Human Action Recognition: Paper Review

Sourabh Daptardar, Minh Hoai Nguyen

September 16, 2014

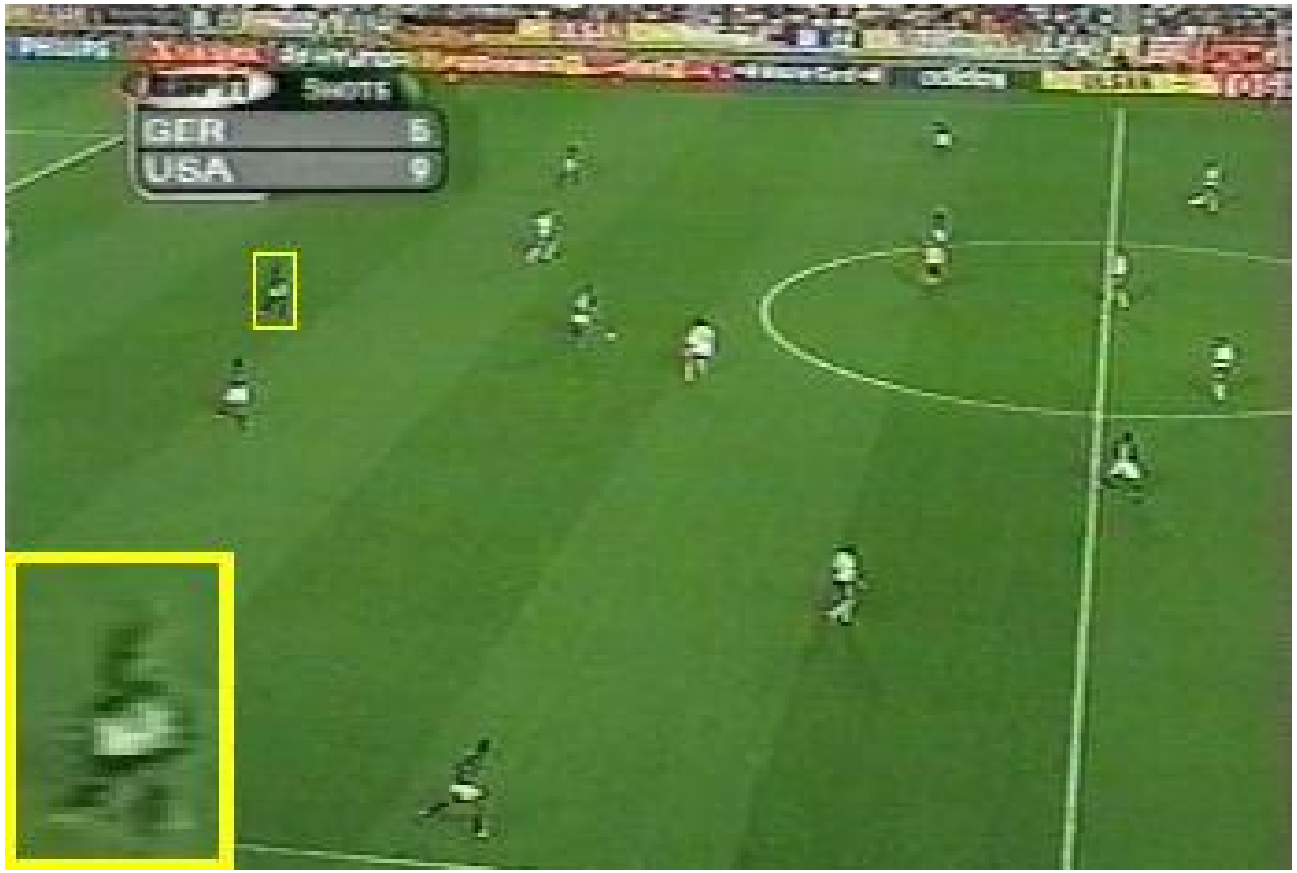
Outline

1. Recognizing Action at a Distance: Efros et al, ICCV 2003

2. Learning realistic human actions from movies: Laptev et al, CVPR 2008

NTSC World Cup broadcast: action recognition from "medium" field view

Goal: recognize human actions at a distance



Algorithm

- Track and stabilize moving figure
- Compute motion descriptors which compute "residual" caused by body parts
- Classification: nearest neighbor matching with a dataset

What do we desire in feature descriptors?

