Finding Group Interactions in Social Clutter

Ruonan Li, Parker Porfilio, Todd Zickler

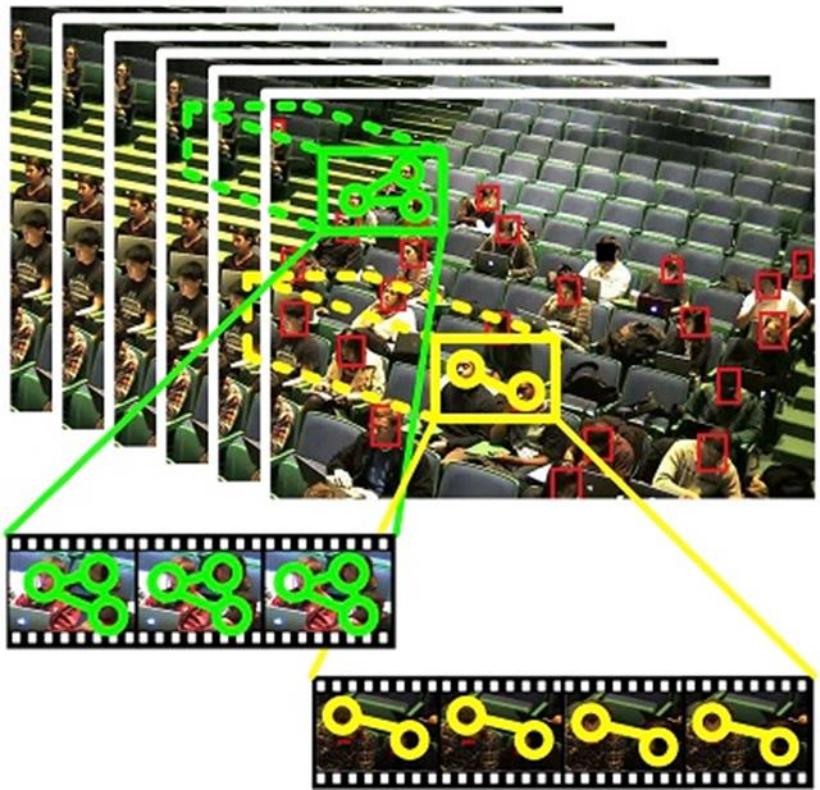
Emine Gül DANACI N12129719

Related Works

- M. Amer and S. Todorovic. A chains model for localizing participants of group activities in videos. In ICCV, 2011.
- M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, Apr. 2011.
- K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS, 2005.
- C. Lampert, M. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. PAMI, 31(12):2129–2142, 2011.

Approach

- Voting based approach
  - Matching between temporal units
  - Voting for participant identification
  - Branch-and-bound temporal localization



Matching and Localizing Interactions

- Video: A sequence of T temporal units
- M track-space
- T frame

$$\{\mathbf{f}_{m,t}\} \qquad \{\mathbf{g}_{m,m',t}\}$$

$$\mathcal{Q}_t \triangleq \{\mathbf{f}_{m,t}, \mathbf{g}_{m,m',t}\}$$

$$\mathcal{D}_s \triangleq \{\mathbf{f}^D_{n,s}, \mathbf{g}^D_{n,n',s}\}$$

Matching Between Temporal Units

Similarity

Mahalonobis distance

$$\hat{D}(\mathcal{Q}_t, \mathcal{D}_s, W) = \sum_{nm} w_{nm} d_I(\mathbf{f}_{m,t}, \mathbf{f}^D_{n,s}) + \sum_{nmn'm'} w_{nm} w_{n'm'} d_P(\mathbf{g}_{m,m',s}, \mathbf{g}^D_{n,n',t}),$$

Optimization

$$\min_{\mathbf{w}} \mathbf{c}^T \mathbf{w} + \mathbf{w}^T H \mathbf{w}, \text{s.t. } w_{nm} \in \{0,1\}, W\mathbf{1} = \mathbf{1}, W^T\mathbf{1} \leq \mathbf{1}$$

c is a MNx1 vector of distances between individual descriptors
H is a MNxMN matrix of distances between pairwise descriptors

