Large volumes of video

flickr

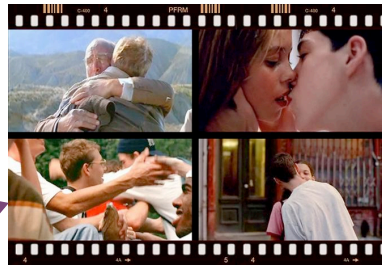
YAHOO!
SCREEN

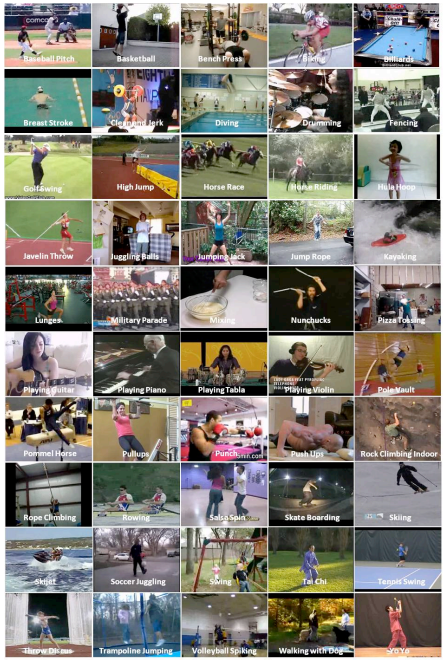
- For YouTube alone
  - More than 1 billion unique user visits each month
  - Over 6 billion hours of video are watched each month
  - 100 hours of video are uploaded every minute

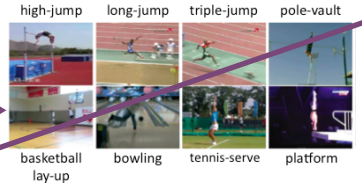
http://www.youtube.com/yt/press/statistics.html

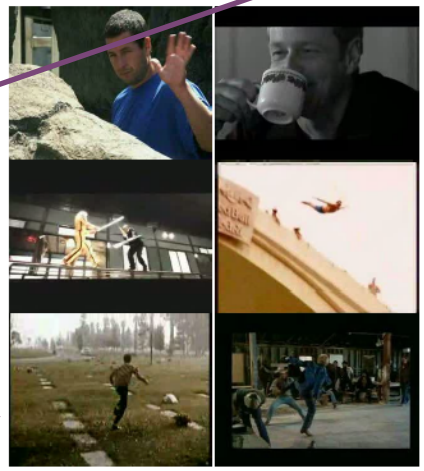
Available Datasets

Dataset	#Classes
KTH	6
Weizmann	9
IXMAS	11
Hollywood	8
UCF Sports	9
Hollywood2	12
UCF YouTube	11
MSR	3
Olympic	16
UCF50	50
HMDB51	51









http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

Videos in the wild

- Unrestricted type of events with various activities

Harlem Shake : http://www.youtube.com/watch?v=4hpEnLtqUDg

Our attempts

- Videos as sequence of frames
  - Detect concepts in each frame
  - Utilize image search engines

- Discover important knowledge from videos itself
  - Discriminate parts

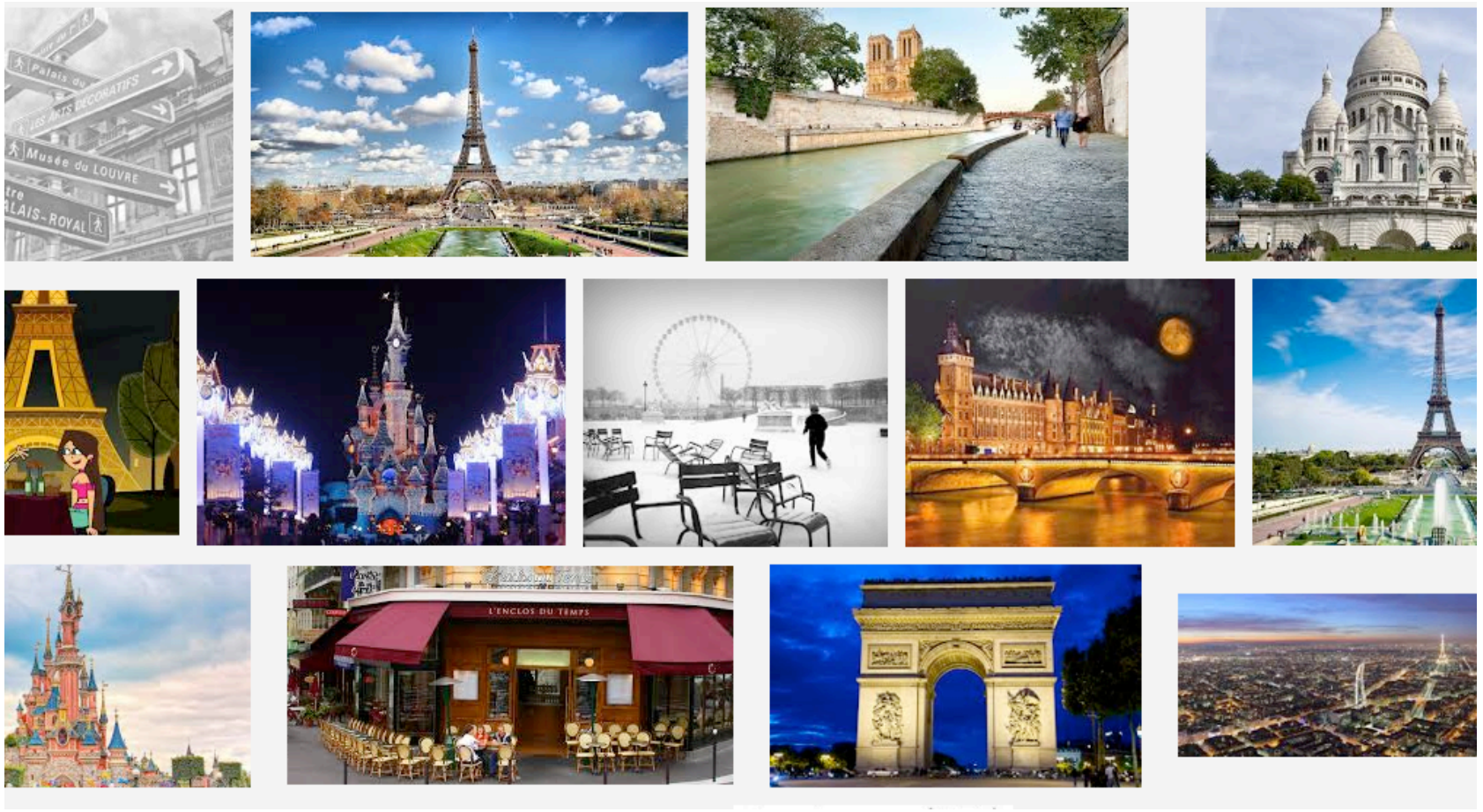
- Understand actions in videos
  - Simple but effective descriptors

Utilizing large volumes of weakly labeled images

Pinar Duygulu, January 13, 2014, CMU

Utilize image search results

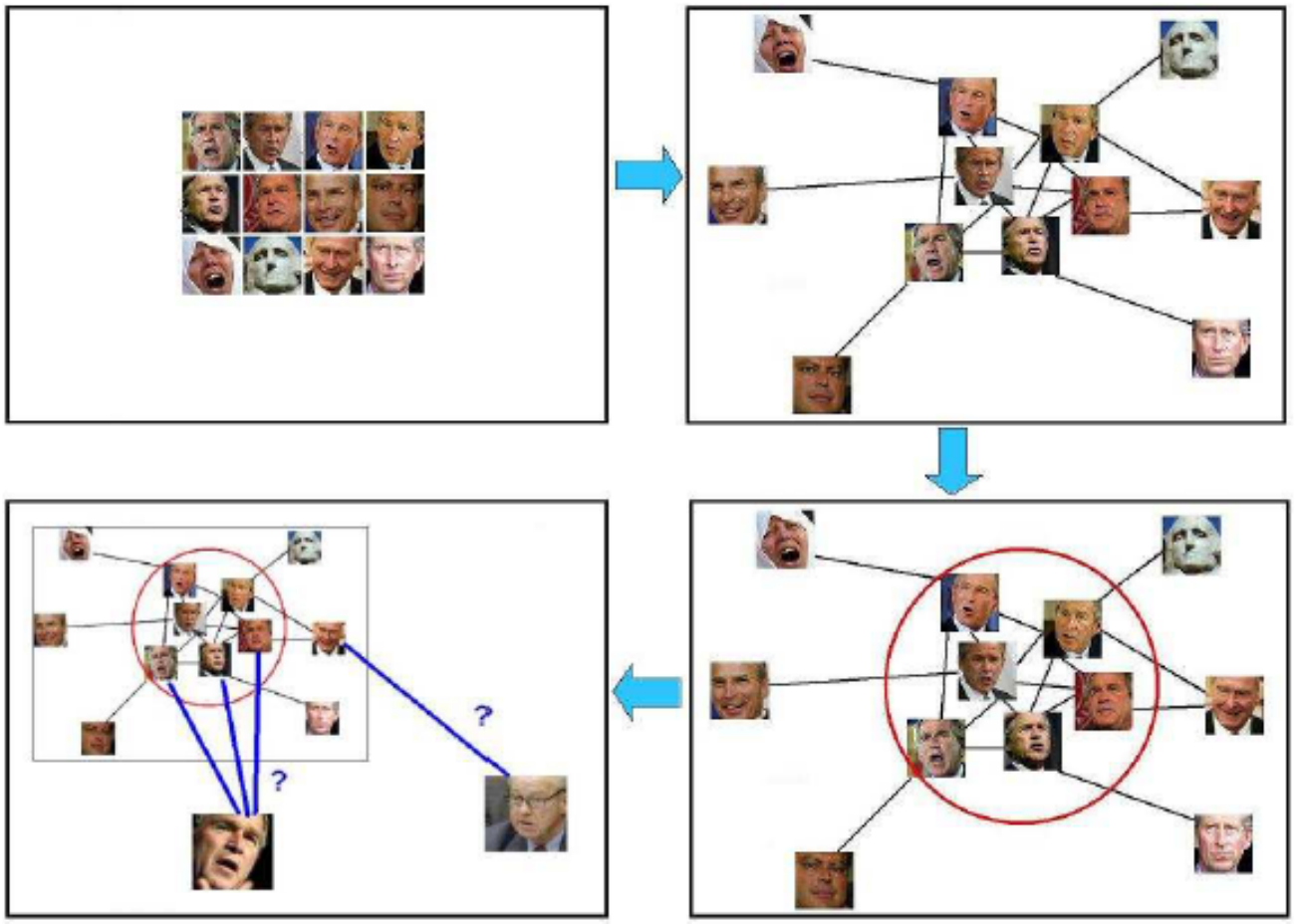
Query : Paris



Single Dominant Category

Query : George W. Bush

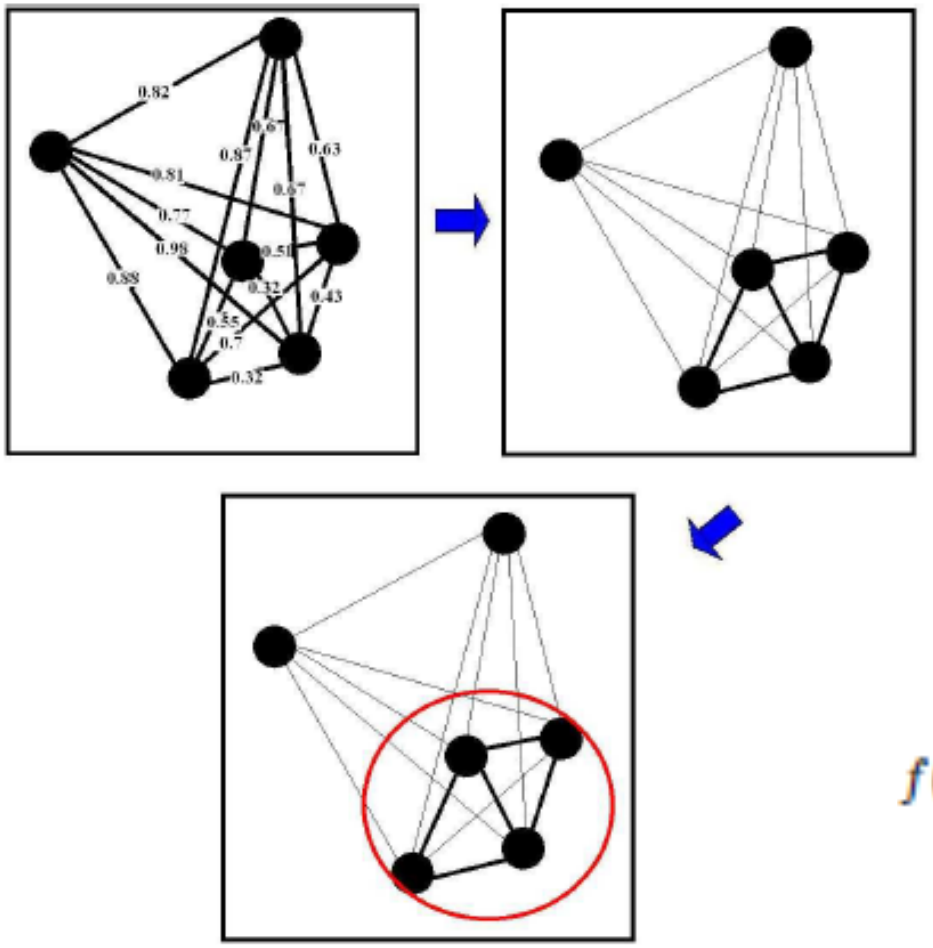
Naming faces



Among the faces associated with a name find the correct subset : The most similar subset of faces

Ozkan, D., Duygulu, P., "Interesting Faces: A Graph Based Approach for Finding People in News", Pattern Recognition, 2010
Ozkan, D., Duygulu, P., "A Graph Based Approach for Naming Faces in News Photos", CVPR, 2006
Ozkan, D., Duygulu, P., "Finding People Frequently Appearing in News", CIVR, 2006

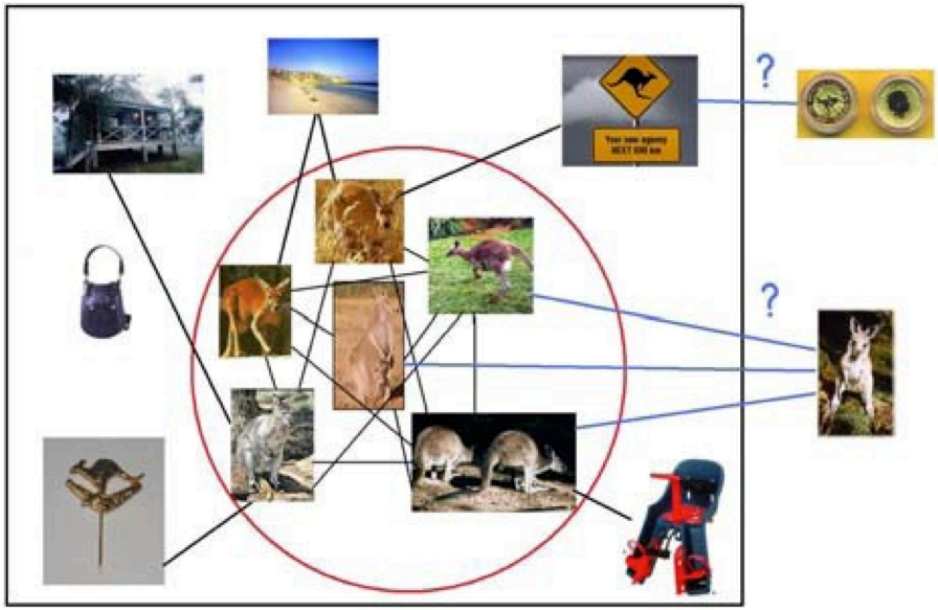
Finding Densest component



$$f(S) = \frac{|E(S)|}{|S|},$$

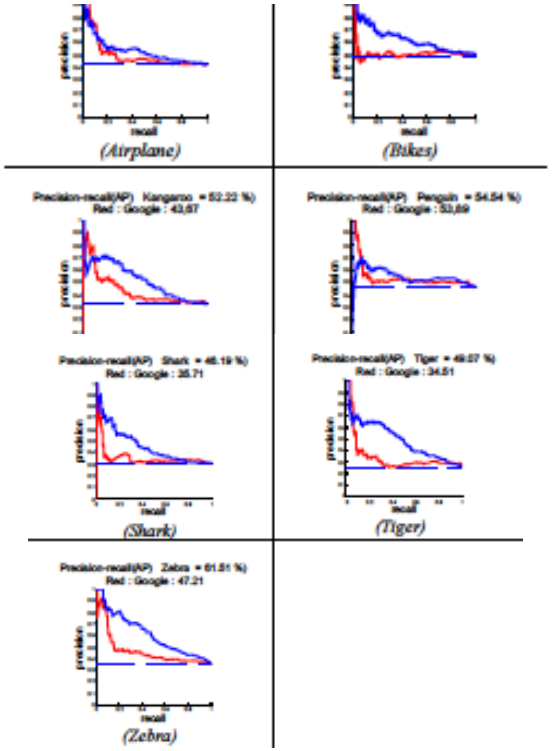
Node with the minimal degree is removed at each iteration (Charikar, 2000)