- Motion independent of appearance.
- Reliability of matching in case of noisy data.
- Discriminative enough

Motion Descriptor Matching

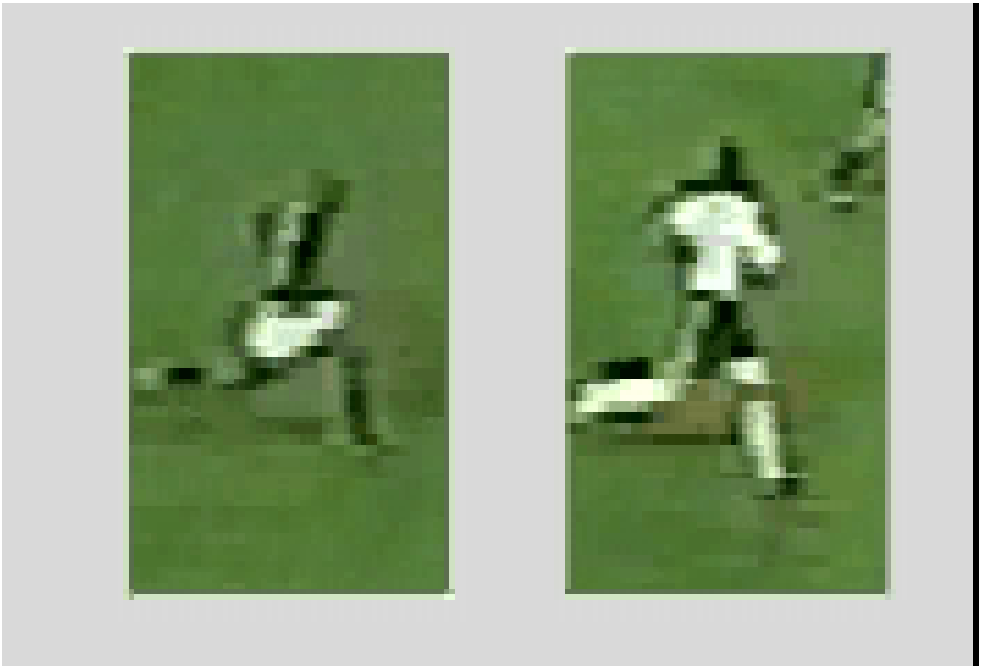
Given a stabilized figure-centric sequence:

- Compute optical flow at each frame.
- Split the optical flow field $F = (F_x, F_y)$ into 4 half wave rectified components F_x+, F_y+, F_x-, F_y-
- Blur each of these with Gaussian.
- Descriptors are compared by normalized cross correlation