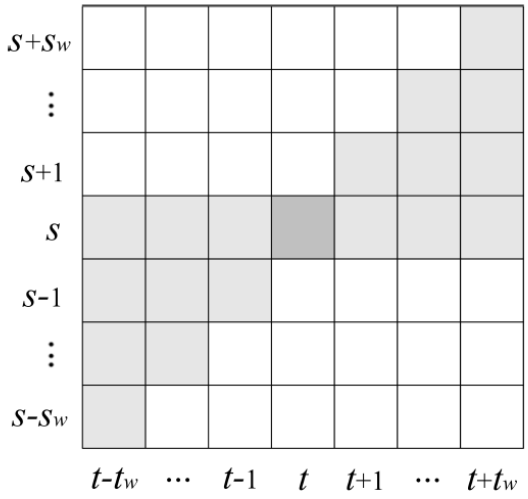
Convex unit (CVX Toolbox)

$$\min_{\mathbf{w}} (\mathbf{c} + \hat{\mathbf{c}})^T \mathbf{w} + \mathbf{w}^T (H + \hat{H}) \mathbf{w}, \text{s.t. } w_{nm} \in \{0,1\}, W\mathbf{1} = \mathbf{1}, W^T\mathbf{1} \leq \mathbf{1}$$

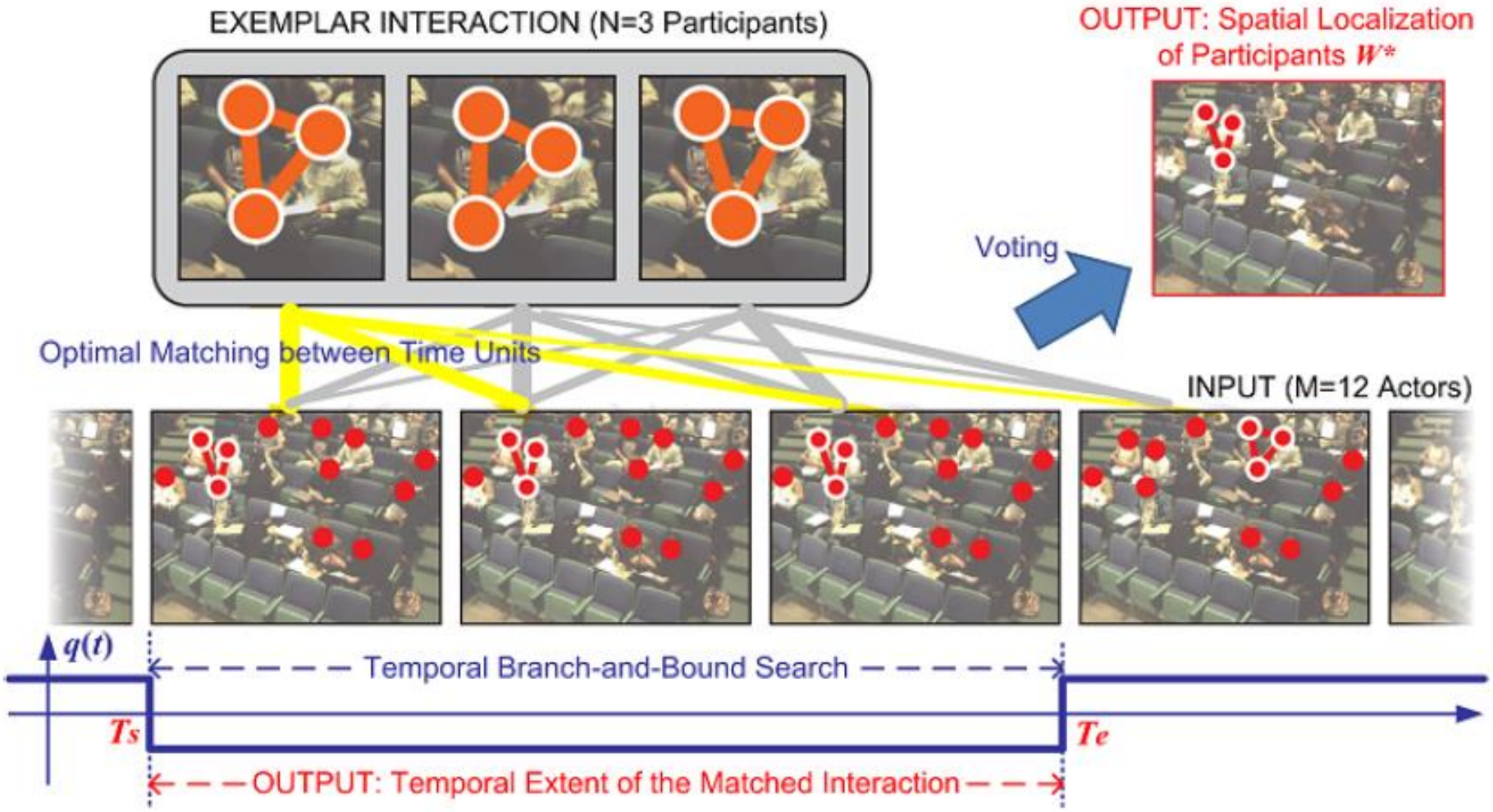
$$\hat{\mathbf{c}} = [\sigma_1, \sigma_2, \cdots, \sigma_{MN}]^T, \hat{H} = \text{diag}\{-\sigma_1, -\sigma_2, \cdots, -\sigma_{MN}\}$$