Image Re-ranking

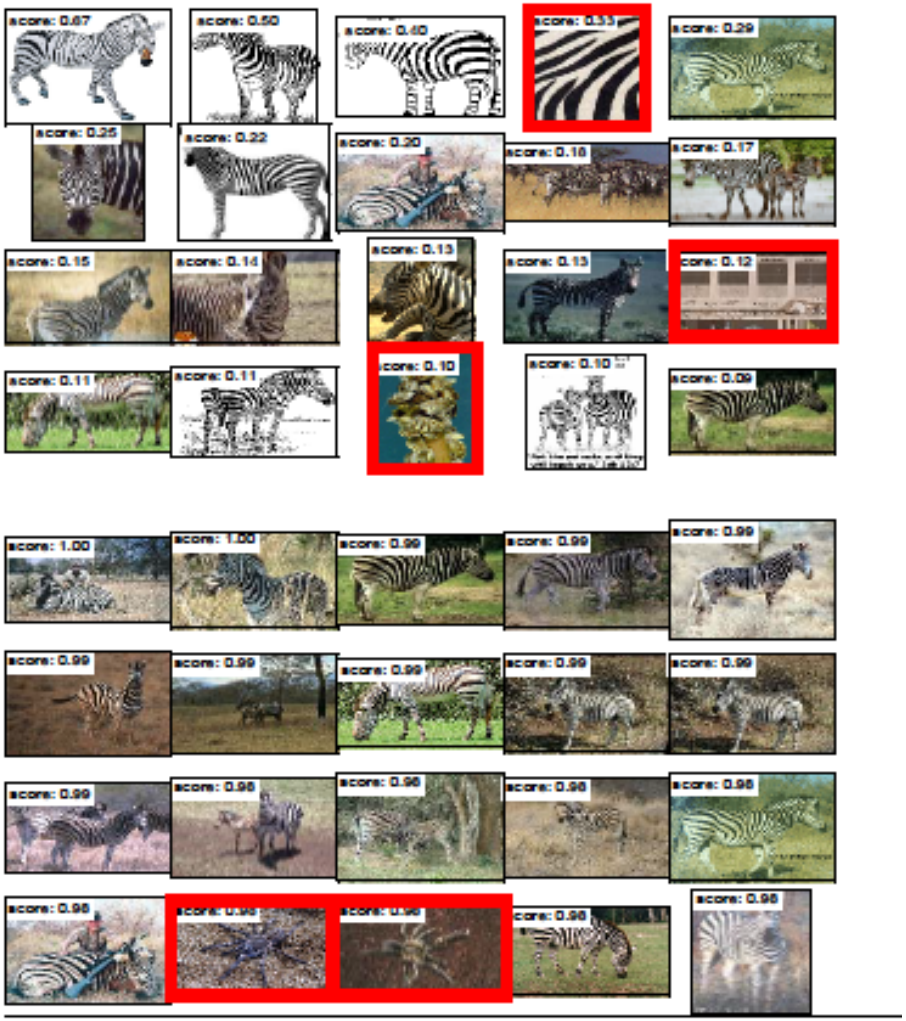


Zitouni, H., Sevil, S. G., Ozkan, D., Duygulu, P., "Re-ranking of Image Search Results using a Graph Algorithm", ICPR 2008

Multiple Instance Learning for re-ranking



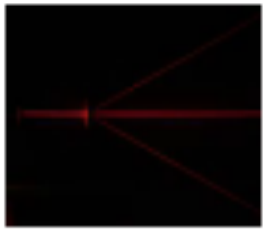
On the dataset by Schroff, F., ICCV 2007 "Harvesting Image Databases from the Web".



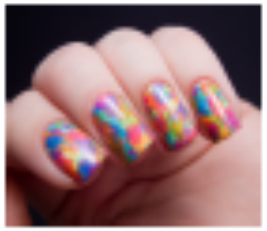
Sener, F., Ikizler-Cinbis, N., Duygulu, P., "Multiple Instance Learning for re-ranking of Web image search results", SIU 2012

Multiple meanings/variations





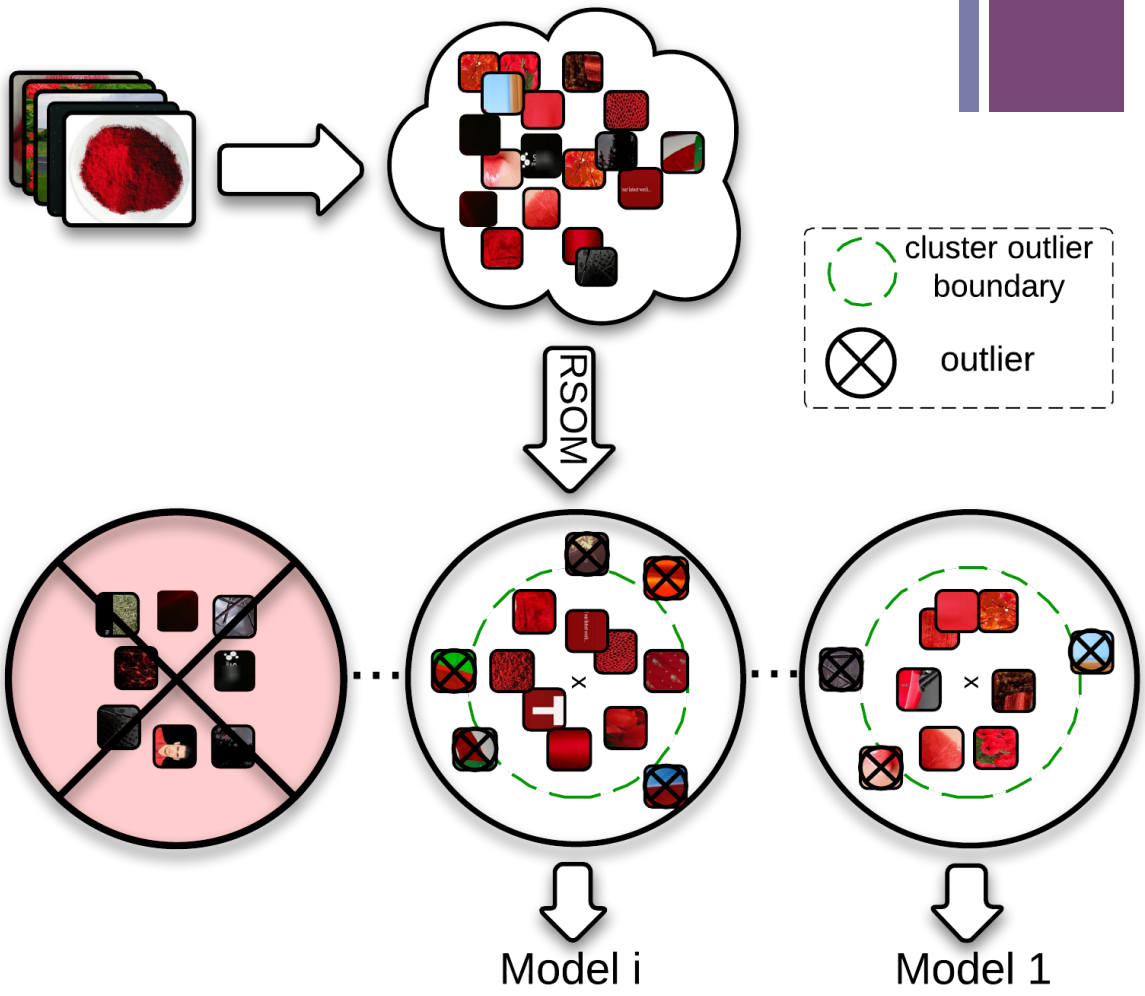



the attributes are observed in different forms and in small portions requiring grouping and non-attribute parts to be eliminated.

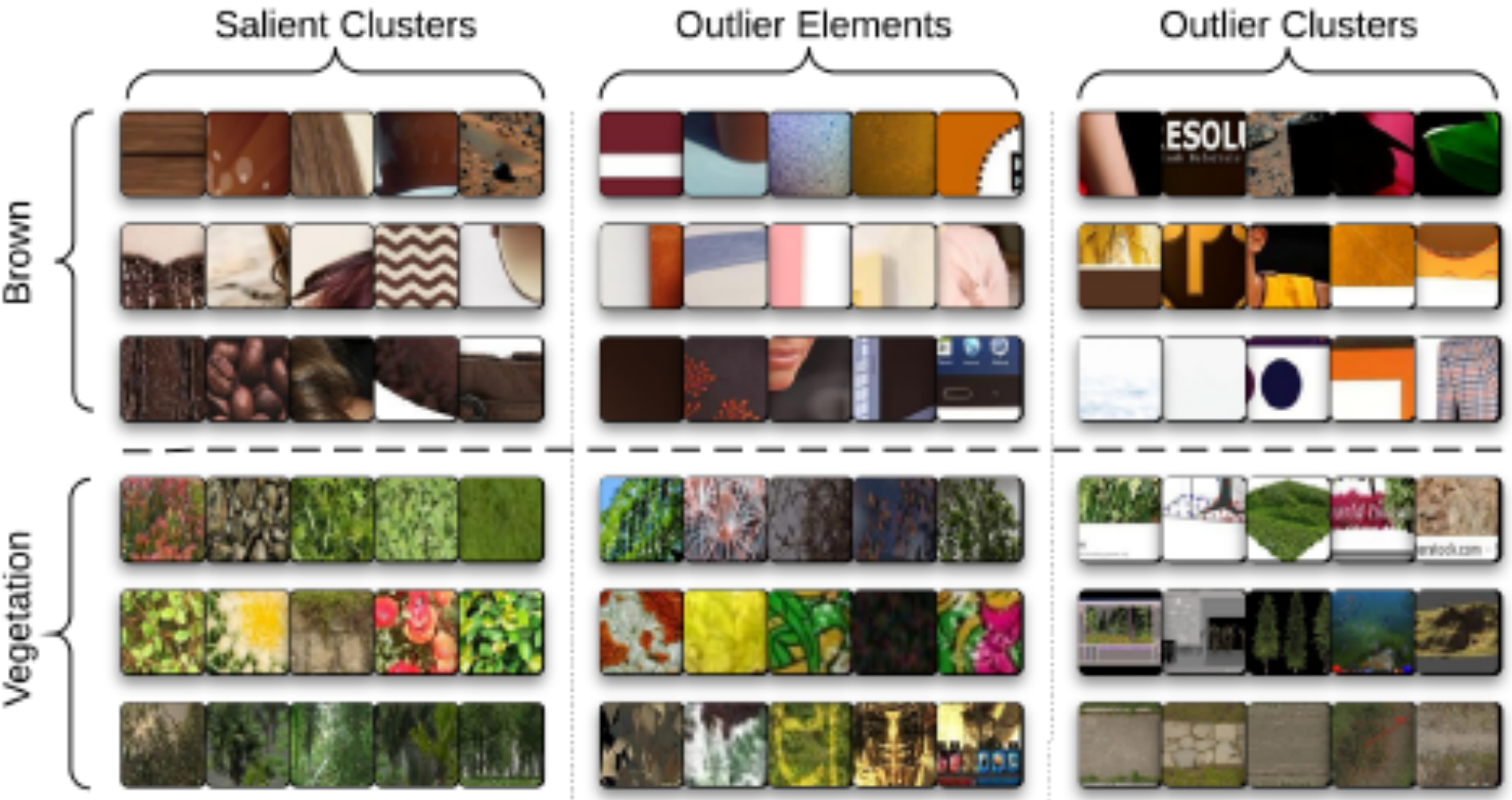
RSOM for Concept Learning

- Collect images from web for a keyword
- Clustering and outlier detection
- RSOM (Rectifying Self Organizing Maps)
- Learn a model for each cluster

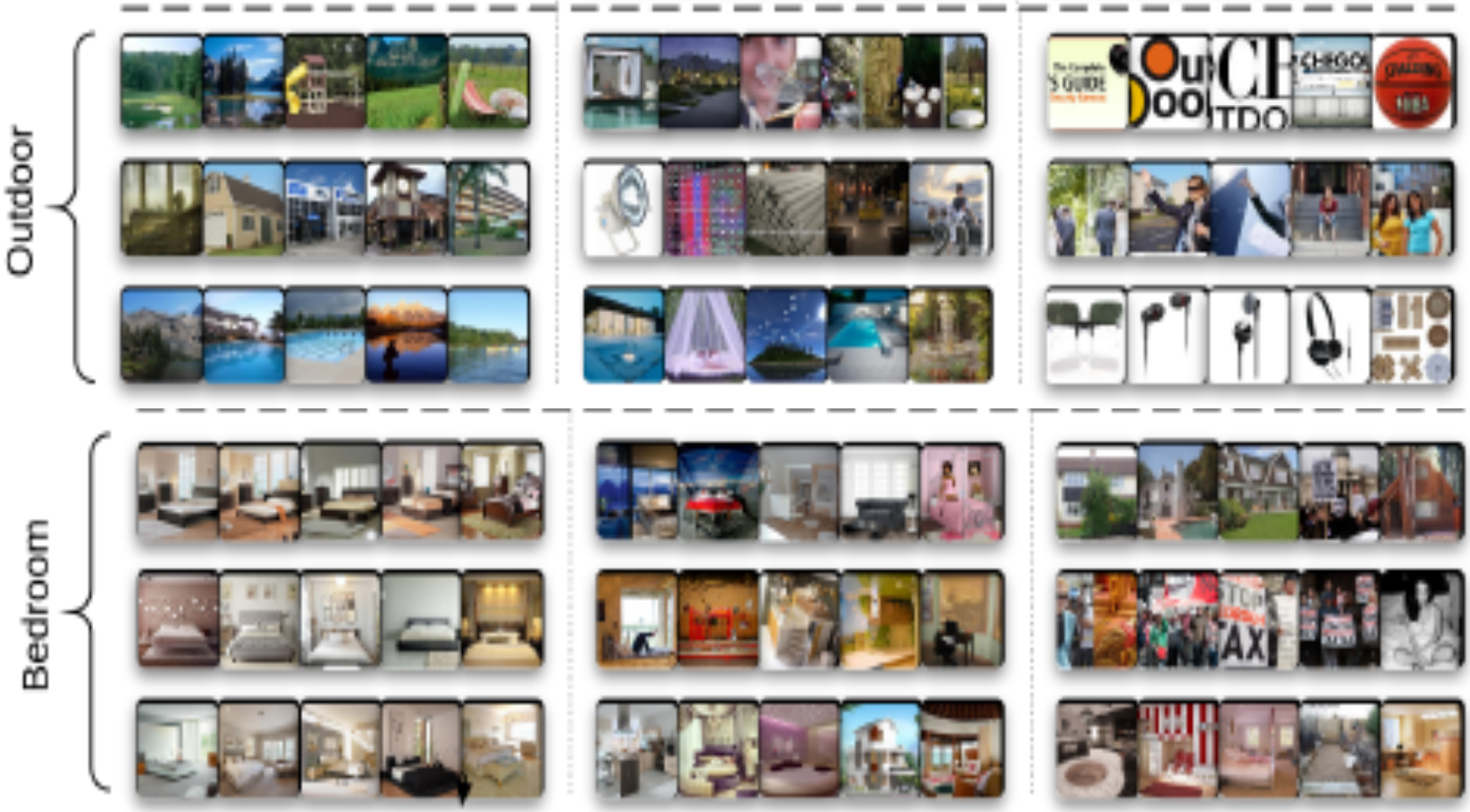


Golge, E., Duygulu, P., "Rectifying Self Organizing Maps for Automatic Concept Learning from Web Images ", submitted to CVPR 2014