Motion Descriptor Matching: 16 ballet actions dataset

Motion Descriptor Matching: Tennis dataset

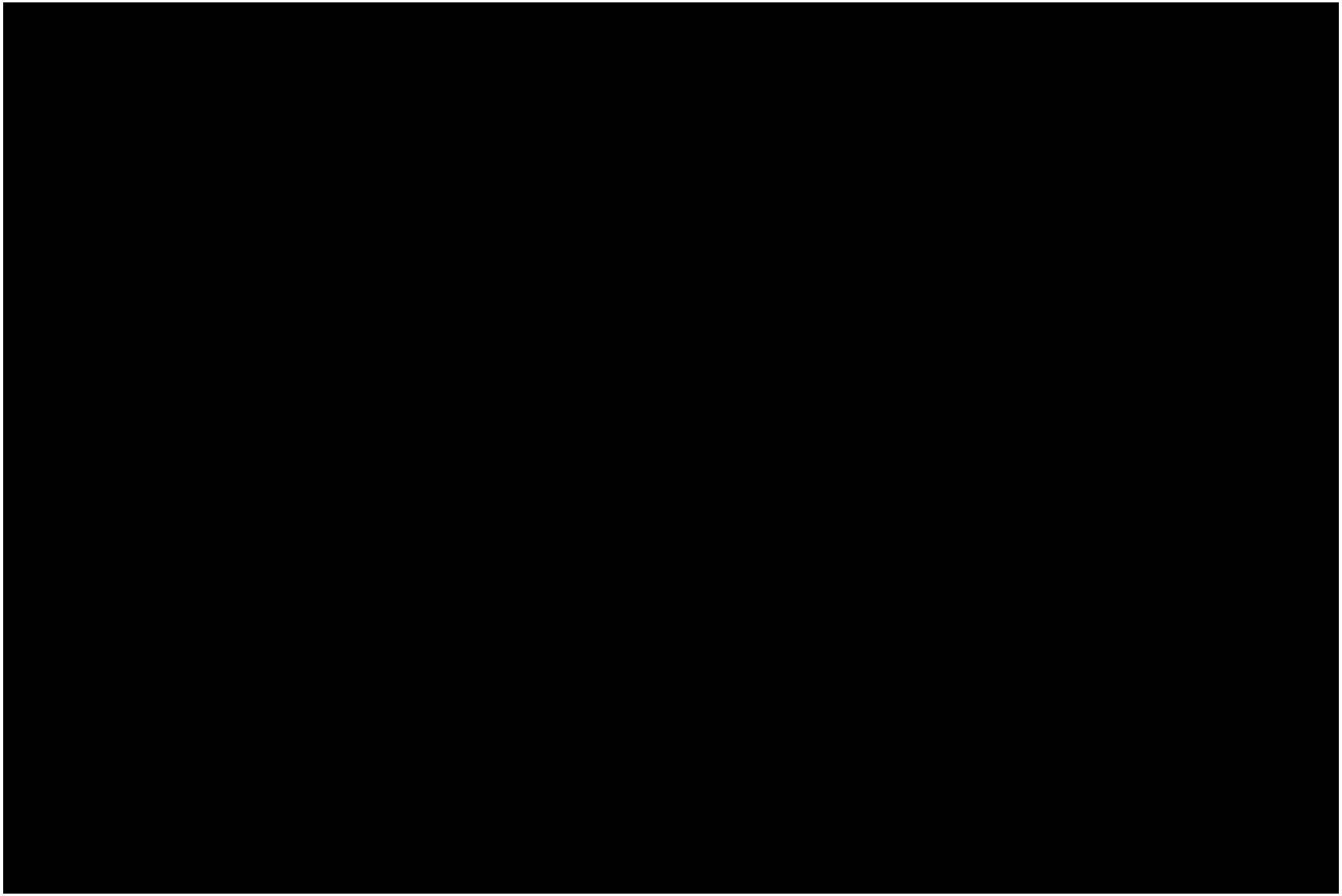
Motion Descriptor Matching: Best Match



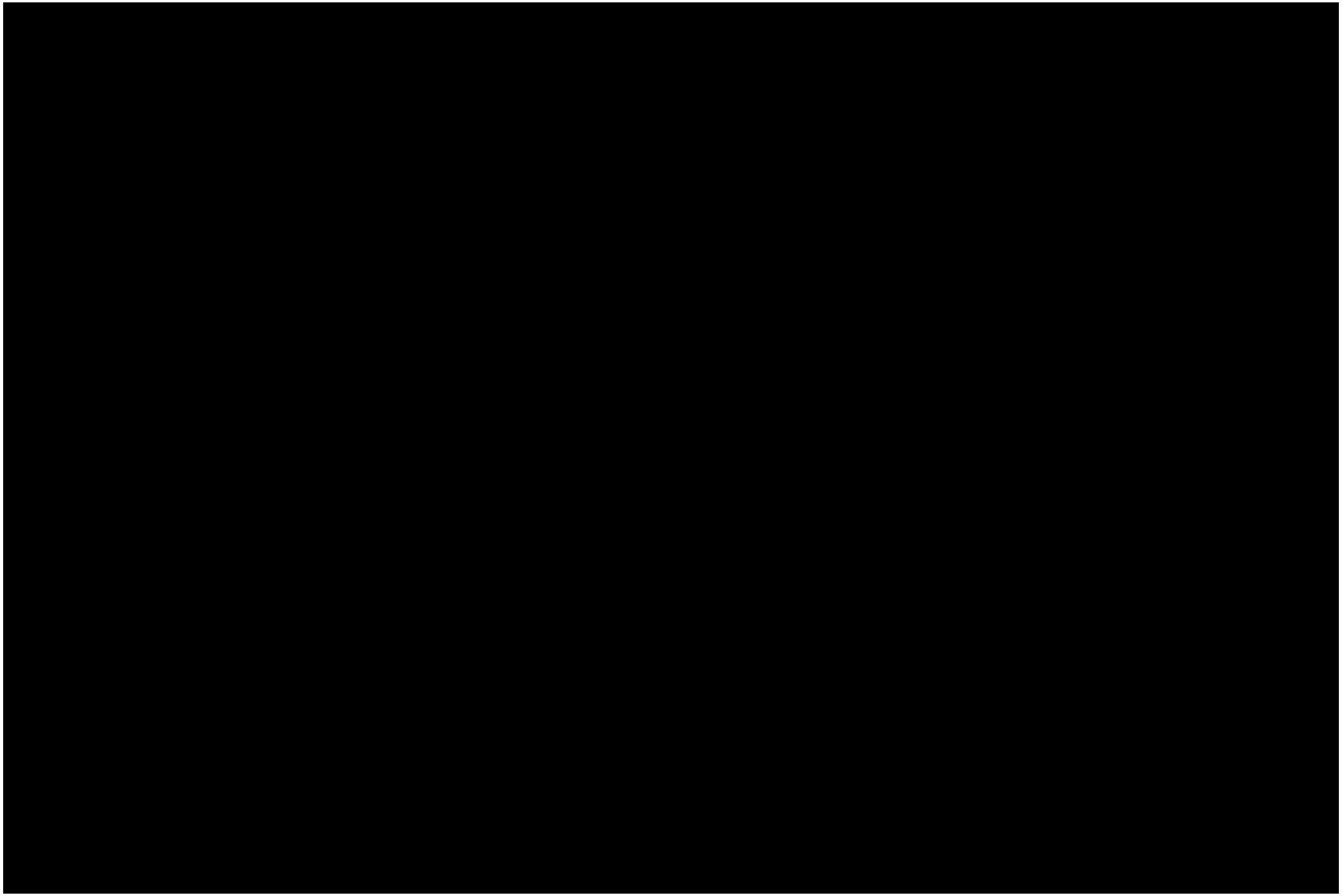
Motion Descriptor Matching: Best Match (Single Player)