Voting for Participant Identification

1. Clear both accumulator arrays;
2. For each $t \in [1, T], s \in [1, S]$, increment the count for the matching matrix $W^{t,s}$ by 1, and increase the sum of weights
   in the companion array corresponding to $W^{t,s}$ by $v(W^{t,s})$;
3. Identify a subarray of matrices receiving more than $\frac{S}{2}$ counts,
   and normalize the sum of weights in the companion subarray
   by corresponding counts;
4. Report the matching matrix W^* to be the one in the subarray receiving the minimum normalized sum of weights.



Voting for Participant Identification



Branch-and-Bound Temporal Localization

- Temporal pyramid

$$k(t, T_s, T_e, s, 1, S) \triangleq \sum_{l=0}^{L-1} \sum_{i=1}^{2^l} \mathbf{1}(t \in \mathcal{C}(T_s, T_e, l, i)) \mathbf{1}(s \in \mathcal{C}(1, S, l, i))$$

1. Initialize: Let $T_{s,low} = 1, T_{s,upp} = T - T_{min} + 1, T_{e,low} = T_{min} + 1$, and $T_{e,upp} = T$; Initialize priority queue Q as empty;

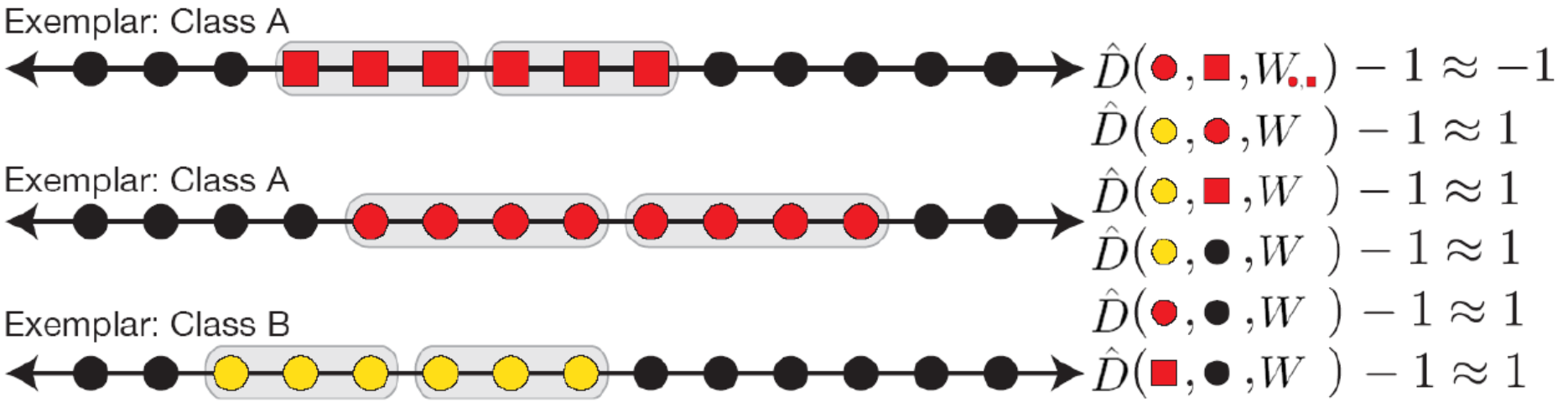
2. Do

   - If $T_{s,upp} - T_{s,low} \geq T_{e,upp} - T_{e,low}$
     $T^{(1)}_{s,low} \leftarrow T_{s,low}, T^{(1)}_{s,upp} \leftarrow T_{s,low} + \frac{T_{s,upp} - T_{s,low}}{2}, T^{(1)}_{e,low} \leftarrow T_{e,low}, T^{(1)}_{e,upp} \leftarrow T_{e,upp},$