Color and Texture Attributes



Scene Concepts



Attribute and Scene Learning

Attribute learning

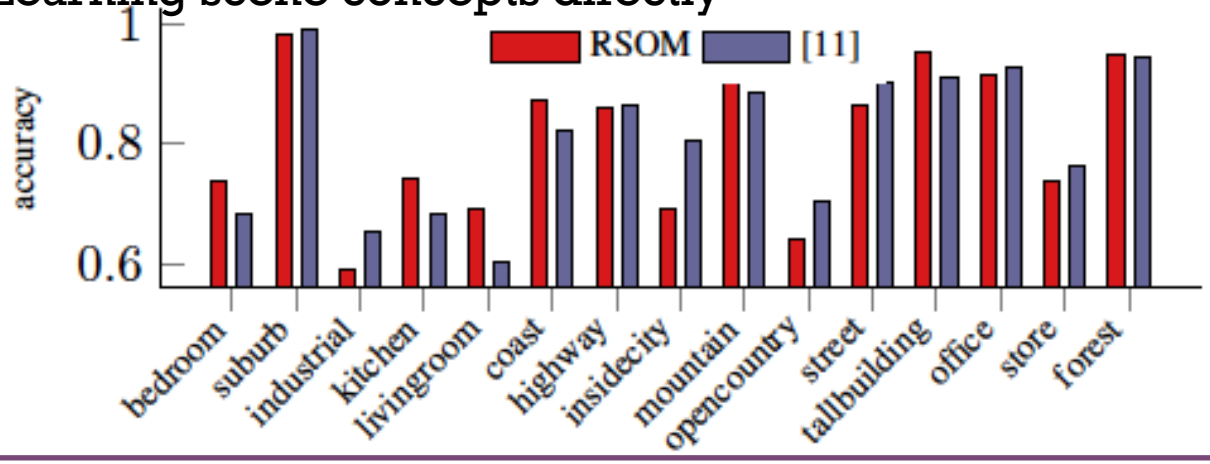
Method	RSOM-M	RSOM	PLSA-reg [22].
cars	0.97	0.92	0.93
shoes	1.0	0.97	0.99
dresses	1.0	1.0	0.99
pottery	0.98	0.92	0.94
overall	0.99	0.95	0.96

Attribute based scene recognition

Method	MIT-indoor [17]	Scene-15 [11]
RSOM-A	46.2%	82.7%
RSOM-S	-	80.7%
RSOM-S+HM	-	81.3%
Li *et al.* [12] VQ	47.6%	82.1%
Pandey *et al.* [16]	43.1%	-
Kwitt *et al.* [9]	44%	82.3%

On ImageNet: 37.4% (RSOM), 36.8% (Russakovsky & Fei-Fei, 2012)

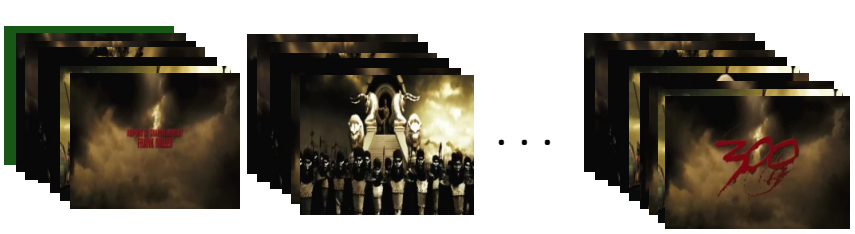
Learning scene concepts directly



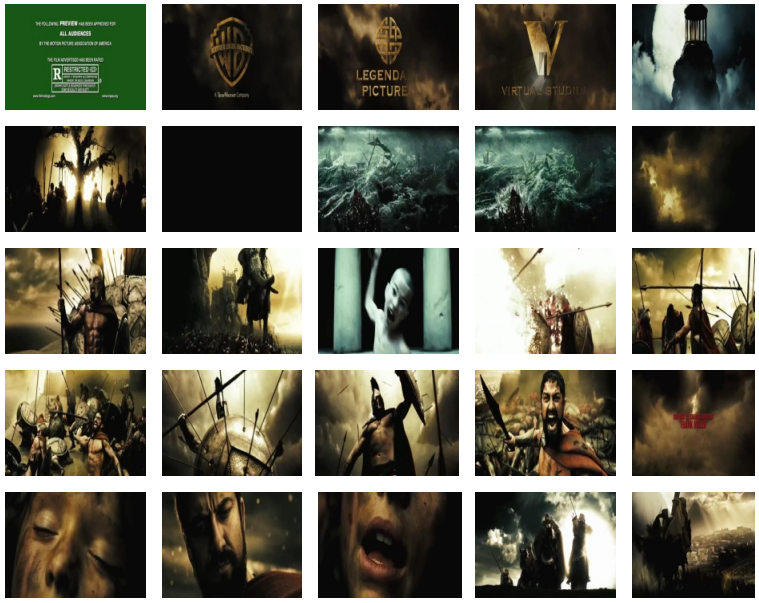
[17] Quattoni and Torralba,"Recognizing Indoor Scenes". 2009
[11] Lazebnik, Schmid, Ponce, "Beyond Bags of features: Spatial pyramid matching for recognizing natural scene categories", CVPR 2006
[22] Van de Weijer, Schmid, Verbeek, Larlus, "Learning Color Names for Real-world Applications", 2009

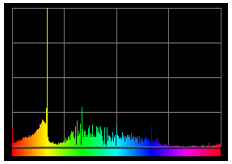
Utilizing videos

Movie Genre & MPAA Ratings

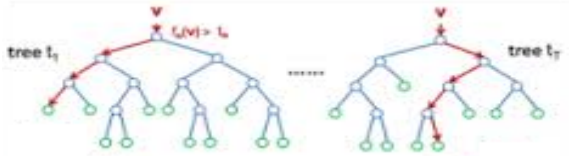


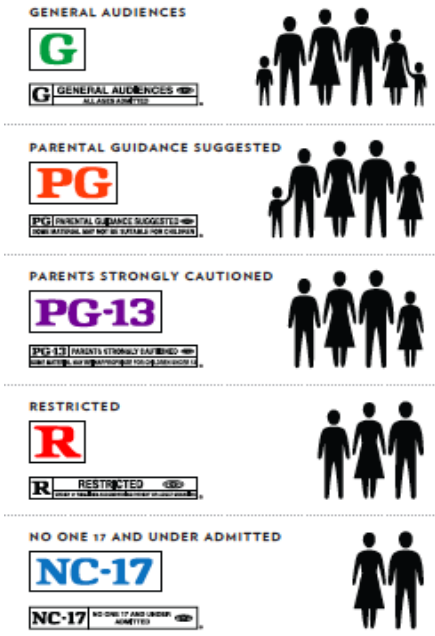
300 Spartans – Movie Trailer
(represented as frames)





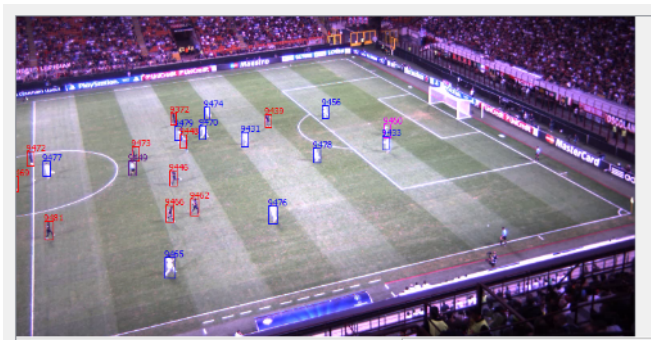
- Statistical features :
  - Key frame count, Black and White frame counts, Avg. in shot frame difference, Total number of frames, Avg. shot length, Avg. key frame HSV color histogram, Avg. key frame edge histogram
- Visual features:
  - SIFT, LAB histogram, Local Binary Patterns, GIST

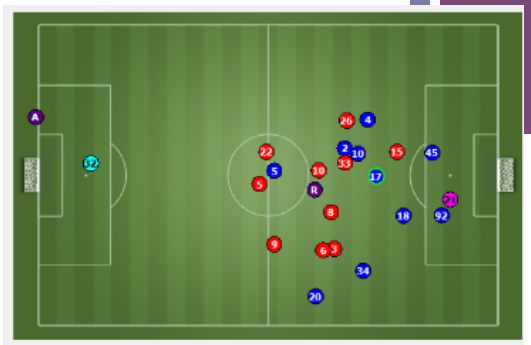




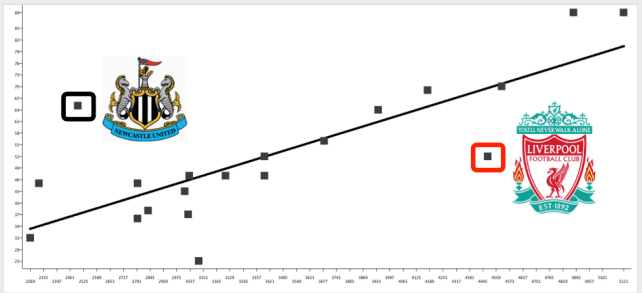
Genre: action, horror, animation, comedy, drama

Video Analysis for Sports

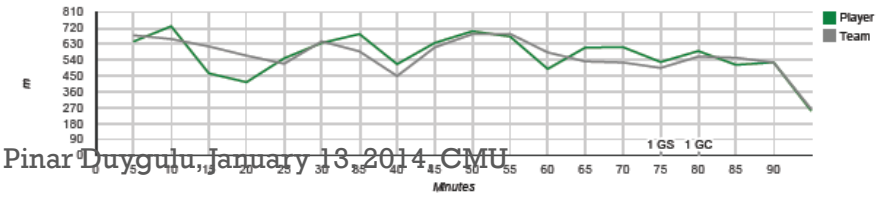


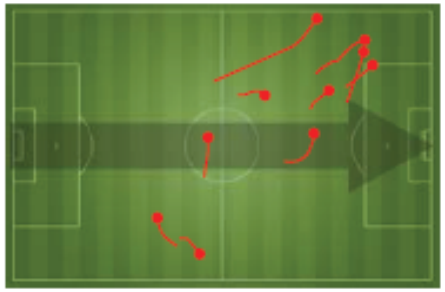


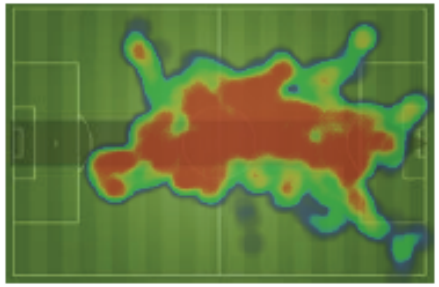
- Track soccer players in real-time in all matches
  - Extract XY coordinates and event data
- Using the big data
  - Player/team performance analysis/comparison
  - Extract game features that is correlated with seasonal success
  - Performance indexing and market value evaluation
  - Fatigue and injury prediction



Successful Passes in the Final 3rd







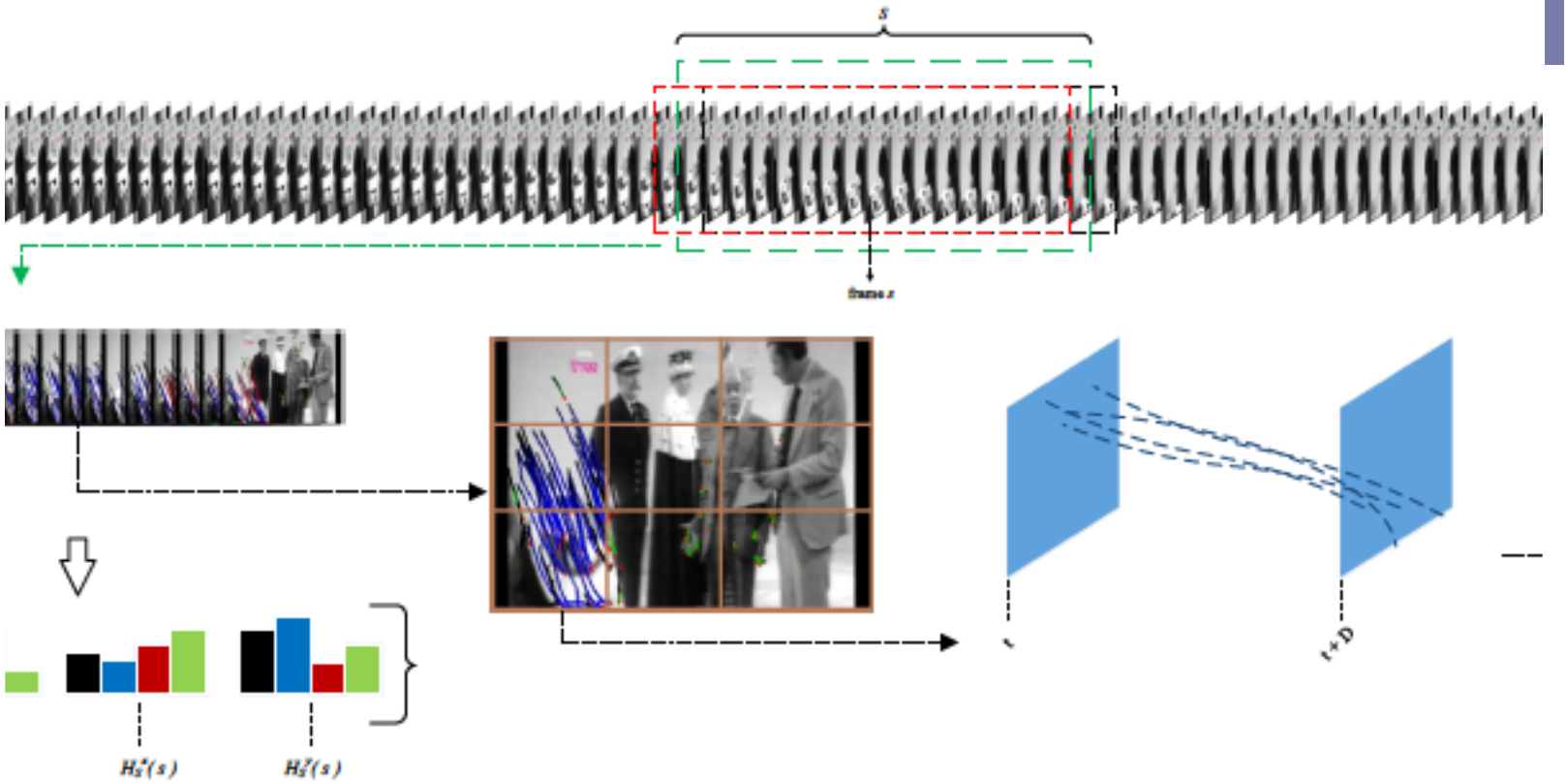
Utilize video data: Capture usualness in unusual videos

