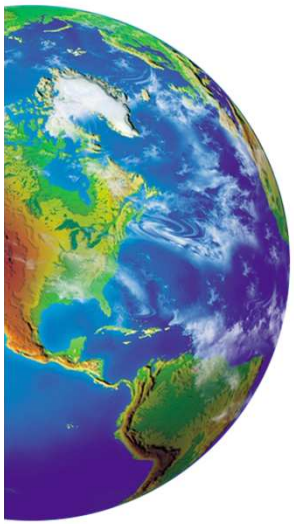
Spatial-Temporal Phrases for Activity Recognition



ECCV 2012

Yimeng Zhang, Xiaoming Liu, Ming-Ching Chang, Weina Ge, Tsuhan Chen
Computer Vision Lab.

Cornell University
School of Electrical and Computer Engineering

GE Global Research

NIJ #2009-SQ-B9-K013

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Overview

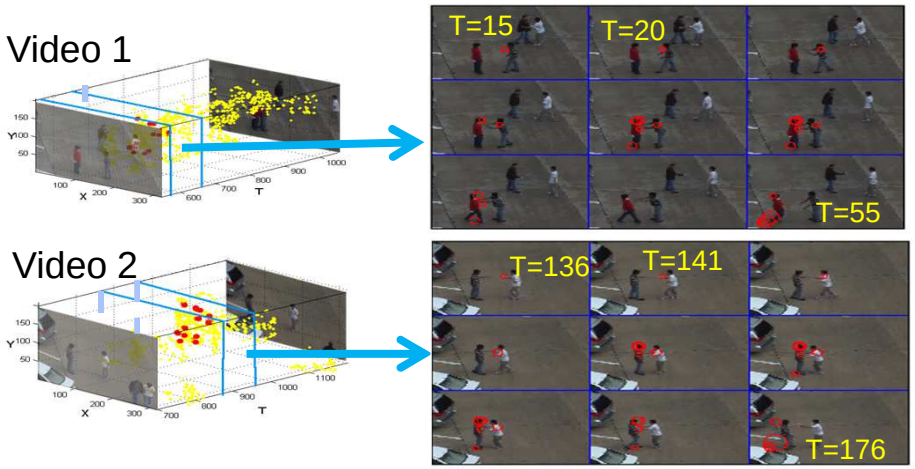
Goal: algorithm that models the spatial and temporal information of the local features, which capture the local movements of the body parts.

Limitation of Prior Work: either 1) not translation or temporal-shift invariant, 2) only capture neighboring information of local movements, 3) only encode weak/limited spatial or temporal information

Proposed: **efficient** algorithm that identifies both local and **long-range** shift-invariant motion interactions, eg. the causality relationship of the hand movement of one person and the foot movement of the other person several frames later.

Main Idea: Bag of Spatio-Temporal Phrases

ST-Phrases: a combination of local features in a certain spatial and temporal layout.



Runtime of proposed algorithm: detect all ST-phrases in linear time

Example co-occurring ST-Phrases detected with our algorithm from two videos including the *push* activity at different time stamp

Main Contribution:

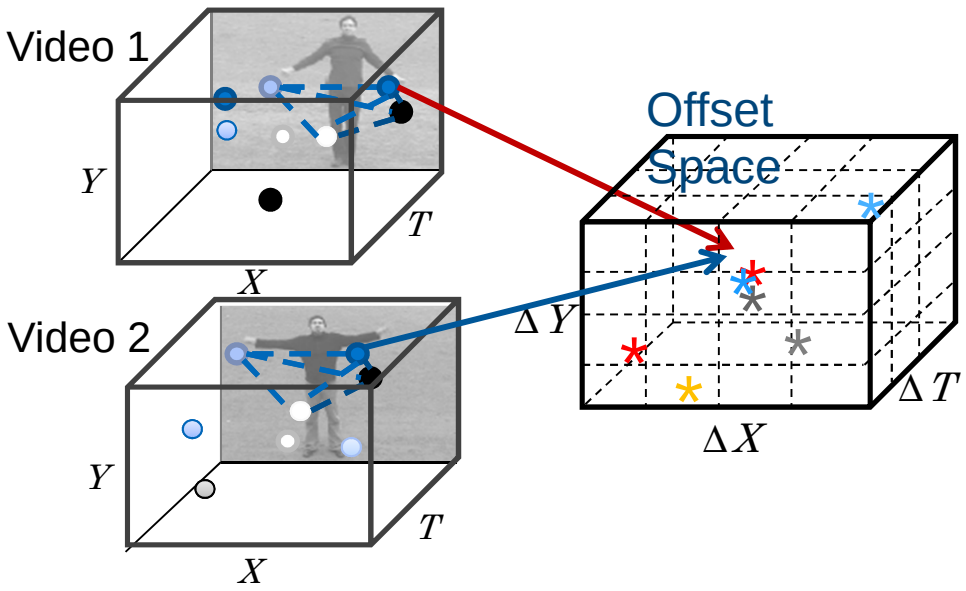
1) extend the Bag-of-Phrase algorithm [Zhang et al. CVPR'11] from 2D static images to videos;
2) an efficient algorithm for online activity detection

