Modeling Transition Patterns Between Events for Temporal Human Action Segmentation and Classification

Yelin Kim[1], Jixu Chen[2], Ming-Ching Chang[2], Xin Wang[3], Emily Mower Provost[1], and Siwei Lyu[3]

University of Michigan, Ann Arbor[1]

GE Global Research[2]

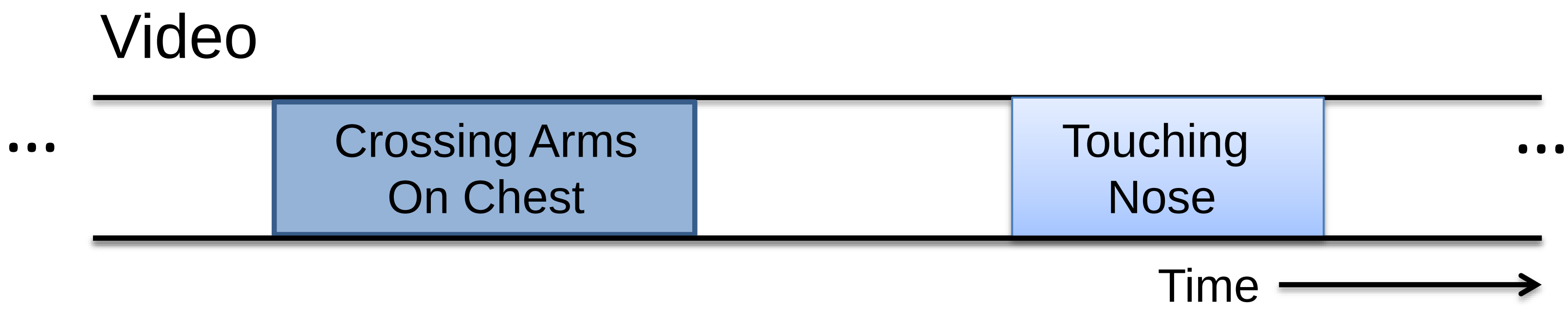
State University of New York, Albany[3]

Photo by Ludovic Bertron

- Why **Event Recognition**?
  - Gigantic amount of video data
    - ➔ need to identify events of interest
  - Indexing/retrieval of video collections

- How **Event Recognition?**
  1. Localization of events (*when happened?*)
  2. Classification of events (*what happened?*)

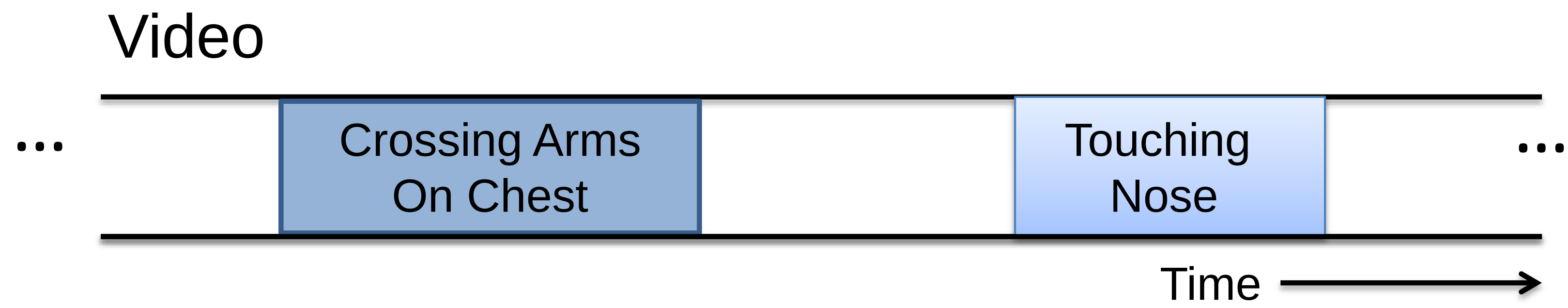
Previous Work



- Most of the previous methods: treat localization and classification as separate problems[13, 15]

