HAND DETECTION AND TRACKING FOR FINE-GRAINED ACTION RECOGNITION

Nga H. Do and **Keiji Yanai**

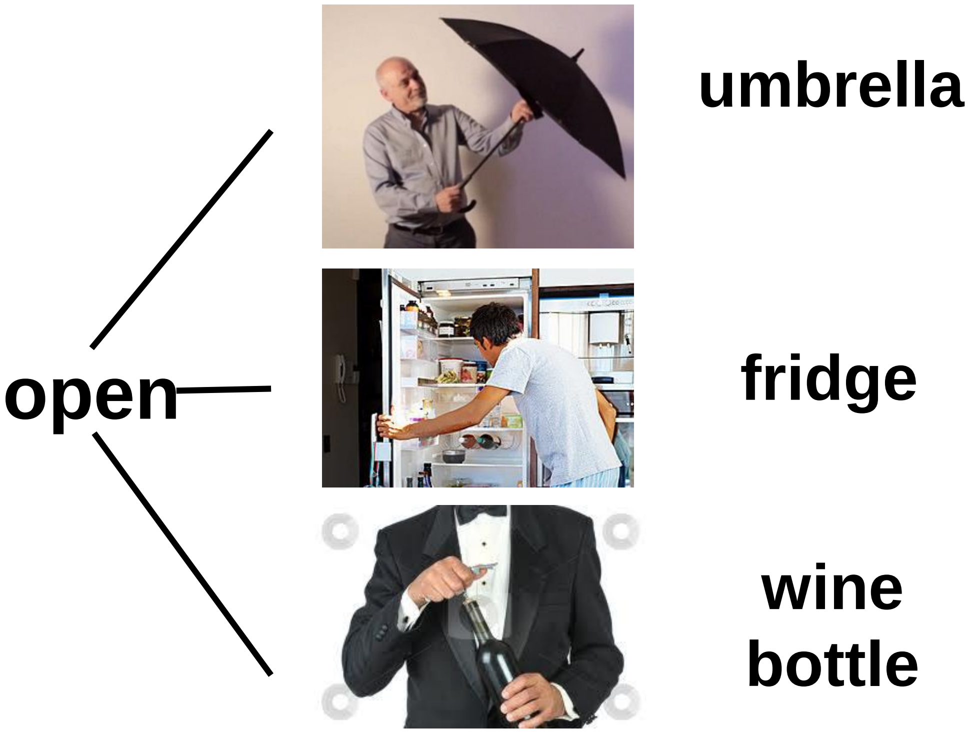
The University of Electro-Communications, Tokyo

Introduction and Related Work

- General recognition vs. fine-grained recognition

UCF Sports dataset

Large intra-class variability



open

umbrella

fridge

wine bottle

Introduction and Related Work

- Fine-grained activity recognition
  - *Database: Cooking fine-grained activities (Rohrbach et al. CVPR2012)*
  - *Activity = composite of multiple actions*
    - Eg.: take X from fridge = open fridge + take X
  - *Recognition accuracies by state-of-the-art:*
    - Rohrbach et al. ECCV2012: 21.3%
    - Ni et al. CVPR2014: 28.6%

tough problem!

Should we consider a simpler problem?

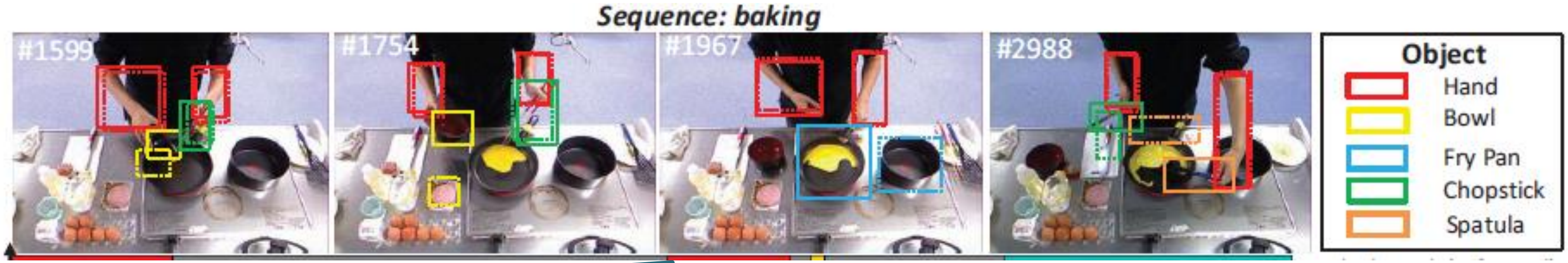
Introduction and Related Work

- Fine-grained action recognition
  - *Sub-problem of fine-grained activity recognition*
  - *Target: single actions*

Can we apply fine-grained activity recognition approaches to fine-grained action recognition?