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Approach

Detecting co-occurring ST-phrases from two videos

(extend algorithm of [Zhang et al. CVPR11])



Activity Speed or Scale Variations:
add a temporal or spatial scaling dimension to the offset space

ST-Kernel for the SVM:
Inner-product of the bag-of-ST-phrase histograms = the number of co-occurring ST phrases in two videos

Decision Score for a test image *V*:

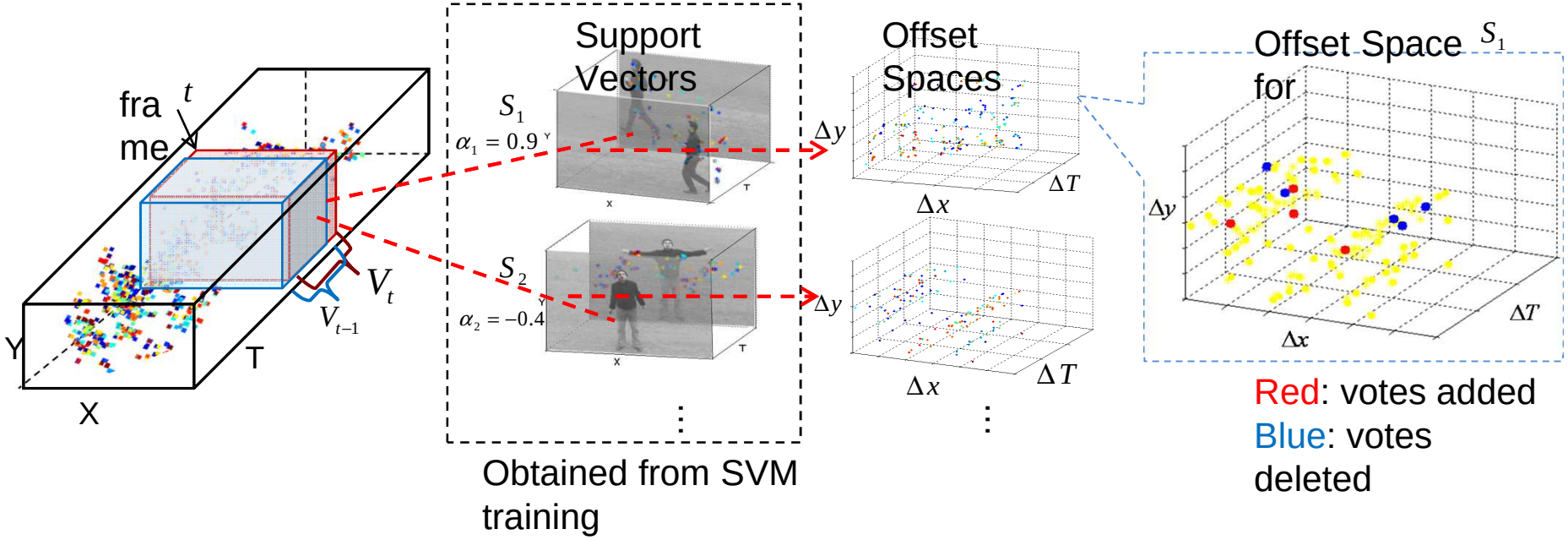
$$Score(V) = \sum_j \alpha_j K(V; S_j)$$

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Approach

imagination at work

Online activity detection: classify every video frame (using the previous T seconds)
Algorithm: update the kernel values by avoiding the overlapping part of each frame