[13] I. Laptev, CVPR, 2008
[15] J. C. Niebles, ICCV, 2010

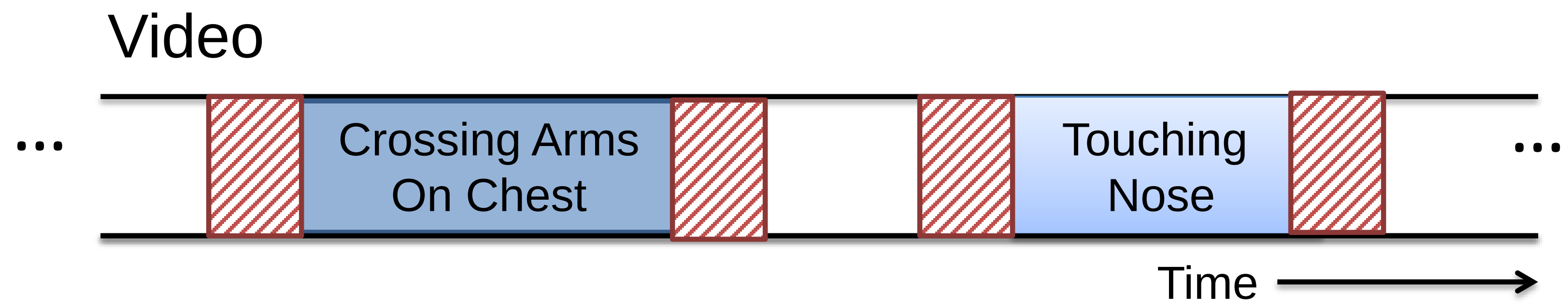
Previous Work



- Most of the previous methods: treat localization and classification as separate problems[13, 15]
- Recent work: jointly localize and classify events of interest in videos[1,2]

[13] I. Laptev, CVPR, 2008
[15] J. C. Niebles, ICCV, 2010
[1] M. Hoai et al., CVPR, 2013
[2] Y. Cheng et al., CVPR, 2014

Novelty of Proposed Method



The novelty of our work is that we introduce the modeling of two kinds of event transition information:

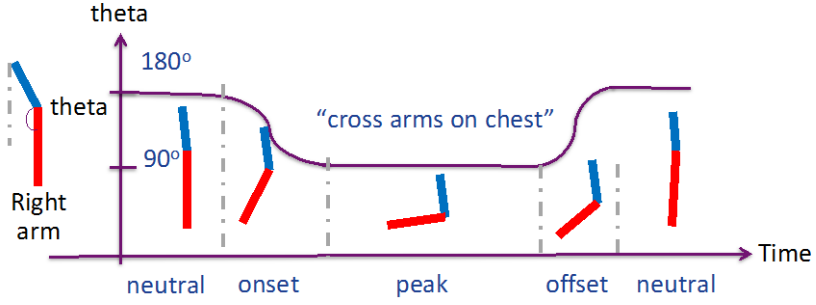
1. Event transition segments
2. Event transition probabilities

Proposed Method: Event Transition

- Event Transition Segments: capture the occurrence patterns between two consecutive events of interest
- Event Transition Probabilities: model the transition probability between the two events

Why Event Transition Information?