Introduction and Related Work

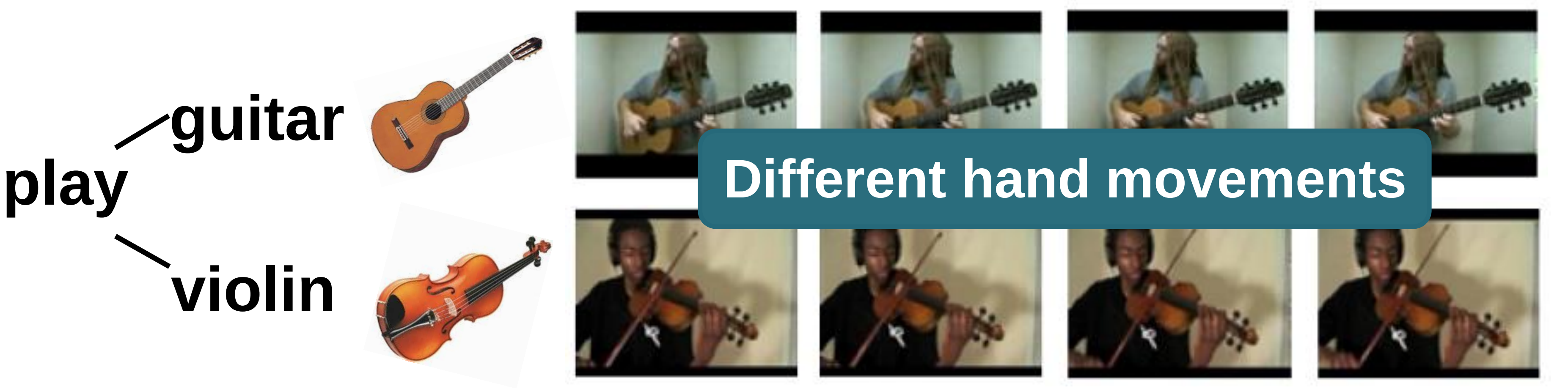
- Most popular approaches for fine-grained activity recognition:
  - *Object detection + Human-object interactions*



B. Ni, V. R. Paramathayalan, and P. Moulin. Multiple granularity analysis for fine-grained action detection. CVPR 2014.

Introduction and Related Work

- Most popular approaches for fine-grained activity recognition:
  - *Object detection + Human-object interactions*
  - *Disadvantages:*
    - requirement of training data for object detectors
    - inefficiency in case of similar objects