Usual versus unusual



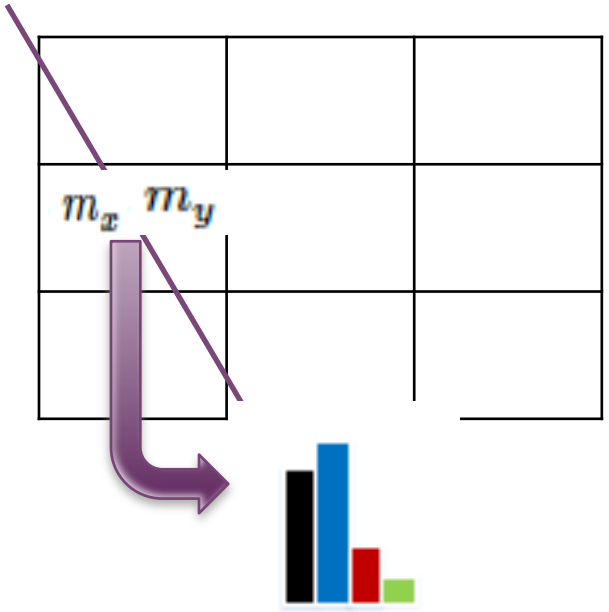



Trajectory Snippet Histograms



Iscen, A., Armagan, A., Duygulu, P., "What is usual in unusual videos? Trajectory snippet histograms for discovering unusualness", submitted to CVPR 2014

Representation



$$H_S^l = \sum_{t=s-(\|S\|/2)}^{s+(\|S\|/2)} H_S^l(t) \qquad H_S = (H_S^l, H_S^x, H_S^y)$$

$$H_S^l(t) = (H_S^l(t)_{[1,1]}, \ldots H_S^l(t)_{[1,N]}, \ldots H_S^l(t)_{[N,N]})$$

Velocity and spatial extension of the motion

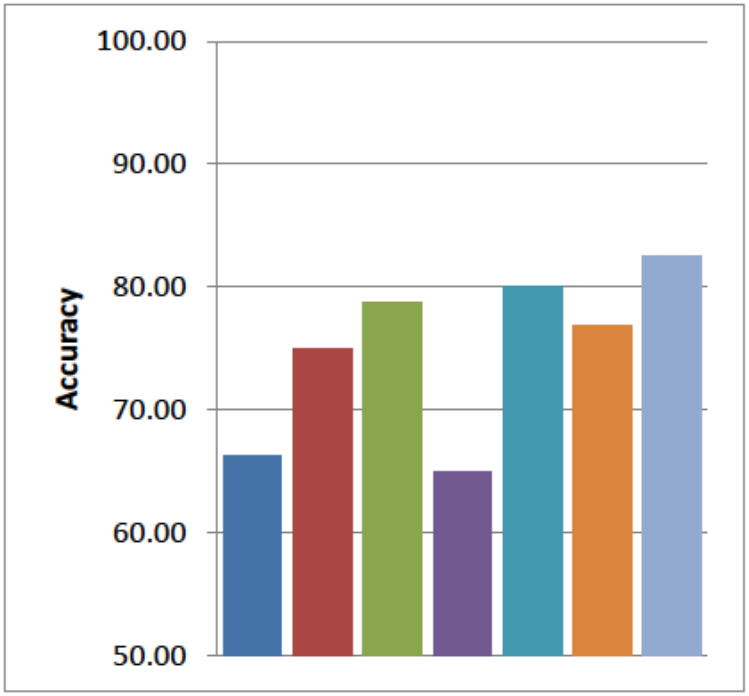
$$T = (P_t, ..., P_{t+D-1}) \quad P_t = (x_t, y_t)$$

$$m_x = \frac{1}{D}\sum_t^{t+D-1} x_t, v_x = \frac{1}{D}\sum_t^{t+D-1} (x_t - m_x)^2$$

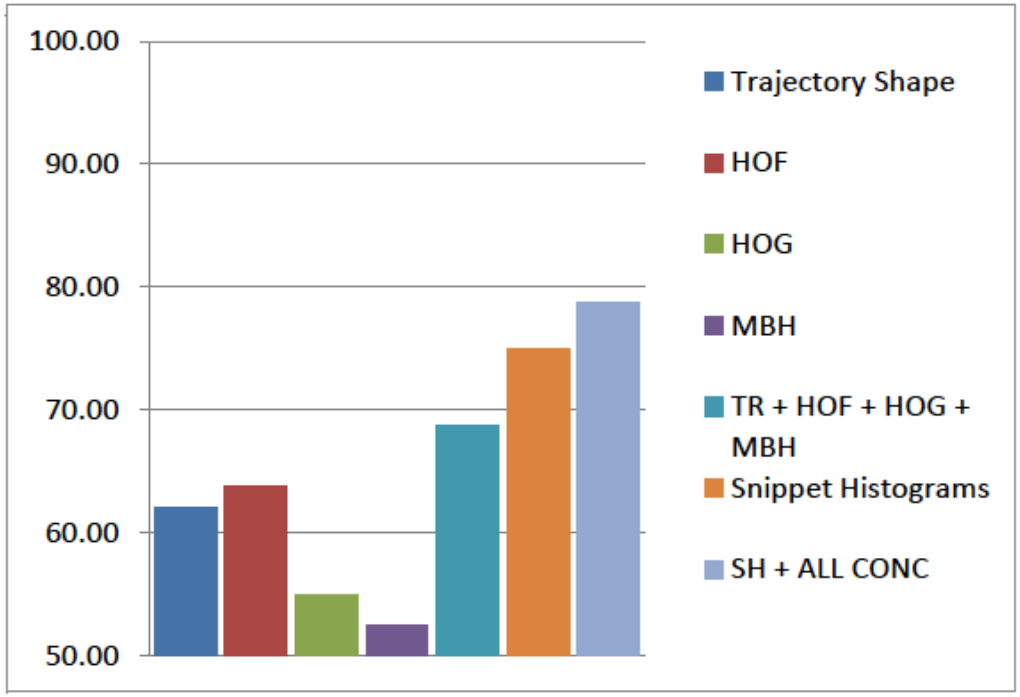
$$m_y = \frac{1}{D}\sum_t^{t+D-1} y_t, v_y = \frac{1}{D}\sum_t^{t+D-1} (y_t - m_y)^2,$$

$$l = \sum_t^{t+D-1} \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2}$$

Classification



People Falling

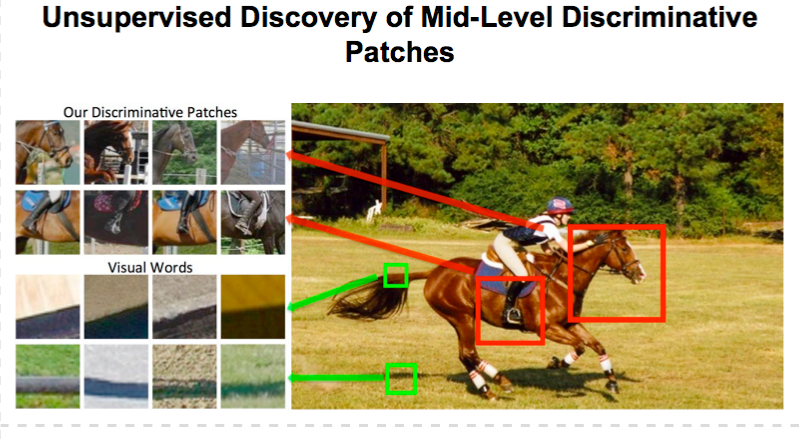


Funny videos

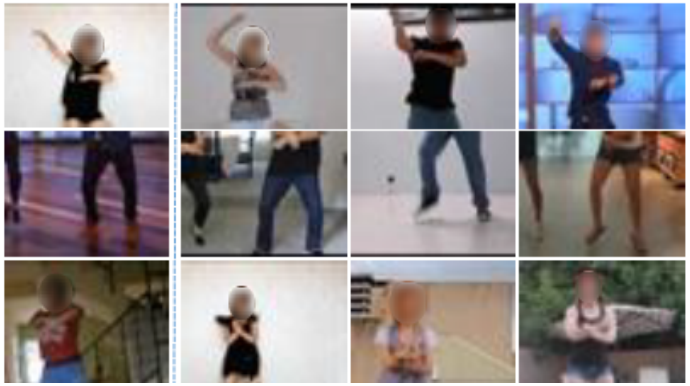
Snapshot Discovery

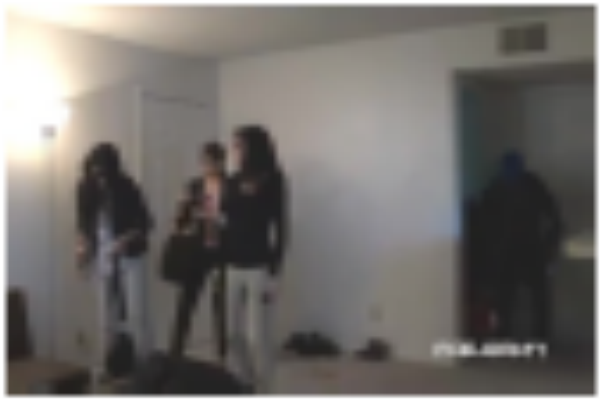
- Discriminative video patch idea over snippets (short video sequences)

Singh ECCV 2012



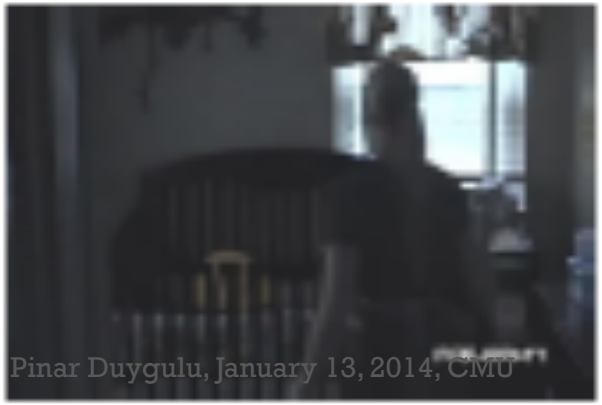
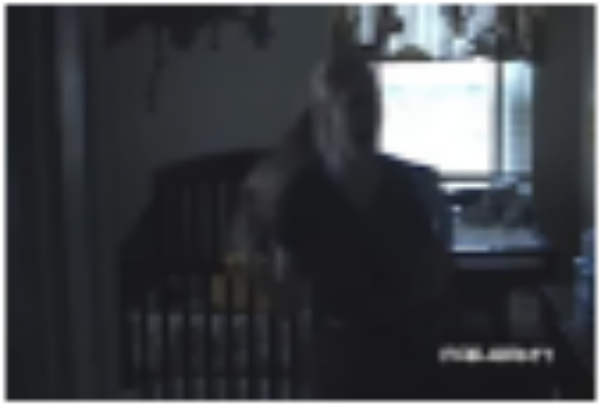
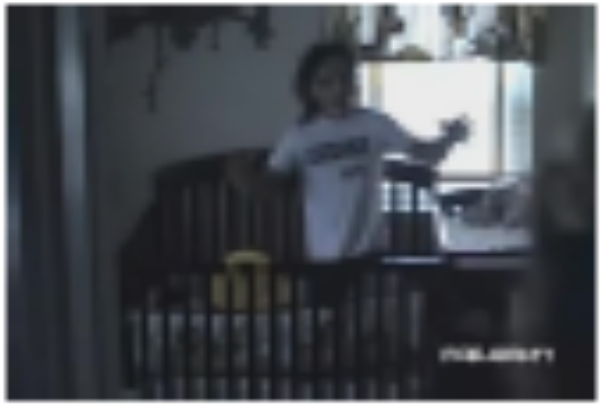
Jain CVPR 2013

