Pattern Analysis & Computer Vision

Istituto Italiano di Tecnologia

# Groups and Crowds:
# Detection, Tracking and Behavior Analysis of People Aggregations

Vittorio Murino


639-12
03/24/10 04:58:36 PM
FISU
FISU
3687-18
Vittel
© Benoist Sebire
blog.benoa.net
thoughtequity

# Groups and crowds: why?

- Video analytics
  - scene understanding and interpretation
- Video surveillance
  - beyond normal/abnormal, events, activity recognition
- Social robotics, human-robot interaction
  - advanced interaction models
- Retailing, marketing
  - customer profiling
- Architectural planning tools

# Analysing groups and crowds …

- Actions and *inter-actions*
- Activities and *collective* activities
- Detection of *abnormal* behaviors, recognition/detection of *specific* behaviors

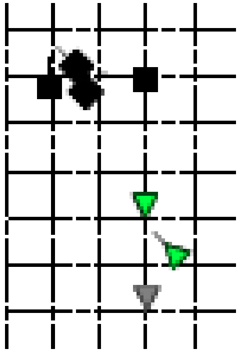
- Groups? Or rather *gatherings*
- Only one class of crowd? Which are the drivers for modeling crowd behavior

- Can Computer Vision do the job alone?
- What about other disciplines such as Sociology, Psychology, Neuro-psychology
- Social Signal Processing paved the way to go

# GROUPS

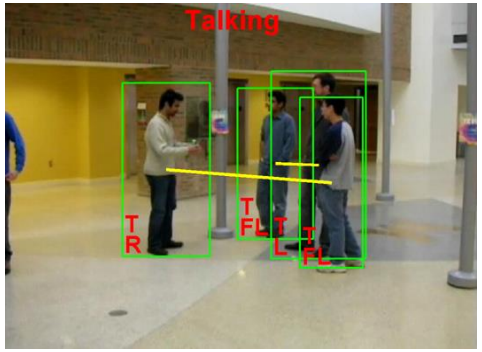
# Types of approaches on group analysis

Group *detection*

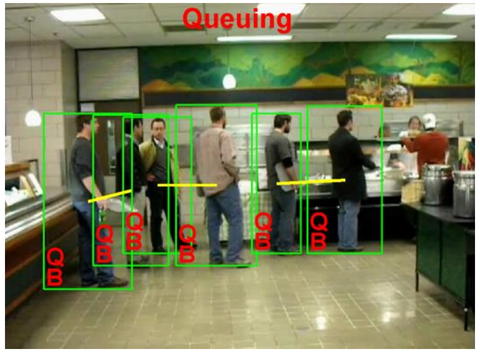



Group *tracking*

Group (*collective*) *activity recognition*

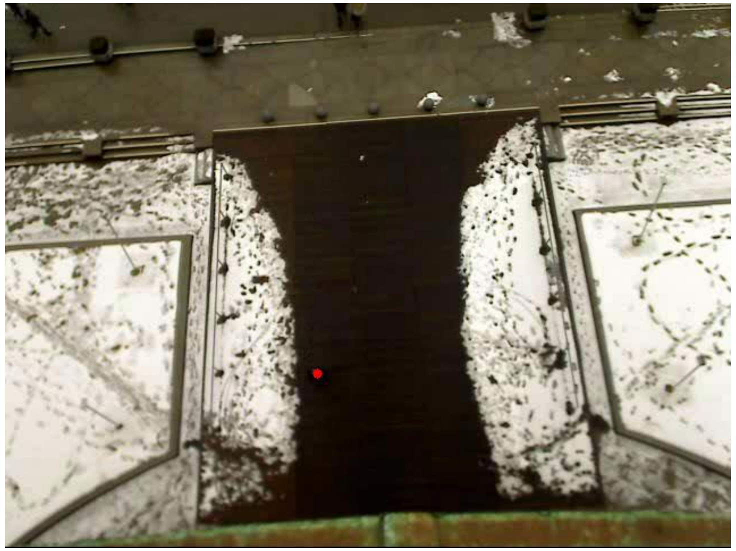






# Common definitions for group analysis

- Group
  - (an entity whose) "members are close to each other, with similar speed, with similar direction of motion" [Ge *et al.* TPAMI '12], and the like [Zeidenberg *et al.* AVSS '12, Pellegrini et al. ICCV '09, Bazzani *et al.* CVPR '12…]



**BIWI Walking Pedestrians dataset**
[Pellegrini et al. ICCV '09]

# Common definitions for group analysis

- What happens in the case of still images?
- Structured group [Choi et al. ECCV 2014]:

  “*consistent spatial configurations of people*” (doing the same activity)


**Structured Group Dataset**
[Choi et al. ECCV ’14]

# Summarizing (for groups) …

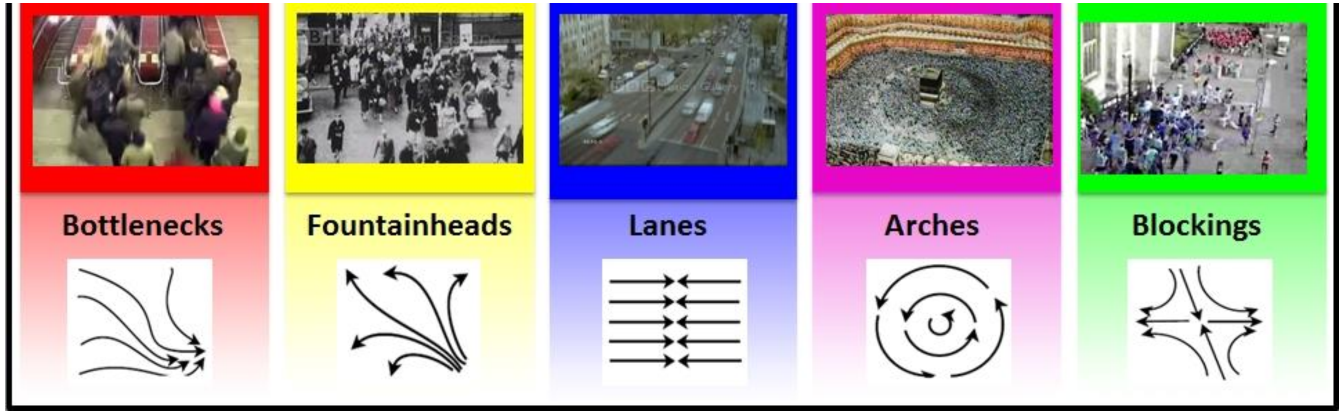
- We can conclude that a group is *an entity formed by more than one person, where its components are close to each other*, and can do the following activities:
  - *moving together, with similar oriented motion*
  - *doing the same activity like crossing, waiting, talking …*

- **Open questions**
  - Is there only one type of group?
  - Is there any maximum number of people that can form a group? When a group(s) becomes a crowd?

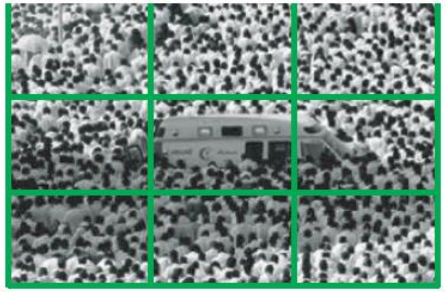
# CROWD

# Classes of approaches on crowds

crowd behavior understanding/ crowd tracking, segmentation, anomaly detection




(ROI, LOI)
people counting/density estimation



| | | |
|---|---|---|
| 45 | 49 | 52 |
| 43 | 18 | 45 |
| 44 | 42 | 46 |

tracking individuals in the crowd

# Common definitions for crowd analysis

- Crowd
  - (is identified when) "the density of the people is sufficiently large to disable individual and group identification"
    [Jacques *et al.* SPM '10, Boghossian & Velastin ICECS 1999…]
  - *"a collection of individuals obeying a set of analytical rules"* [Still 2000, Moore et al. ACM 2011], like the ones listed by the *Social Force Model (repulsion, attraction)* [Mehran et al. '09]

# Some crowd datasets

**Crowd Segmentation Data Set**
[Ali CVPR '07]

**Web Dataset: Abnormal/Normal Crowd activities** [Mehran CVPR '09]


# Summarizing (for crowds) …

- There is only one kind of crowd
  - *that can exhibit collective motion*
  - *whose activities can be normal or abnormal*

- **Open questions**
  - Are there *different types of crowd*, whose recognition may be of interest for computer vision?
  - Is there a way to drive/control crowd behavior?
  - How can we approach crowd behavior modeling?