play
guitar
violin

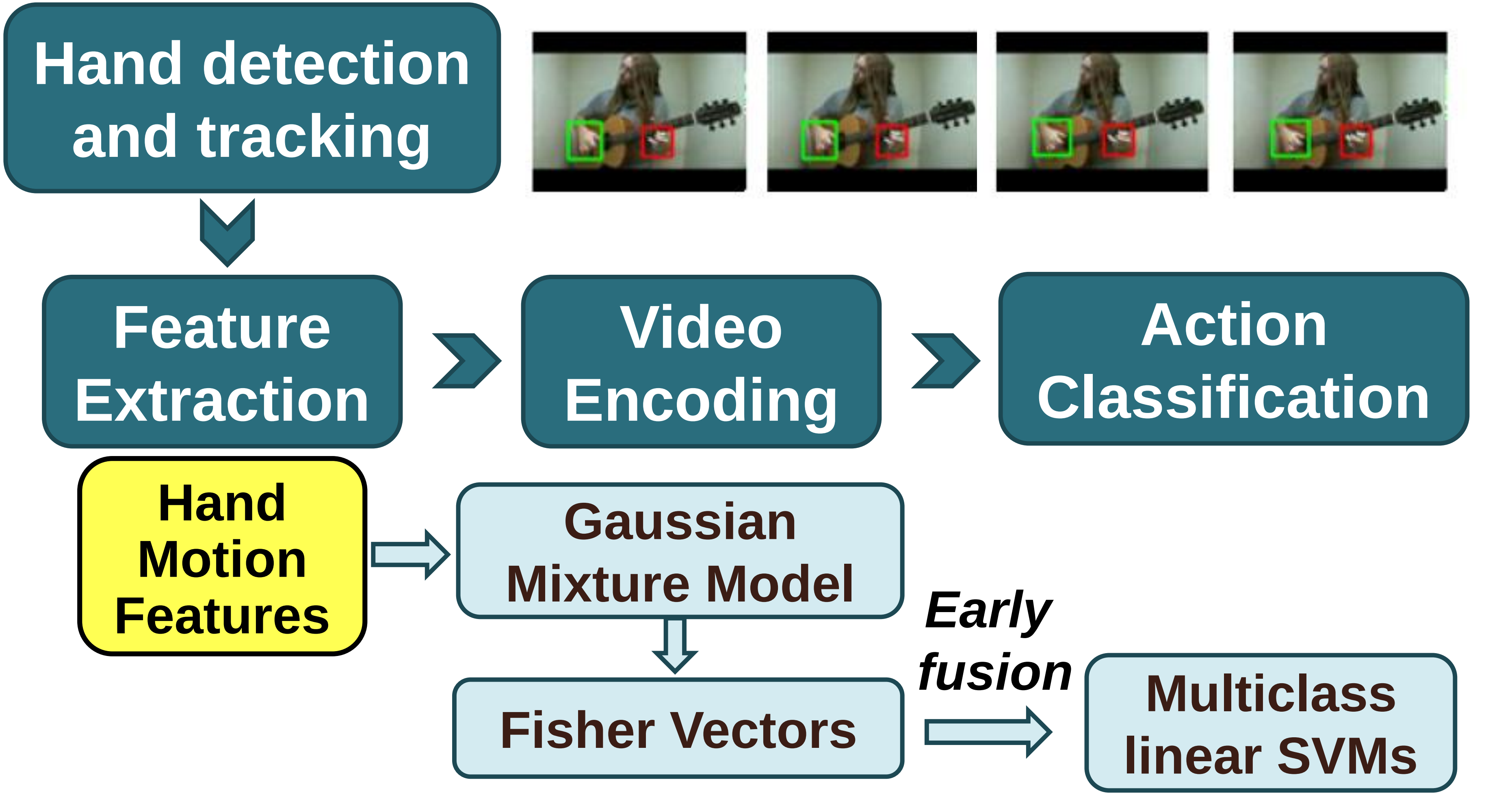
Different hand movements

Introduction and Related Work

- This work:
  - fine-grained action classification based on how human use their hands to operate the actions
    - *Hand motion features for fine-grained action recognition*
  - A system of hand detection and tracking in uncontrolled videos

