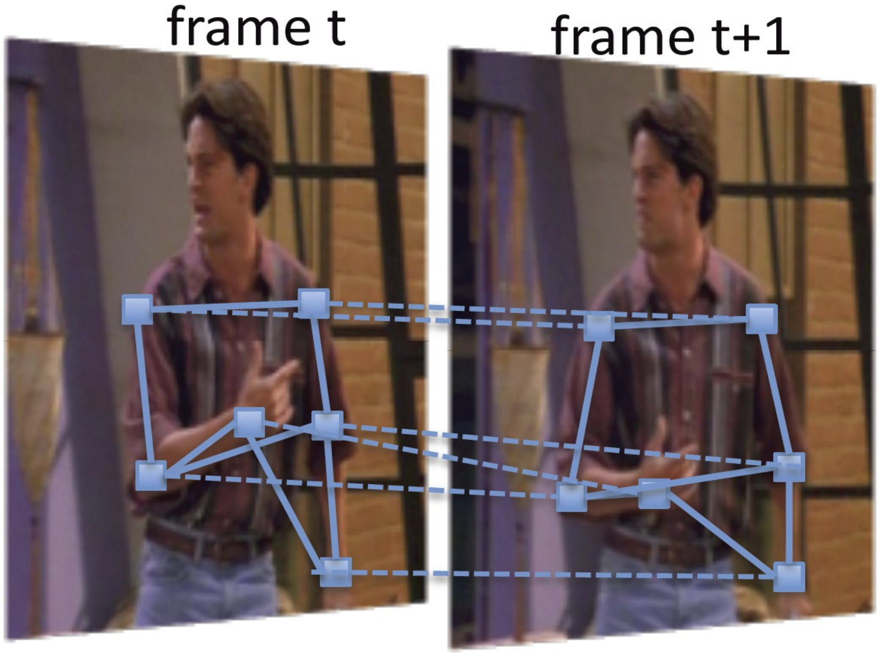
Parsing Human Motion with Stretchable Models

from CVPR2011 by Sapp, Weiss, Taskar at University of Pennsylvania
interpreted by Magnus Burenius at KTH

- Estimate 2D motion of arms in video.
- Similar to Pictorial Structures with graph edges over time.
- Part detectors based HOG, color, contours, optical flow, etc.
- Full graph handles kinematic connections, left and right symmetry and temporal continuity.
- Various approximations of the full graph for tractable inference.