# Summarizing (for crowds) …

- Recent trends propose that crowd behavior is driven by small groups and that social relations influence the way people behave in crowds

- Crowd models should consider both local behavior of pedestrians/small groups during interactions, and the global dynamics of the crowd at high density

- Newtonian mechanics models have limitations, need of embed cognitive processes (heuristics) used by pedestrians (collision avoidance, physical and social interactions, imitation)

# Analysing groups and crowds *(from a sociological standpoint)*

- Group:
  - a social unit whose members stand in status and relationships with one another (Forsyth 2010)
  - it entails some durable membership and organization (Goffman 1961)
  - two or more people interacting to reach a common goal and perceiving a shared membership, based on both physical (spatial proximity) and social identities (Turner, 1981)

- Gathering: any set of two or more individuals in co-presence having some form of social interaction (Goffman 1966)

- Many types of gatherings, depending on:
  - the number of people being present
  - the form, or kind of social interaction at hand
  - the properties of the setting (private/public, static/dynamic)

- Crowd: a gathering constituted by a "large" number of people [McPhail 1991]


Slide credit:
M. Cristani, C. Bassetti

# Analysing groups and crowds *(from a sociological standpoint)*

- Gatherings (2 to N)
  Two or more persons in co-presence in a given space-time

| **Small gathering (2 to 6)** | **Medium gathering (7 to 12/30)** | **Large gathering (13/31 to N)** |
|---|---|---|
| *Occurring in private, semi-public and public places* | *Occurring in private but mostly semi-public/public places* | *Occurring in semi-public but mostly in public places* |

- *private places:* home, private garden, car
- *semi-public places:* classroom, office, club, party area
- *public places:* open plaza, transportation, station, walkway, park, street


Slide credit:
M. Cristani, C. Bassetti

# Analysing groups and crowds *(from a sociological standpoint)*

- Kinds of *social interaction* (Goffman 1961, 1966; Kendon 1988)
  - unfocused interaction: whenever two or more individuals find themselves by circumstance in the immediate presence of others (forming a queue, crossing the street...)
  - focused interaction: whenever two or more individuals willingly agree to sustain for a time period a single focus of attention.

- It may be further specified into:
  - common focused interaction: the focus of attention is common and not reciprocal (watching a movie at the cinema, attending a lecture with your colleagues...)
  - jointly focused interaction: entails the sense of a mutual activity, participation is not peripheral but engaged (conversation, board game…)


Slide credit:
M. Cristani, C. Bassetti

# Small gathering (2 to 6)

Occurring in private, semi public and public places

Line at the shop register, watching timetables, eating at a cantine (without knowing the neighborhood)
(unfocused)



television-watching group,
(common-focused)

conversational group,
game players,
fight
(jointly-focused)



Slide credit:
M. Cristani, C. Bassetti

# Medium gathering (7 to 12-30)

Occurring *in private but mostly in semi public and public places*

Line at the
post office
(**unfocused**)

classroom group,
touring group
at the museum
(**common-focused**)

meeting group,
extended family
commensal
(**jointly-focused**)

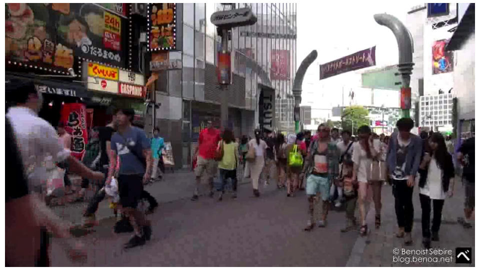
*In these cases, small gatherings of other typologies of gathering may be present:*
*difficult to catch/model*
*but important to individuate*

Slide credit:
M. Cristani, C. Bassetti

# Large gathering (13-31 to N)

*Occurring in semi public but mostly in public places*

line at the airport check in,
walking in a street
(unfocused)
**Prosaic [3] or**
**Casual [10,11] crowd**



sport/theatre/cinema spectators
(common-focused)
**Spectator [3] or Conventional [10,11] Crowd**

mob/riot/sit-in/march participants
(common and jointly-focused)
**Demonstrations/Protest [3] or**
**Acting [10,11] crowd**

*In these cases, small gatherings of other typologies of gathering may be present:*
*difficult to catch/model*
*but important to individuate*

# Analysing groups and crowds *(from a sociological standpoint)*

| | Unfocused | Common focused | Jointly focused |
|---|---|---|---|
| Static |  |  |  |
| Dynamic |  |  |  |



Slide credit:
M. Cristani, C. Bassetti

# Gatherings

Two or more persons in co-presence in a given space-time

## Small gathering

*private, semi public and public places*

- Line at the shop register, watching timetables (unfocused)
- television-watching (common-focused)
- free-standing conversational group, game players (jointly-focused)

## Medium gathering

*private but mostly semi public and public places*

- Line at the post office (unfocused)
- classroom, touring group at the museum (common-focused)
- meeting, extended family commensal (jointly-focused)

## Large gathering

*semi public but mostly public places*

- line at the check-in (unfocused - Prosaic or Casual crowd)
- sport/theatre/cinema spectators (common-fo. Spectator crowd)
- flash-mob, Mass, sport supporters (jointly-focused - Expressive crowd)
- mob/riot/sit-in/march (common&jointly-foc. - Protest/Acting crowd)

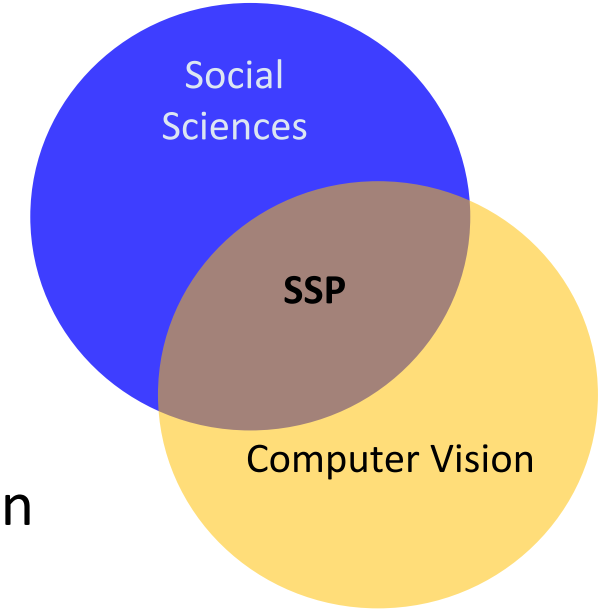
stronger social relations

harder to model


Slide credit:
M. Cristani, C. Bassetti

# The point of view of *Social Signal Processing*

- *SSP cues*:
  - distance (from being far to physical contact)
    → social relationship
  - body pose/posture → facing, symmetry
  - head/gaze orientation/eye contact → focus of visual attention
  - gesture & posture → kind of interaction






A.Vinciarelli, M.Pantic and H.Bourlard, "Social Signal Processing: Survey of an Emerging Domain", Image and Vision Computing Journal, Vol. 27(12), 2009.