- Facial expression recognition [7, 11, 23]:
  - Onset, offset modeling
- Human action event recognition:
  - Unique, distinguishable transitions
  - Importance of transition patterns



[7] X. Ding et al., ICCV, 2003
[11] S. Koelstra et al., IEEE PAMI, 2010
[23] M. Valstar and M. Pantic, IEEE SMC, 2012

Contributions of Our Work

- Proposed a temporal segmentation and classification method using transition patterns between events of interest
- Improved the human action detection accuracy of two datasets: Smartroom and CMU-MAD[10]
- Demonstrated the importance of transition patterns to detect human action events

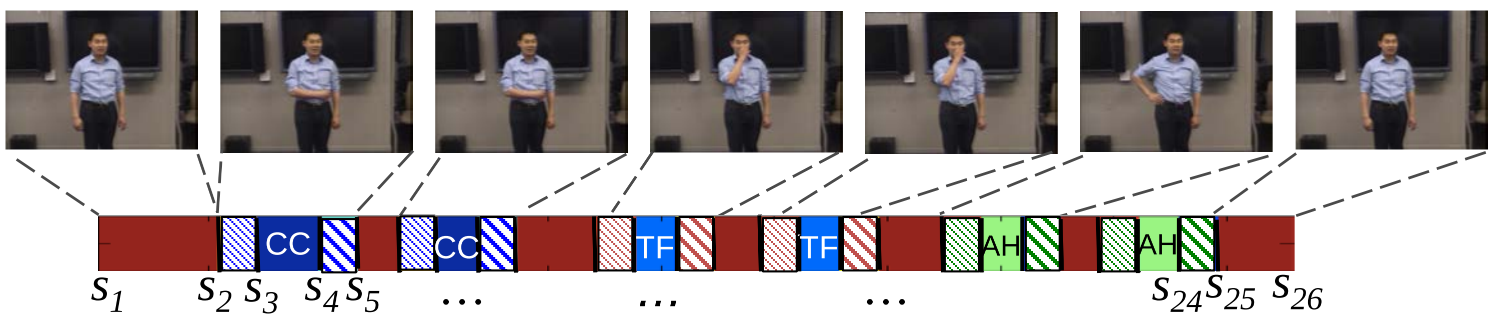
[10] CMU-MAD Database: D. Huang et al., ECCV, 2014.

PROPOSED METHOD

Proposed Method

Training Phase

i-th training video:



1. Extract per-frame human pose cues
2. Calculate variable-length segment-level features
3. Train segment-SVM[1] → learned weights w_y

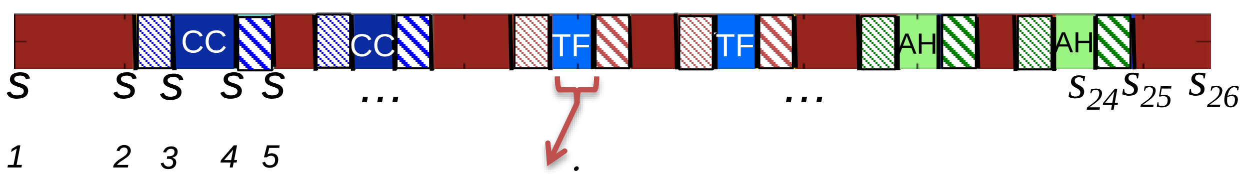
Proposed Method

Testing Phase

Input:

Video: X

Estimated Segmentation



$$X^i_{(s^i_t, s^i_{t+1}]}$$

Output:

Number of segments: k

Segment points: $s_t \in \{1, \dots, k+1\}, s_1 = 0, s_{k+1} = len(X)$

Segment labels: $y_t \in \{N, CC, TF, AH, CC_{onset}, TF_{onset}, \dots, AH_{offset}\}$

- Normal (N)
- Crossing arms on chest (CC)
- Touching face (TF)
- Arms on hips (AH)
- Onset of CC,TF,AH
- Offset of CC,TF,AH

Proposed Method

Testing Phase (Hoai et al. CVPR 2013)

(Our Proposed Method)

– Segmentation goal:

$$\max_{k,s_t,y_t} \sum_{t=1}^{k} w_{y_t}^T \varphi(x_t) + (1+\gamma)\log P(y_t|y_{t-1})$$

$$\min_{k,s_t,y_t,\xi_t\geq 0} \sum_{t=1}^{k} \xi_t$$

$$\text{s.t. } l_{min} \leq s_{t+1} - s_t \leq l_{max}, \forall t,$$
$$s_1 = 0, s_{k+1} = len(X),$$