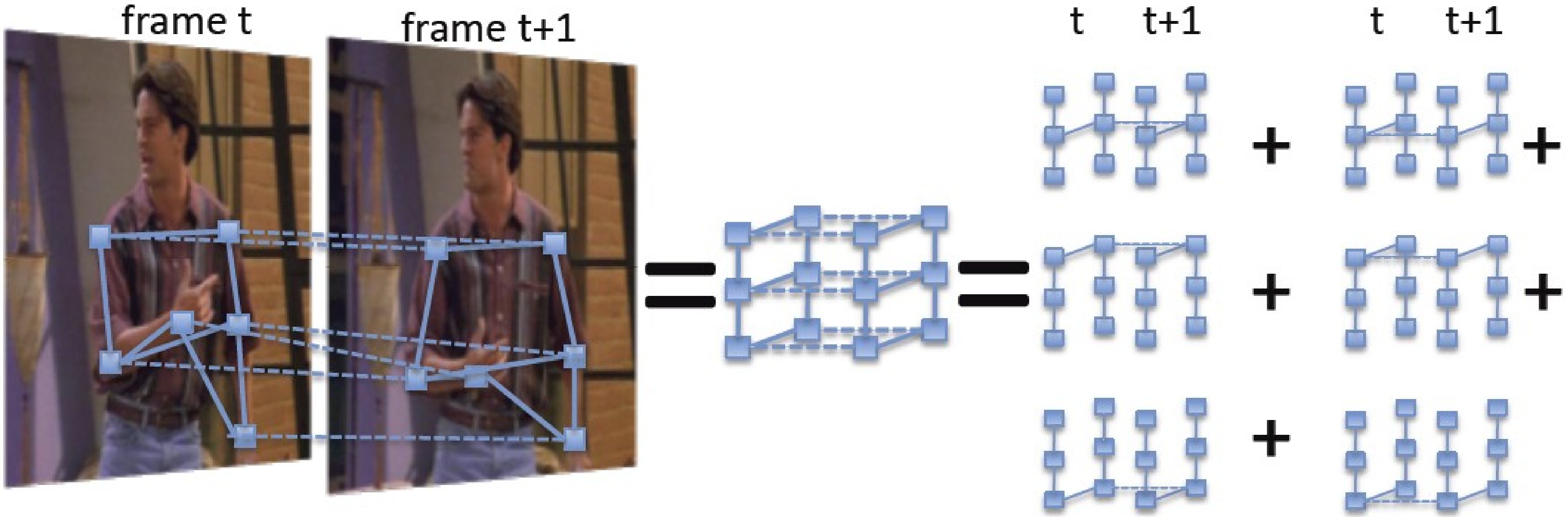


Parsing Human Motion with Stretchable Models

- They have focused on arms since they are the most difficult part to estimate, while being very important for action recognition.
- They run a single frame pictorial structures implementation to get 500 candidate positions for each joint and each frame.
- They want to find the optimal motion from the candidates, but doing exact optimization over the full graph is intractable.
- Decomposes the full graph model into tree models which make exact inference possible.
- Explore ways to couple these models.

One Model per Joint

- Each tree sub-model handles the kinematic connections of the joints and the left-right-symmetry and temporal continuity of a single joint.



Joint Optimization over Time

x input video
y estimated part positions for all frames

Score:

$$\theta(x, y) = \sum_{i \in V} \theta_i(x, y_i) + \sum_{(i, j) \in E} \theta_{i, j}(x, y_i, y_j)$$

$$\underset{y}{argmax} \quad \theta(x, y)$$

Not tractable to solve exactly

Decompose Graph into Trees

Instead of finding the maximum score defined over the full graph:

$$\underset{y}{argmax} \ \theta(x, y)$$

They use the decomposition of the graph, with a tree for each joint, and try to find total maximum score over these:

$$\underset{y}{argmax} \ \sum_{m} \theta_m(x, y)$$

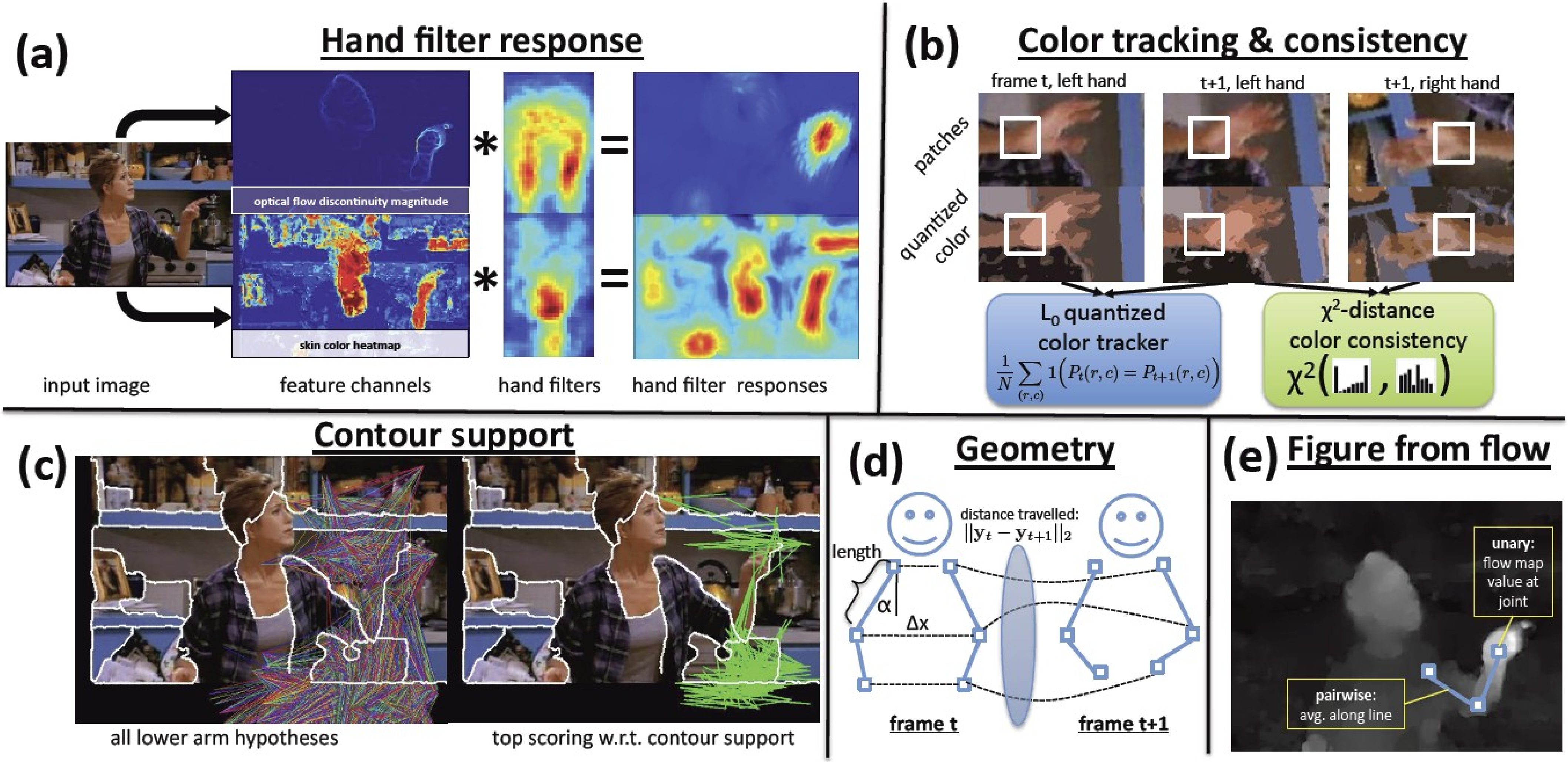
The models θ are trained discriminatively using a large-margin loss.

Four Inference Approaches

$$\underset{y}{argmax} \sum_{m} \theta_m(x, y)$$

- Full Agreement via Dual Decomposition (iterative approximation, 500x slower).
- Single Frame Agreement (exact).
- Single Variable Agreement (exact).
- Independent (exact).
  Use each model only for the joint it connects over time.