# Gatherings and SSP cues

## Unfocused small gath.

- People are close to each other
- not common body/head/feet orientation
- no unique/coincident focus of visual attention
- Semi-static dynamics


## Comm.-foc. small/medium gathering

- close to each other
- similar and often symmetrical posture
- all people looking at the same target
- semi-static dynamics

## Joint-foc. medium gath.

- close to each other
- facing each other
- no intruders between participants
  → F-formations
  → Many datasets available


## Unfocused large gath.

*(casual crowd, also protest crowd)*

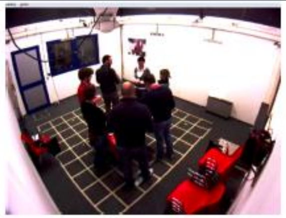
- no unique focus of visual attention
- no unique motion dynamics
- normally, people walk

## Common focused large gath.

(spectator crowd)

- single focus of visual attention
- mostly common head feet orientation
- normally, people stand or sit

# Detection of jointly focused gatherings: Datasets

| **Dataset** | Easy frames | | Hard frames | |
|---|---|---|---|---|
| *IDIAP Poster* |  |  |  |  |
| *Cocktail Party* |  |  |  |  |
| *Coffee Break* |  |  |  |  |
| *GDet* |  |  |  |  |

# Challenges

– Importance of detecting different typologies of gatherings …
but also their evolution!

# In conclusion …

- Sociology provides a taxonomy for people gatherings and a way about how to approach them
- Sociologists may help in labeling gatherings, specifying if they are
  - *unfocused*
  - *common focused*
  - *jointly focused*

- Recognizing these typologies of gatherings and their temporal evolution may help the surveillance field to do better profiling, activity analysis, event recognition, etc.

# Group detection

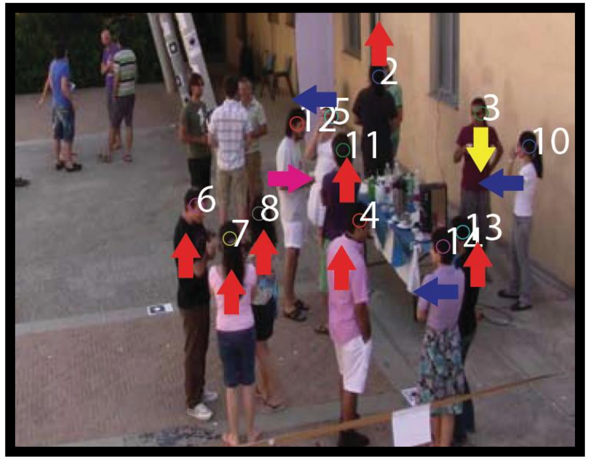
## Hough-based Approach

# The scenario

Detection of groups in cocktail party situations

M.Cristani, L.Bazzani, G.Paggetti, A. Fossati, A.Del Bue, D.Tosato, G.Menegaz, V.Murino, ***Social interaction discovery by statistical analysis of F-formations***, BMVC 2011

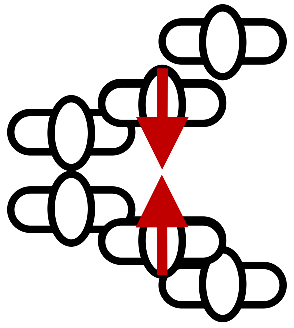
# The scenario




- Our unconstrained, ecological scenario:
  - A **full-calibrated camera**
  - People **tracking**
  - **Head orientation classification**, with at least 4 orientations

# F-formations

- Our approach **detects interactions** by considering
  - the **spatial layout of people**
  - the **head/body orientation**



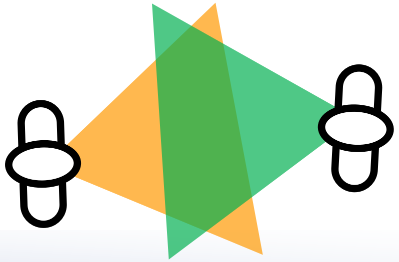
- In sociology, these cues naturally define an **F-formation**

# State of the art: Computer Vision

- **Tracking** as classic element for detecting interactions

- **[Robertson *et al.*, ECCV06, Orozco *et al.*, BMVC09, Tosato *et al.*, ECCV10]** estimated the head direction as key cue (visual focus of attention, VFOA)

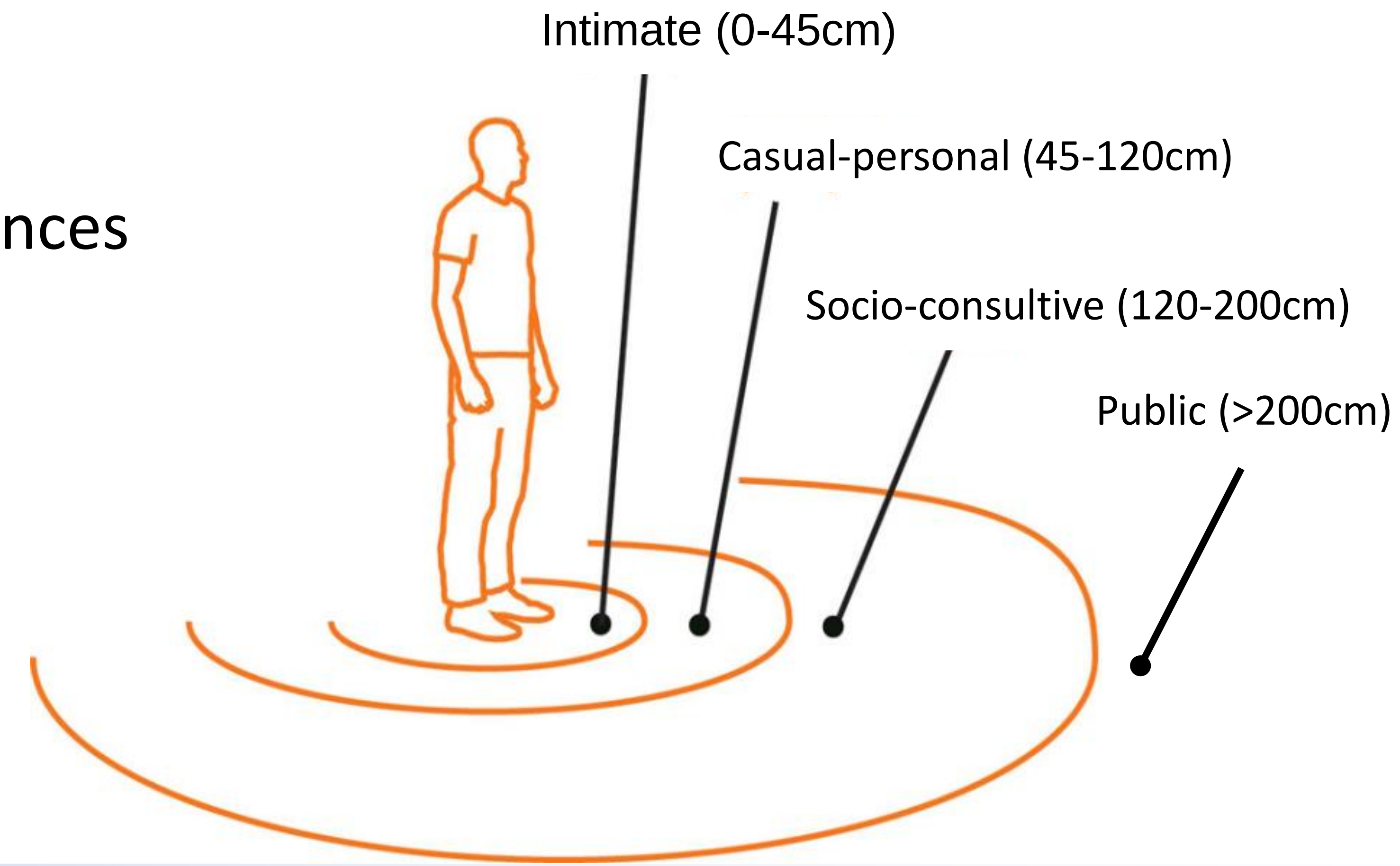
- Interaction = VFOA + position + velocity **[Robertson *et al.*, EURASIP '11]**
- Interaction = VFOA and position in a 3D environment, the **IRPM** approach **[Bazzani et al., Expert Systems '11]**