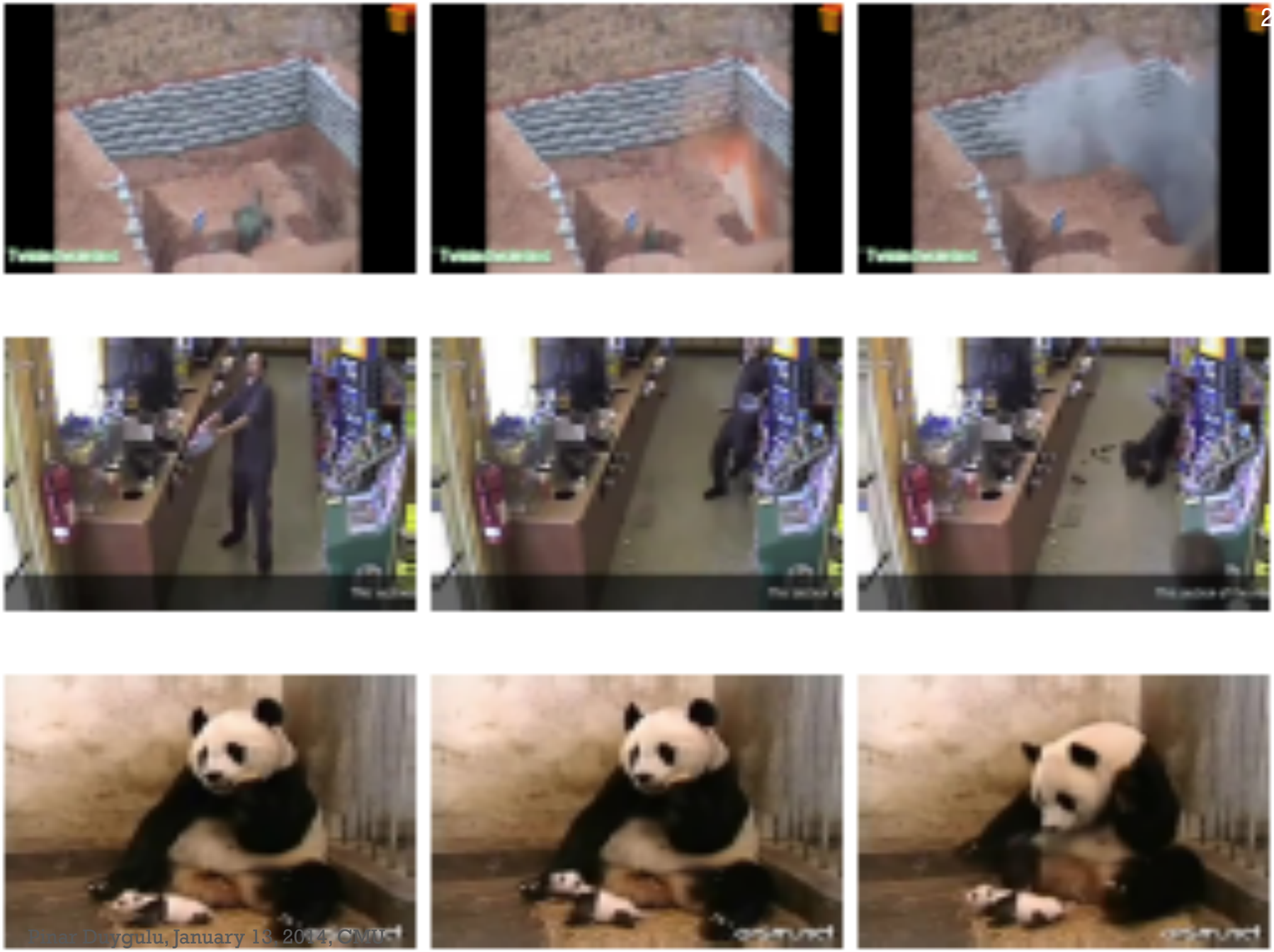




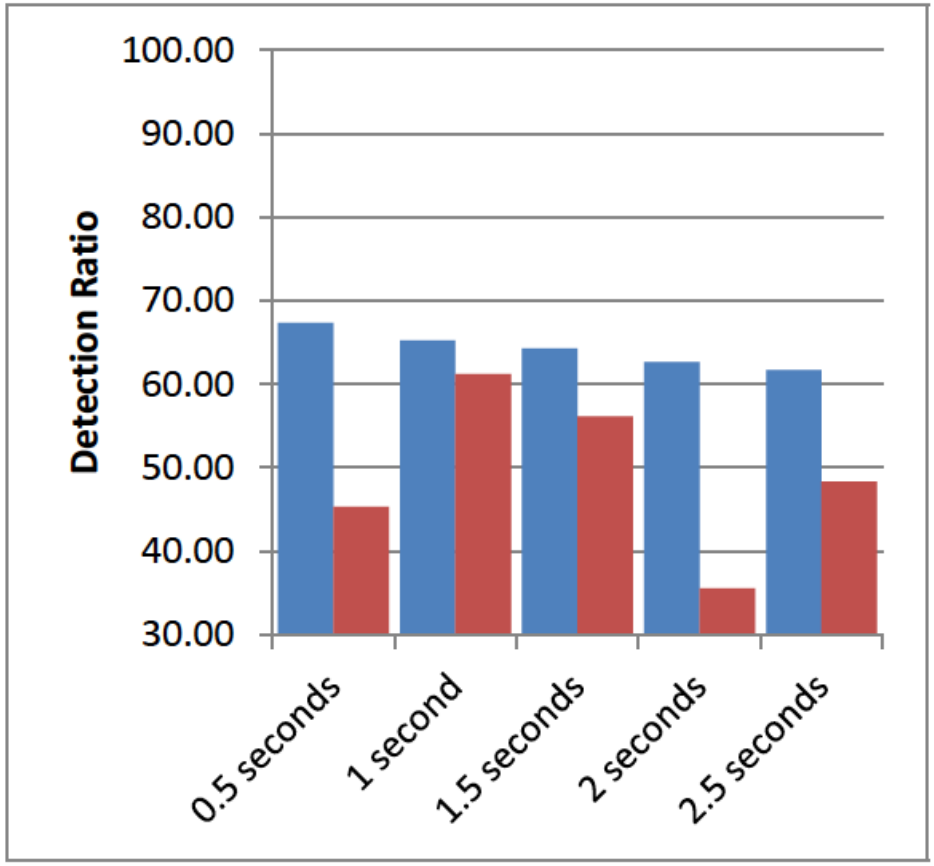




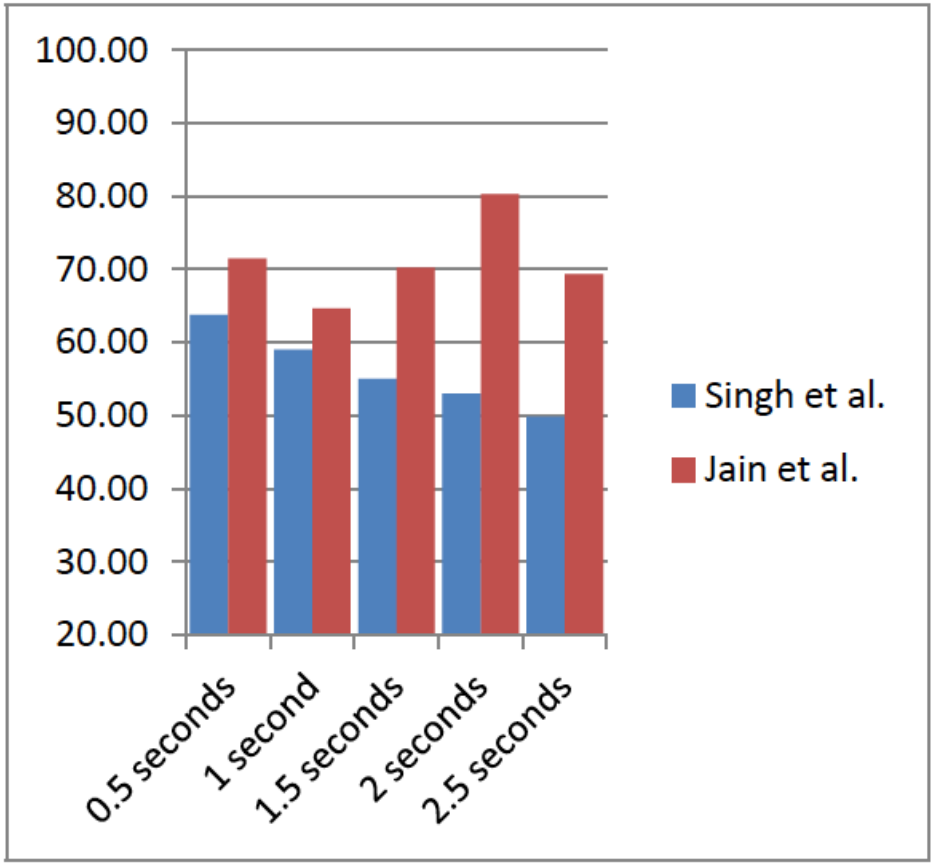






Snapshot discovery



People Falling

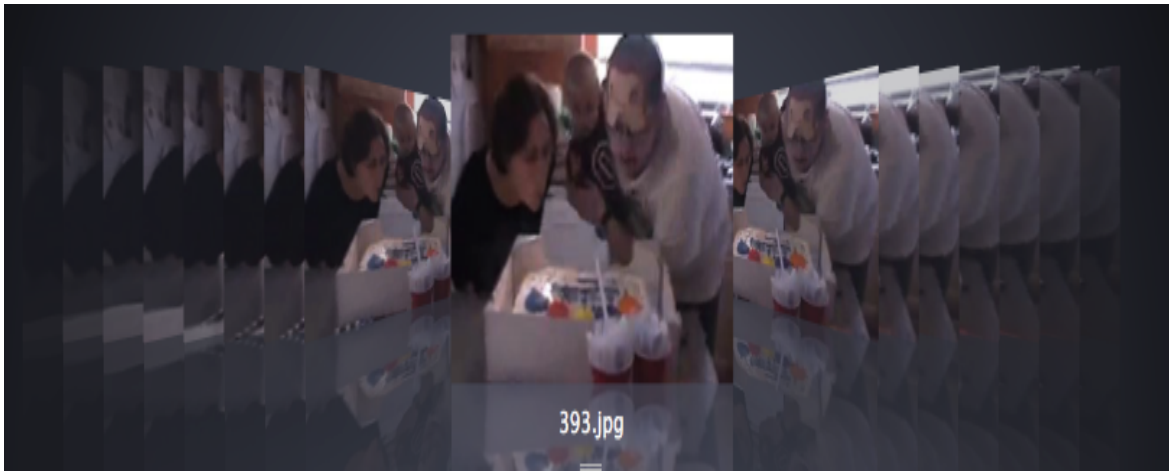


Funny videos

HOG3D --- people falling 25 %, funny videos 32 %

Birthday event

Flowing candles



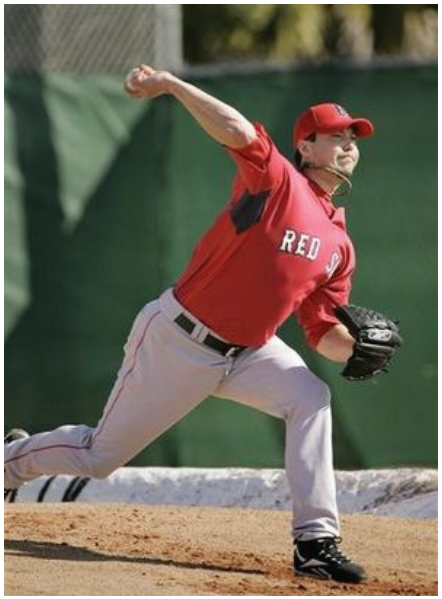
Human Activity Analysis

Pinar Duygulu, January 13, 2014, CMU

What do these people do?

running

walking



throwing

crouching

- Pose tells a lot about the actions.
- How can we describe the pose?

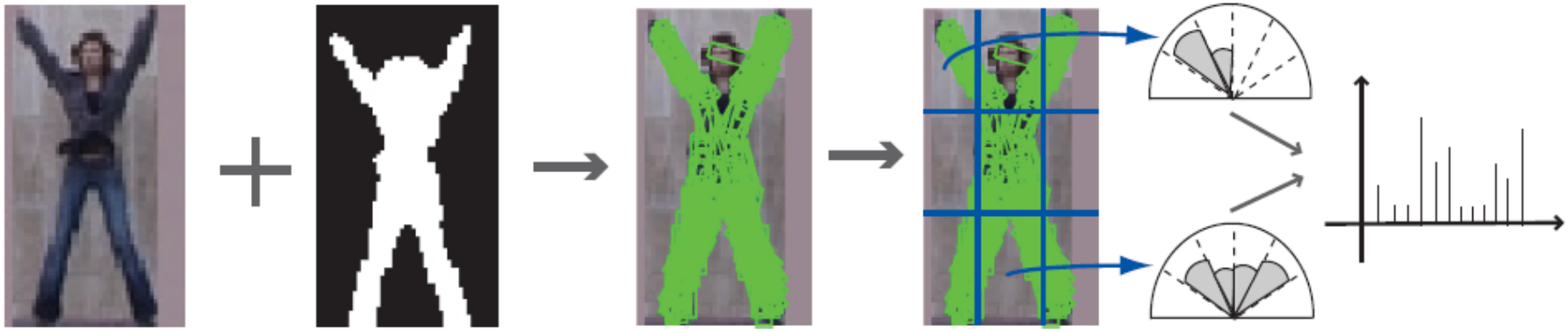
Pose as a Collection of Rectangles

- Human body is composed of cylindrical parts.
- The projection of a cylinder on 2D is a rectangle.
- Body can be thought as a collection of rectangular regions
- We can represent the pose based on the orientation of these rectangles

Ikizler, N. Duygulu, P. ''Human Action Recognition Using Distribution of Oriented Rectanguar Patches'', Proc. 2nd Workshop on Human Motion: Understanding, Modeling, Capture and Animation, In conjunction with ICCV2007

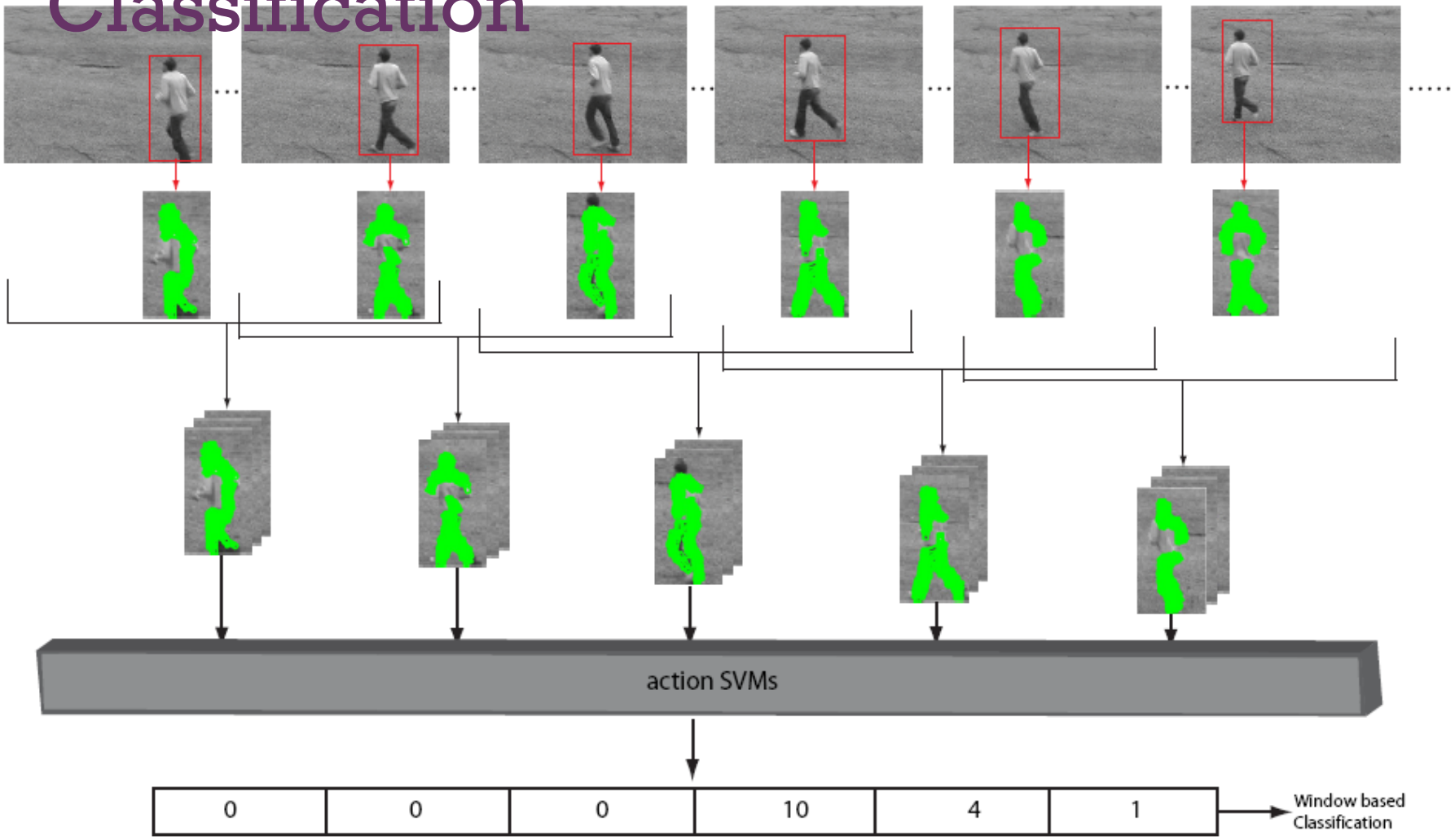
Ikizler, N. ve Duygulu P., ''Histogram of Oriented Rectangles: A New Pose descriptor for Human Action Recognition'', Image and Vision Computing, volume 27, Issue 10, pages 1515-1526, September 2009

Histogram of Oriented Rectangles (HOR)



- Rectangular regions are extracted over silhouettes using convolution of a zero-padded rectangular 2D Gaussian on different orientations and scales
  - 12 angles 15° apart

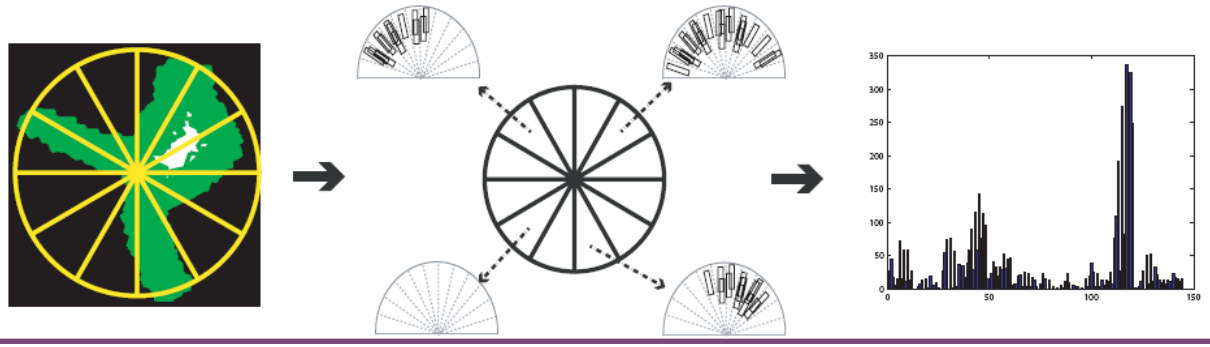
Classification



- Use snippets of frames and form histogram of oriented rectangles over a window (HORW)