     $T^{(2)}_{s,low} \leftarrow T_{s,low} + \frac{T_{s,upp} - T_{s,low}}{2}, T^{(2)}_{s,upp} \leftarrow T_{s,upp}, T^{(2)}_{e,low} \leftarrow T_{e,low}, T^{(2)}_{e,upp} \leftarrow T_{e,upp};$
     else
     $T^{(1)}_{s,low} \leftarrow T_{s,low}, T^{(1)}_{s,upp} \leftarrow T_{s,upp}, T^{(1)}_{e,low} \leftarrow T_{e,low}, T^{(1)}_{e,upp} \leftarrow T_{e,low} + \frac{T_{e,upp} - T_{e,low}}{2}, T^{(2)}_{s,low} \leftarrow T_{s,low},$
     $T^{(2)}_{s,upp} \leftarrow T_{s,upp}, T^{(2)}_{e,low} \leftarrow T_{e,low} + \frac{T_{e,upp} - T_{e,low}}{2}, T^{(2)}_{e,upp} \leftarrow T_{e,upp};$
   - If $T_{min} \leq T^{(1)}_{e,upp} - T^{(1)}_{s,low}$, push $(T^{(1)}_{s,low}, T^{(1)}_{s,upp}, T^{(1)}_{e,low}, T^{(1)}_{e,upp}, \hat{f}(T^{(1)}_{s,low}, T^{(1)}_{s,upp}, T^{(1)}_{e,low}, T^{(1)}_{e,upp}))$ into Q;
   - If $T_{min} \leq T^{(2)}_{e,upp} - T^{(2)}_{s,low}$, push $(T^{(2)}_{s,low}, T^{(2)}_{s,upp}, T^{(2)}_{e,low}, T^{(2)}_{e,upp}, \hat{f}(T^{(2)}_{s,low}, T^{(2)}_{s,upp}, T^{(2)}_{e,low}, T^{(2)}_{e,upp}))$ into Q;
   - Let $(T_{s,low}, T_{s,upp}, T_{e,low}, T_{e,upp})$ be the tuple in Q achieving the minimal $\hat{f}$;

   Until $T_{s,low} = T_{s,upp}, T_{e,low} = T_{e,upp}$.