# State of the art: Social sciences

In our study, we consider proxemics principles:

Hall's social distances
*[Hall66]*



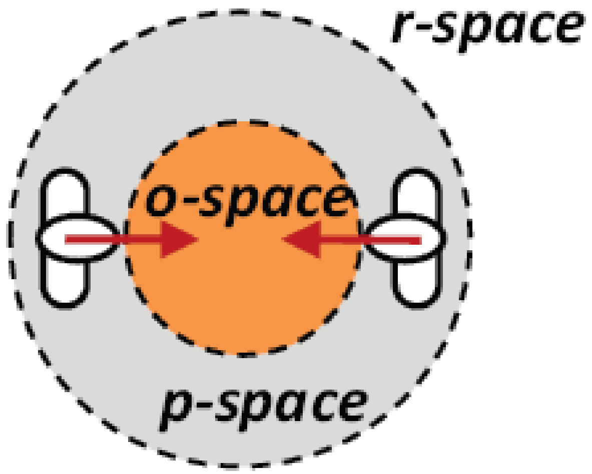
# State of the art: Social sciences

How people are placed when interacting

→ F-formations **[Kendon et al., 1977-’10]**

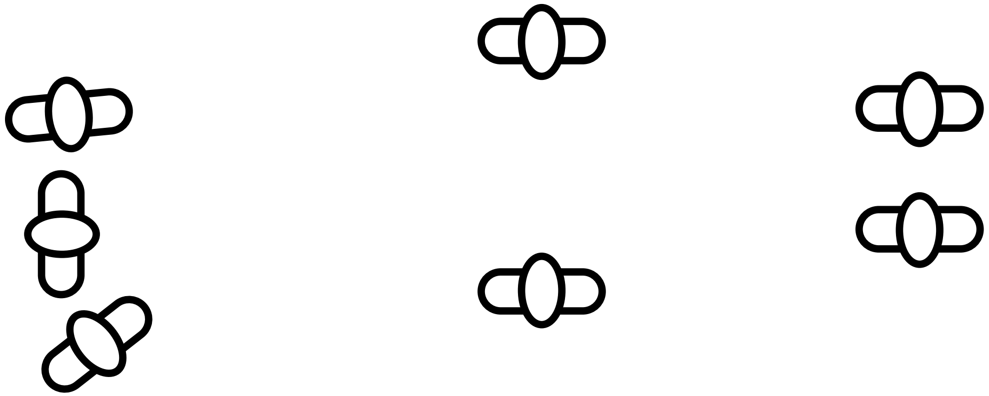
# F-formation

- Three concentric regions…
  - o-space: a convex empty space surrounded by the people involved in a social interaction, where every participant looks inward into it, and no external people is allowed
  - p-space: a narrow stripe that surrounds the o-space, and that contains the bodies of the talking people
  - r-space is the area beyond the p-space



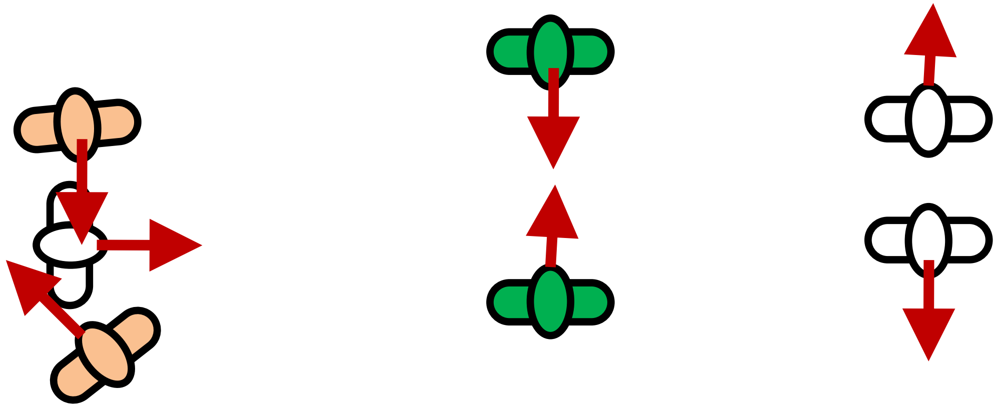
# Our approach: the idea

- F-formation definition
  - *F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access*



# Our approach: the idea

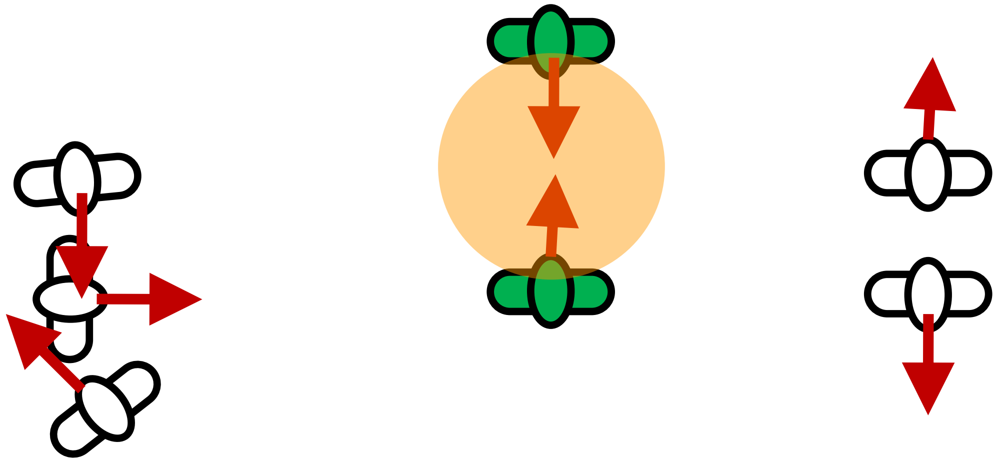
- F-formation definition
  - *F-formation arises whenever two or more people sustain a* ***spatial*** *and* ***orientational relationship*** *in which* ***the space between them is one to which they have equal, direct, and exclusive access***



The range of distances is suggested by Hall!

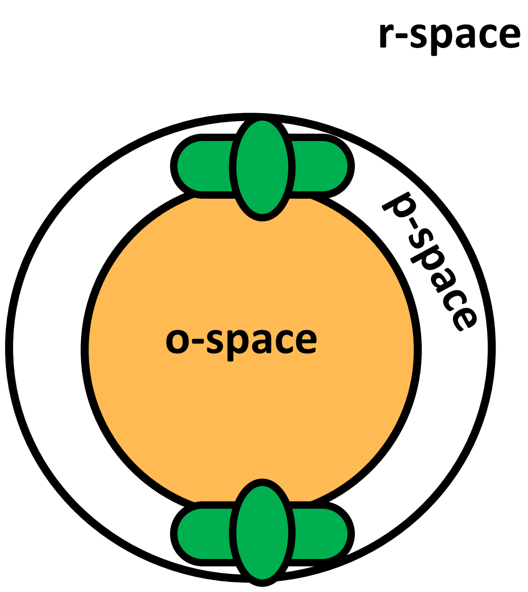
# Our approach: the idea

- F-formation definition
  – *F-formation arises whenever two or more people sustain a **spatial** and **orientational relationship** in which **the space between them is one to which they have equal, direct, and exclusive access***

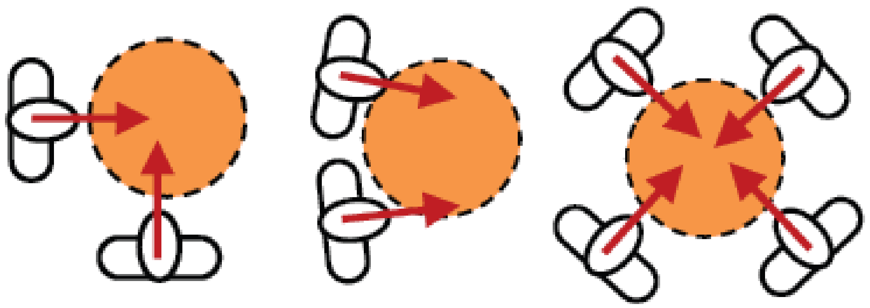


# Our approach: the idea

- Modelling F-formations
  - Three “spaces”: **o-space**, p-space, r-space
  - The **o-space** can be thought as a circular area