Features

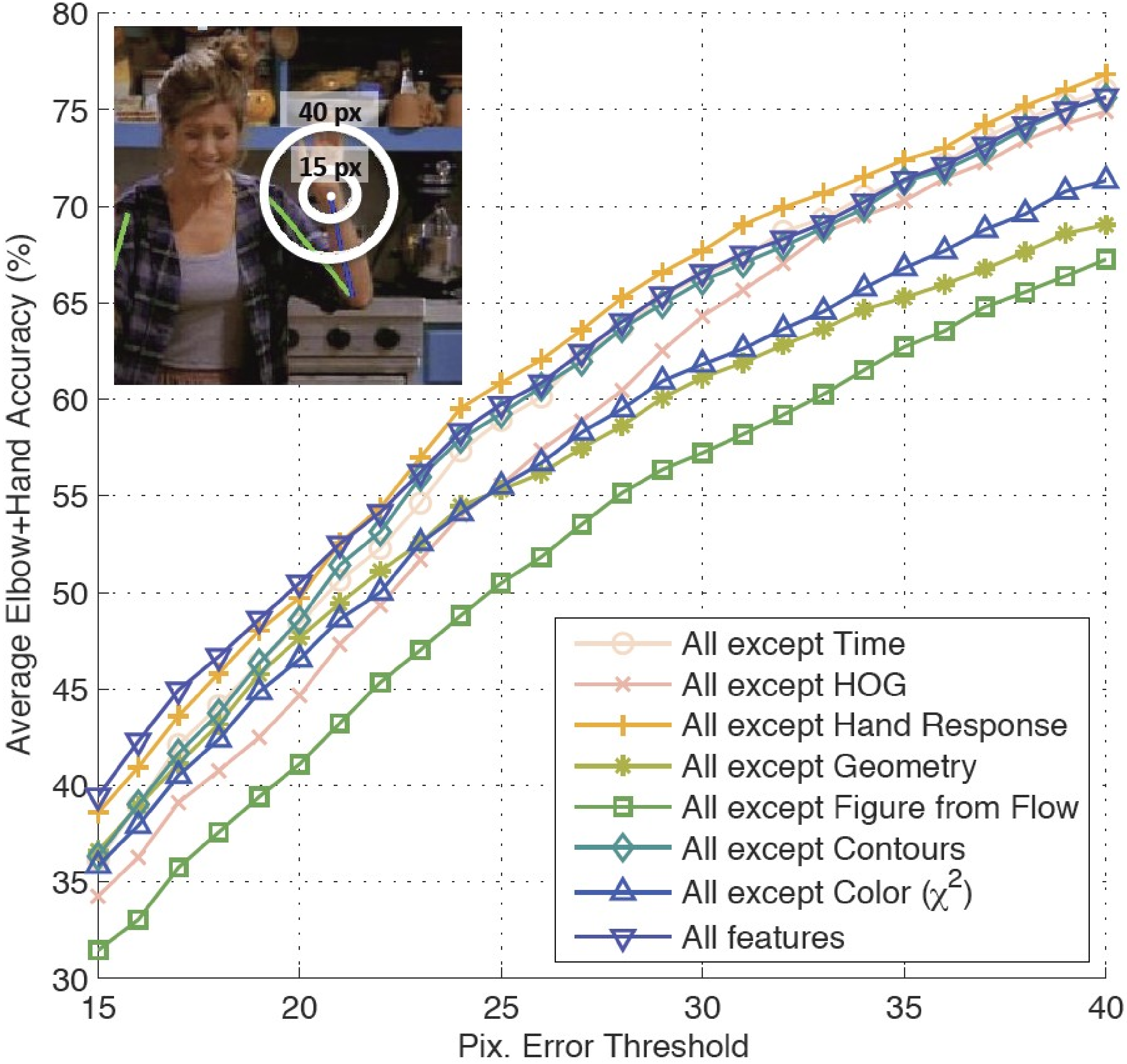

(a)
Hand filter response
optical flow discontinuity magnitude
skin color heatmap
*
*
=
=
input image
feature channels
hand filters
hand filter responses
(b)
Color tracking & consistency
frame t, left hand
t+1, left hand
t+1, right hand
patches
quantized color
L_0 quantized color tracker
$\frac{1}{N}\sum_{(r,c)} 1\big(P_t(r,c) = P_{t+1}(r,c)\big)$
χ^2-distance color consistency
$\chi^2($, $)$
(c)
Contour support
all lower arm hypotheses
top scoring w.r.t. contour support
(d)
Geometry
distance travelled:
$||\mathbf{y}_t - \mathbf{y}_{t+1}||_2$
length
α
Δx
frame t
frame t+1
(e)
Figure from flow
unary: flow map value at joint
pairwise: avg. along line

