Detecting Human Actions in Surveillance Videos

Ming Yang, Shuiwang Ji, Wei Xu, Jinjun Wang, Fengjun Lv, Kai Yu, Yihong Gong

NEC Laboratories America, Inc., Cupertino, CA, USA

Mert Dikmen, Dennis J.Lin, Thomas S.Huang

Dept. of ECE, UIUC, Urbana, IL, USA

Online

- Introduction
- NEC's System
  - Human detection and tracking
  - BoW features based SVM
  - Cube based Convolutional Neural Networks
- Experiments
- UIUC's System
- Conclusions

Motivation

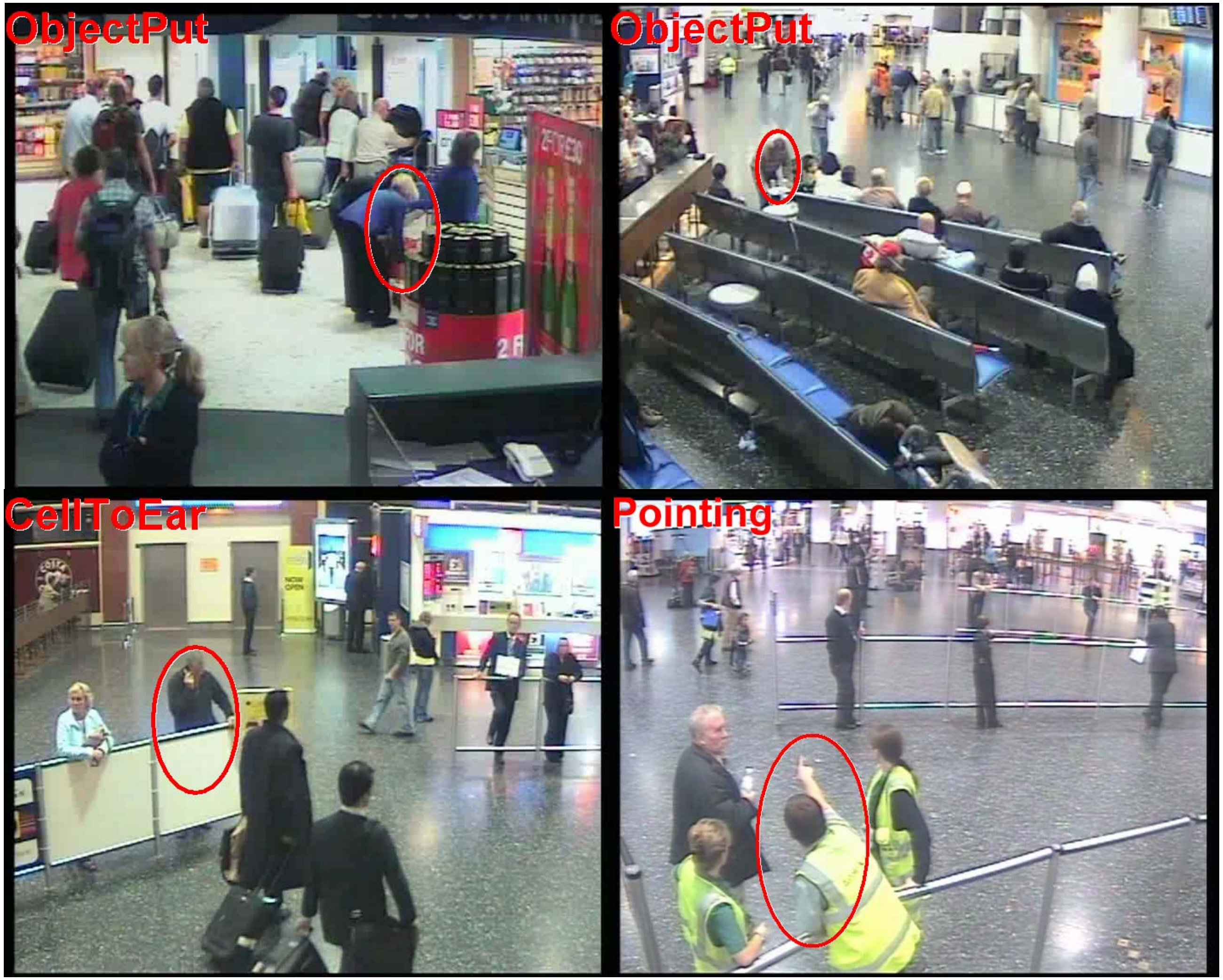
- Huge advances in action recognition in controlled environment or in movie or sports videos.
  – Known temporal segments of actions
  – One action occurs at a time
  – Little scale and viewpoint changes
  – Static and clean background
  – Actions are less natural in staged environments
- How is the performance of action detection in huge amount of real surveillance videos?

TRECVid 2009 Event Detection

NEC Laboratories America
Relentless passion for innovation

- Real surveillance videos recorded in London Gatwick Airport.
  - Crowded scenes with cluttered background
  - Large variances in scales, viewpoints and action styles
- Huge amount of video data:
  - ~144 hours of videos with image resolution 720×576
  - Computational efficiency is very critical!
- 10 required events:
  - *CellToEar*, *Objectput*, *Pointing*, PersonRuns, PeopleMeet, PeopleSplit, OpposeFlow, Embrace, ElevatorNoEntry, TakePicture.