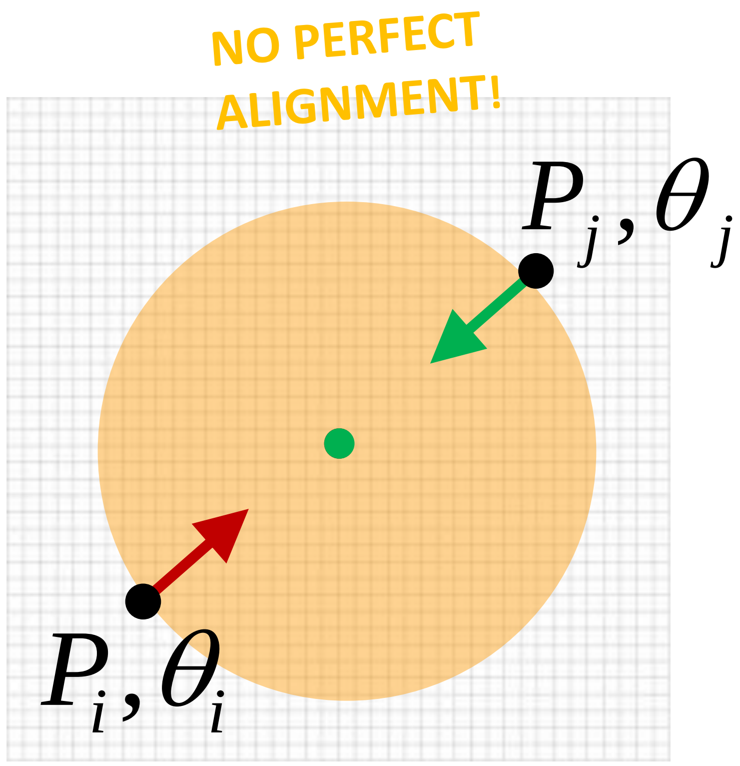

- Different kinds of F-formations



# Our approach: the algorithm

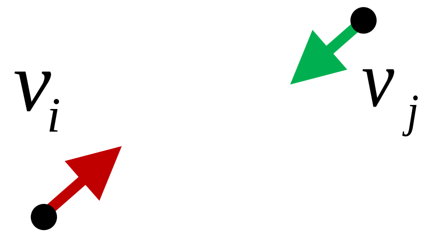
- A 3-step Hough voting approach
- Each person **votes for a o-space center location** considering the head orientation and a distance
- The center location that gets the highest number of votes is a potential o-space

- PROBLEM!