Proposed Method

Fine-grained action recognition based on hand movements



Hand detection and tracking

Feature Extraction

Video Encoding

Action Classification

Hand Motion Features

Gaussian Mixture Model

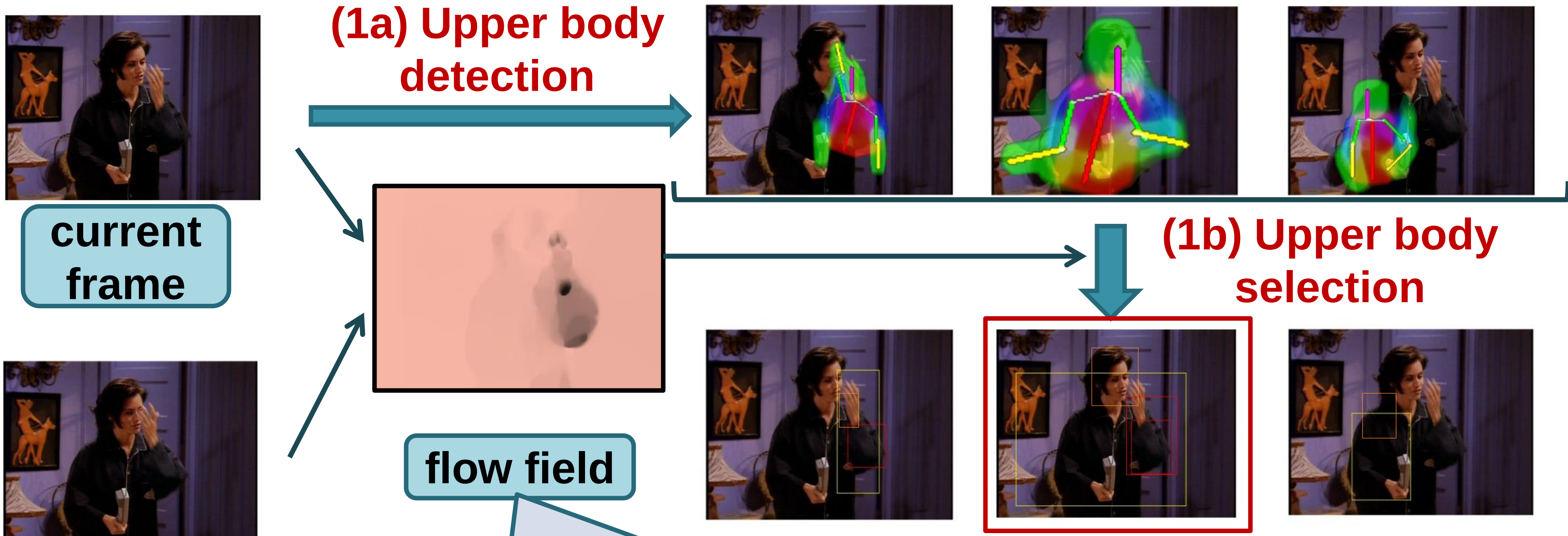
Fisher Vectors

Early fusion

Multiclass linear SVMs

Hand Detection

http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector/



current frame

next frame

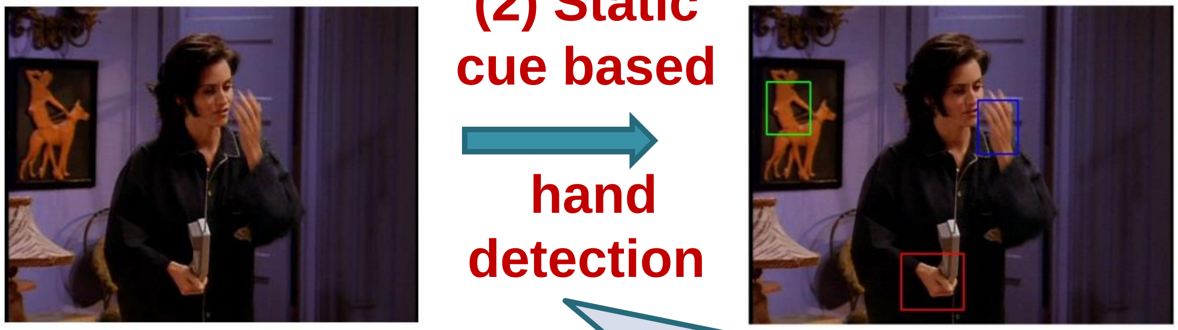
flow field

(1a) Upper body detection

(1b) Upper body selection

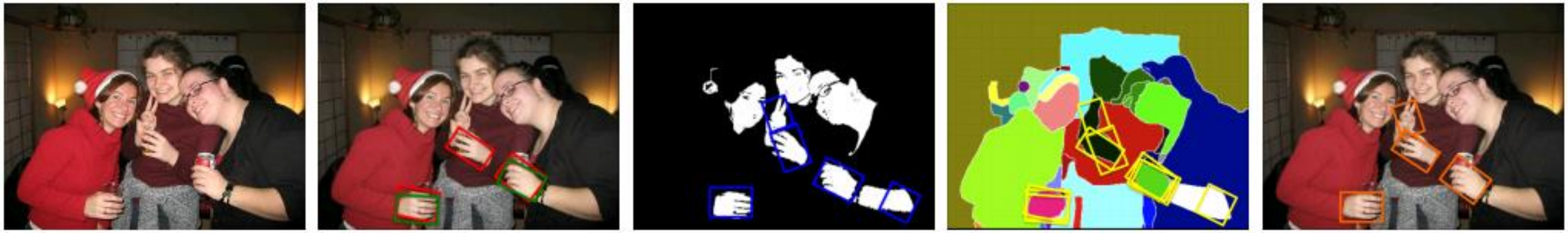
P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. ICCV2013.