TRECVid 2009 Event Detection



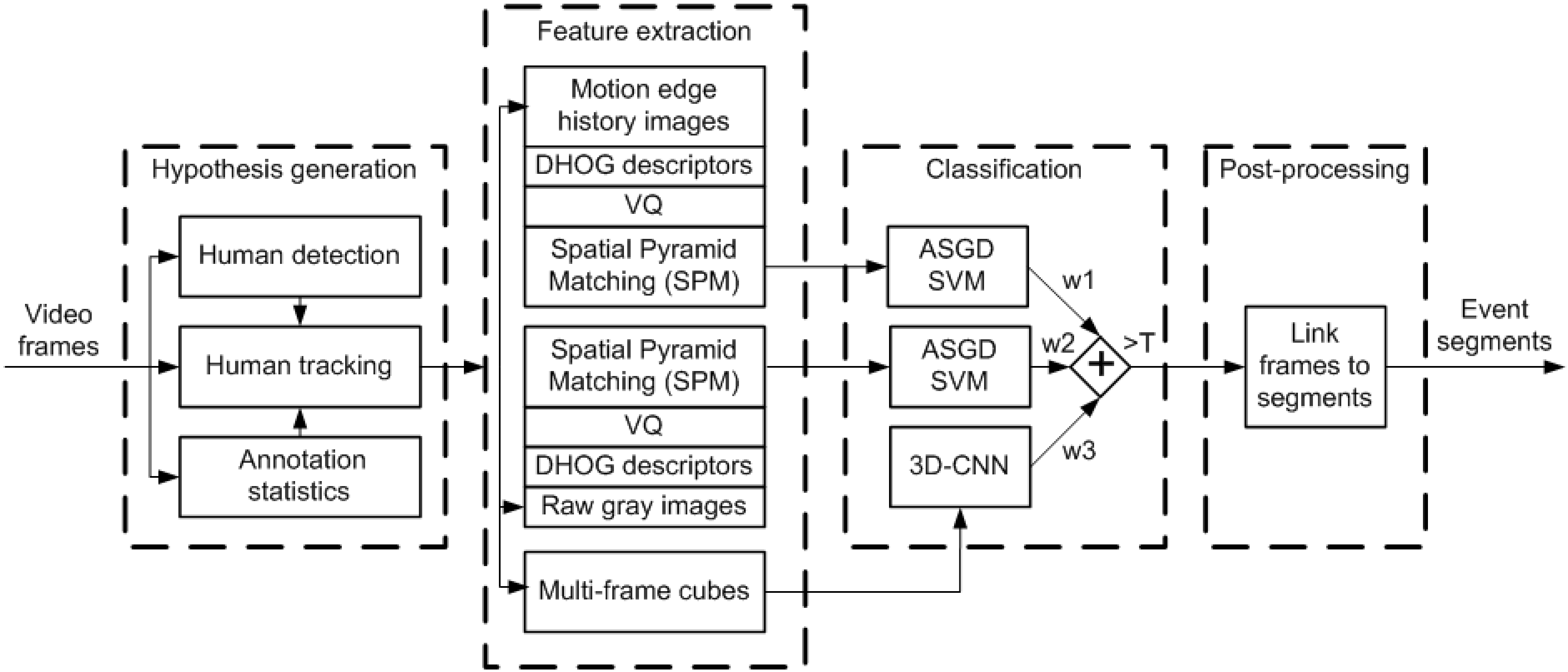
A formidably challenging task !

Related Work

- Action representations:
  – graphical models of key poses or examplars
  – holistic space-time templates
  – bag-of-words models of space-time interest points
  – A vast pool of spatio-temporal features
- How to locate actions:
  – sliding window/volume search
  – efficient subwindow/subvolume search
  – human detection and tracking

NEC's System

NEC Laboratories America
Relentless passion for innovation



Human Detection and Tracking

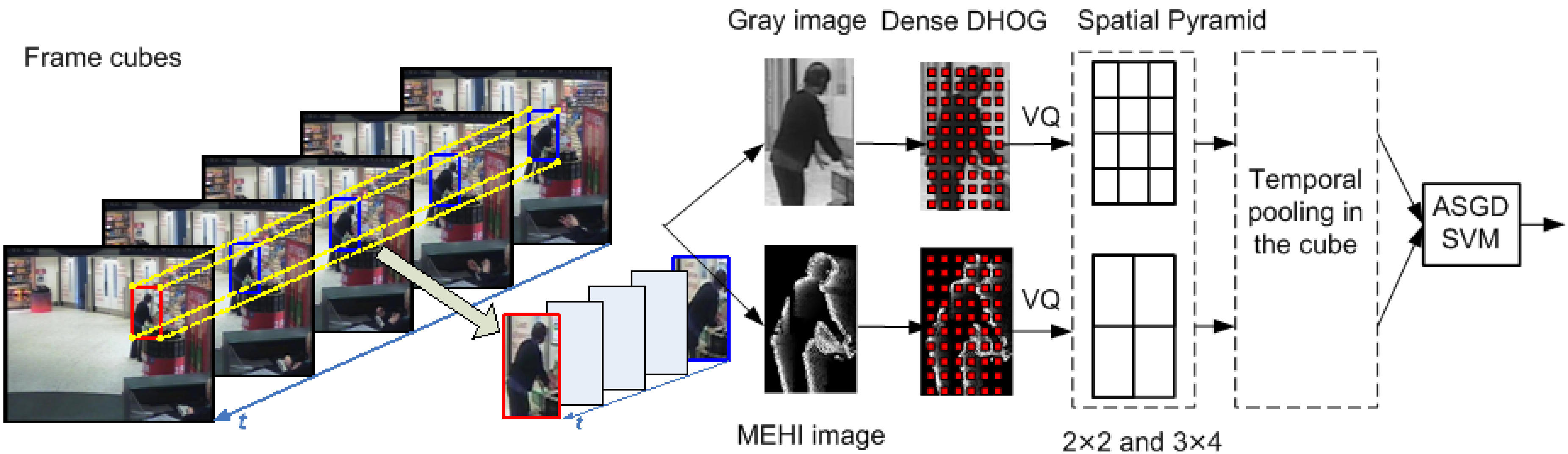
NEC Laboratories America
Relentless passion for innovation