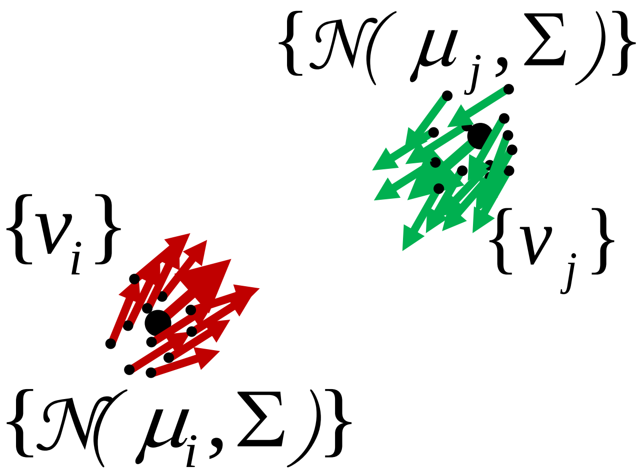


# Our approach: the algorithm

- Steps:
  1. Given some subjects

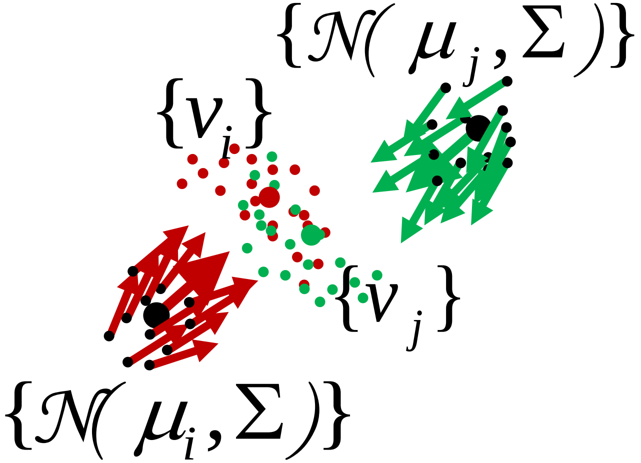



  2. Sample a set of positions

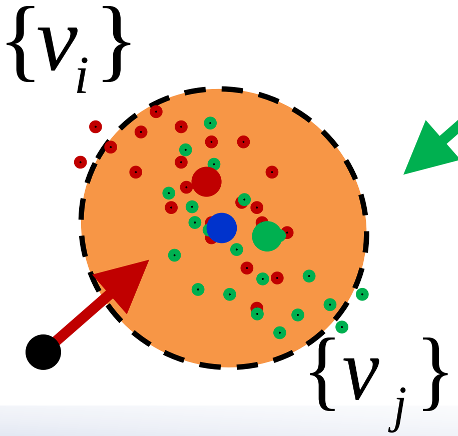


# Our approach: the algorithm

3. Each position votes for a possible center



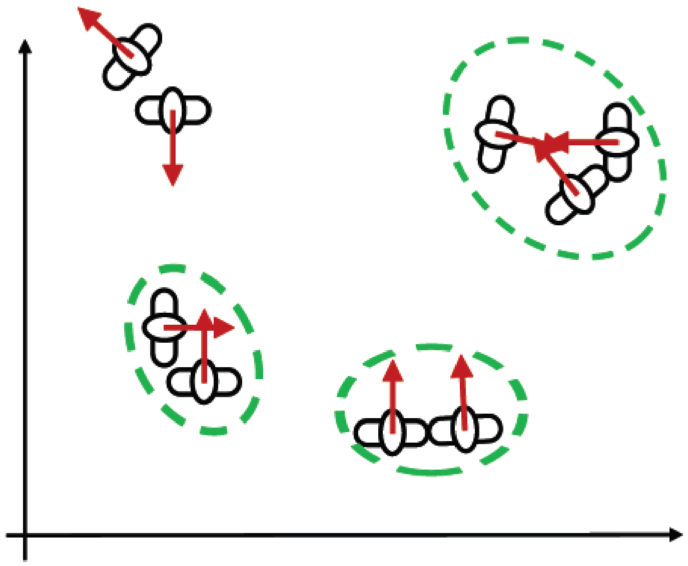

4. The location with the max of votes determines the center of a o-space

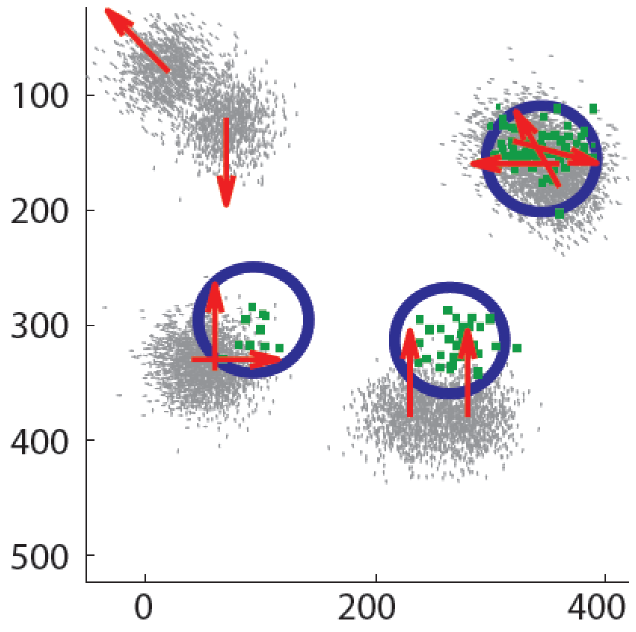


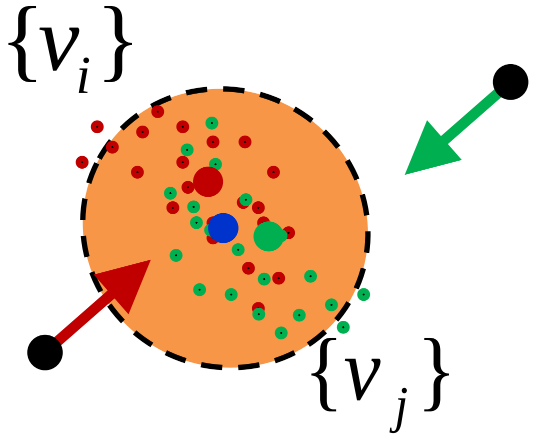
# Our approach: the algorithm

5. Check if none is present in the o-space, and you get the F-formation

*Example*

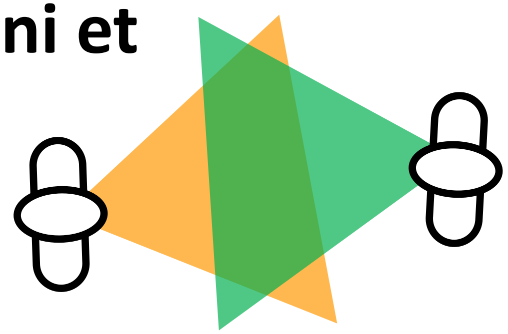




# Experiments

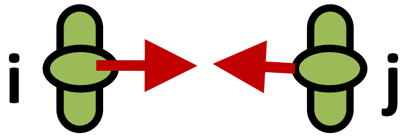
- Three datasets have been taken into account, for a total of 447 frames:
  - a synthetic dataset
  - two real datasets
- Each dataset has a ground truth, created from psychologists that annotated the interactions
- As competitive approach, we consider **IRPM [Bazzani et al.11 Expert]** (position + VFOA intersection)



# Experiments: accuracy measures

- How *effective* is the method?
  - A group is matched if $\lceil 2/3 \cdot |G| \rceil$ of their individuals have been selected.
  - Compute **precision** and **recall**

- Considering the entire sequence
  - Relation matrix (from IRPM) + Mantel test




$$\begin{array}{c|c|c|c|} \multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{\mathbf{i}} & \multicolumn{1}{c}{} \\ \hline & & & \\ \hline \mathbf{j} & & \mathbf{+1} & \\ \hline & & & \\ \hline \end{array}$$

# Experiments

- The **CoffeeBreak** dataset
  - 2 sequences have been annotated indicating the groups present in the scenes, for a total of 45 frames for Seq1 and 75 frames for Seq2.
  - Tens of people, different groups

# Experiments: CoffeeBreak dataset

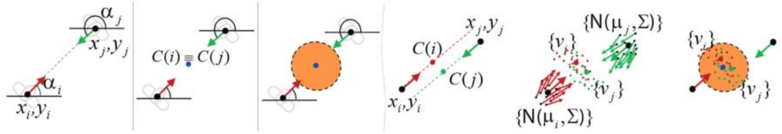

| Method | ***precision*** | ***recall*** | ***Mantel test*** |
|---|---|---|---|
| **IRPM** | 0.55 | 0.19 | 0.67 |
| **Our approach** | **0.85** | **0.76** | **0.76** |

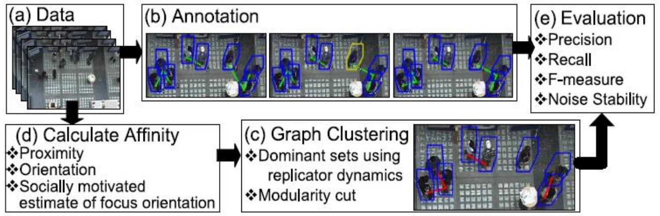
# Group detection

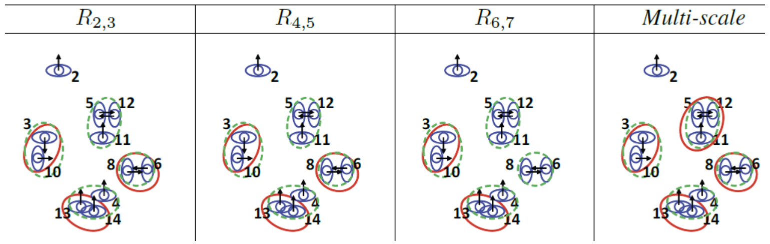
## Game-theoretic Approach

# State of the art

- F-Formation detection algorithms:
  - Hough voting [2]
    - Samples vote for an o-space
    - O-space with the majority of votes is taken.
  - Dominant Set [3]
    - A scene is represented as a weighted graph G.
    - An F-F is represented as a Dominant Set (a clique)
    - Find maximal cliques in G for finding the FFs.
  - Multi-Scale [4]
    - Based on [2] Hough Voting schema but for different F-F sizes.
    - Select for each location the F-F having the highest weighted Boltzmann entropy.

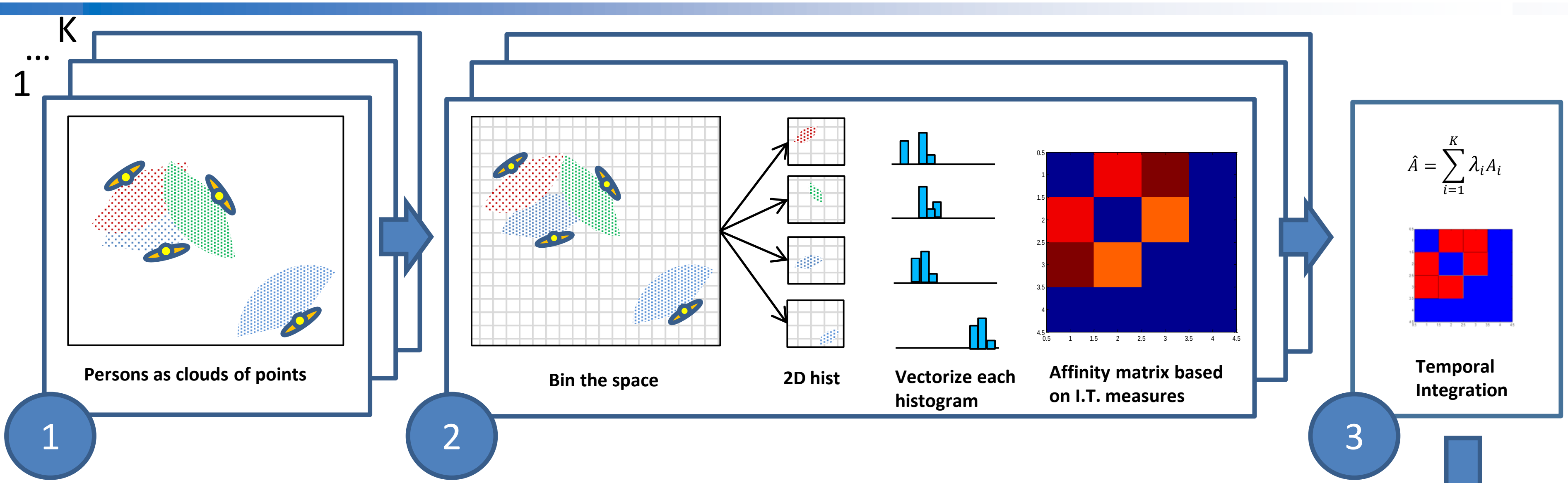


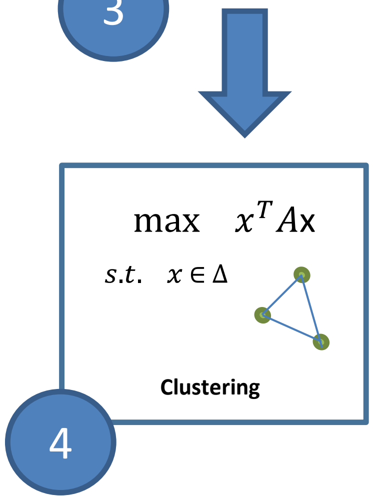



[2] Cristani et al: Social interaction discovery by statistical analysis of F-formations. In: Proc. Of BMVC, BMVA Press (2011)
[3] Hung, H., Krose, B.: Detecting F-formations as dominant sets. In: ICMI. (2011)
[4] Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M.: Multi-Scale F-Formation Discovery for Group Detection. In: ICIP. (2013)