Hand detection



(2) Static cue based hand detection

A. Mittal, A. Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. BMVC 2011.



Image

HOG based hand detection

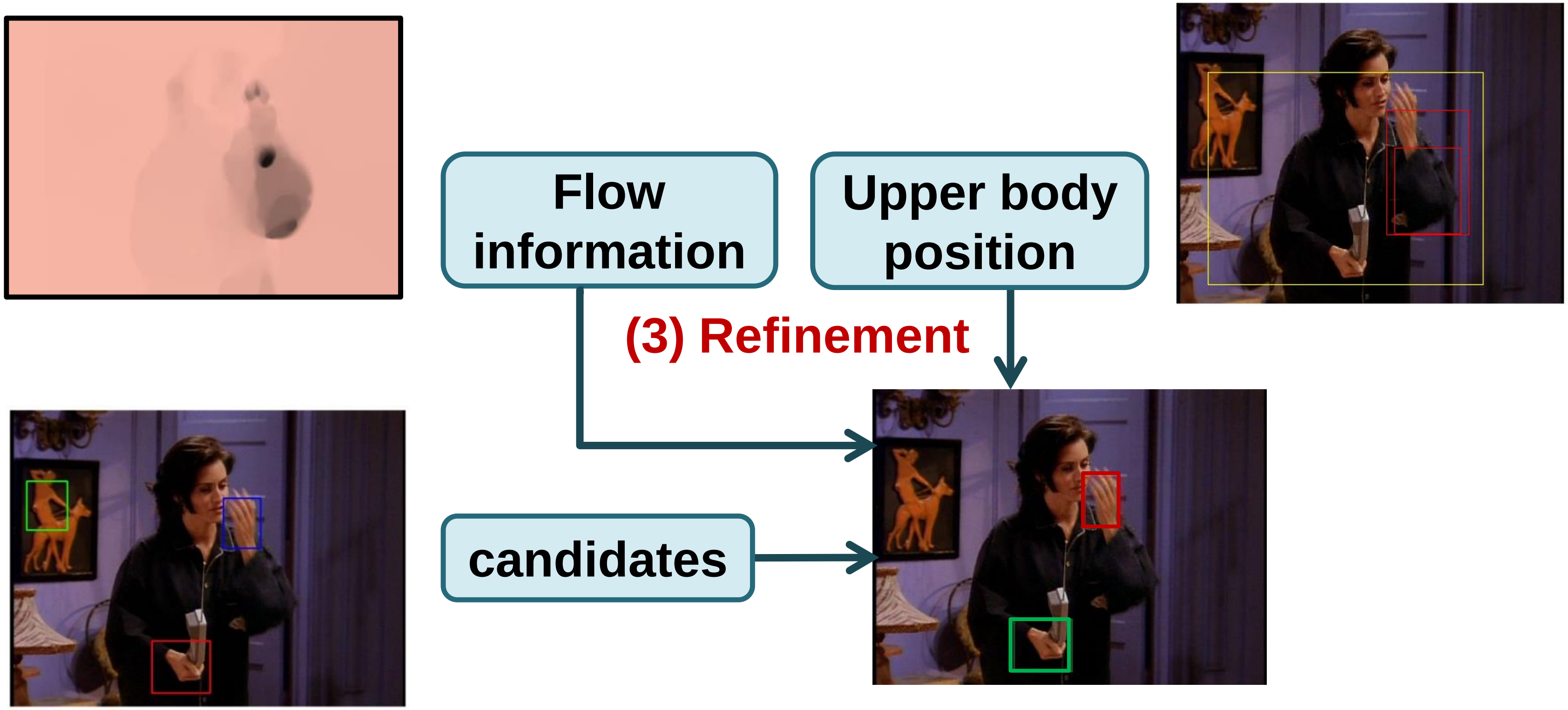
Skin based hand detection

Super-pixel based NMS

Post-processing

- PASCAL VOC
- Buffy Stickman
- INRIA Pedestrian
- etc

Hand detection



Flow information

Upper body position

(3) Refinement

candidates