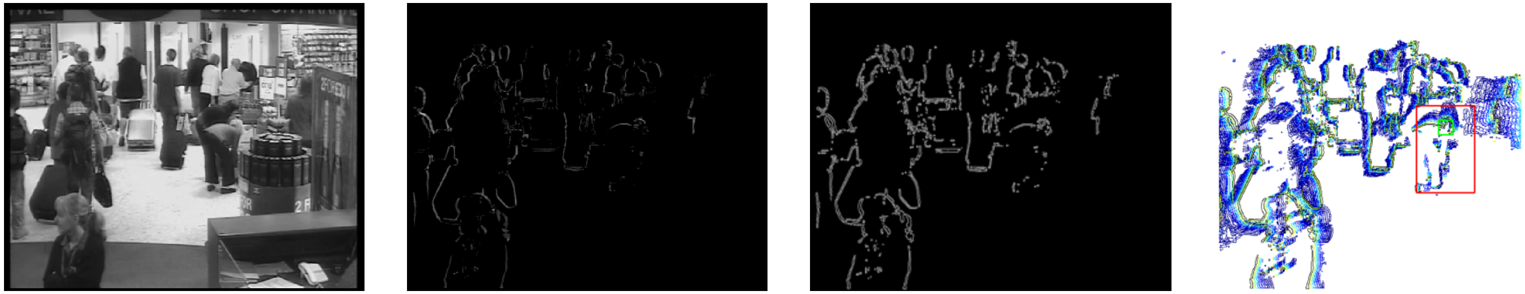
- The human detector
  - Based on Convolutional Neural Networks (CNN)
- The human tracker
  - A new multi-cue based head tracker

Table 1. Performance of detection and tracking (per head)

Per frame	CAM1	CAM2	CAM3	CAM5	Overall
# of frames	3775	3774	3774	3772	15095
avg. # of labels	5.505	24.315	11.486	7.330	12.159
avg. # of detected heads	3.349	16.122	7.236	5.459	8.042
avg. # of tracked heads	4.120	21.545	8.940	8.070	10.668
recall of the detector	43.53%	46.25%	42.58%	45.25%	**44.81%**
precision of the detector	74.40%	67.37%	66.09%	62.21%	**66.99%**
recall of the tracker	51.68%	56.76%	48.66%	54.11%	**53.65%**
precision of the tracker	70.80%	62.42%	61.19%	51.03%	**60.80%**

BoW features based SVM

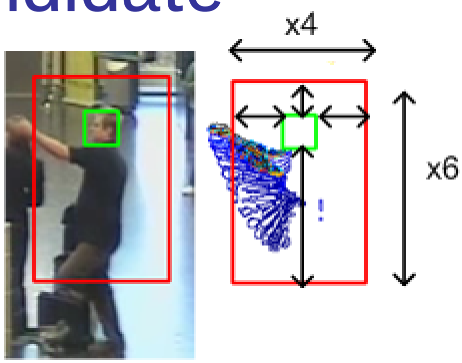




Motion edge history image (MEHI)

Implementation

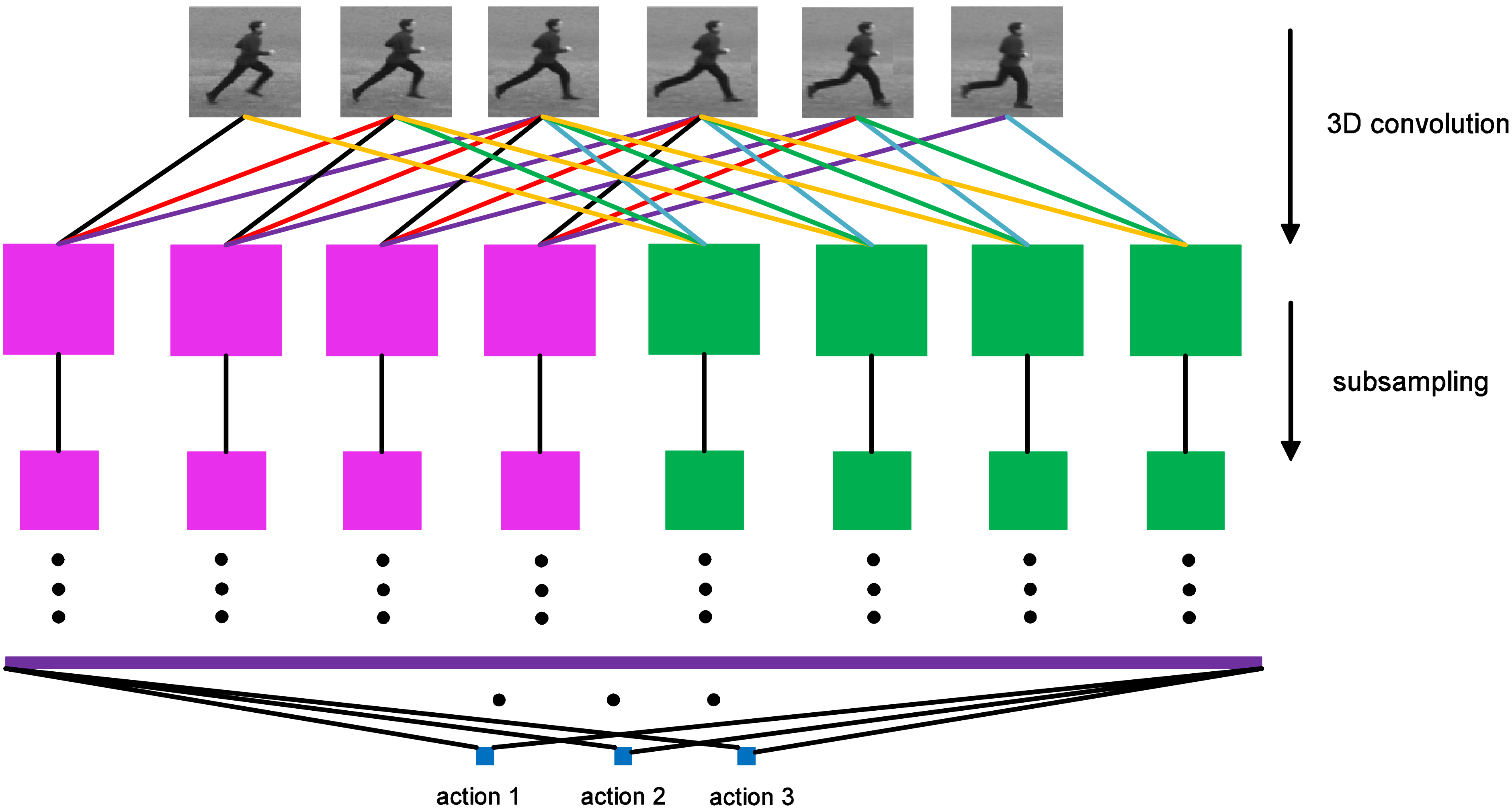
- Dense DHOG features
  - Every 6 pixels from 7×7 and 16×16 patches
  - Soft quantization using a 512-word codebook
- Spatial pyramids
  - 2×2 and 3×4 cells
- Frame based or cube based
  - 1 frame or 7 frames (-6, -4, -2, 0, 2, 4, 6)
- The feature vector for one candidate
  - $512 \times (2 \times 2 + 3 \times 4) = 8192\text{D}$