# The method




1. **Probabilistic model of Frustum of Visual Attention**
2. **Quantify interactions in a pairwise matrix using Information-Theoretic measures**
3. **Multiple Payoff Games to integrate the K-consecutive frames**
4. **Game-theoretic clustering for finding groups**


Torsello, A., Rota Bulo, S., Pelillo, M.: Grouping with asymmetric affinities: A game theoretic perspective. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1. (2006) 292–299

# The method – Step 1
## Frustum

- A person in a scene is described by his/her position *(x,y)* and the head orientation *ϑ*
- The frustum represents the area in which a person can sustain a conversation and is defined by an aperture and a length

$(x, y, \theta)$

All Samples

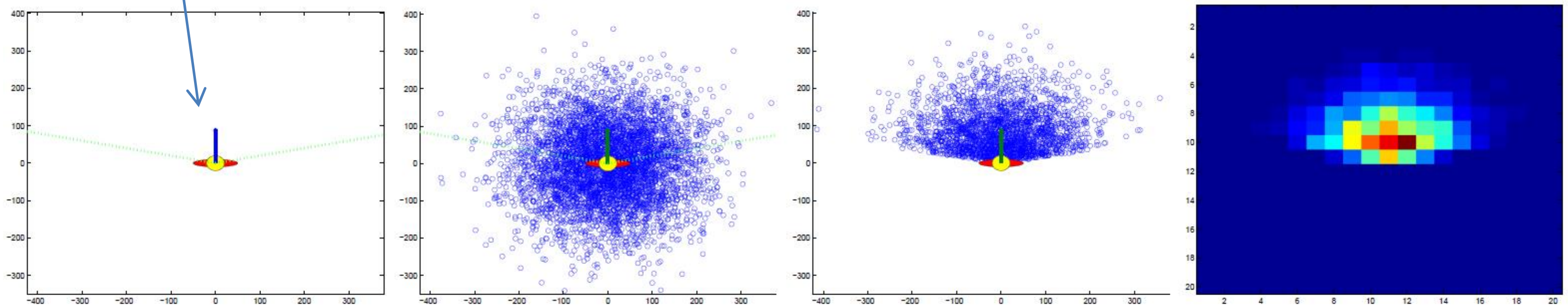
$\mathcal{N}_x(0, l)\ \mathcal{N}_y(0, l)$

$l$ Is the length of the Frustum

Valid Samples

VFoV: 120º
Other senses: 20º for each side

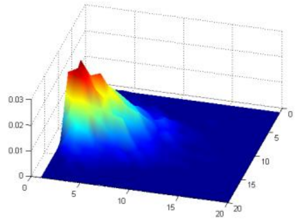
Normalized 2D histogram of the samples.
20x20 grid



iit PAVIS

# The method – Step 1
## Frustum

- A frustum implicitly embeds:
  - Spatial position of each person
  - Biological area in which interactions may occurs
  - Each histogram’s cell represents the probability of having a conversation in that location



# The method – Step 2
## Quantify Pairwise Interaction

- A frustum is a normalized 2D histogram representing the density of the feasible samples of a person in a scene.

- Given two persons in a scene the intersection of their frustum gives us a measure of the probability of having an interaction between them.

- Distances from Information-theory domain provides a measure to evaluate it.

# The method – Step 2
## Quantify Pairwise Interaction

- Given two histograms $P$ and $Q$ their distance is:

| Kullback-Leibler divergence (A-Sym) | Jensen-Shannon divergence (Sym) |
|---|---|
| $KL(P \,\|\|\, Q) = \sum_{i=1}^{n} \left( \log(p_i) \frac{p_i}{q_i} \right)$ | $JS(P,Q) = \frac{KL(P\|\|\,M) + KL(Q\|\|\,M)}{2}$ <br> $M = \frac{1}{2}(P+Q)$ |

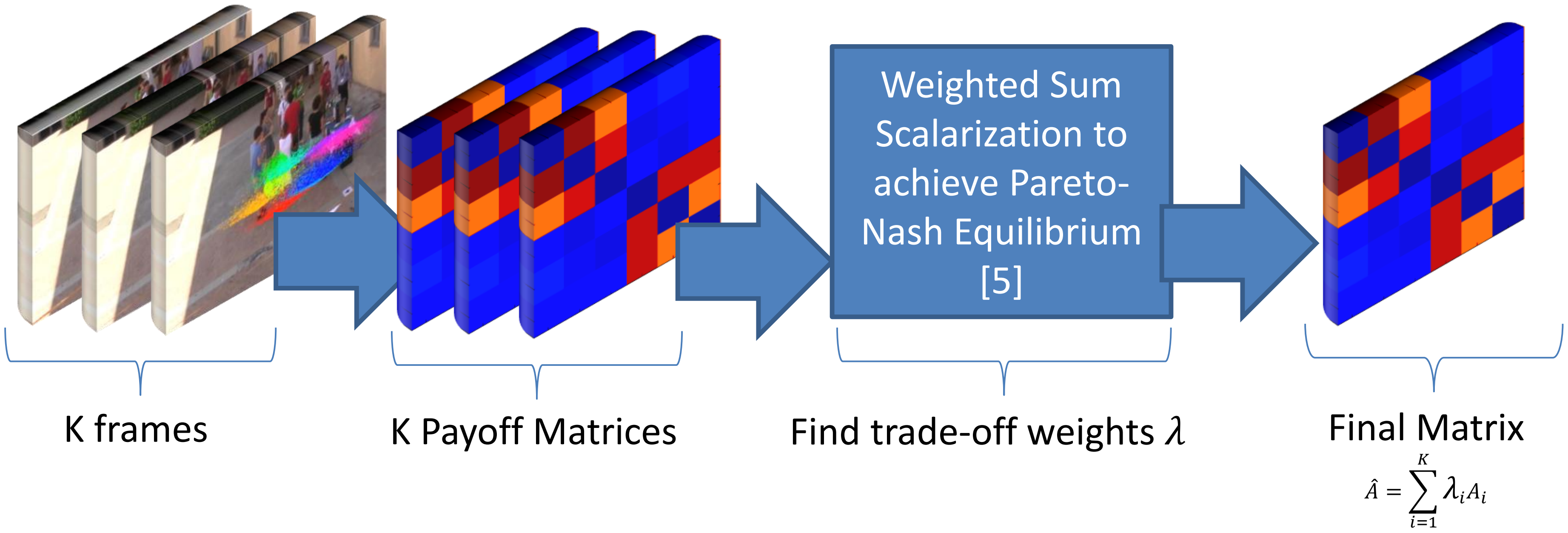
- A measure of affinity is obtained through a Gaussian Kernel

$$a_{P,Q} = exp\left\{ -\frac{d(P,Q)}{\sigma} \right\}$$

where P,Q are the frustum of two persons, *d*(…) could be either KL or JS and σ act as normalization term.

# The method – Step 3
## Temporal integration as a Multi-Payoff Games

- Integrate different temporal instants (frames) to **smooth unreliable detections**. Each frame is represented as a Payoff Matrix. If K frames are available the game has Multiple-Payoff.