3. Output: $T_s \leftarrow T_{s,low}, T_e \leftarrow T_{e,low}$.

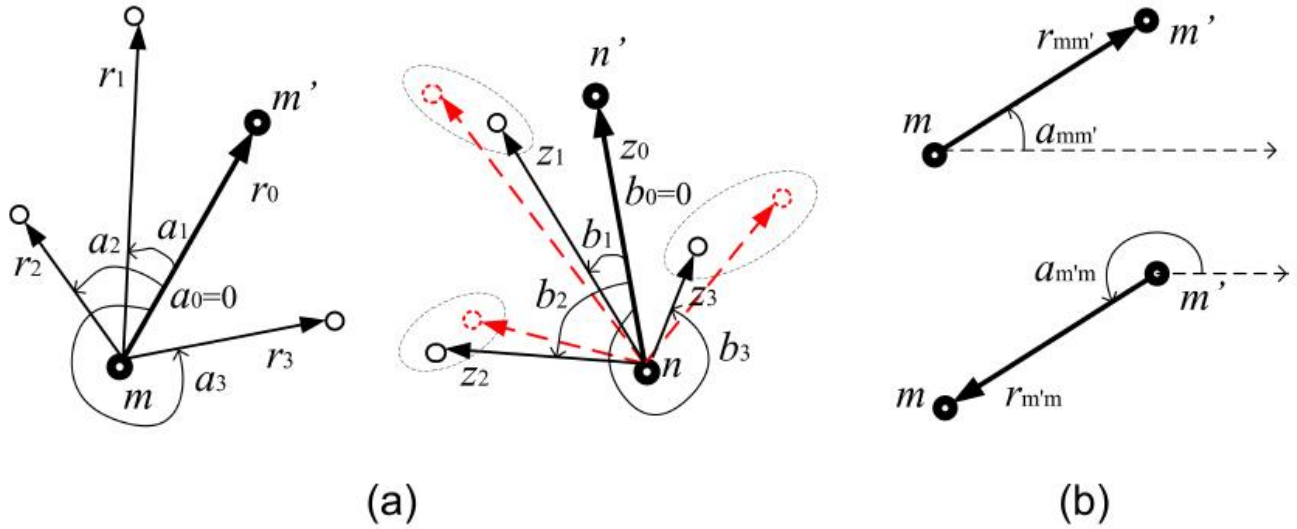
Descriptor Metric Learning



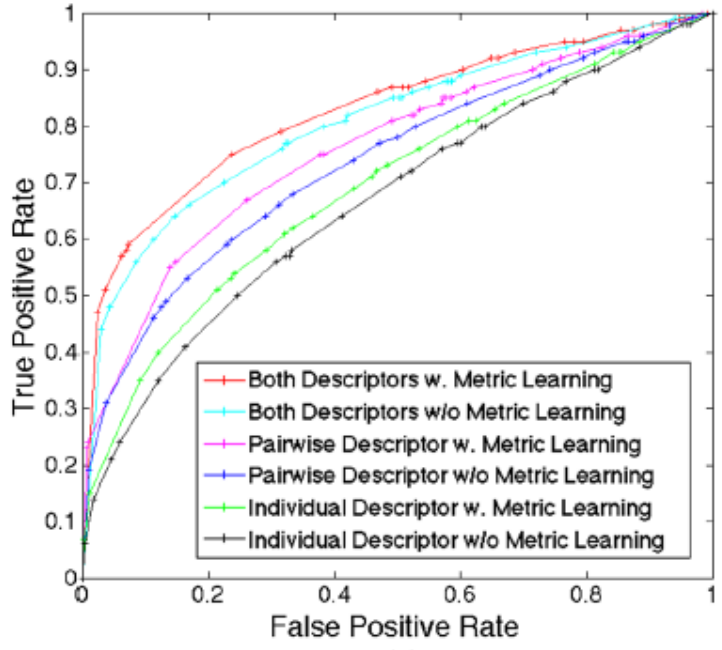
Experiments

Classroom Interaction Database

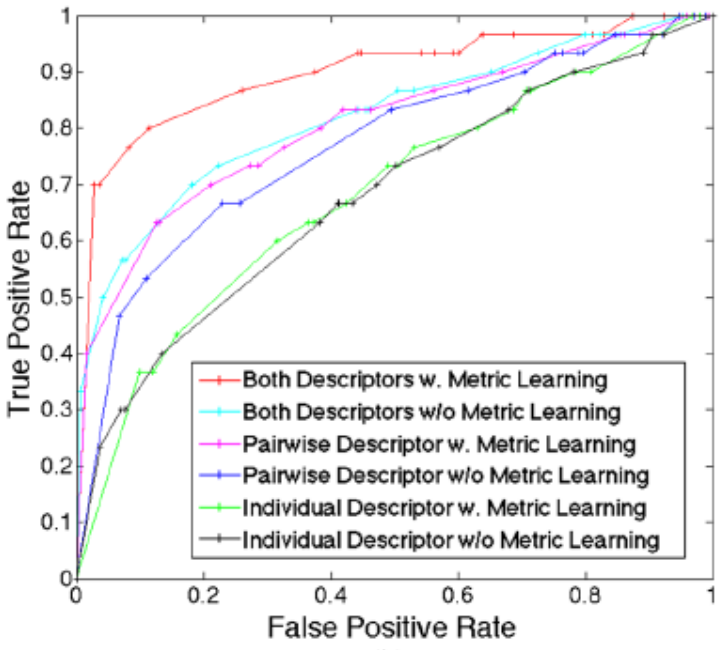
- 254 two-person, 112 three-person, and 16 four-person interactions in total
- Histogram of Oriented Gradient (HOG) feature within each temporal unit
- Train nine one-versus-all SVM (to estimate the likelihood of nine head poses)