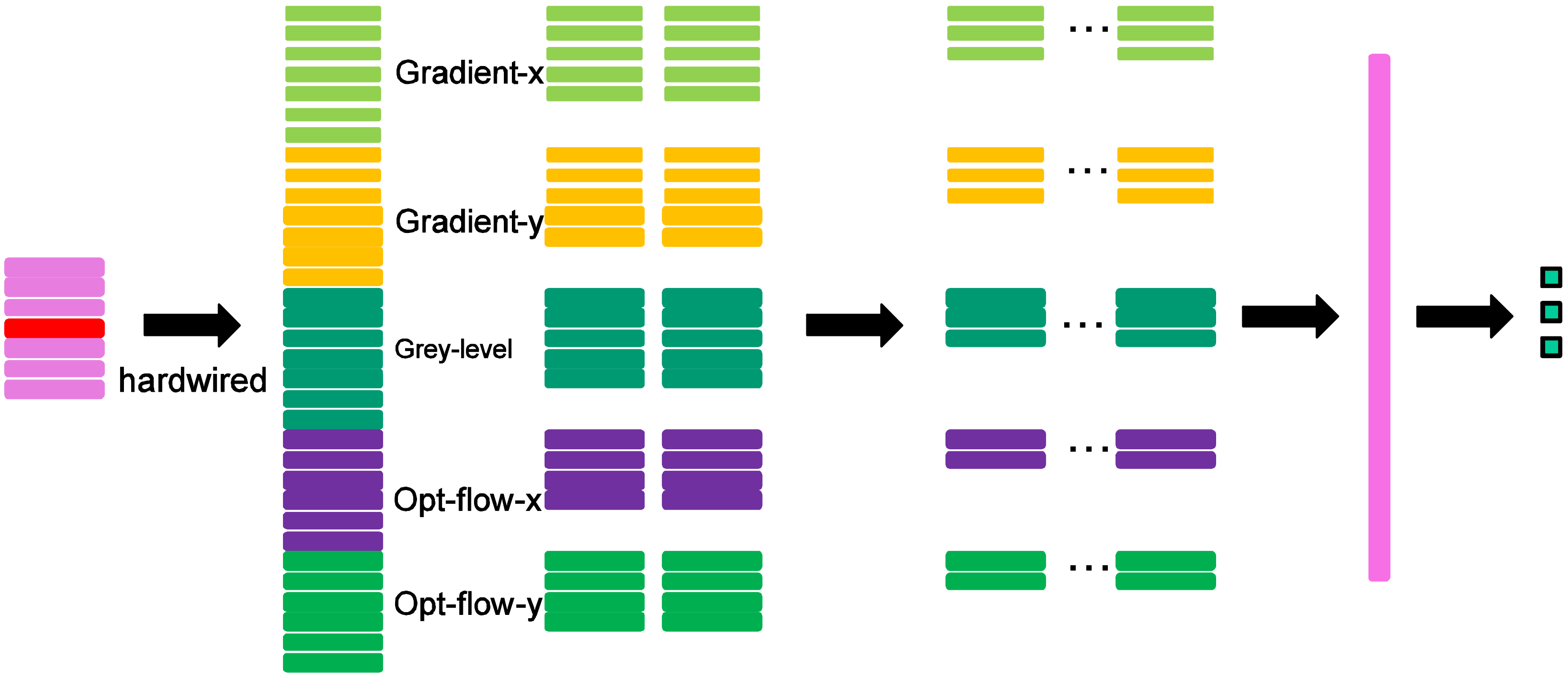
Training of SVM Classifiers

- Binary SVM classifiers for each action category
- One set of training features: 520K in total
  - 520K × 8192 × 4 (float)=17G bytes
- SVM classifiers trained by averaged stochastic gradient descent (ASGD)
- Highly efficient for training on large scale datasets
  - 2.5 mins to train 3 SVM classifiers on a 64bit blade server
  - CPU Intel Xeon 2.5GHz (8 cores)
  - 16GB RAM