Videos

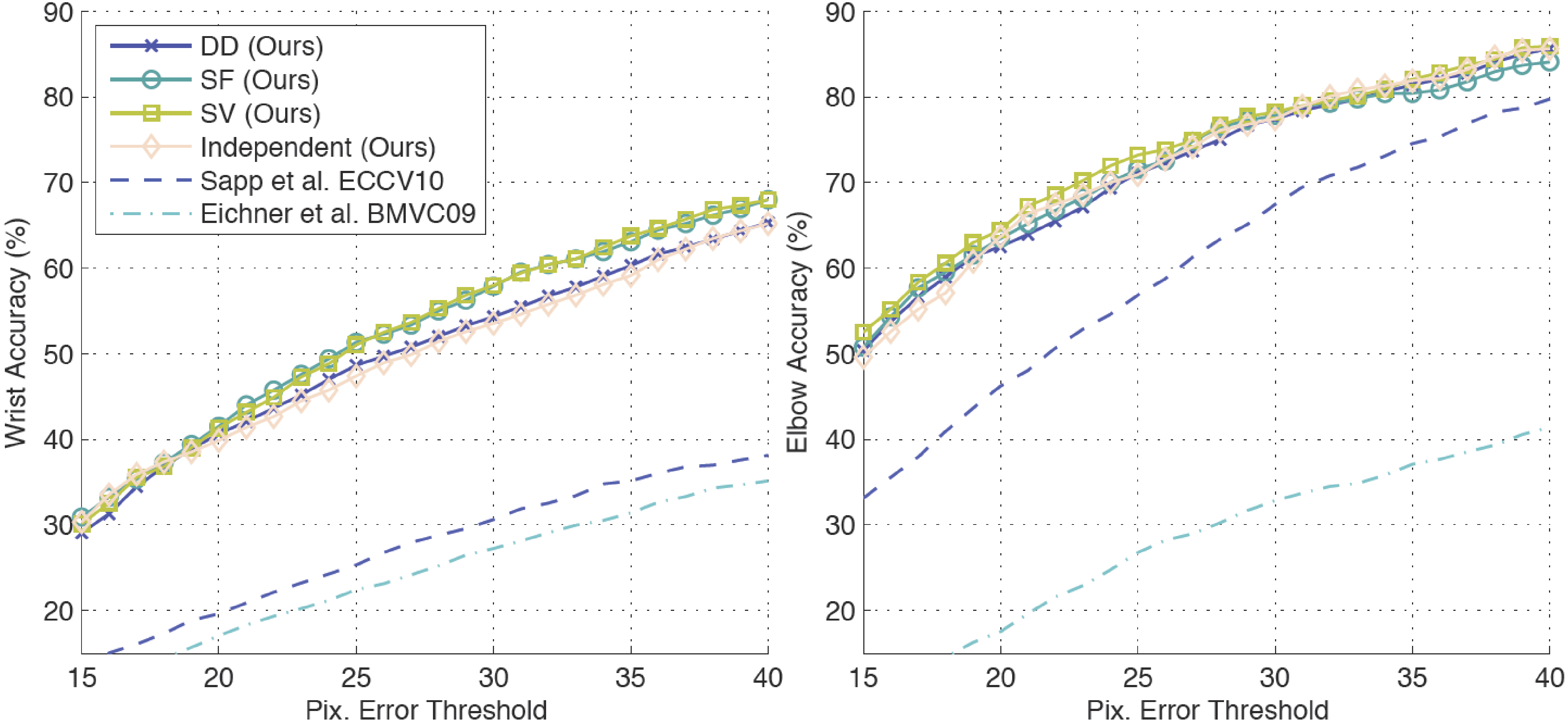
http://www.seas.upenn.edu/~dwe/

http://www.seas.upenn.edu/~bensapp/

Feature Importance



Inference Approaches



Single Frame Baselines

Sapp, Toshev, Taskar.	Cascaded models for articulated pose estimation.	ECCV 2010
Eichner, Ferrari.	Better appearance models for pictorial structures. (They also tried adding temporal continuity with loopy belief propagation but got worse results)	BMVC 2009