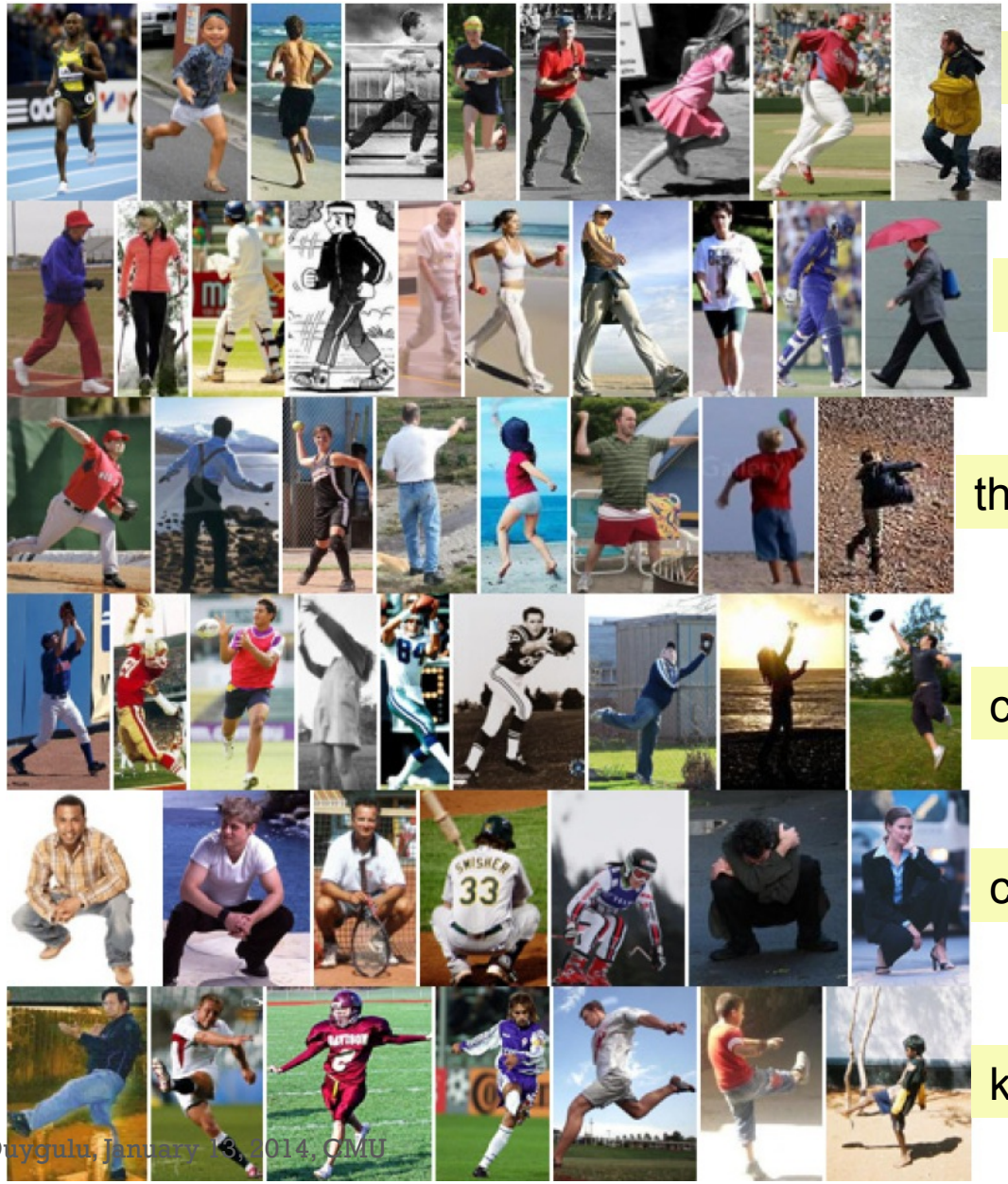
Action Recognition in Still Images

- Pose estimation by Ramanan's method based on CRFs.
- Form Circular HORs (CHORs)
- Classification based on LDA+SVM



Ikizler, N., Cinbis, R. G., Pehlivan, S., Duygulu, P., "Recognizing actions from still images",Proc. 19th International Conference on Pattern Recognition (ICPR 2008)

+ Still Image Results



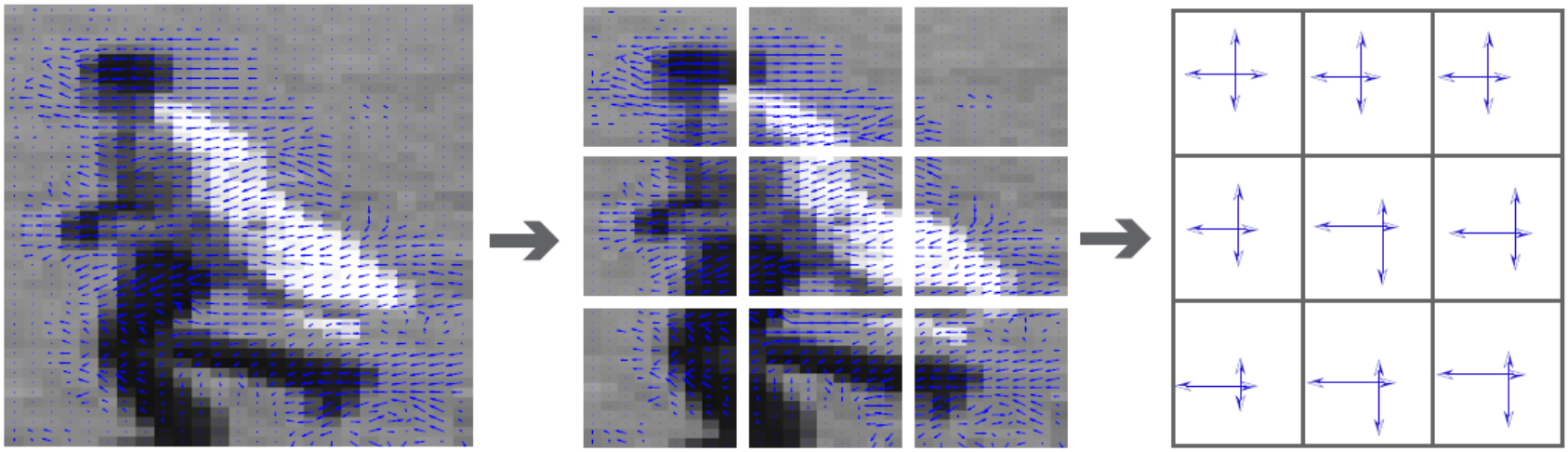
ActionWeb dataset - 467 images collected from the web

Correctly classified action images

Boundary-fitted Lines

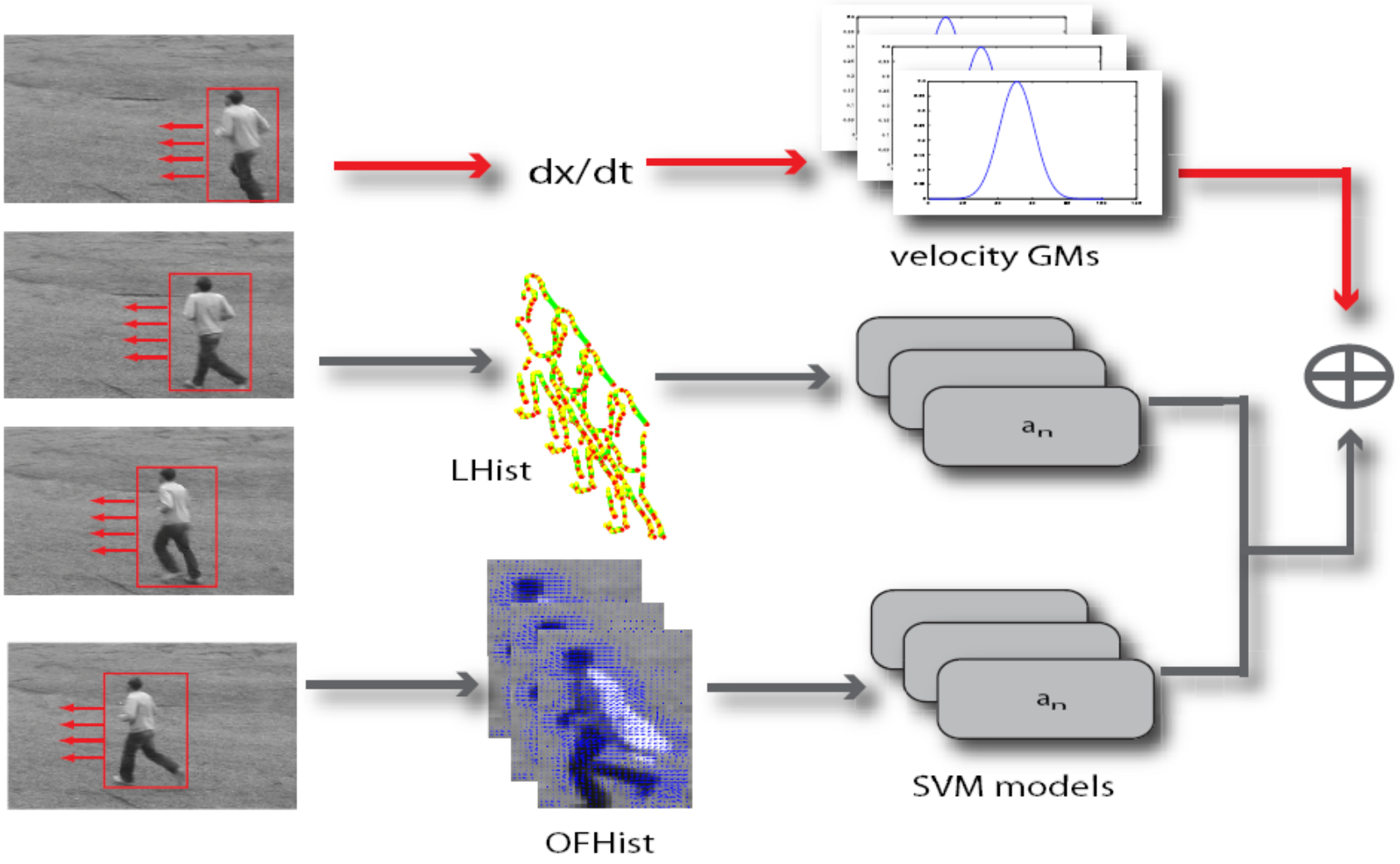
- In the absence of silhouettes, we can use lines fitted to the boundaries (Pb) (Martin PAMI2004) of human figures

..and Optical Flow



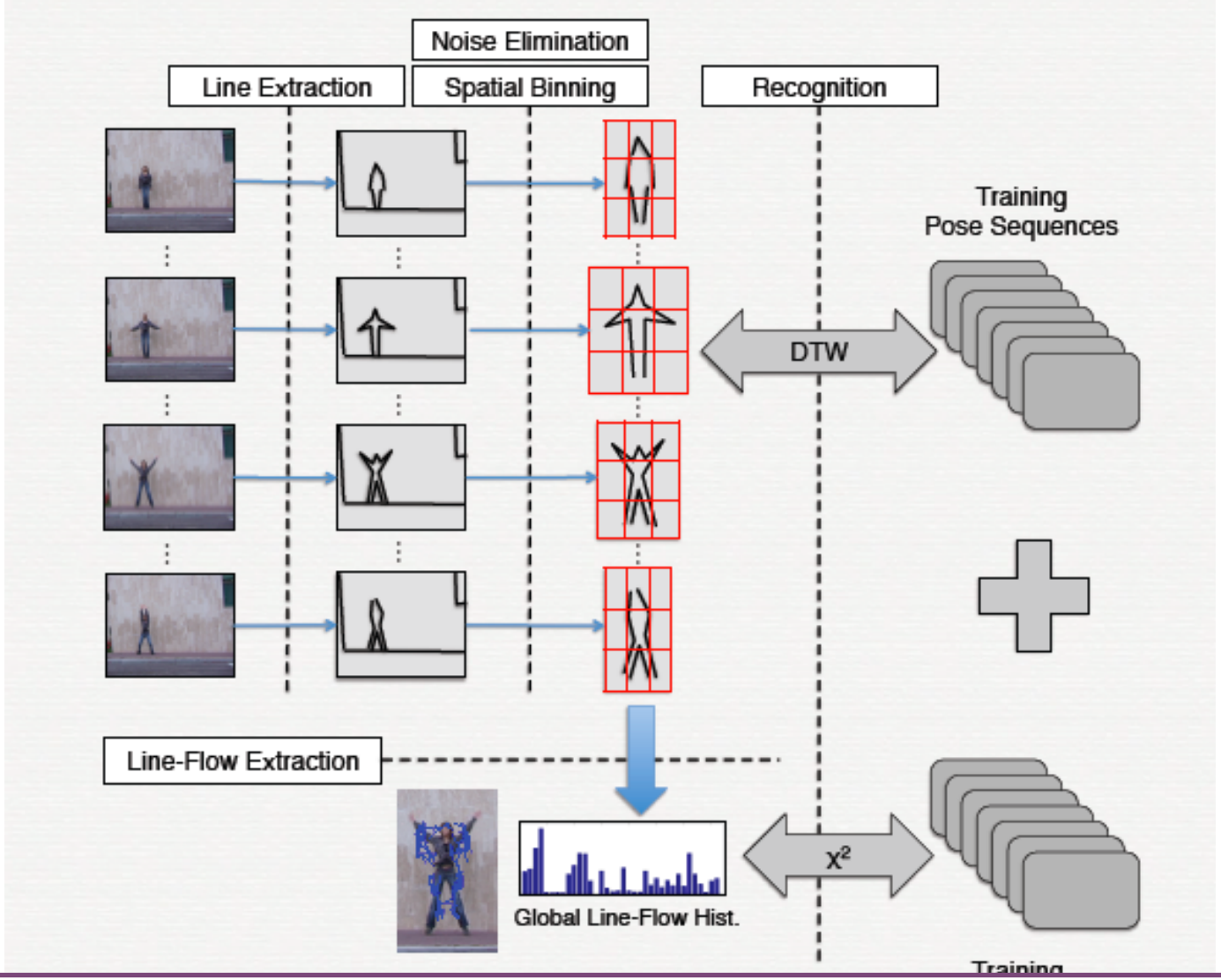
- Dense block-based optical flow calculation
  - L_1 block distance
  - 5x5 template size with a window size of 3

Recognition with LHist and OFHist



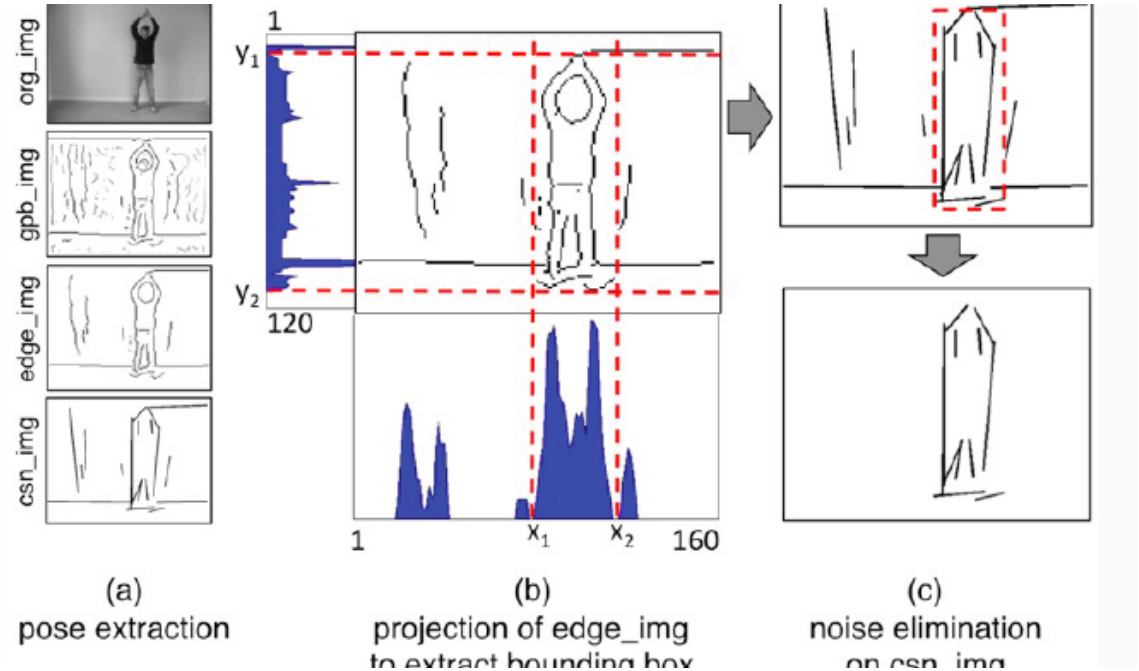
Ikizler, N., Cinbis, R. G., Duygulu, P., "Human action recognition with line and flow histograms", Proc. 19th International Conference on Pattern Recognition (ICPR 2008),