Cube based CNN

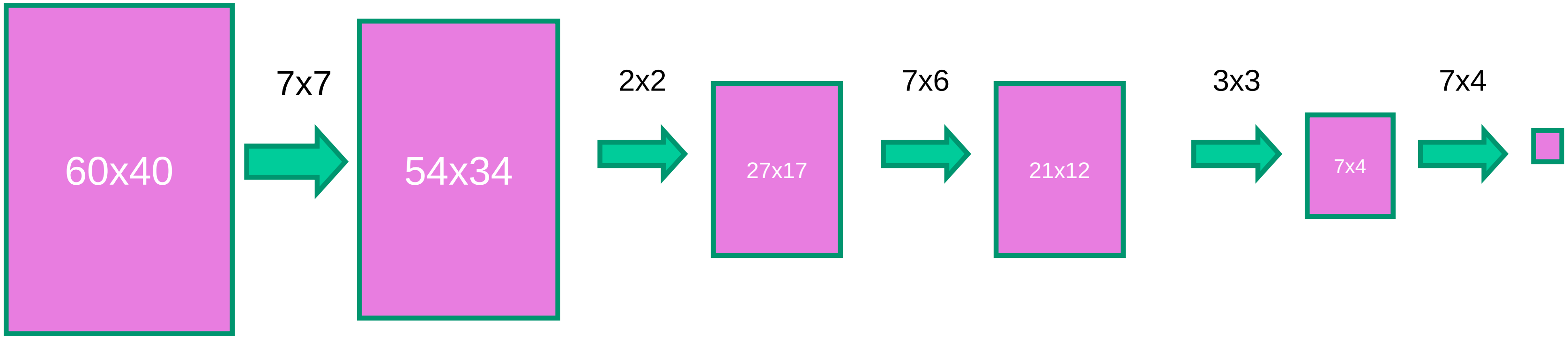


CNN Architecture



- Each candidate is a cube of 7 frames
- 5 different types of input features

CNN Configuration



- Input image patches: 60x40
- Use 3 frames before and 3 frames after current frame with step size 2
  – i.e., -6, -4, -2, 0, 2, 4, 6
- Compute N*3+(N-1)*2 feature maps from N=7 input frames using hardwired weights
  – Grey, x-gradient, y-gradient, x-optical-flow, y-optical-flow

What Else We Tried?

NEC Laboratories America
Relentless passion for innovation

- Sparse coding of DHOG features
  - The computations are unaffordable.
- Gaussian Mixture Model (GMM)
  - The storage and memory requirements are unaffordable.

Experiments

- Criteria: Normalized Detection Cost Rate (*NDCR*)
- Training set: ~100 hours of videos
- Test Set: ~14 hours out of 44 hours
  - The subset of 14 hours videos used in testing is unknown to participants
- The entire system is implemented with C++
  - 64bit blade servers with Intel Xeon 2.5GHz CPU (8 cores) and 16GB RAM.

Training Sample Preparation

NEC Laboratories America
Relentless passion for innovation

- Positive samples
  - Label the person performing the action every 3 frames
  - Generate 6 additional samples by some perturbations
- Negative samples
  - The same person performing the actions in two 30-frame intervals before and after the action occurs.
  - The detected persons that are not performing the actions when the action occurs.

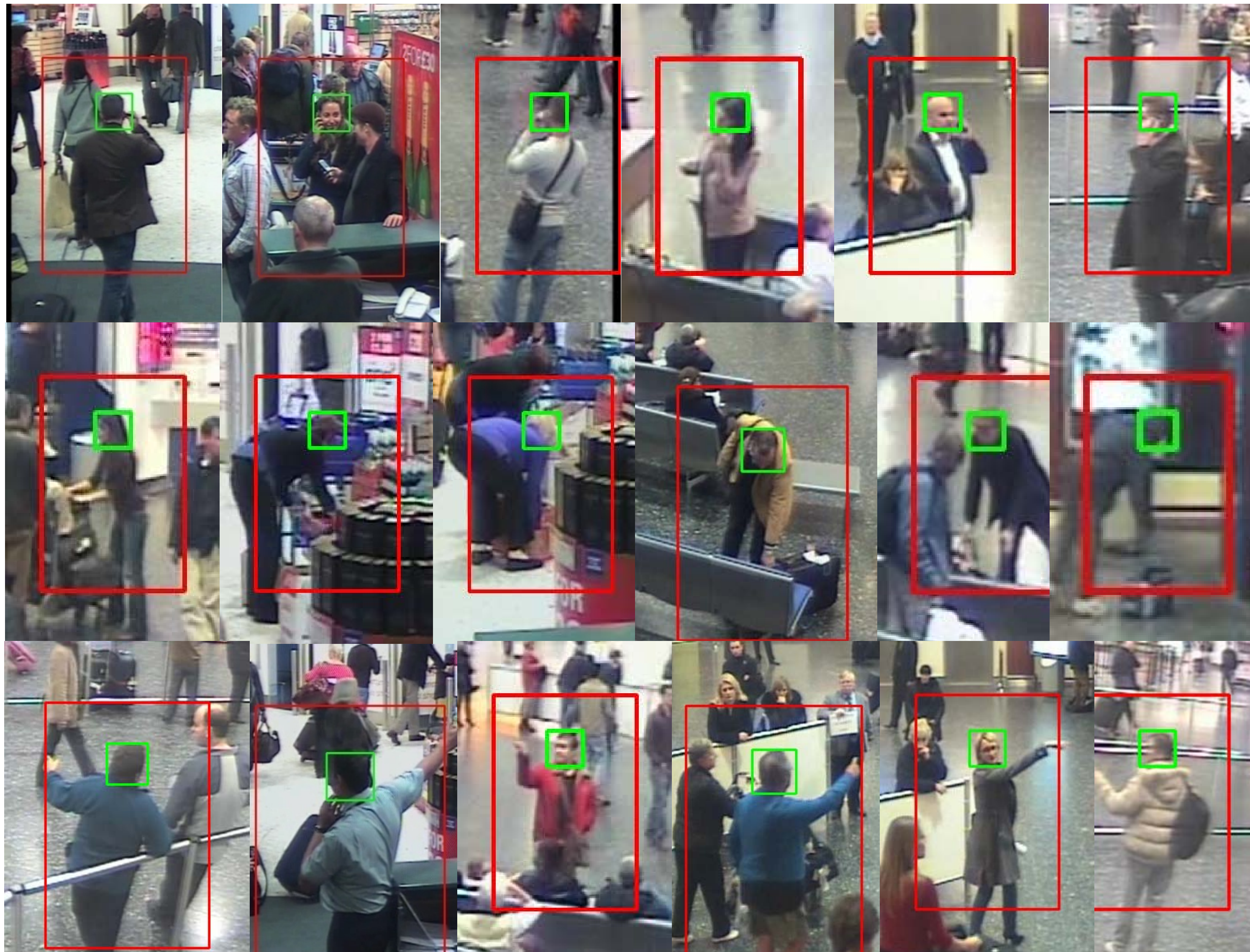
CellToEar	ObjectPut	Pointing	Negative	Total
25.2K	39.3K	152.2K	303K	520K

Sample of Positive Samples

CelltoEar

ObjectPut

Pointing



Feature Extraction

- Training of the codebook using K-Means based on 8 hours videos on 11/12/2007
- 4 set of BoW features:
  - Gray-Frame
  - Gray-Cube
  - MEHI-Frame
  - MEHI-Cube
- 3D-CNN
- Evaluation on a 2-hour video may take 1-2 days.