– Dynamic Programming-based Inference: for $X_{(0,w]}$

$$f(u, y_k) = \max_{l, y_{k-1}} f(u-l, y_{k-1}) + \eta(u, l, y_k)$$

$$f(u) = \min_{l} f(u-l) + \xi(u, l)$$

$$+ (1+\gamma)\log P(y_k|y_{k-1})$$

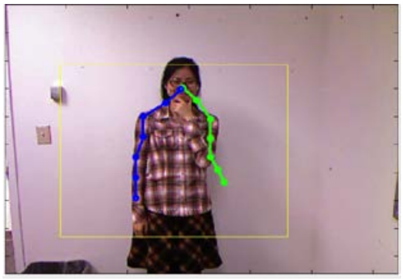
- $\xi(u,l) = \max\{0, 1 - (w_{\hat{y}} - w_{\tilde{y}})^T \varphi(X_{(u-l,u]})\}$
  $\eta(u,l,y) = w_y^T \varphi(X_{(u-l,u]})$

EXPERIMENTAL RESULTS

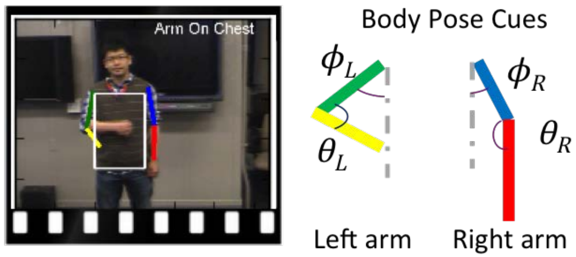
Data: Smartroom and CMU-MAD

- Smartroom Dataset: suspicious behavior recognition

Clean

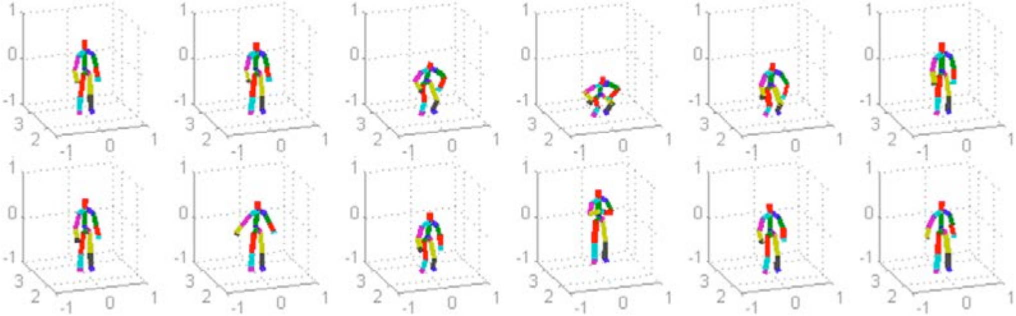


Noisy



Body pose cues[19]

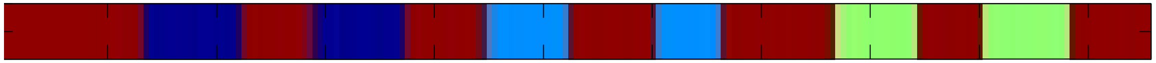
- CMU-MAD Dataset[10]: Human action dataset



[10] D. Huang et al., ECCV, 2014. [19] B. Sapp and B. Taskar, CVPR, 2013.

Performance Measure

- Compare the performance to [Hoai et al., 2011]



Performance Measure

- Frame-level and event-level recognition rates:
  1. Frame-level recognition rate

  2. Event-level recognition rate[10]: the ratio of event segments that are correctly identified, by counting the number of correct frames that overlaps with 50% of a segment.

[10] D. Huang et al., ECCV, 2014.