$$final\ score = Mittal's\ detector\ score + w_f \alpha_f + w_u \alpha_u$$

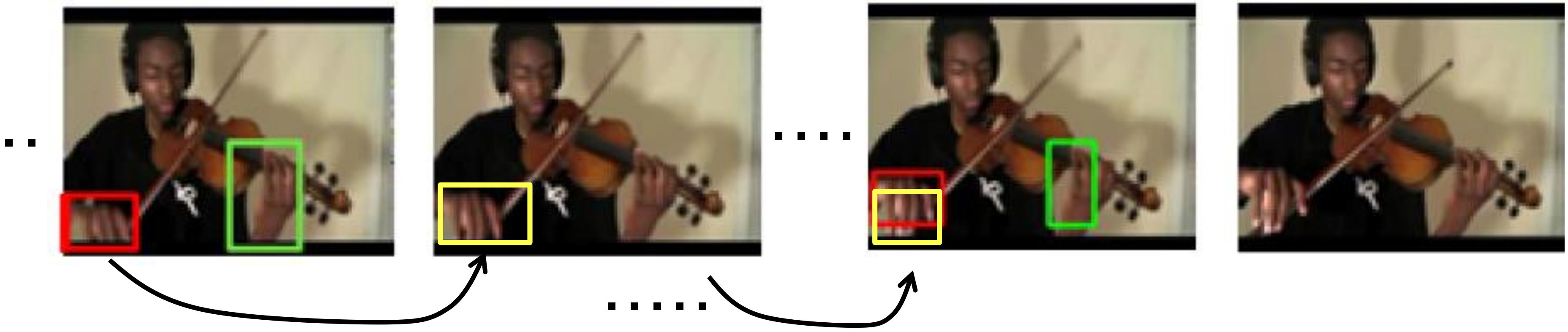
$$flow\ score\colon\ \alpha_f = \frac{1}{N}\sum_{i=1}^{N}\sqrt{{u_i}^2 + {v_i}^2} \qquad upperbody\ score\colon\ \alpha_u = \frac{area(B_d \cap B_u)}{area(B_d)}$$

$$w_f = 0.7, w_u = 0.2$$

Proposed Method

Hand tracking

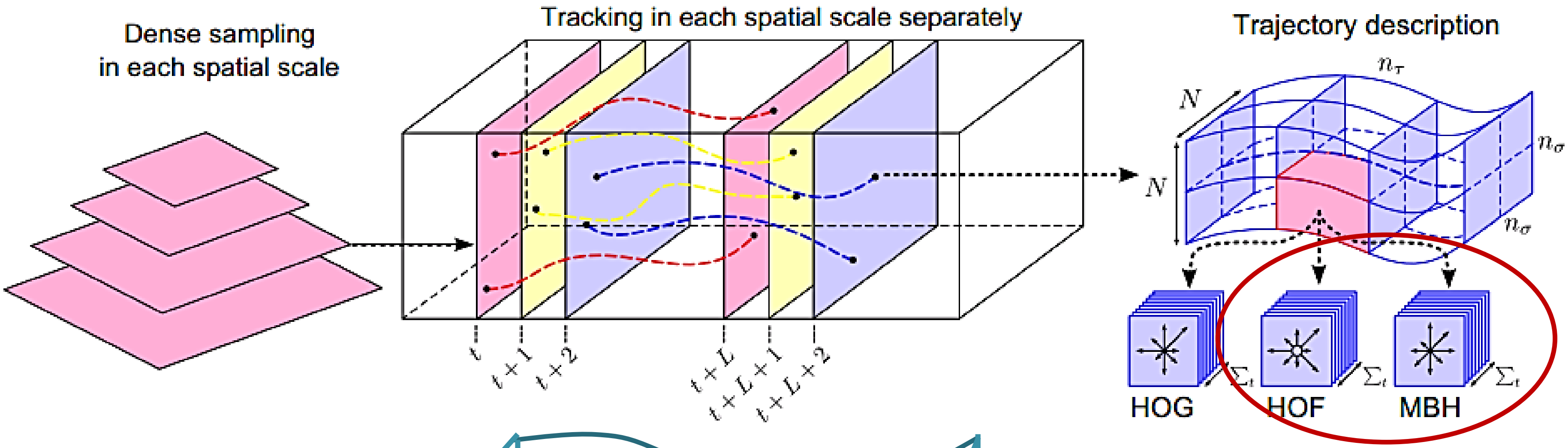
Select detections which overlap with many others



