Social behavior recognition in continuous video

Xavier P. Burgos-Artizzu[1] **Piotr Dollár**[2] **Dayu Lin**[3] **David J. Anderson**[4] **Pietro Perona**[1]

[1] Electrical Engineering, California Institute of Technology, {xpburgos,perona}@caltech.edu

[2] Interactive Visual Media, Microsoft Research, Redmond, pdollar@microsoft.com

[3] Dept. of Phsychiatry and Neuroscience, NYU Medical Center, dayu.lin@nyumc.org

[4] Division of Biology, California Institute of Technology, wuwei@caltech.edu

Jane Coffin Childs Memorial Fund for Medical Research

Office of Naval Research

Summary

- Largest behavior dataset to date.
- Social behavior.
- Segmentation and classification of actions in continuous video.
- Novel trajectory features.
- Temporal context in behavior analysis.

Caltech Resident-Intruder Mouse Dataset (CRIM13)

http://www.vision.caltech.edu/Video_Datasets/CRIM13

Largest behavior dataset to date: 88 hours and 8 million frames.

Continuous, unsegmented videos: ~10 min, 140 behavior instances/video.

Real-world scenario: high variance in behavior frequencies and durations, both inter/intra-class.

Frame-by-frame expert annotation: experts spent 350 hours annotating videos.

13 Meaningful behaviors: basis for a state-of-the-art neuroscience study of behavior, published in Nature.

Focused in social interaction: dataset collected to study neurophysiological mechanisms involved in aggression and courtship in mice.

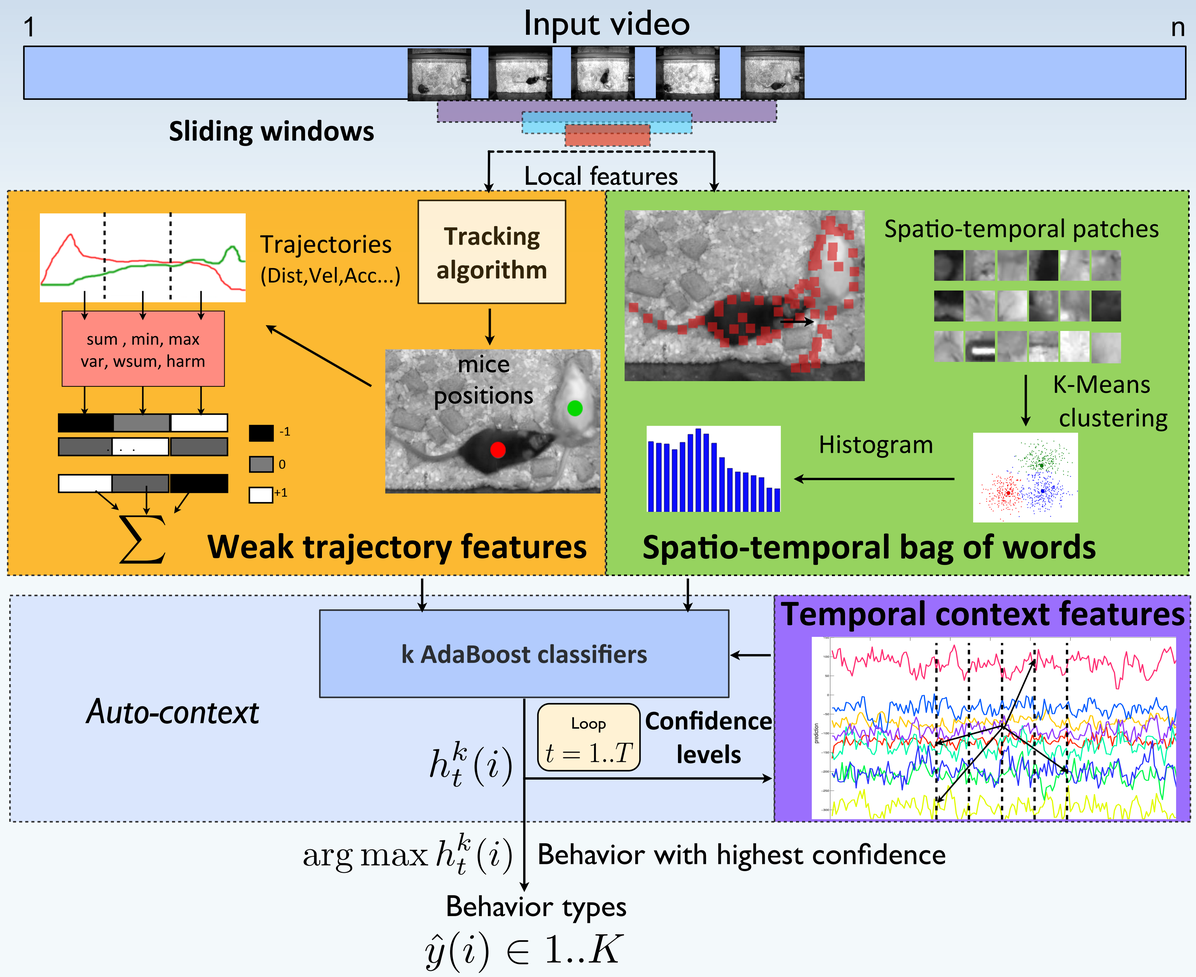
Synchronized top and side views: allow to study behavior from different views.