Classification (Football)

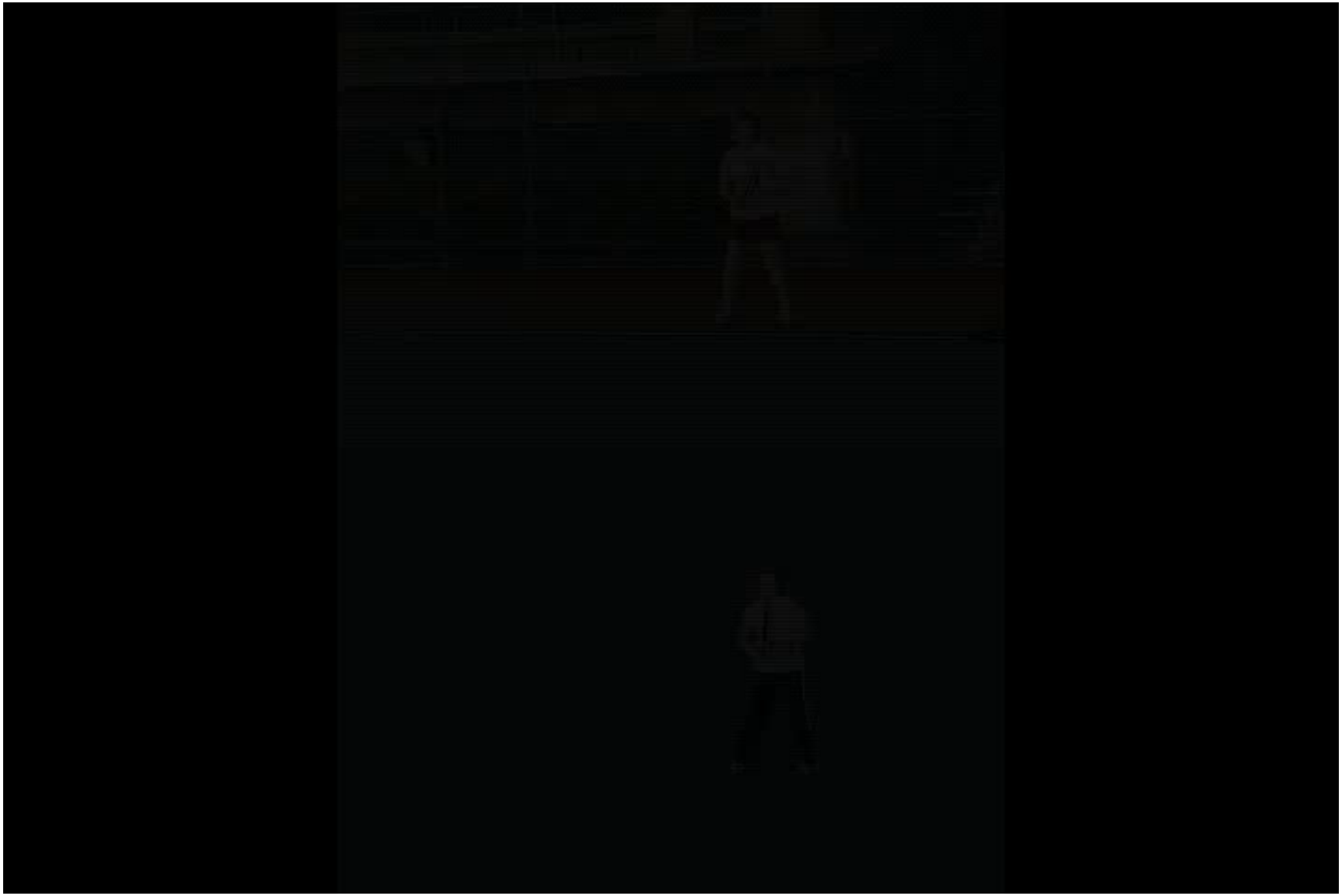
Classification (Tennis)



