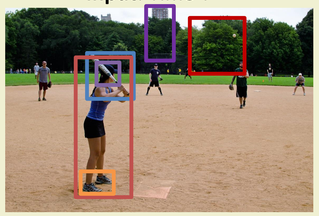
Predicting Object Dynamics in Scenes

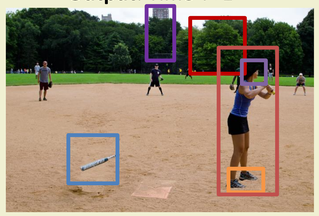
David F. Fouhey[1], Larry Zitnick[2]
[1]Robotics Institute, Carnegie Mellon University, [2]Microsoft Research

Overview

Input: Time T



Output: Time T+1



Goal: Learn mapping

Input: bounding boxes at time T

Output: bounding boxes at time T+1

How can we learn spatiotemporal common-sense to predict that:

- The woman will probably move right?
- The woman's hat will probably go with her?
- The trees will stay still?

Problems:

- How can we model and satisfy these constraints?
- How can we do this if we cannot detect people and hats reliably?

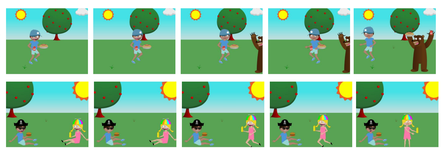
Datasets

Prediction models were trained on Abstract Scenes Data and applied to both abstract and natural scenes.

Abstract Scenes Data:

5,000 sequences of 5 scenes
Gathered via Amazon Mechanical Turk (AMT)

Interface

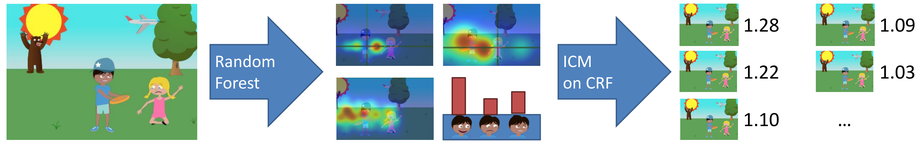


Sample Stories

Natural Images:

225 natural images from flickr.com with bounding-boxes labeled
Only one image; ground-truth predictions labeled by AMT

Method

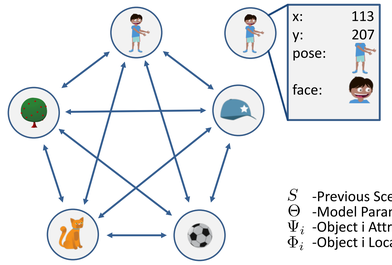


Object Locations (Time T)

Potentials for Objects (Time T+1)

Scenes + Scores

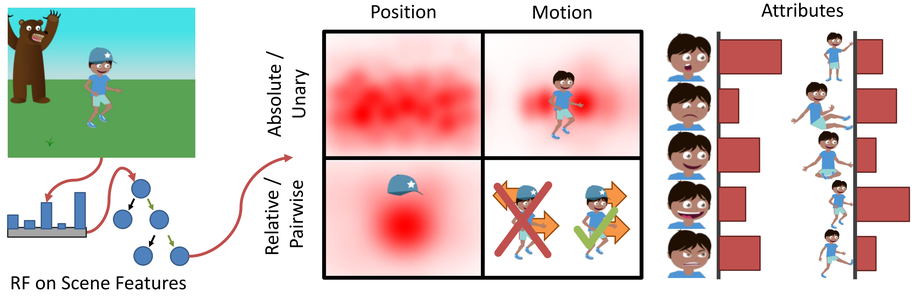
CRF Scene Model



S -Previous Scene
Θ -Model Parameters
Ψ_i -Object i Attributes
Φ_i -Object i Location

$$\log P(\Phi, \Psi | S, \Theta) =$$
$$\sum_i \left(\overbrace{\lambda(\Phi_i; \theta_\lambda)}^{abs.\ location} + \overbrace{\omega_i(\Phi_i, S; \theta_\omega)}^{abs.\ motion} + \overbrace{\pi_i(\Psi_i, S; \theta_\pi)}^{attributes} \right)$$
$$+ \sum_{i,j} \left(\overbrace{\phi_{i,j}(\Phi_i, \Phi_j; \theta_\phi)}^{rel.\ location} + \overbrace{\varphi_{i,j}(\Phi_i, \Phi_j, S; \theta_\varphi)}^{rel.\ motion} \right)$$
$$+ \overbrace{\alpha(\Phi, S; \Theta_\alpha)}^{motion\ prior} + Z(S, \Theta)$$