Experimental Results: Smartroom

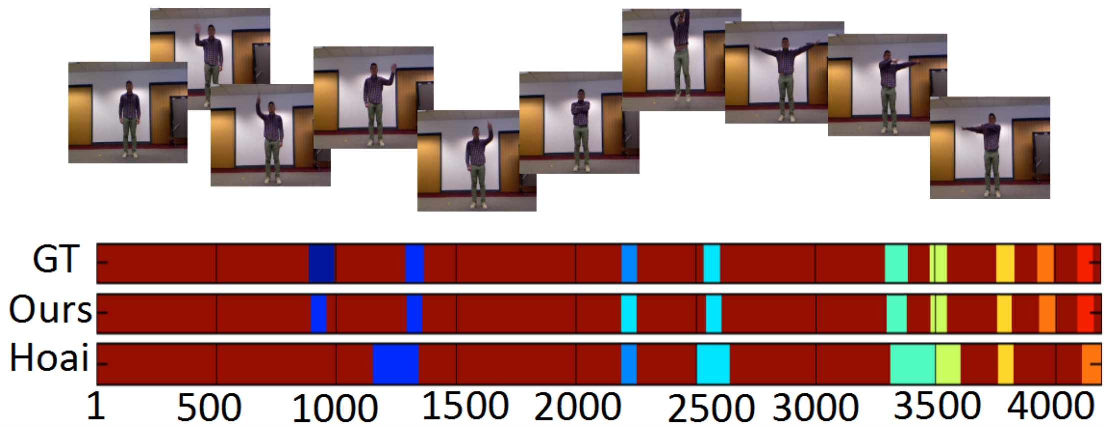
- Clean

	Frame-level						Event-level					
	Prec		Rec		F-mea		Prec		Rec		F-mea	
Method	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Ours	**83.84**	7.45	**80.41**	12.18	**81.95**	9.52	**86.67**	11.55	**89.63**	10.02	**88.07**	10.54
Hoai	56.19	5.32	60.50	7.98	58.15	5.74	71.11	7.70	67.41	12.24	68.32	3.86
Diff	**27.65**		**19.91**		**23.79**		**15.55**		**22.22**		**19.75**	

- Noisy

	Frame-level						Event-level					
	Prec		Rec		F-mea		Prec		Rec		F-mea	
Method	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Ours	**44.41**	18.85	**40.38**	18.20	**41.33**	17.09	**25.36**	16.36	**54.45**	15.91	**33.51**	17.93
Hoai	24.39	11.54	13.60	6.88	17.26	8.33	14.33	14.93	11.20	6.81	11.75	10.56
Diff	**20.02**		**26.78**		**24.07**		**11.03**		**43.24**		**21.76**	

Experimental Results: CMU-MAD



	Frame-level						Event-level					
	Prec		Rec		F-mea		Prec		Rec		F-mea	
Method	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Ours	**85.00**	8.82	**71.41**	7.25	**77.41**	7.01	**74.40**	15.02	**85.02**	12.17	**78.83**	12.95
Hoai	73.79	9.62	70.57	9.96	71.87	8.70	73.45	15.84	83.88	13.06	77.85	14.23
Diff	**11.21**		**0.84**		**5.54**		**0.95**		**1.14**		**0.98**	

1. CMU-MAD: the transition segments were not explicitly labeled (performance gain: frame-level > event-level)
2. Difference in visual features (w/ vs. without depth)

Conclusions

- Explicitly model event transition segments
- Improve the state-of-art performance on the joint localization and classification of video events
- Future Work:
  - automatic methods that can learn the transition probabilities of the full set of pairwise event transitions.
  - Segmentation and classification of audio-visual cues

Acknowledgements

- Morpho Detection LLC.
- Dr. Peter Tu
- Dr. Guiju Song
- Computer Vision Lab, GE Global Research

Thank You!

Questions?

yelinkim@umich.edu