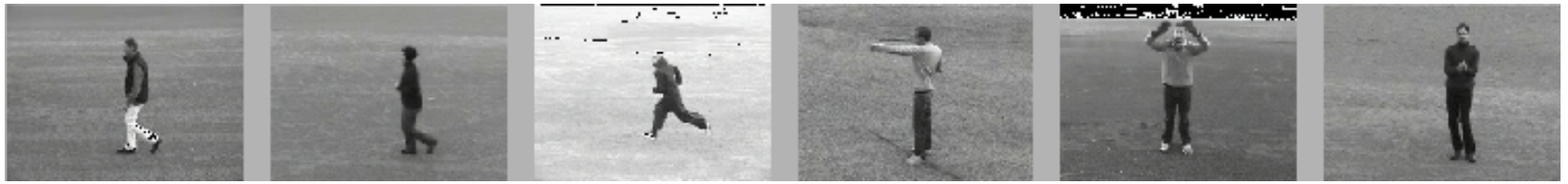
Only need to update the offset space by processing votes from two frames, thus we can include more context frames when processing each frame

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Experiments

Single Person Activity: KTH Dataset

Walking Jogging Running Boxing Hand waving Hand clap



2391 videos of 25 subjects

Setting	BoW	**BoP**	SPM [1]	HT [2]	Correlaton [3]	ST-relation [4]
16/9	91.5	**94.0**	91.8	n/a	n/a	91.1
5-folder	92.9	**94.6**	n/a	92.0	n/a	n/a
leave one out	91.9	**95.5**	n/a	n/a	94.2	93.8

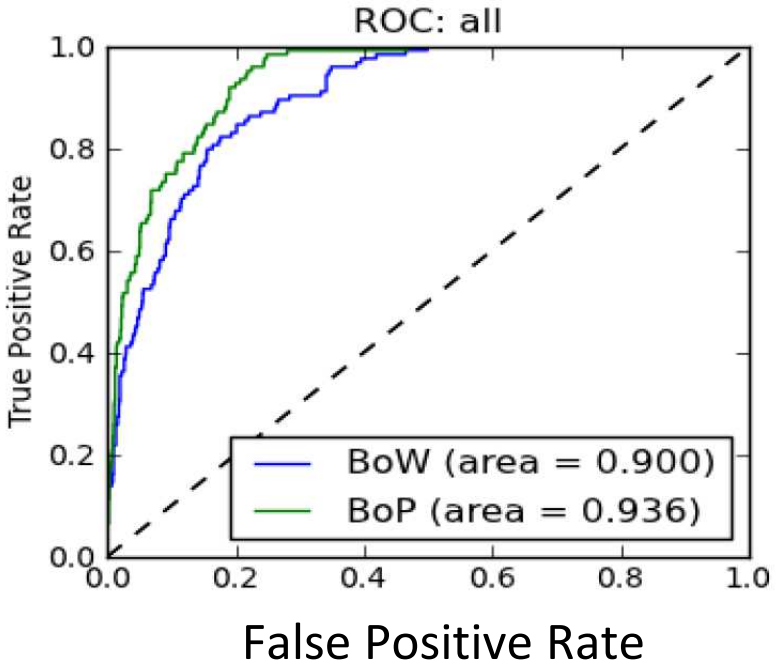
Accuracies (%) using different train-test settings: [1] Laptev et al. CVPR08 [2] Yao et al., CVPR10 [3] Savarese et al. WMVC08 [4] Ryoo et al. ICCV09

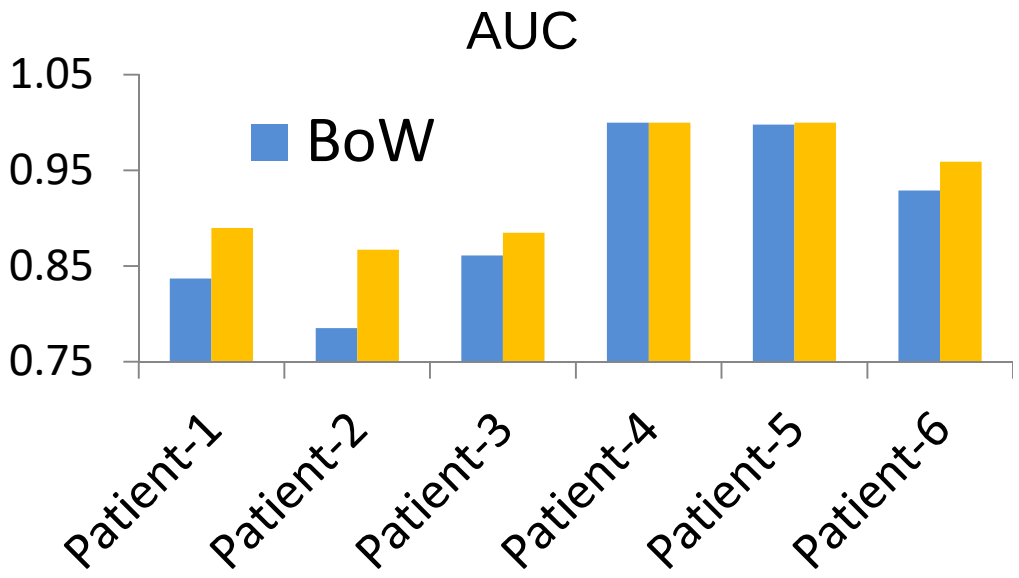
This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Experiment