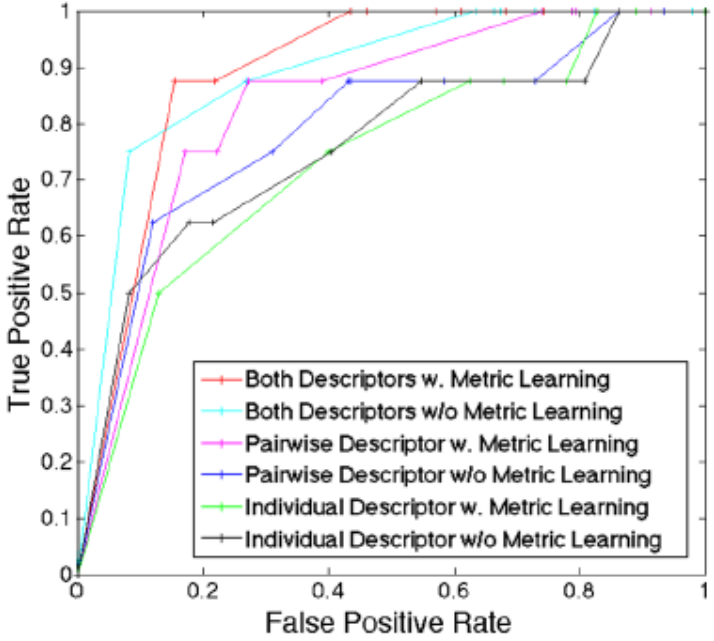

(a)
(b)
(c-1)
(c-2)
(c-3)


True Positive Rate
False Positive Rate
Both Descriptors w. Metric Learning
Both Descriptors w/o Metric Learning
Pairwise Descriptor w. Metric Learning
Pairwise Descriptor w/o Metric Learning
Individual Descriptor w. Metric Learning
Individual Descriptor w/o Metric Learning

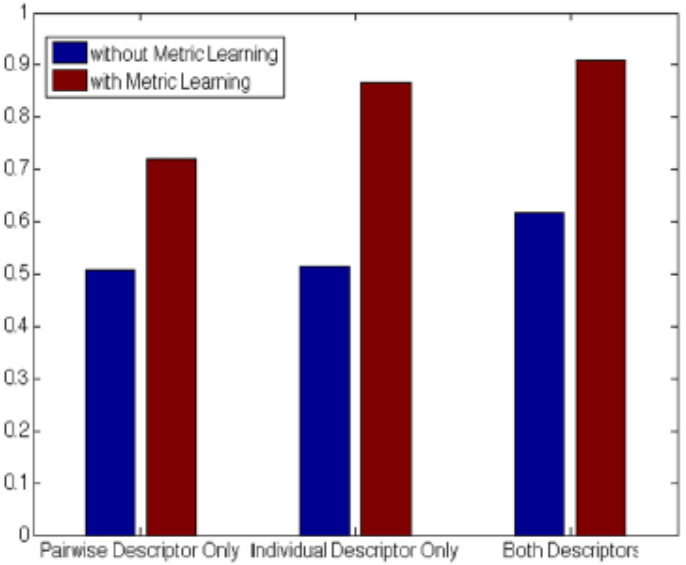
(a)


True Positive Rate
False Positive Rate
Both Descriptors w. Metric Learning
Both Descriptors w/o Metric Learning
Pairwise Descriptor w. Metric Learning
Pairwise Descriptor w/o Metric Learning
Individual Descriptor w. Metric Learning
Individual Descriptor w/o Metric Learning

(b)