Action Synthesis: Do As I Do 1



Action Synthesis: Do As I Do 2



Action Synthesis: Do As I Say 1

Action Synthesis: Do As I Say 2

So far: the datasets looked like this

Learning realistic human actions from movies

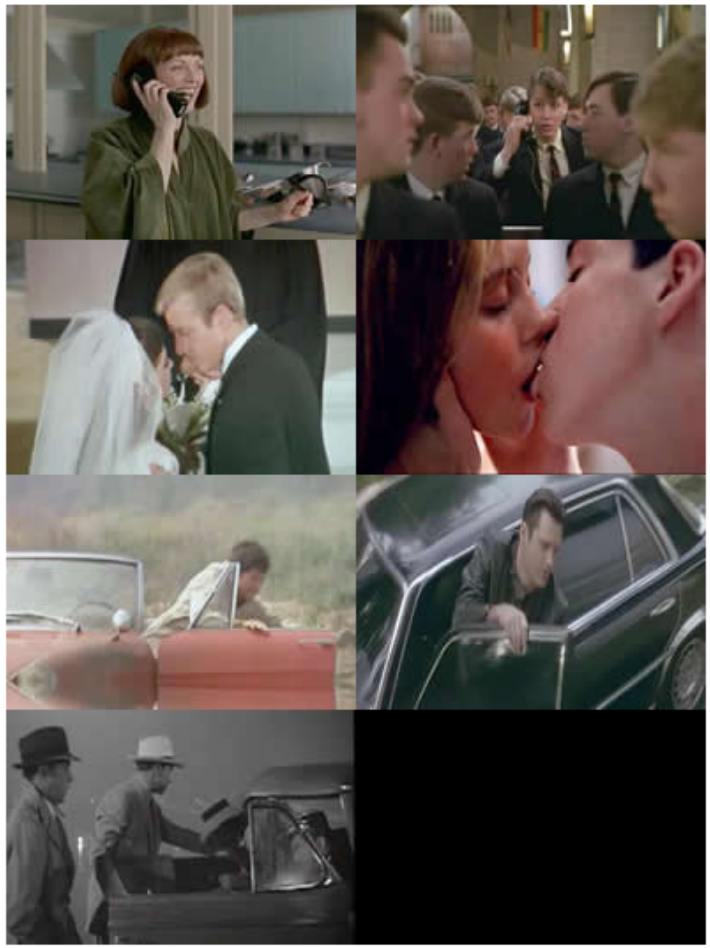
Movies

- American Beauty
- Being John Malkovich
- Big Fish
- Casablanca
- .....

Actions (Classes)

- AnswerPhone
- GetOutCar
- HandShake
- HugPerson
- Kiss
- SitDown
- SitUp
- StandUp