Spatio-temporal features benchmark

Detector+ Descriptor	Performance	Codebook size	fps
Harris3D+ Pca-Sift	20.9%	250	2.7
Harris3D+ Hog3D	18.7%	250	4.0
Harris3D+ Hog/Hof	15.5%	500	1.1
Cuboids+ Pca-Sift	**24.6%**	**250**	**4.5**
Cuboids+ Hog3D	18.2%	250	8.7
Cuboids+ Hog/Hof	19.8%	500	1.6
Cuboids+ Pca-Sift multi-scale	16.4%	1000	0.8
LTP	22.2%	-	15

- Differences are small.
- We use Cuboids+Pca-Sift.

Proposed method



Trajectory features

Position	$x_{m_{i\in[1,2]}}(t), y_{m_{i\in[1,2]}}(t)$
Distance	$\mathbf{Dist(t)} = \sqrt{(x_{m_1}(t) - x_{m_2}(t))^2 + (y_{m_1}(t) - y_{m_2}(t))^2}$
Dist change	$\mathbf{CDist(t)} = Dist(t) - Dist(t-1)$
Direction	$\mathbf{Dir_i(t)} = atan^{-1}(\frac{y_{m_i}(t) - y_{m_i}(t-1)}{x_{m_i}(t) - x_{m_i}(t-1)})$
Dir change	$\mathbf{CDir_i(t)} = Dir_i(t) - Dir_i(t-1)$
Dir difference	$\mathbf{DDir(t)} = \lvert Dir_1(t) - Dir_2(t)\rvert$
Velocity	$\begin{bmatrix} \mathbf{Vx_{m_i}(t)} \\ \mathbf{Vy_{m_i}(t)} \end{bmatrix} = \frac{1}{\Delta t} \begin{bmatrix} x_{m_i}(t-1) - x_{m_i}(t-2) & x_{m_i}(t) - x_{m_i}(t-1) & x_{m_i}(t+1) - x_{m_i}(t) \\ y_{m_i}(t-1) - y_{m_i}(t-2) & y_{m_i}(t) - y_{m_i}(t-1) & y_{m_i}(t+1) - y_{m_i}(t) \end{bmatrix} * \begin{bmatrix} 0.25 \\ 0.5 \\ 0.25 \end{bmatrix}$
Acceleration	$\begin{bmatrix} \mathbf{Ax_{m_i}(t)} \\ \mathbf{Ay_{m_i}(t)} \end{bmatrix} = \begin{bmatrix} \frac{Vx_{m_i}(t+1) - Vx_{m_i}(t-1)}{2\Delta t} \\ \frac{Vy_{m_i}(t+1) - Vy_{m_i}(t-1)}{2\Delta t} \end{bmatrix}$