Parameter Selection

NEC Laboratories America
Relentless passion for innovation

- Linear combination of scores from 3 methods
- Exhaustive search of the weights and threshold to minimize the NDCR directly.
- NDCR calculation is implemented with C++.
- 5-fold cross-validation to evaluate the performance
- Search the best parameters for 2 combinations
  - Gray-Frame + Gray-Cube + MEHI-Cube
  - Gray-Frame + MEHI-Frame + 3D-CNN

Cross-validation (1)

Table 2. 5-fold cross-validation performance of *Gray-Frame* + *Gray-Cube* + *MEHI-Cube*

	CellToEar	ObjectPut	Pointing
CAM1	1.0000 (40/0/0)	0.9979 (706/9/38)	0.9973 (926/5/9)
CAM2	1.0015 (265/0/6)	0.9937 (1122/7/11)	0.9990 (999/3/8)
CAM3	1.0053 (262/0/21)	1.0010 (843/1/9)	1.0023 (1056/0/9)
CAM5	0.9526 (239/21/217)	1.0000 (432/0/0)	1.0070 (1048/2/36)
Overall	**0.9896** (806/21/244)	**0.9981** (3103/17/58)	**1.0014** (4029/10/62)

Table 4. Parameter selection of *Gray-Frame* + *Gray-Cube* + *MEHI-Cube*

	CellToEar	ObjectPut	Pointing
CAM1	1.0003 (40/0/0) (0.3,0.3,0.4,1)	0.9871 (706/17/33) (0.3,0,0.7,0.67)	0.9971 (926/6/10) (0,0.6,0.4,0.77)
CAM2	1.0003 (265/0/0) (0.3,0.3,0.4,1)	0.9944 (1122/7/12) (0.7,0,0.3,0.72)	0.9968 (999/6/10) (0.2,0.2,0.6,0.65)
CAM3	1.0003 (262/0/0) (0.3,0.3,0.4,1)	1.0002 (843/1/3) (0,0.3,0.7,0.74)	1.0001 (1056/0/1) (0,0.5,0.5,0.92)
CAM5	0.9591 (239/20/218) (0,0.3,0.7,0.15)	0.9991 (432/1/0) (0.2,0.3,0.5,0.86)	0.9963 (1048/8/11) (0.2,0,0.8,0.81)
Overall	**0.9892** (806/20/218)	**0.9946** (3103/26/48)	**0.9970** (4029/20/32)

Cross-validation (2)

Table 3. 5-fold cross-validation performance of *Gray-Frame* + *MEHI-Frame* + *3D-CNN*

	CellToEar	ObjectPut	Pointing
CAM1	1.0000 (40/0/0)	0.9915 (706/13/33)	0.9978 (926/4/7)
CAM2	1.0000 (265/0/0)	1.0059 (1122/2/34)	1.0000 (999/2/8)
CAM3	1.0313 (262/0/125)	1.0010 (843/0/4)	1.0033 (1056/3/25)
CAM5	0.9507 (239/17/132)	1.0003 (432/0/1)	1.0088 (1048/9/71)
Overall	**0.9954** (806/17/257)	**0.9997** (3103/15/72)	**1.0025** (4029/18/111)

Table 5. Parameter selection of *Gray-Frame* + *MEHI-Frame* + *3D-CNN*

	CellToEar	ObjectPut	Pointing
CAM1	1.0003 (40 0 0) (0.3,0.3,0.4,1)	0.9866 (706/16/30) (0.5,0.3,0.2,0.54)	0.9961 (926/6/6) (0.6,0.3,0.1,0.74)
CAM2	1.0003 (265/0/0) (0.3,0.3,0.4,1)	0.9974 (1122/9/32) (0.1,0.5,0.4,0.55)	0.9982 (999/4/6) (0.3,0.7,0,0.73)
CAM3	1.0000 (262/0/0) (0.3,.3,0.4,1)	1.0000 (843/1/2) (0.5,0.5,0,0.67)	0.9994 (1056/3/7) (0.5,0.1,0.4,0.59)
CAM5	0.9529 (239/17/152) (0,0.6,0.4,0.41)	0.9994 (432/1/1) (0.4,0.4,0.2,0.67)	0.9968 (1048/18/59) (0,0.6,0.4,0.46)
Overall	**0.9877** (806/17/152)	**0.9953** (3103/27/66)	**0.9970** (4029/31/78)

Submissions

- NEC-1:
  - Gray-Frame + Gray-Cube + MEHI-Cube
  - CelltoEar: 118; ObjecPut: 21; Pointing: 27
- NEC-2
  - Gray-Frame + MEHI-Frame + 3DNN
  - CelltoEar: 63; ObjecPut: 26; Pointing: 19
- NEC-3
  - Combination of NEC-1 and NEC-2 on per camera per event basis according to the cross-validation
  - CelltoEar: 63; ObjecPut: 13; Pointing: 27
- UIUC-1

