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)

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

Method

CRF Scene Model  
\[
\log P(\Psi, S | \Theta) = \sum_{i} \log(\lambda(\Psi_i, \Theta_i)) + \omega(\Psi_i, S_i, \Theta_i) + \pi(\Psi_i, S_i, \Theta_i) + \sum_{i,j} \gamma_{ij}(\Psi_i, \Psi_j, S_i, S_j) + \sum_{i,j} \psi_{ij}(\Psi_i, S_i, S_j) + Z(S, \Theta)
\]

CRF Potentials learned by Random Forest

Learned Motion Potentials  
Distribution on Jenny’s motion vector in the same scene with and without a bear present.

Results

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.

Quantitative Evaluation

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

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.