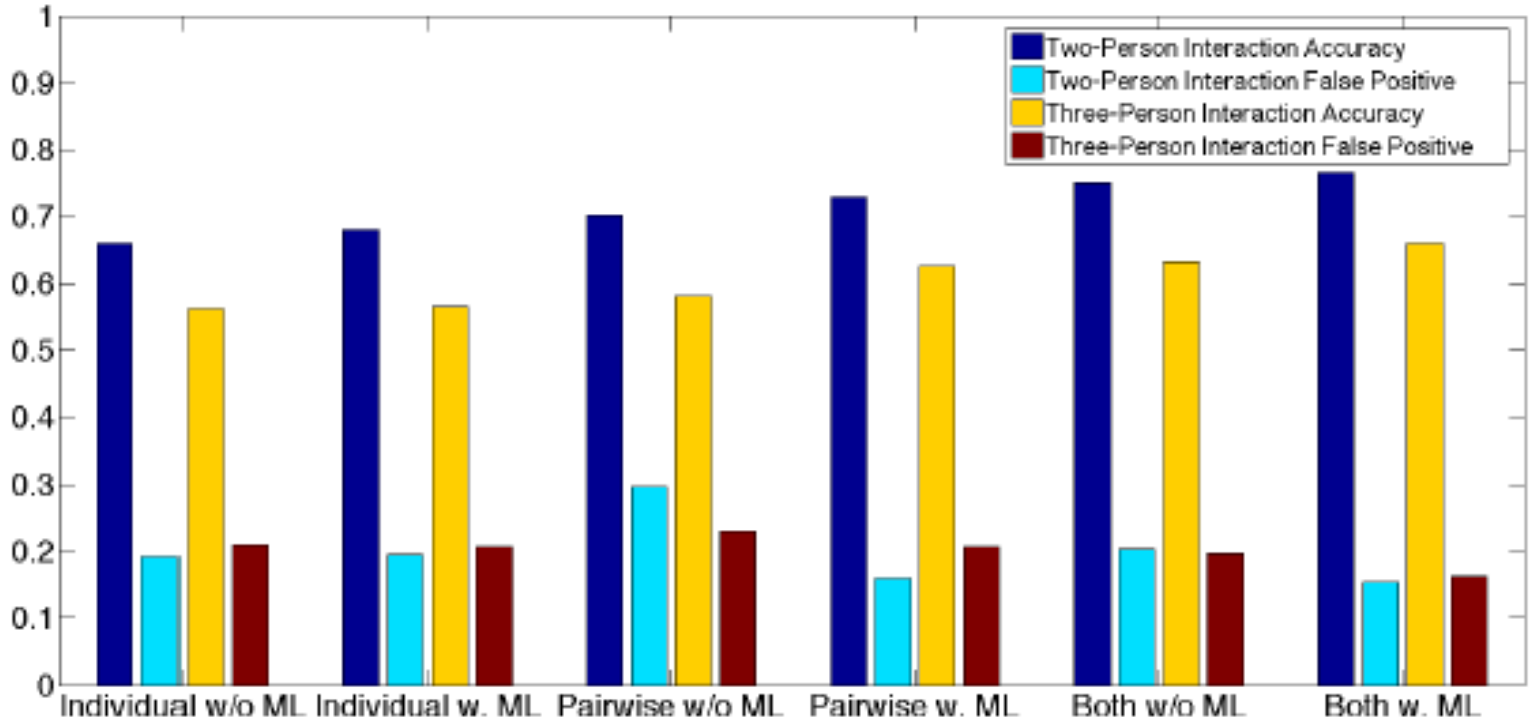

True Positive Rate
False Positive Rate
Both Descriptors w. Metric Learning
Both Descriptors w/o Metric Learning
Pairwise Descriptor w. Metric Learning
Pairwise Descriptor w/o Metric Learning
Individual Descriptor w. Metric Learning
Individual Descriptor w/o Metric Learning

(c)


without Metric Learning
with Metric Learning
Pairwise Descriptor Only
Individual Descriptor Only
Both Descriptors

(d)