Performance

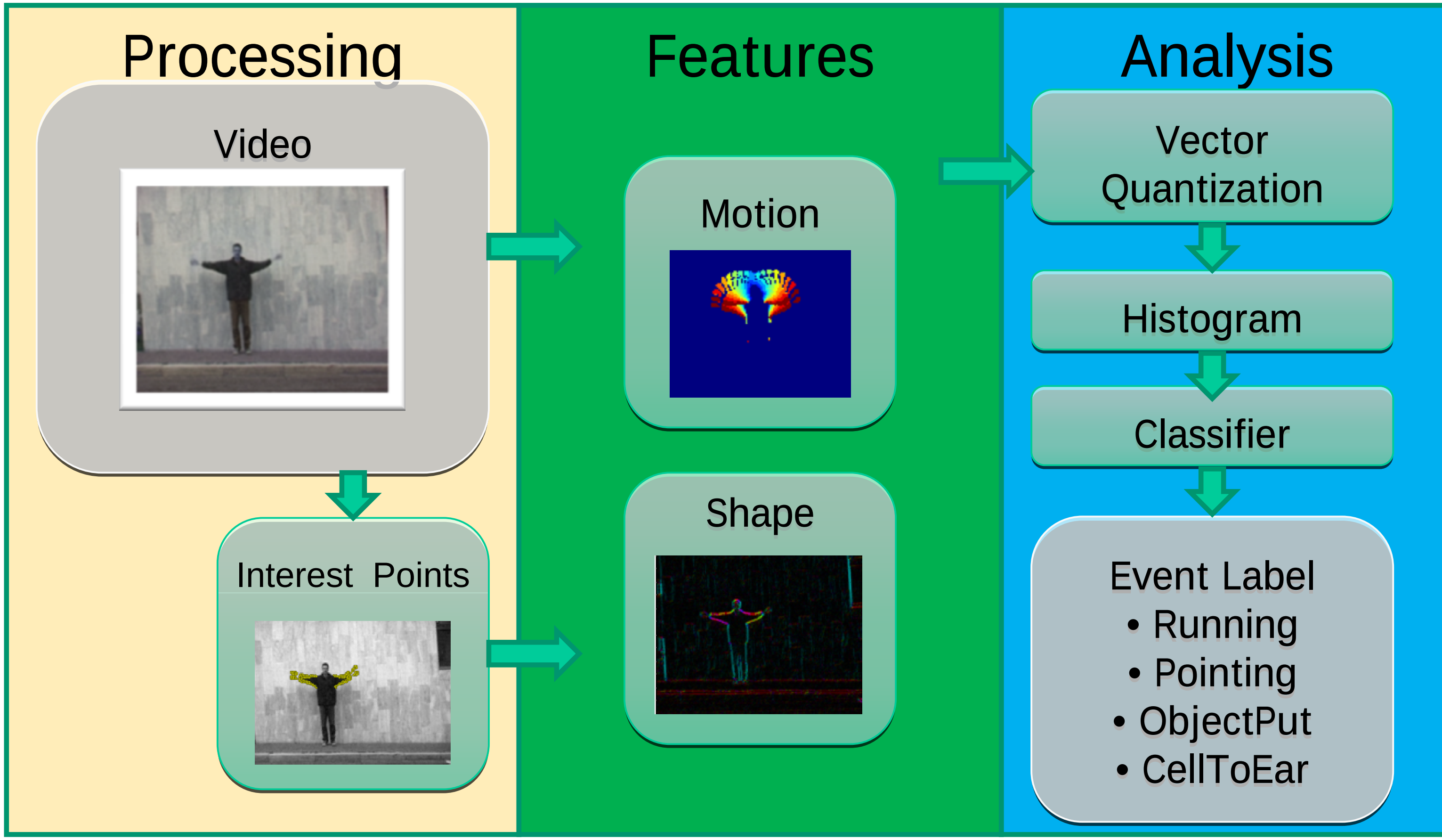
CellToEar	#Ref	#Sys	#CorDet	#FA	#Miss	Act.DCR	Min.DCR
NEC-1	194	35	3	32	191	**0.995**	**0.991**
NEC-2	194	20	1	19	193	1.001	0.998
NEC-3	194	20	1	19	193	1.001	0.998
UIUC-1	194	183	0	58	194	1.019	1.060
ObjectPut	#Ref	#Sys	#CorDet	#FA	#Miss	Act.DCR	Min.DCR
NEC-1	621	10	2	8	619	0.999	0.997
NEC-2	621	11	3	8	618	0.998	0.998
NEC-3	621	5	2	3	619	**0.998**	**0.997**
UIUC-1	621	555	1	190	620	1.061	1.020
Pointing	#Ref	#Sys	#CorDet	#FA	#Miss	Act.DCR	Min.DCR
NEC-1	1063	6	2	4	1061	0.999	0.999
NEC-2	1063	5	2	3	1061	**0.999**	**0.998**
NEC-3	1063	6	2	4	1061	0.999	0.999
UIUC-1	1063	774	13	225	1050	1.062	1.006

Act.DCR: 0.999X (2008) -> 0.99X (2009)