propagate

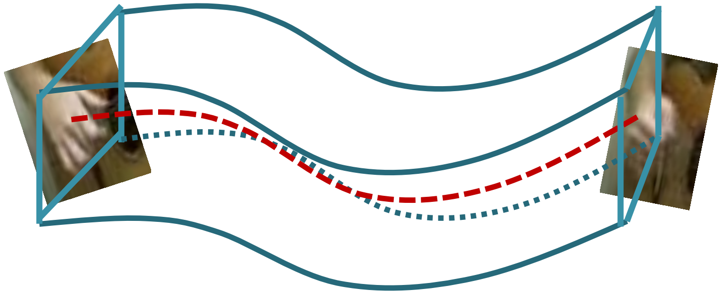
Extraction of Hand Motion Features

Wang et al. Action recognition with improved dense trajectories. CVPR 2013.



Only hand regions

Hand tracks



$$S' = \frac{(\Delta P_t, \dots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1}\|\Delta P_j\|}$$

5.4 Experiments and Results

Data and Evaluation Methods

Exp.	Dataset	Evaluation method
Hand detection	Video Pose 2.0 • 14 video shots • 2453 frames • 3814 hands	• Average precision • Overlap score > 0.5 $O = \frac{area(B_g \cap B_d)}{area(B_g \cup B_d)}$
Fine-grained action recognition	Playing Instruments (UCF101) • 10 types of instruments • 1428 videos	• Average precision • 3 training/test splits [1]