Two and three person interaction



Computational Cost Comparison

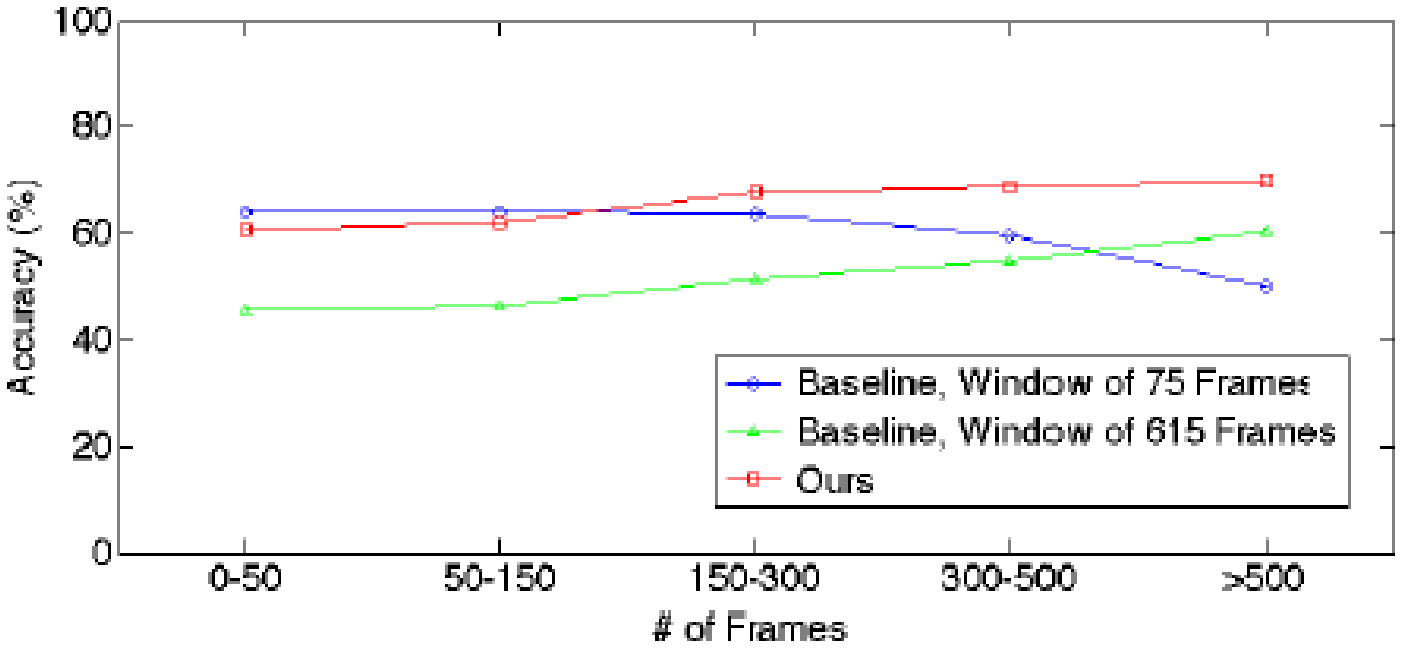
# of Participants	2	3	4
Exhaustive+Sliding Window	17.2	60.4	253.2
Exhaustive+Branch and Bound	12.6	27.6	59.7
Optimal Pairing+Sliding Window	12.4	23.2	40.8
Proposed	8.0	19.8	32.3

- UT-Interaction Dataset
  - 32-dimensional histogram of spatio-temporal features
  - 32-dimensional histograms computed for each of the two humans

	Accuracy ([21], [1], ours)	FP Rate ([21], [1], ours)
Hug	(0.875, 0.904, 1.00)	(0.075, 0.055, 0.00)
Kick	(0.750, 0.775, 0.875)	(0.138, 0.108, 0.063)
Point	(0.625, 0.663, 0.750)	(0.025, 0.025, 0.088)
Punch	(0.500, 0.632, 0.750)	(0.201, 0.154, 0.138)
Push	(0.750, 0.782 , 0.750)	(0.125, 0.101, 0.138)
Shake Hands	(0.750, 0.789, 1.00)	(0.088, 0.060, 0.00)
Average	(0.708, 0.758, 0.854)	(0.108, 0.083, 0.071)

	Individual only	pairwise only	Both
Accur. w. ML	0.688	0.813	0.854
Accur. w/o ML	0.647	0.750	0.771
FP Rate w. ML	0.125	0.096	0.071
FP Rate w/o ML	0.163	0.113	0.083

- Caltech Resident-Intruder Mouse Dataset
  - Spatiotemporal interest points (STIP) based appearance features



X. Burgos-Artizzu, P. Dollar, D. Lin, D. Anderson, and P. Perona. Social behavior recognition in continuous videos. In *CVPR*, 2012.

THANKS