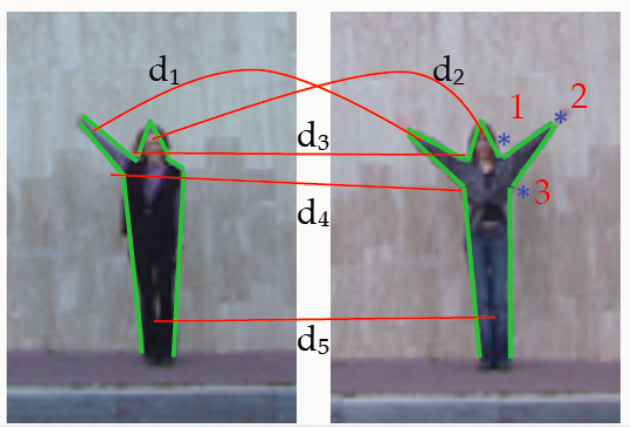
Pose as line segments

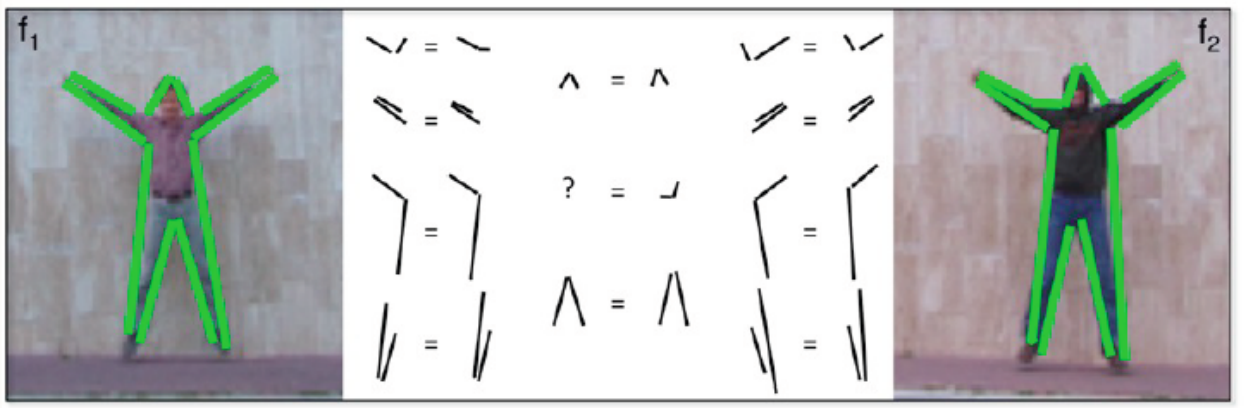


Baysal, S., Duygulu, P., "A Line Based Pose Representation For Human Action Recognition", Signal Processing: Image Communication, Volume 28, Issue 5, Pages 458-471, May 2013

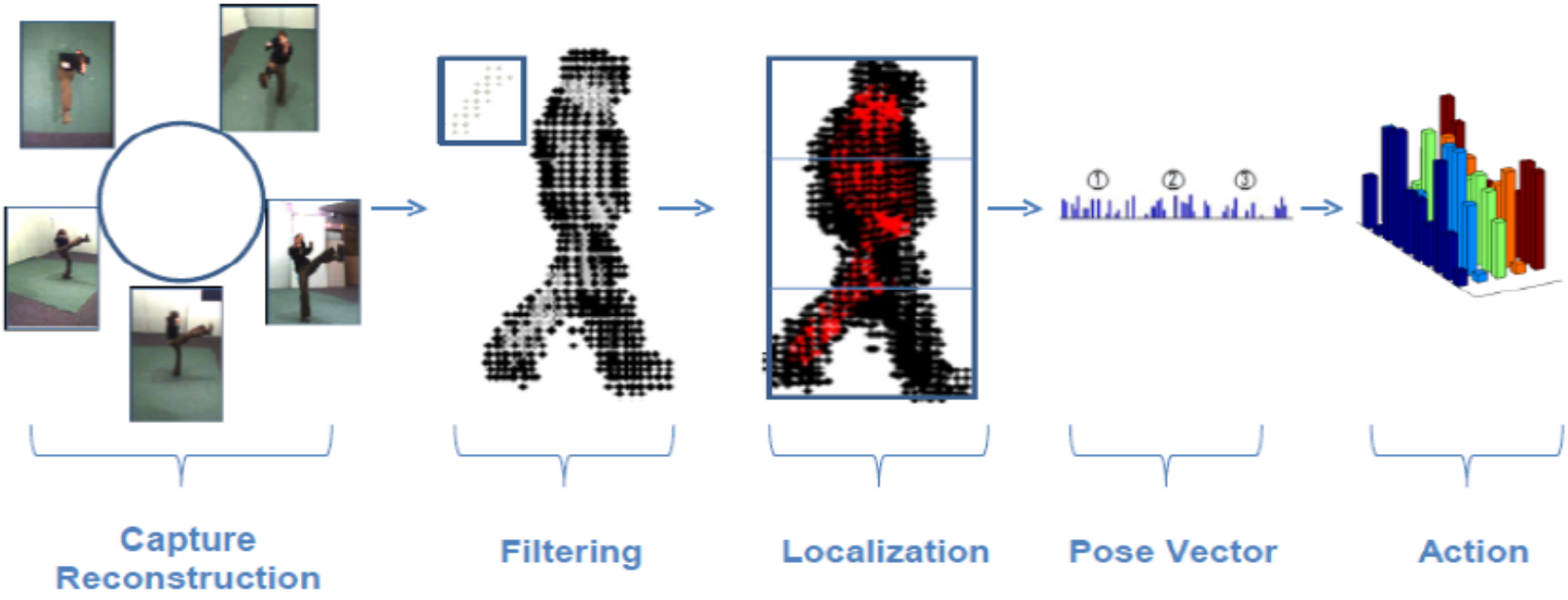
Line pairs


org_img
gpb_img
edge_img
csn_img
1
y1
y2
120
1
x1
x2
160
(a)
pose extraction
(b)
projection of edge_img
(c)
noise elimination

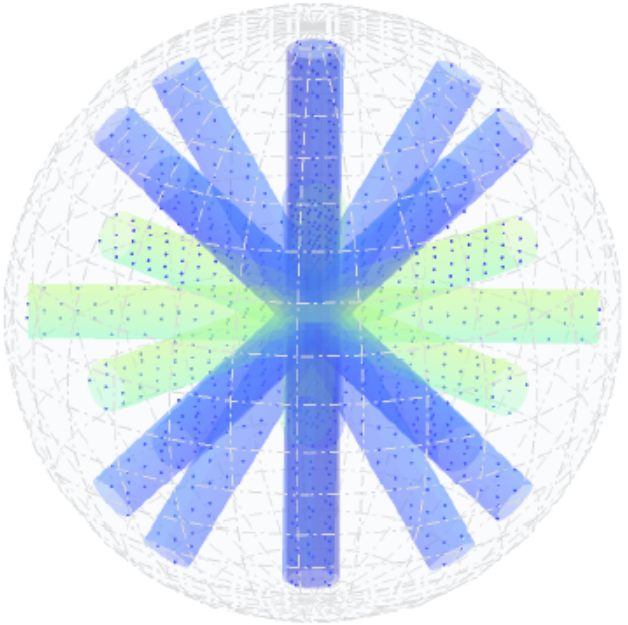

d1
d2
d3
d4
d5
1
2
3

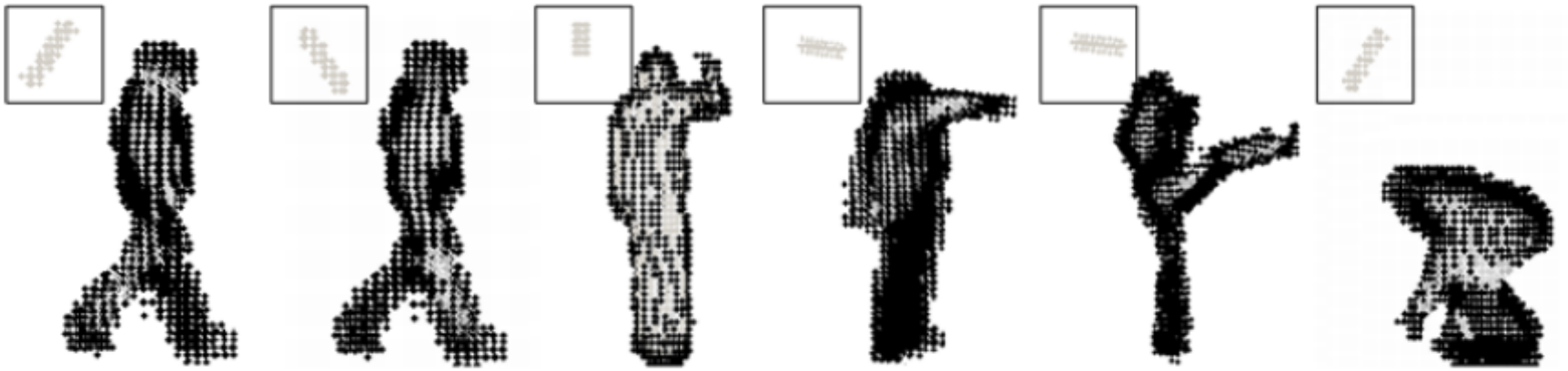

f1
f2

Multiple camera views



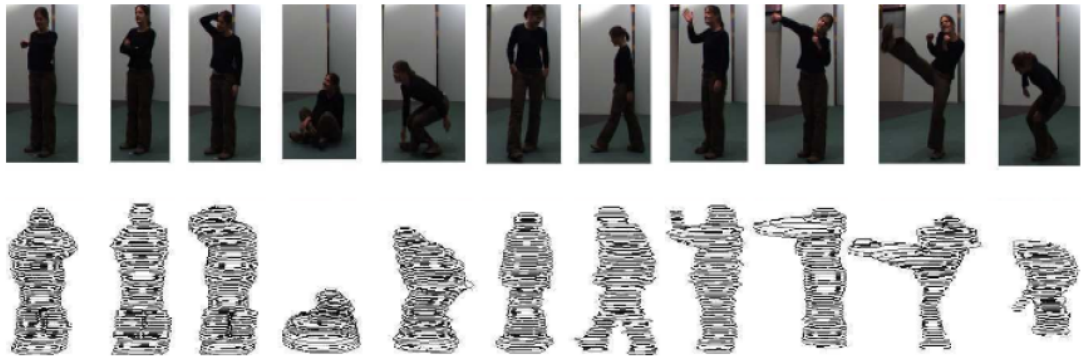
Oriented cylinders

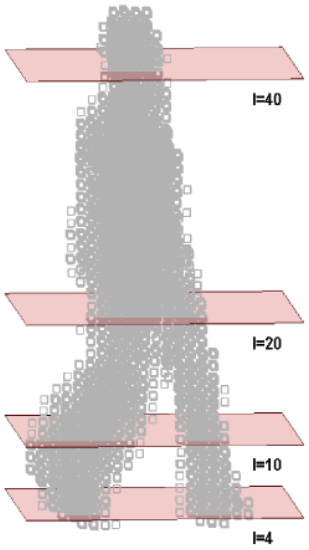


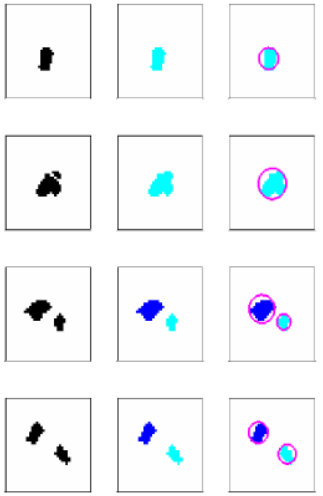


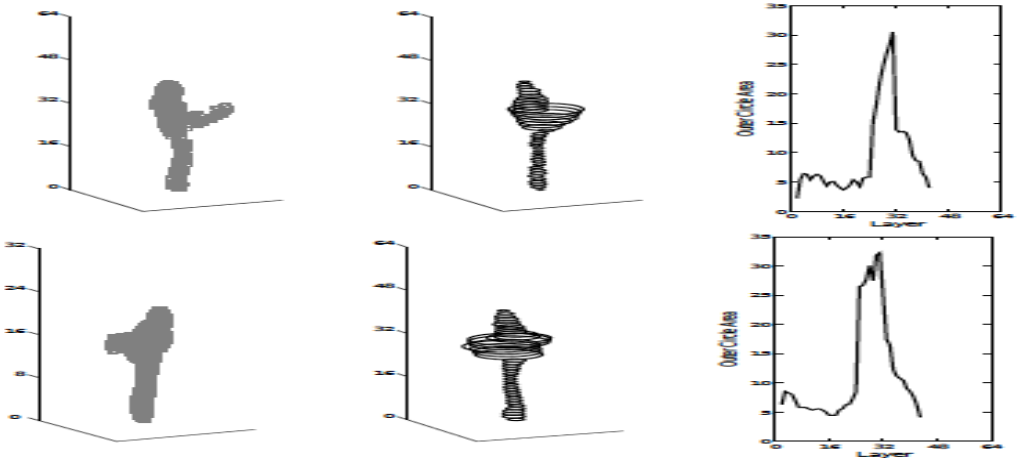
Pehlivan, S., . Duygulu, P. ”3D Human Pose Search using Oriented Cylinders”, IEEE Workshop on Search in 3D and Video (S3DV), in conjunction with ICCV 2009

Projections as circles



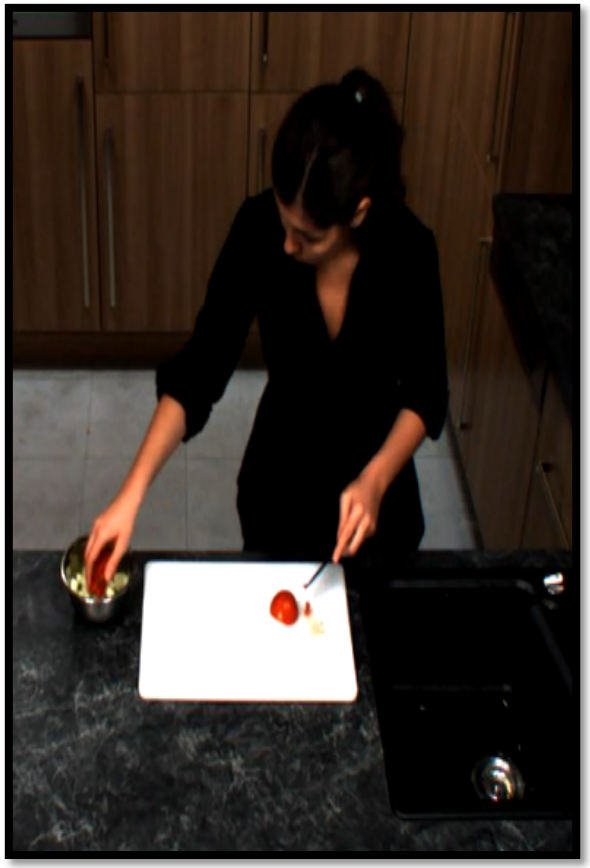






Pehlivan, S., Duygulu, P., "A new pose-based representation for recognizing actions from multiple cameras", Computer Vision and Image Understanding, volume 115, number 2, pages 140-151, February 2011