Temporal context features

Confidence levels	$h_{t-1}^k(i)$
Conf. differences between behaviors	$\mathbf{D(k_1, k_2, i)} = h_{t-1}^{k_1}(i) - h_{t-1}^{k_2}(i), \forall k1, k2 \in 1..K$
Mean behavior confidence (all, past, post)	$\begin{bmatrix} \mu(\mathbf{i} - \frac{\mathbf{sz}}{2}, \mathbf{i} + \frac{\mathbf{sz}}{2}) & \mu(\mathbf{i} - \frac{\mathbf{sz}}{2}, \mathbf{i}) & \mu(\mathbf{i}, \mathbf{i} + \frac{\mathbf{sz}}{2}) \end{bmatrix}$ where, $\mu(\mathbf{start}, \mathbf{end}) = \frac{\sum_{start}^{end} h_{t-1}^k}{(end - start)}$
1st order derivative (all, past, post)	$\begin{bmatrix} \frac{\partial \mathbf{h_{t-1}^k}}{\partial \mathbf{t}}(\mathbf{i} - \frac{\mathbf{sz}}{2}, \mathbf{i} + \frac{\mathbf{sz}}{2}) & \frac{\partial \mathbf{h_{t-1}^k}}{\partial \mathbf{t}}(\mathbf{i} - \frac{\mathbf{sz}}{2}, \mathbf{i}) & \frac{\partial \mathbf{h_{t-1}^k}}{\partial \mathbf{t}}(\mathbf{i}, \mathbf{i} + \frac{\mathbf{sz}}{2}) \end{bmatrix}$ where, $\frac{\partial \mathbf{h_{t-1}^k}}{\partial \mathbf{t}}(\mathbf{start}, \mathbf{end}) = h_{t-1}^k(end) - h_{t-1}^k(start)$

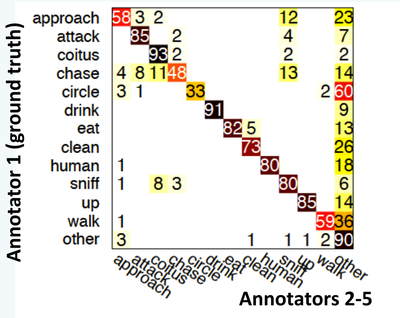
Results

Features used	No Temporal Context	With Temporal Context
TF	52.3%	58.3%
STF (both)	29.3%	43.0%
STF (top)	26.6%	39.3%
STF (side)	28.2%	39.1%
(Full method) TF + STF	53.1%	**61.2%**

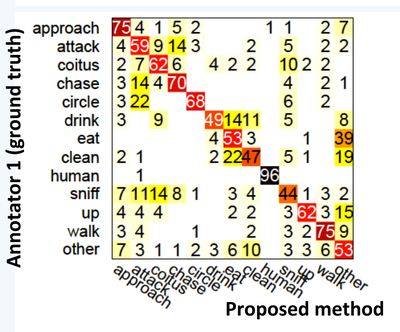
- Spatio-temporal features are not sufficient.
- Novel weak trajectory features (TF) perform 20% better than STF.
- Combination of both features achieves best results.
- Temporal context is crucial, improving final classification 8.1% (14% in average).

Comparison with human annotations

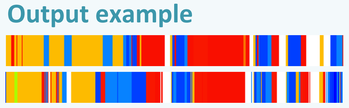
Human agreement 69.7%



Proposed approach 62.6%



Output example



Conclusions

- CRIM13 is a realistic and challenging dataset.
- Novel trajectory features are more discriminative than spatio-temporal features.
- Temporal context is crucial in continuous videos.
- Our method's performance is not far from that of trained human annotators.