Single Person Activity: Hospital Surveillance Dataset

Real surveillance videos of 6 patients in the private sickrooms.
Goal: detect abnormal behaviors of the patients





This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Experiment

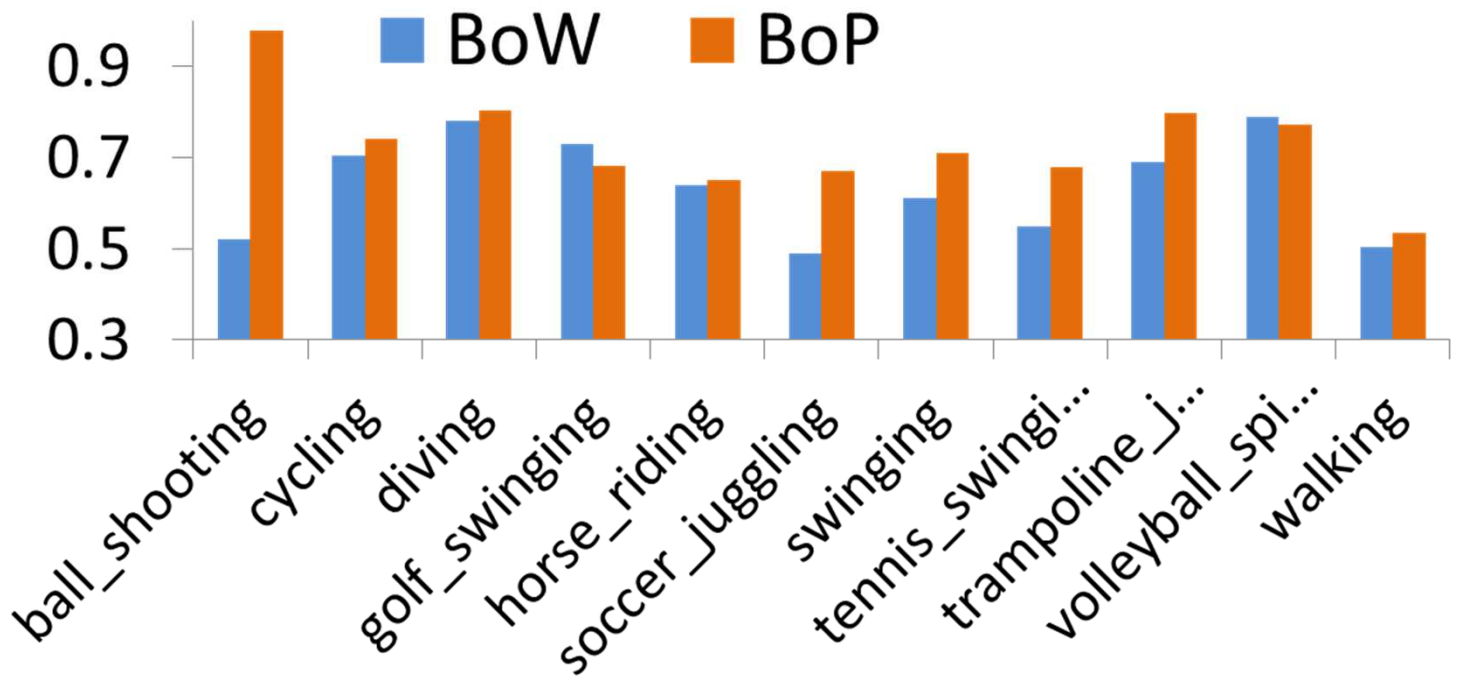
Single Person Activity: YouTube Action Dataset

1168 videos of 11 categories

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Experiment

Single Person Activity: YouTube Action Dataset



Average: 63.7% (BoW) → 72.9% (BoP)