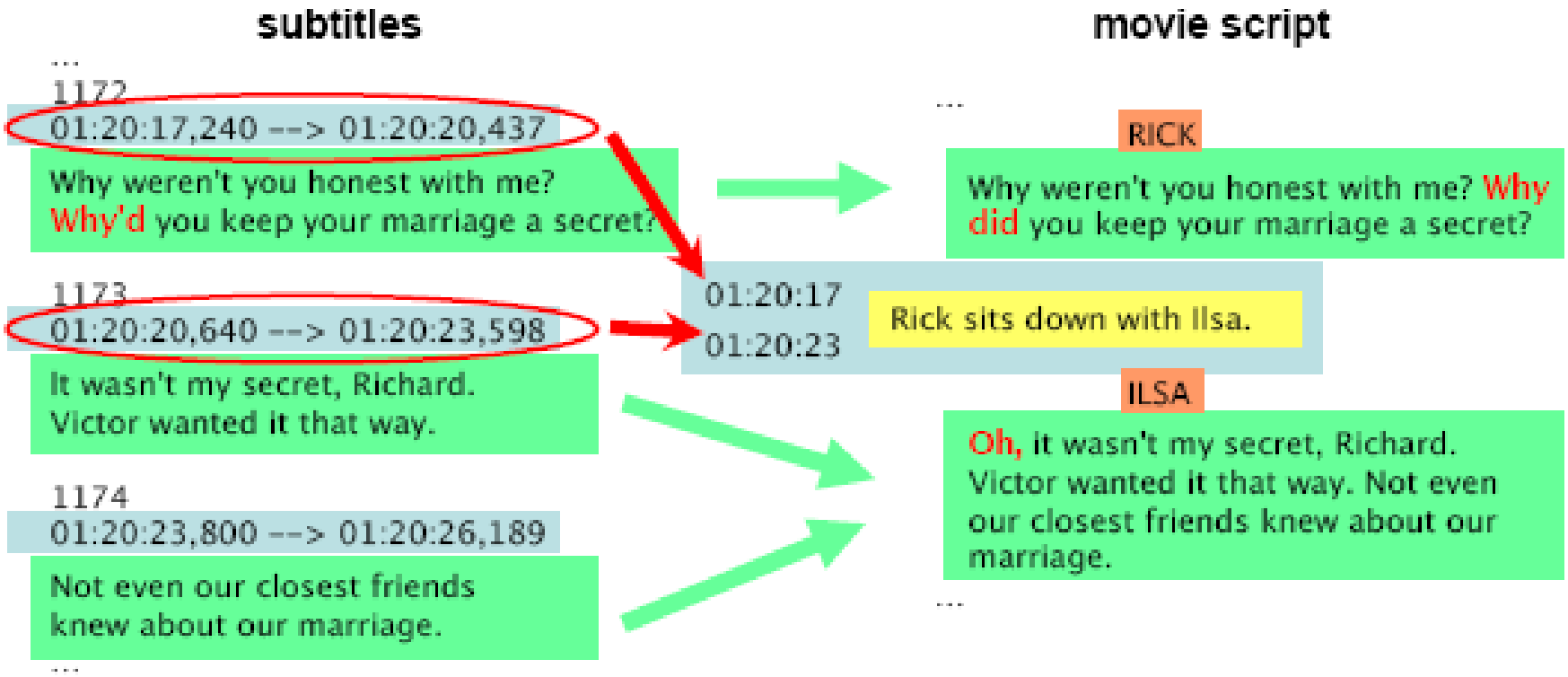
Figure : Variability in Actions



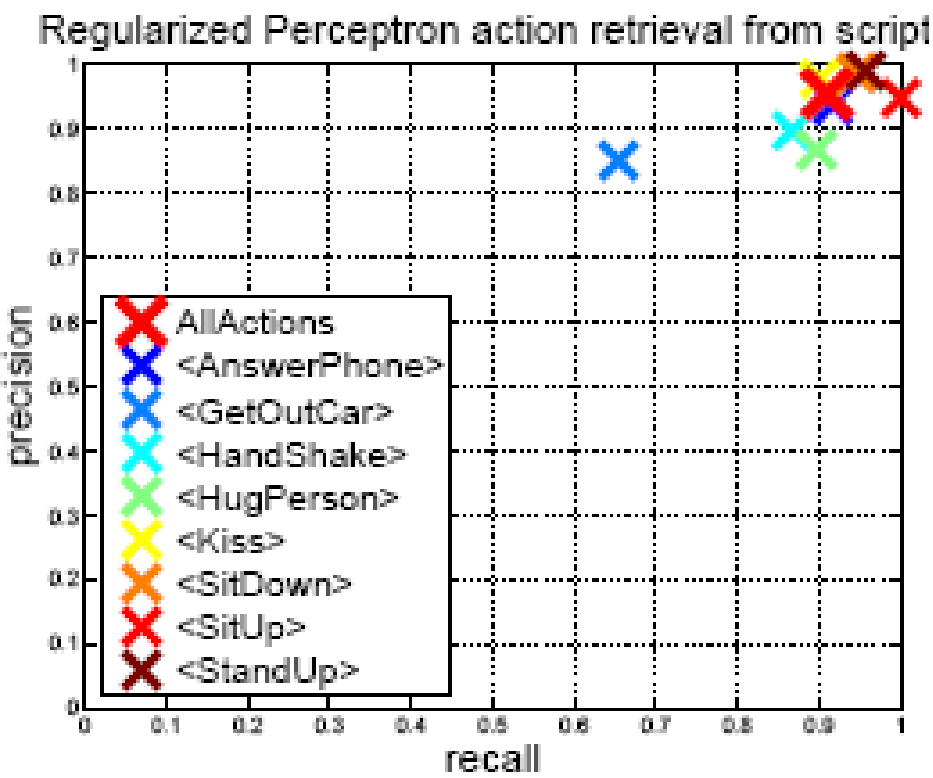
Automatic annotation of human actions

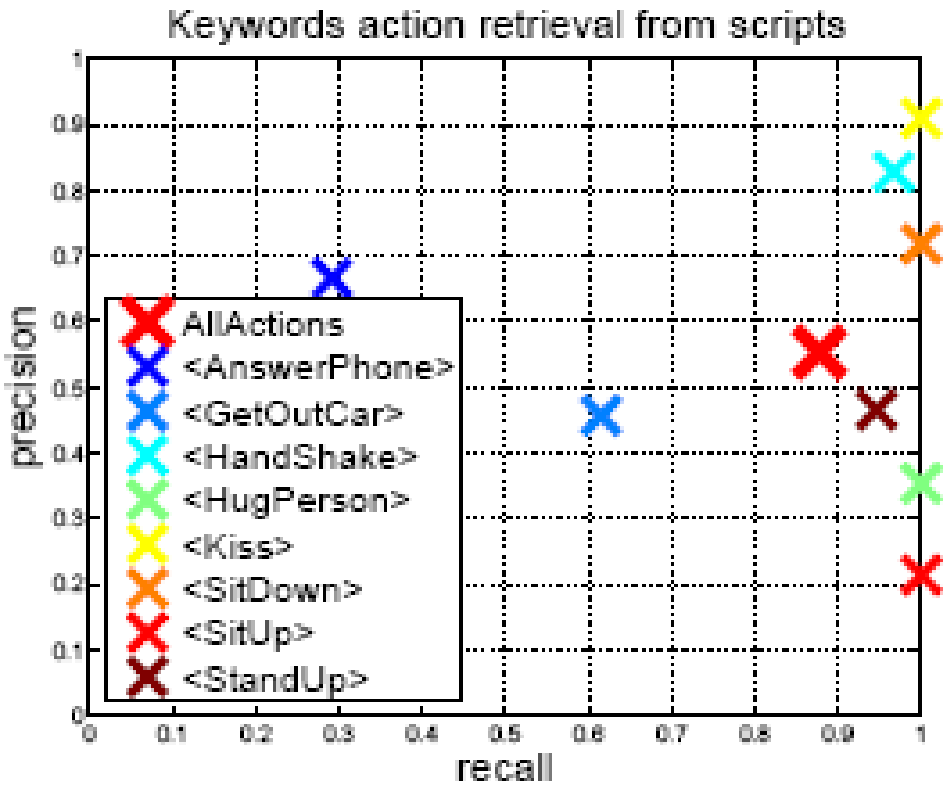
- Match speech sections in subtitles and scripts
- Transfer time information from subtitles to scripts
- alignment score
- Bag-of-features model of text classification.
- Features: words, adjacent pair of words, non-adjacent pair of word in a window
- Classifier: Regularized perceptron (equivalent to SVM)