[1] ***ICCV2013 Workshop on Action Recognition with a Large Number of Classes***

Data Thumbnails

play cello

play daf

play tabla

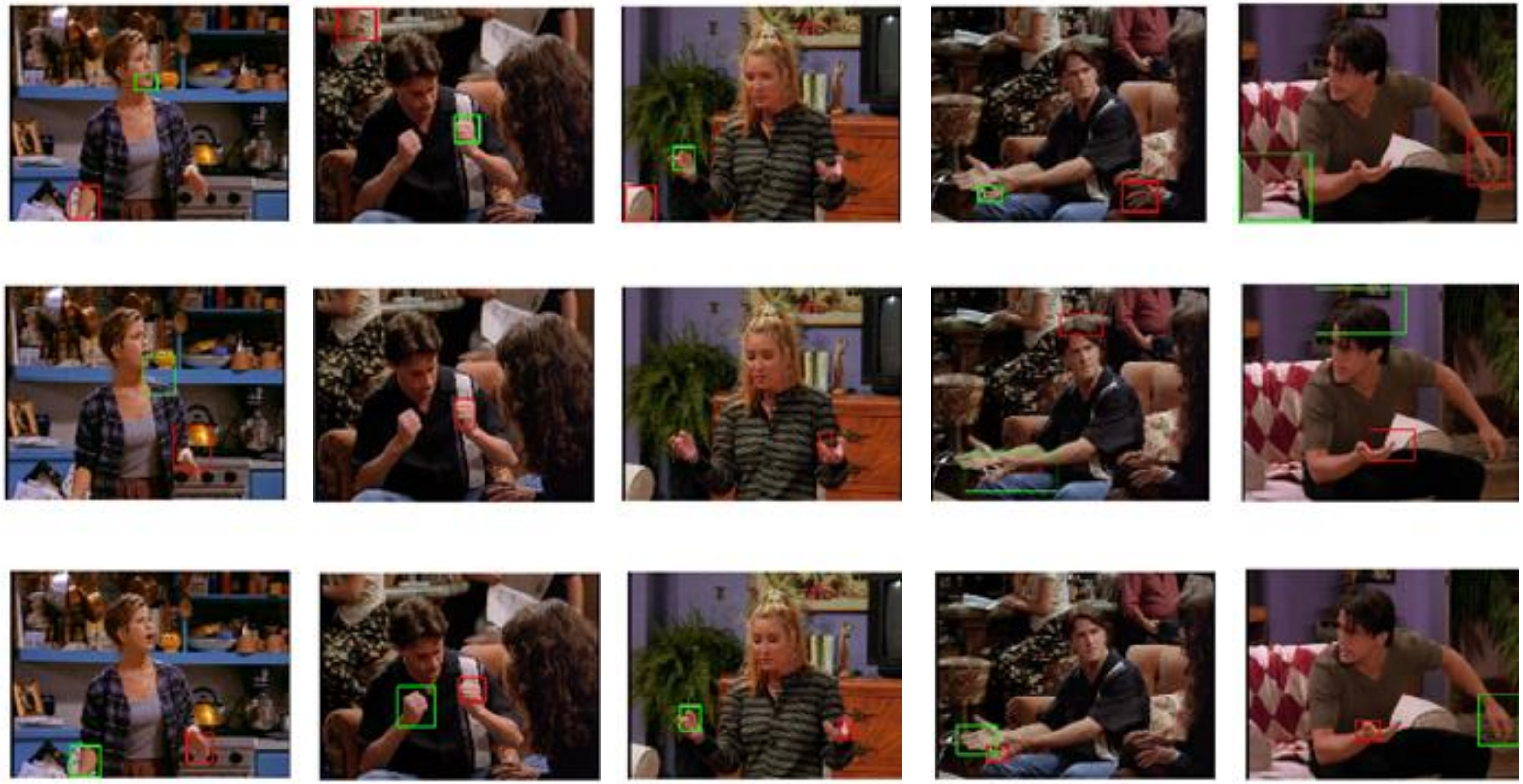
Experiments and Results

Hand Detection Performance

Method	Precision	
Mittal et al.	41.7%	***still features***
Sapp et al.*	18.6%	***motion features***
Our (upper body)	42.6%	***still features + motion features***
Our (flow)	45.5%	
Our (flow+body)	46.3%	

** Sapp et al. Parsing human motion with stretchable models. CVPR2011.*
(flow based hand detector only)

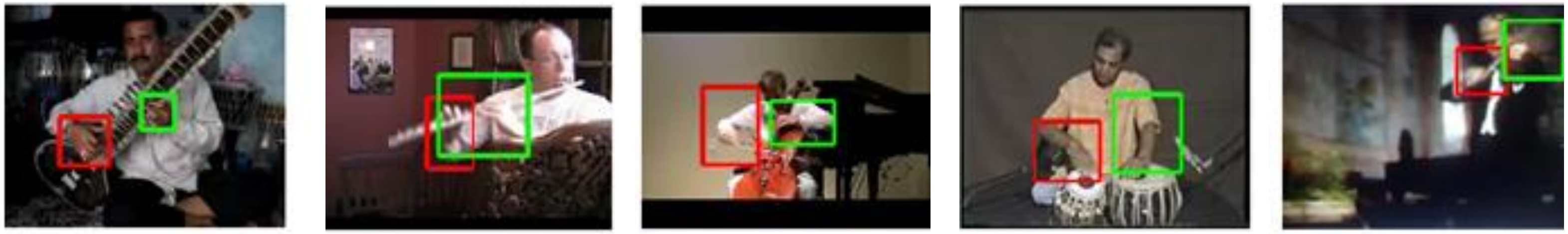
Examples of Detection Results



Mittal et al.

Sapp et al.

Our



UCF101 Playing instruments

Action recognition performance

DT: Dense Trajectory
HDT: Dense Trajectory on Hands
HT: Hand Tracks

Method	Precision
DT	66.7%
HOF_{dt}	83.8%
MBH_{dt}	86.6%
$MBH_{dt} + HOF_{dt} + DT$	87.3%
HDT	66.1%
HOF_{hdt}	81.4%
MBH_{hdt}	85.7%
$MBH_{hdt} + HOF_{hdt} + HDT$	86.2%
HT	36.0%
$MBH_{dt} + HOF_{dt} + DT + HT$	87.6%
$\boldsymbol{MBH_{hdt} + HOF_{hdt} + HDT + HT}$	**88.5%**

whole frame (more features)

only hand regions

Conclusions

- **Proposed a system of hand detection and tracking in videos**
- **Applied the system to fine-grained action recognition**
  - ***Action recognition based on only hand motion features***