Cooking Activities: High Intra-class Variance




Low Inter-class Variance

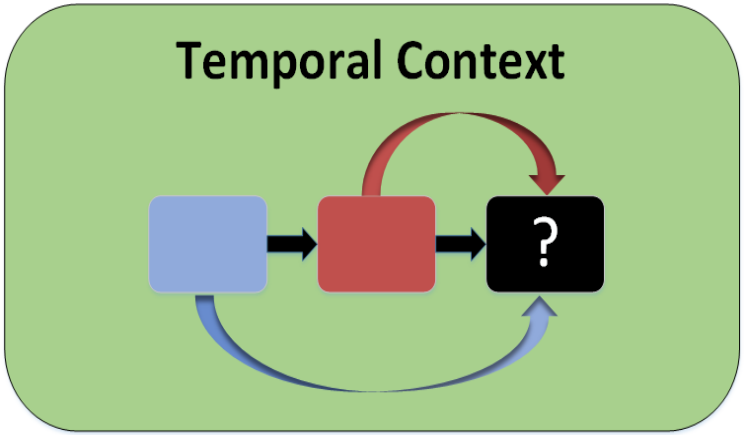


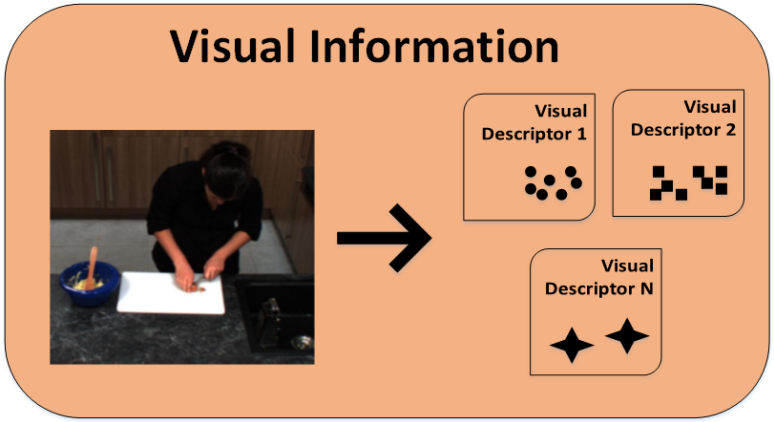



Solution

$$y = \underset{i}{argmax}\, P(c_i|x)$$

$$P(c_i|x) = T(c_i)\,.\,A(c_i, x)$$





Put in Pan or Put in Bowl?

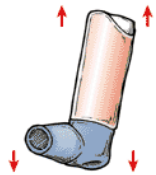
P("put in bowl" | "cut dice") > P("put in pan" | "cut dice")

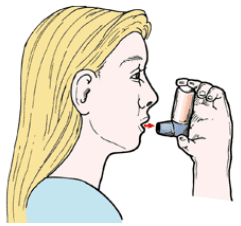
P("put in pan" | "spread") > P("put in bowl" | "spread")

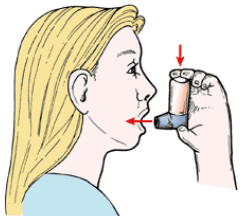
Asthma Inhaler use

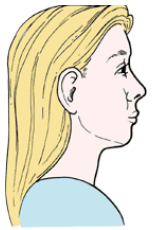
Correct steps to use an inhaler:

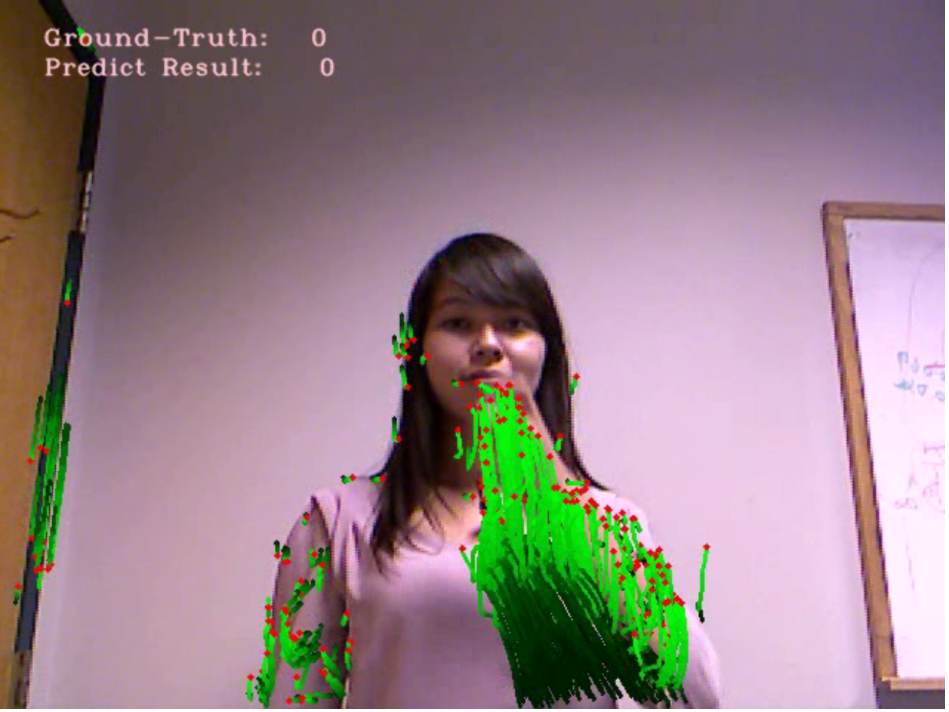
- Shake the inhaler (for 5 second)
- Breathe out
- Put the inhaler about 2 inches in front of your mouth
- Breathe in and push down the button at the same time
- Hold your breath for 10 seconds
- Breathe out slowly



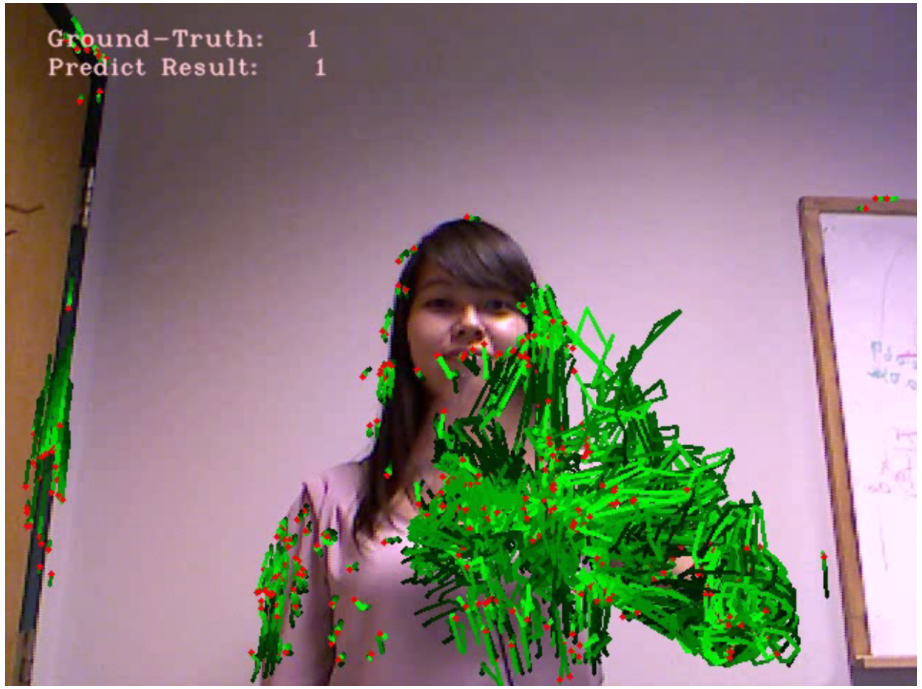








Reaching mouth



shaking

Contributors

- Ahmet Iscen
- Eren Golge
- Anil Armagan
- Sermetcan Baysal
- Fadime Sener
- Hilal Zitouni
- Sare Gul Sevil
- Selen Pehlivan
- Gokberk Cinbis
- Derya Ozkan
- Nazli Ikizler

TagSuggestr

- Given a few initial tags

predict more

Give more weights to the

visually similar images

$$W = \begin{array}{c|cccccc} & t_1 & t_2 & t_3 & \cdots & t_n & \\ I_1 & 1 & 0 & 1 & \cdots & 0 & \times\ \omega_1 \\ I_2 & 1 & 1 & 0 & \cdots & 0 & \times\ \omega_2 \\ & & & & & & \\ I_m & 0 & 0 & 0 & \cdots & 1 & \times\ \omega_m \end{array} \Bigg\} \sum = \begin{array}{|c|c|c|c|c|c|} t_1 & t_2 & t_3 & & & t_n \end{array}$$

Initial tags: casa, mila
Original tags: barcelon

Tags given by users participated in user-study:
barcelona, spain, architecture, catalonia, gaudi, building, casamila, catalunya, espana, house, antonigaudi, architect, arquitectura, art, catalan

Suggestions of the method (using RGB CH):
spain, gaudi, pedrera, catalunya, casamila, architecture, house, espana

1

Retrieve relevant photos & tags from Flickr

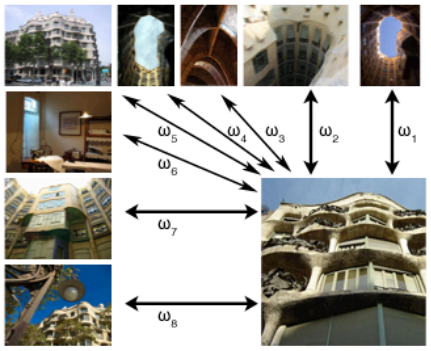
2

Create a unique tag list and eliminate stopwords

Unique Tags	-	Stopword List	=	Candidate Tags
300n, 35mm, aberration, abigfave, antoni, antonigaudi, antonigaudí, aplusphoto, architect, architecture, arquitectura, art,..., travel, unesco, unfound, viewtheworld, works, world		canon, nikon, abigfave, trkiye, hdr, geotagged, urban, bw, aplusphoto, 2007, anawesomeshot, december, diamondclassphotographer, bcn, 2006, eos, goldstaraward, flickrdiamond, panorama, ...		aberration, antoni, antonigaudi, antonigaudí, architect, architecture, arquitectura, art, artnouveau, barcelona, blue, building, casamila, casamilalapedrera, catalan, catalogne, ..., travel, unfound, viewtheworld, works

3

Calculate visual similarities



4

Sort candidate tags according to the total weights

Tag Suggestions:

spain
gaudi
pedrera
catalunya
casamila
architecture
house
espana

Accuracy compared to ground-truth is 87.5%

Sevil, S., Kucuktunc, O., Duygulu, P., Can. F., "Automatic Tag Expansion using Visual Similarity for Photo Sharing Websites", MTAP 2010

Experimental Evaluation of HOR

Method	Accuracy
HOR	**100%**
Blank et al. [12]	99.64%
Jhuang et al. [48]	98.8%
Wang et al. [96]	97.78%
Niebles et al. [63]	72.8%

Comparison to other methods on the Weizzman dataset

Method	Accuracy
Jhuang et al. [48]	91.7%
Wong et al. [100]	91.6%
HORW	**89.4%**
Niebles et al. [64]	81.5%
Dollár et al. [24]	81.2%
Ke et al. [50]	80.9%
Schuldt et al. [84]	71.7%

Comparison to other methods on the KTH dataset

Comparison to HOGs on the KTH

	HOG	HOR	HORW
SVM	76.85%	77.31%	85.65%
DTW	67.59%	74.54%	78.24%
v+SVM	82.41%	81.48%	89.35%

+ Line and Flow Results

	boxing	hclapping	hwaving	jogging	running	walking
boxing	0.97	0.03	0.0	0.0	0.0	0.0
hclapping	0.06	0.89	0.06	0.0	0.0	0.0
hwaving	0.03	0.06	0.92	0.0	0.0	0.0
jogging	0.0	0.0	0.0	0.92	0.0	0.08
running	0.0	0.0	0.0	0.14	0.83	0.03
walking	0.03	0.0	0.0	0.03	0.0	0.94

Method	Accuracy
LFV	**94.0%**
Jhuang [48]	91.7%
Wong [100]	91.6%
Niebles [64]	81.5%
Dollár [24]	81.2%
Ke [50]	80.9%
Schuldt [84]	71.7%

Condition	LFV	Jhuang [48]
s1	**98.2%**	96.0%
s2	**90.7%**	86.1%
s3	88.9%	**89.8%**
s4	**98.2%**	94.8%

- Shape and flow are complimentary to each other.
- Again, this depends on the nature of the actions in mention.

Still Image Results

(a) catch, walk, catch, throw

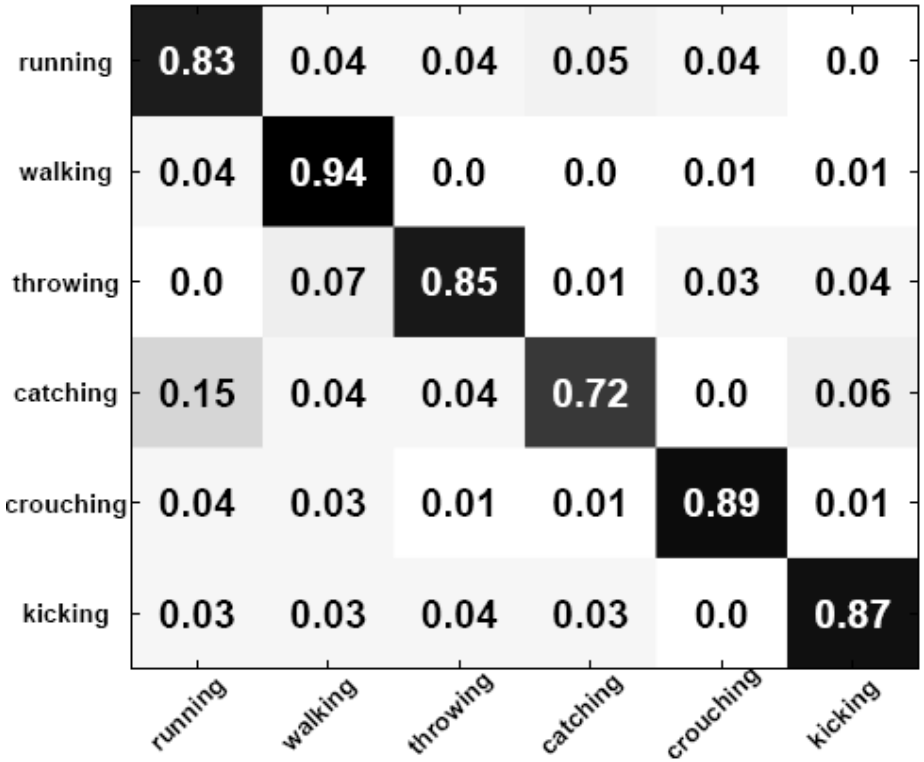
(b) run, run, run, kick

(c) catch, kick, walk, crouch

(d) run, throw, run, run

(e) kick, walk, walk, catch

(f) throw, walk, run, throw



Total accuracy 85.1%

Misclassified action images

Multi-view

Method	Accuracy (over 11 actions)	Accuracy (over 13 actions)
Weinland et al. [61]	93.33%	-
Our method	**90.91%**	**88.63%**
Liu et al. [32]	-	82.8%
Weinland et al. [59]	81.27%	-
Lv et al. [35]	-	80.6%
Yan et al. [62]	78.0%	-