Alignment of actions in script



Text Classification



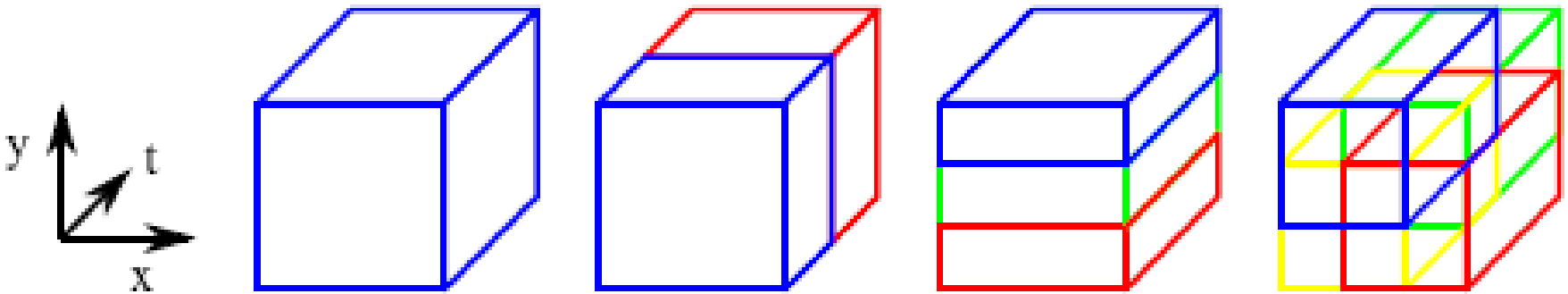


Video Classification

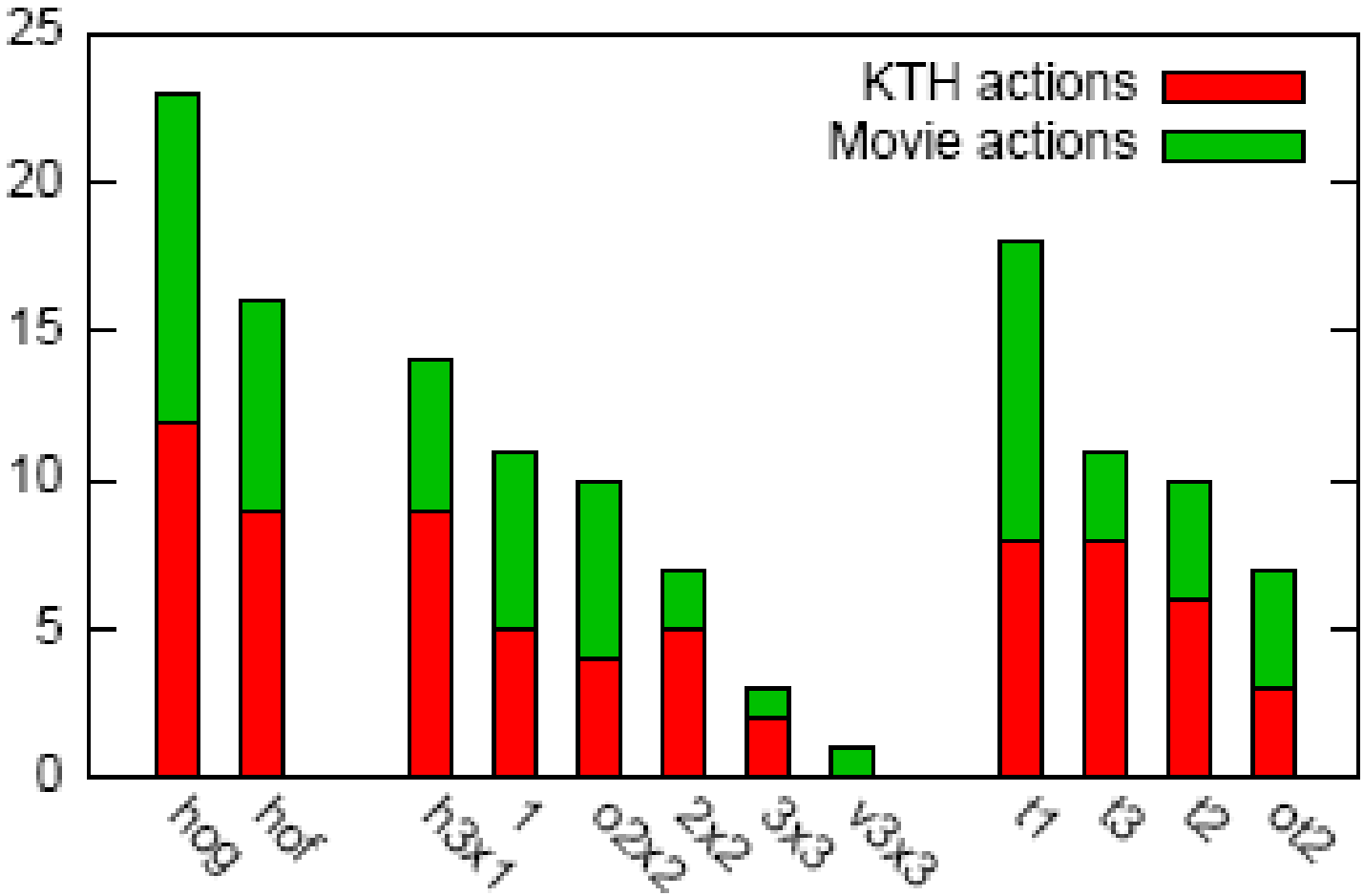
- Interest Points: Space Time Features
  - Harris Corners
  - Multiscale: 6 spatial, 2 temporal
  - Eliminate artifacts due to shot boundaries
  - Calculate histogram of gradients over space time volumes: Histogram of gradient, Histogram of flow descriptors
  - Parameters chosen by experimentation
- Spatio-temporal bag-of-features:
  - 100K features sampled from training video
  - 4000 clusters for K-means: dictionary
  - BoF histogram of visual words
  - binning: spatio-temporal bins
- Classifier: Non-linear support vector machines
  - Comparing histograms: χ^2 kernel

Space-time interest points

Spatio-temporal Grids



Number of occurrences of channel



Average classification accuracy on KTH

Method	Schuldt et al. [icpr04]	Niebles et al. [bmvc06]	Wong et al. [iccv07]	ours
Accuracy	71.7%	81.5%	86.7%	**91.8%**

Average Precision on HOHA dataset

	Clean	Automatic	Chance
AnswerPhone	32.1%	16.4%	10.6%
GetOutCar	41.5%	16.4%	6.0%
HandShake	32.3%	9.9%	8.8%
HugPerson	40.6%	26.8%	10.1%
Kiss	53.3%	45.1%	23.5%
SitDown	38.6%	24.8%	13.8%
SitUp	18.2%	10.4%	4.6%
StandUp	50.5%	33.6%	22.6%

Thank You