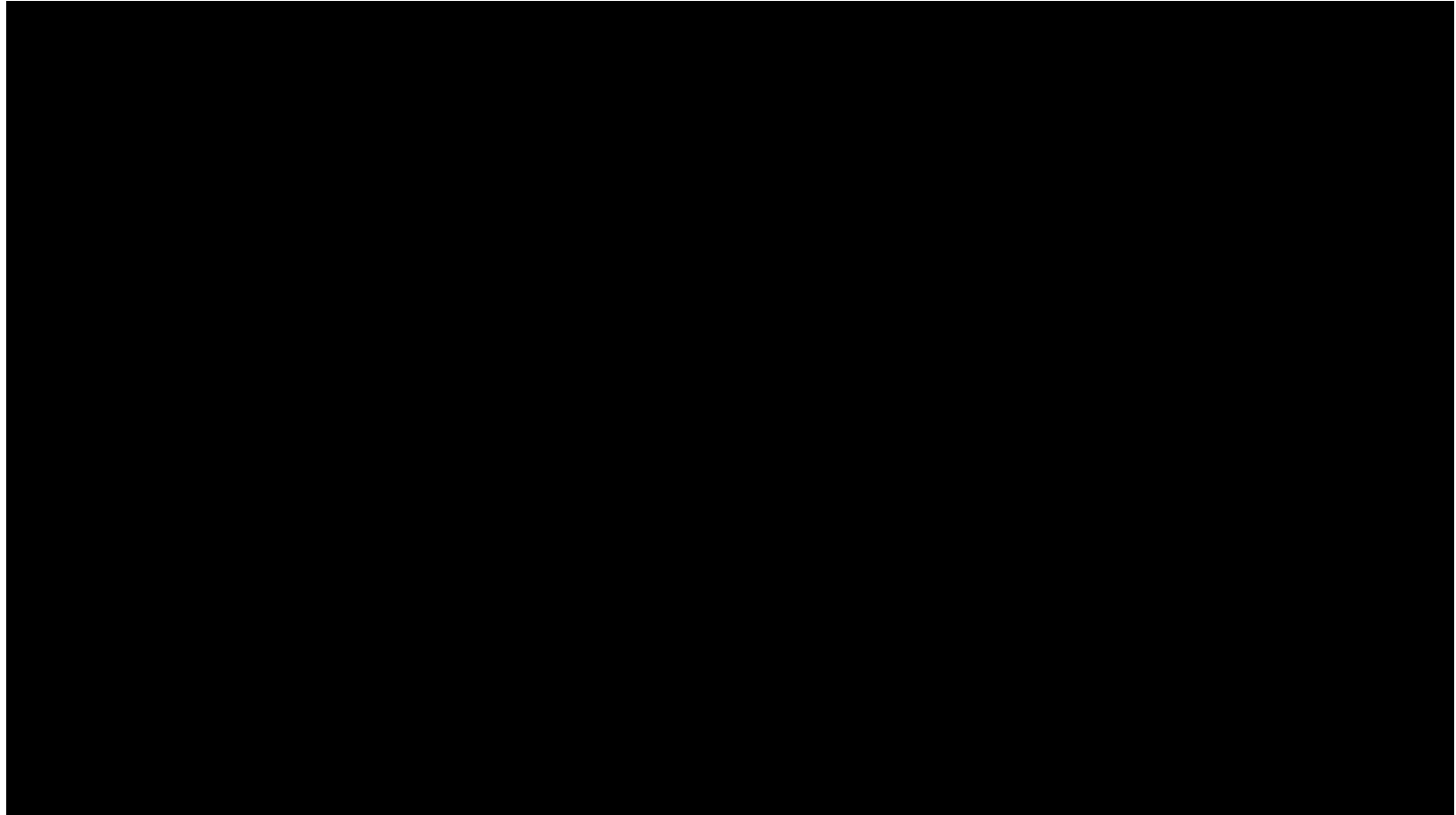
V. Delaitre, D. F. Fouhey,  
I. Laptev, J. Sivic, A. Gupta,  
A. A. Efros

Poster Tomorrow Morning: S7-P5B!

# Conclusions

1. Humans are a valuable cue for understanding scenes.
2. Although pose estimation is not perfect, there's enough signal in the data.

# Thank You



See project page:  
<http://graphics.cs.cmu.edu/projects/peopleWatching/>