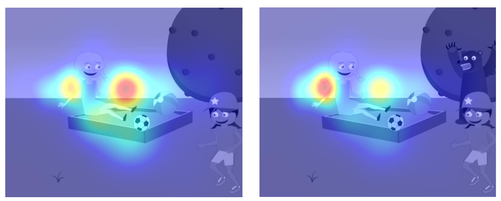
CRF Potentials learned by Random Forest



RF on Scene Features

Learned Motion Potentials

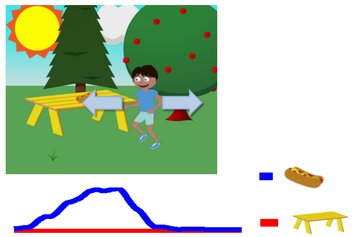
Distribution on Jenny's motion vector in the same scene with and without a bear present.



Without Bear

With Bear

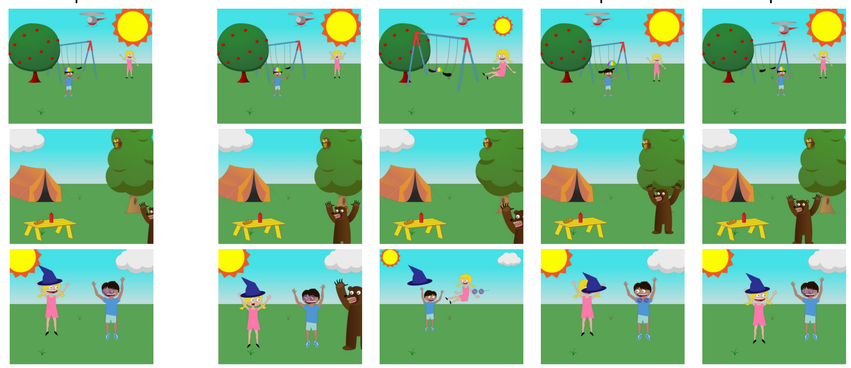
P(motion) as a function of Mike's position



Results

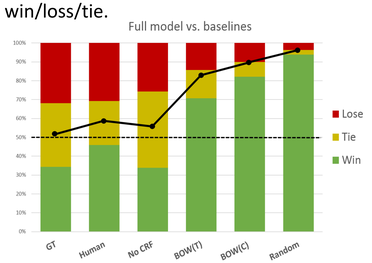
Compare with:

Global Prediction: BoW + Motion Vector Transfer;
Independent Prediction: Use RF on same features;
Humans: Ask another Turker to complete.



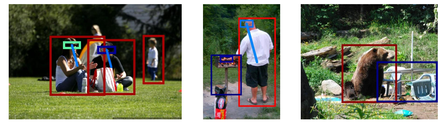
Human Evaluation

Another human rates pairs of predictions win/loss/tie.

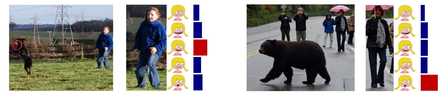


Natural Image Results

Who moves and how?
Hot color: likely to move. Line: moves together.



Predicted attributes: what is Jenny feeling?



Quantitative Evaluation

Ask yes/no question about scene; compare original sequence and prediction.

Abstract Scenes Dataset

F1 Score:	Does X move?	Do X and Y move together?	Does X move Left or Right?	What are X's Attributes?
Proposed	**49.3**	**42.9**	66.6	**61.9**
Indep.	49.0	11.9	**75.1**	61.5
Humans	48.8	31.8	70.0	52.3
Global	39.6	16.2	61.9	29.1

Natural Scenes Dataset

F1 Score:	Does X move?	Do X and Y move together?
Proposed	91.2	45.2
Indep.	87.5	14.4
Humans	**97.8**	**75.3**
Global	69.3	5.7

Conclusions

- Neither global nor independent prediction models produce accurate joint motion: joint motion must be modeled explicitly.
- Our models are short-term common-sense only; long-term prediction with narratives is an interesting future direction.