Sample Results

NEC Laboratories America
Relentless passion for innovation

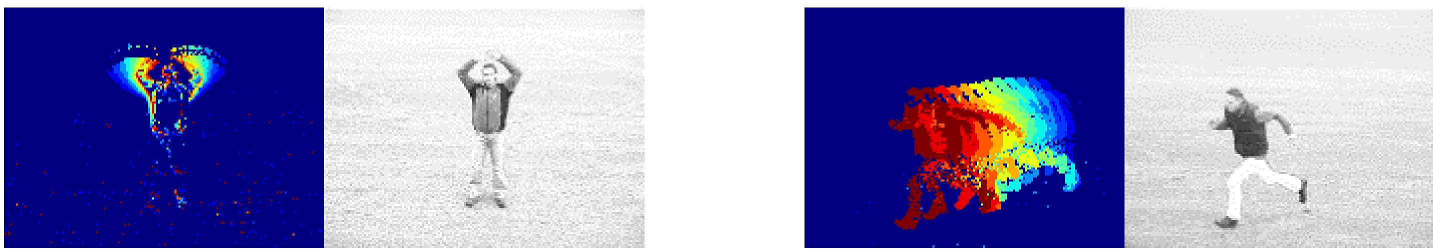
UIUC's System for TRECVid 2009

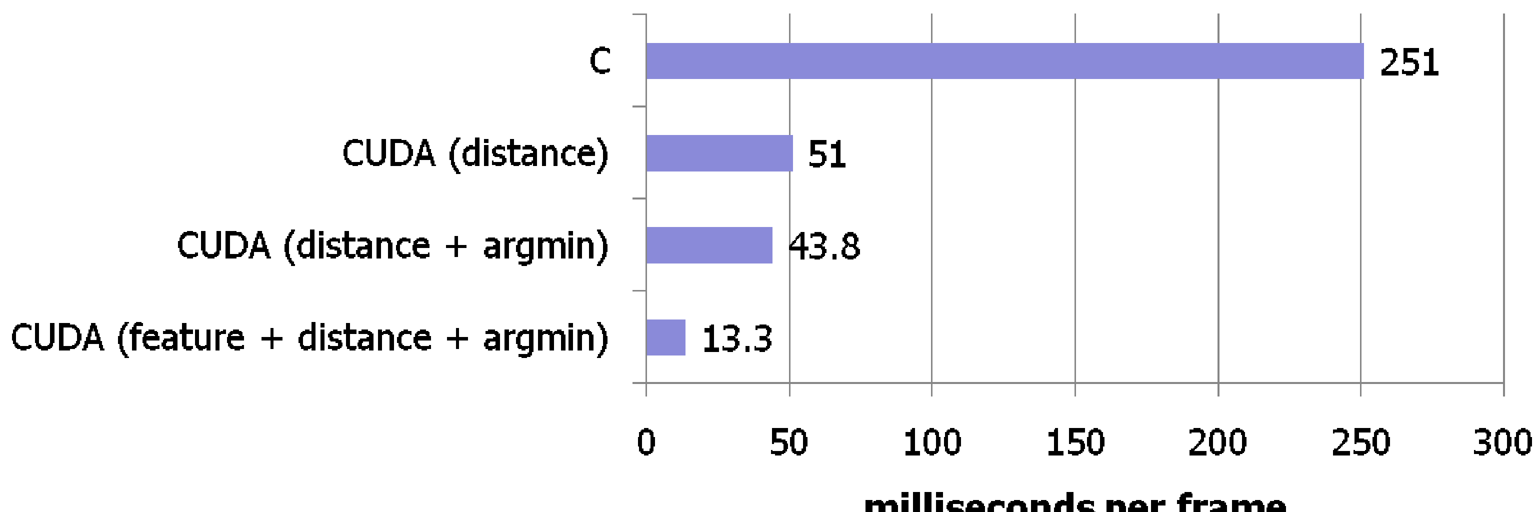


Motion History Images (Bobbick & Davis 2001)

NEC Laboratories America
Relentless passion for innovation

$$H_\tau(x,y,t)=\begin{cases}\tau & \text{if } D(x,y,t)=1\\ \max(0,H_\tau(x,y,t-1)-1) & \text{otherwise}\end{cases}$$





Histograms of
Oriented Gradients
Optical Flow

- Partition the image window into local regions
- Histogram of the {Image Gradient/Optical Flow} based on the direction and magnitude
- Normalize over neighboring regions

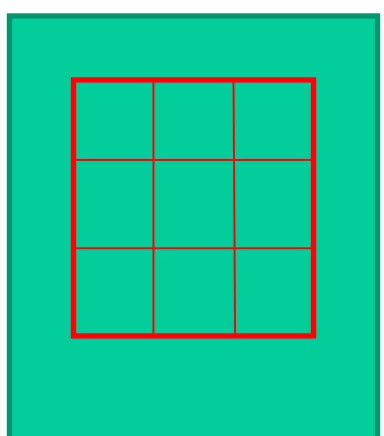
collected from many overlapping

Histograms of
Oriented Gradients
Optical Flow

- Partition the image window into local regions
- Histogram of the {Image Gradient/Optical Flow} based on the direction and magnitude
- Normalize over neighboring regions

Histograms of
Oriented Gradients
Optical Flow

- Partition the image window into local regions
- Histogram of the {Image Gradient/Optical Flow} based on the direction and magnitude
- Normalize over neighboring regions



collected from many overlapping

Results (2009)

	True Positives	False Alarm	Miss	Min DCR
Pointing	13	225	1050	1.006
Cell To Ear	0	58	194	1.060
Person Runs	1	38	106	0.997
Object Put	1	190	620	1.020

Results (2009)

NEC Laboratories America
Relentless passion for innovation

	True Positives	False Alarm	Miss	Min DCR
Pointing	13 (57)	225 (2505)	1050	1.006
Cell To Ear	0 (8)	58 (4005)	194	1.060
Person Runs	1 (0)	38 (314)	106	0.997
Object Put	1 (21)	190 (2703)	620	1.020

(2008 Results)

Video Computer Vision on Graphics Processors -- ViVid

NEC Laboratories America
Relentless passion for innovation

Image / Video Processing	Video Decoder
	2D/3D Convolution
	2D/3D Fourier Transform
	Optical Flow
Feature Extraction	Motion Descriptor (Efros et al.)
	Motion History Descriptor
	Histograms of {Oriented Gradients / Optical Flow}
Analysis	Vector Quantization
	SVM Classifier Evaluation

Download:
http://libvivid.sourceforge.net

Conclusions

NEC Laboratories America
Relentless passion for innovation

- A long way to go for human action detection in real-world conditions!
- A fruitful journey!
  - A new multiple human tracking algorithm
  - A new SVM learning algorithm for large scale datasets
  - Parallel processing on graphics processors
  - Evaluation of different action representations
- Thank you!