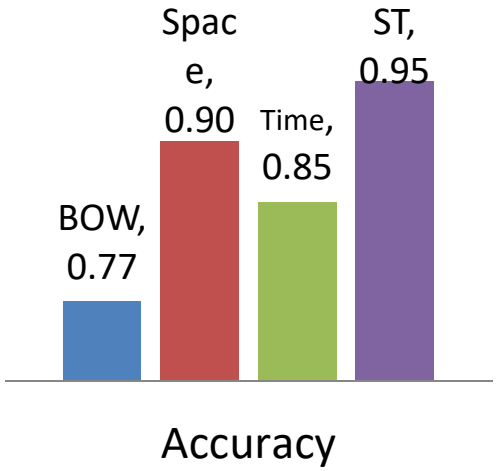
Main improvement**:** different *swing* activities or categories both involve *jump*

Comparison: [Liu et al, CVPR09]: 65.4% (same features as us), 71.2% (add additional features)

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Experiment

Multiple Person Activity: HT-interact Dataset



Dataset	Method	Avg.	Shake	Hug	Kick	Point	Push	Punch
Set 1	BoW	0.77	0.70	0.80	0.90	**1.00**	0.50	0.70
	Hough Voting	0.83	0.50	1.00	1.00	1.00	0.70	0.80
	BoP	**0.95**	**1.00**	**1.00**	**1.00**	0.90	**0.90**	**0.90**
Set2	BoW	0.73	0.70	0.70	0.80	0.80	0.70	0.70
	Hough Voting	0.80	0.70	0.90	1.00	**1.00**	0.80	0.40
	BoP	**0.90**	**0.80**	**1.00**	**1.00**	0.80	**0.90**	**0.90**

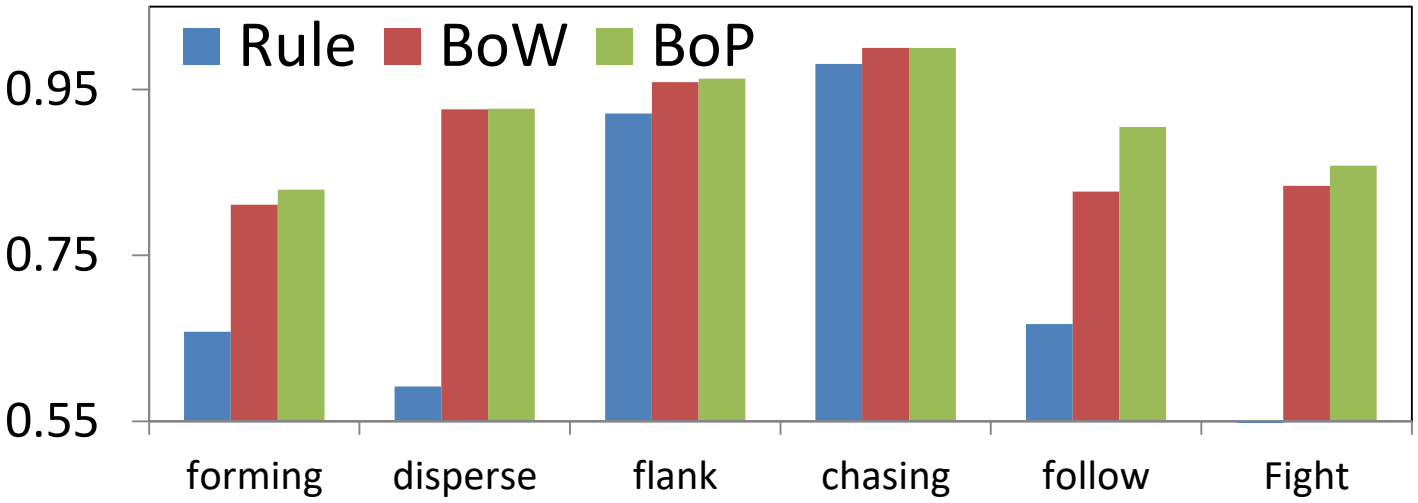
Left: effect of incorporating space and time separately
Right: comparison with Hough Voting (winner of ICPR 10)

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

Experiment

Online Group-Level Activity: MPR Dataset

19 continous surveillance videos in a prison yard



Comparison of AUC scores:
BoW [Zhang et al. WACV12]
Rule [Chang et al. ICCV11]

This project was supported by grant #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.