[5] Somasundaram, K., Baras, J.S.: Achieving symmetric Pareto Nash equilibria using biased replicator dynamics. In: 48th IEEE Conf. Decision Control. (2009) 7000–7005

# The method – Step 4
## Grouping as a non-cooperative game

- A clustering method [6] rooted in the evolutionary game-theory [7].

- Given a set of elements $O = \{1 \dots n\}$ (*pure strategies*), an $n \times n$ affinity matrix $A_{ij}$ (*payoff matrix*) the aim is finding the Evolutionary Stable Strategy $\boldsymbol{x} = (x_1 \dots x_n)^T \in \Delta^n$ that **maximize** the expected payoff $\mathrm{u}(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$

- The ESS is found [6,7] iterating the Replicator Dynamics on the vector $\boldsymbol{x}$ initialized on the barycenter of the $\Delta^n$

$$x_i(t+1) = x_i(t) \frac{(A\boldsymbol{x}(t))_i}{\boldsymbol{x}(t)^T A \boldsymbol{x}(t)}$$

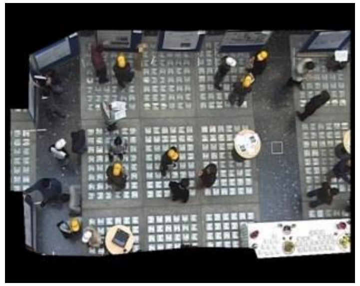
- At convergence of the RD, the support of $\boldsymbol{x}$ correspond to a group.

- The group is removed from the set of elements *O* and the RD are iterated again on the remaining elements.

[6] Torsello, A., Rota Bulo, S., Pelillo, M.: Grouping with asymmetric affinities: A game theoretic perspective. CVPR 2006.
[7] Weibull, J.W.: Evolutionary Game Theory. MIT Press, Cambridge, MA (2005)

# Experiments

| Dataset | #Sequences | #Frames × seq. | Consecutive Frames | Automated Tracking |
|---|---|---|---|---|
| CoffeeBreak | 2 | 45,74 | Y | Y |
| CocktailParty | 1 | 320 | Y | Y |
| GDet | 5 | 132,115,79,17,60 | N | Y |
| PosterData | 82 | 1 | N | N |
| Synth | 10 | 10 | N | N |

- Evaluation criteria:

  A group is correctly detected if at least $\left\lceil \frac{2}{3}|G| \right\rceil$ of its members matches the ground truth [8]

- Metrics: *Precision, Recall, F1-Score* (averaged over the frames)

[8] Setti, F., Hung, H., Cristani, M.: Group Detection in Still Images by F-formation Modeling: a Comparative Study. In: WIAMIS. (2013)

# Results
## Single Frame analysis

- **Aim:** Detect groups in still images

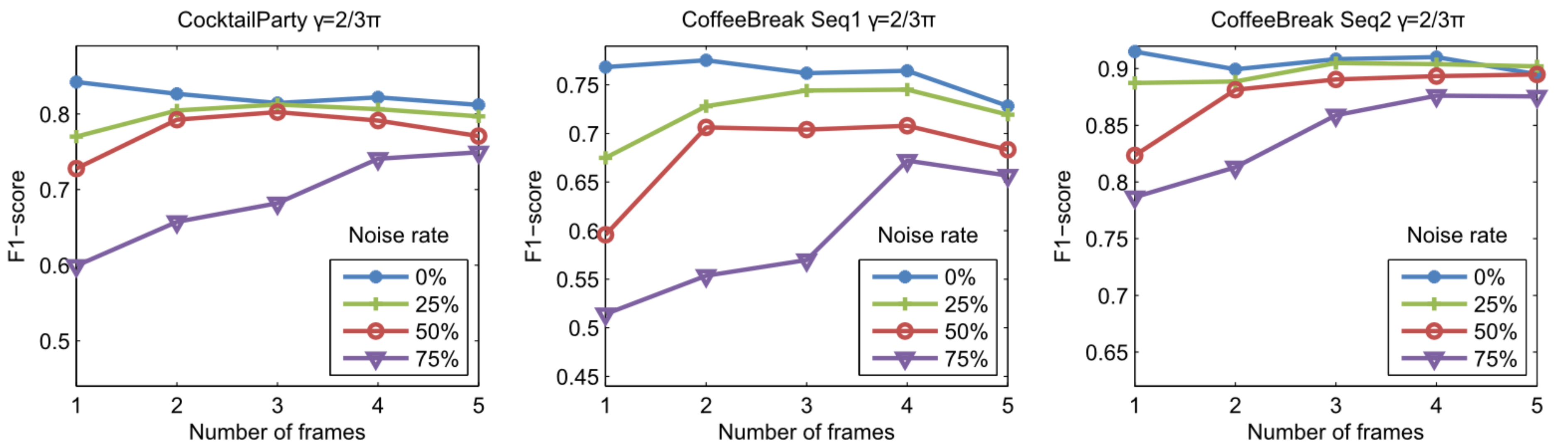
| | CoffeeBreak (S1+S2) | | | PosterData | | | Gdet | | |
|---|---|---|---|---|---|---|---|---|---|
| *Method* | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 |
| HFF | 0,82 | 0,83 | 0,82 | **0,93** | **0,96** | **0,94** | 0,67 | 0,57 | 0,62 |
| DS | 0,68 | 0,65 | 0,66 | **0,93** | 0,92 | 0,92 | - | - | - |
| MULTISCALE | 0,82 | 0,77 | 0,80 | - | - | - | - | - | - |
| *Our* KL | 0,80 | 0,84 | 0,82 | 0,90 | 0,94 | 0,92 | **0,76** | **0,75** | **0,75** |
| | $\sigma$=0.2 , $l$=40 | | | $\sigma$=0.2 $l$=30 | | | $\sigma$=0.5 $l$=80 | | |
| *Our* JS | **0,83** | **0,89** | **0,86** | 0,92 | **0,96** | **0,94** | **0,76** | **0,76** | **0,76** |
| | $\sigma$=0.2 , $l$=50 | | | $\sigma$=0.3 , $l$=25 | | | $\sigma$=0.5 $l$=80 | | |

| | Cocktail Party | | | Synth | | |
|---|---|---|---|---|---|---|
| *Method* | Prec | Rec | F1 | Prec | Rec | F1 |
| HFF | 0,59 | 0,74 | 0,66 | 0,73 | 0,83 | 0,78 |
| MULTISCALE | 0,69 | 0,74 | 0,71 | 0,86 | 0,94 | 0,90 |
| *Our* KL | **0,85** | **0,81** | **0,83** | **1,00** | **1,00** | **1,00** |
| *Our* JS | **0,86** | **0,82** | **0,84** | **1,00** | **1,00** | **1,00** |
| | $\sigma$=0.5 , $l$=60 | | | $\sigma$=0.1 , $l$=30 | | |

- Parameter search: $\sigma$=[0.1 : 0.9] , l=[20 : 150]
- Maximum variance for precision and recall ~0,74%

# Results
## Multi-Frame analysis

- **Aim:** detect groups in a window of K-frames under noise condition.




- Parameter search: $K = \{1,2,3,4,5\}$
- Performance in noisy conditions: $\gamma = \left\{\frac{\pi}{8}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{2\pi}{3}\right\}$ $N = \{0, 25, 50, 75\}\%$

- Mean standard deviation for the precision is 1.61% and for recall is 1.73%

# Conclusions

- Method strengths:
  - Based on sociological and biological constraints
  - No assumption on the size or shape of the F-F
  - Designed to cope with very different realistic scenarios
  - Work on top of any tracker or person detection algorithms
  - Rooted in the Evolutionary Game Theory, a strong mathematical framework to analyze behavior in populations
  - Robust to noise using principled from Multi-Payoff game
  - State of the art in all public available datasets.

- Method weaknesses:
  - Pairwise Affinity matrix does not scale on thousands of detections per frame (but It is an uncommon situation)
  - Groups are detected per frame, no tracking still exploited

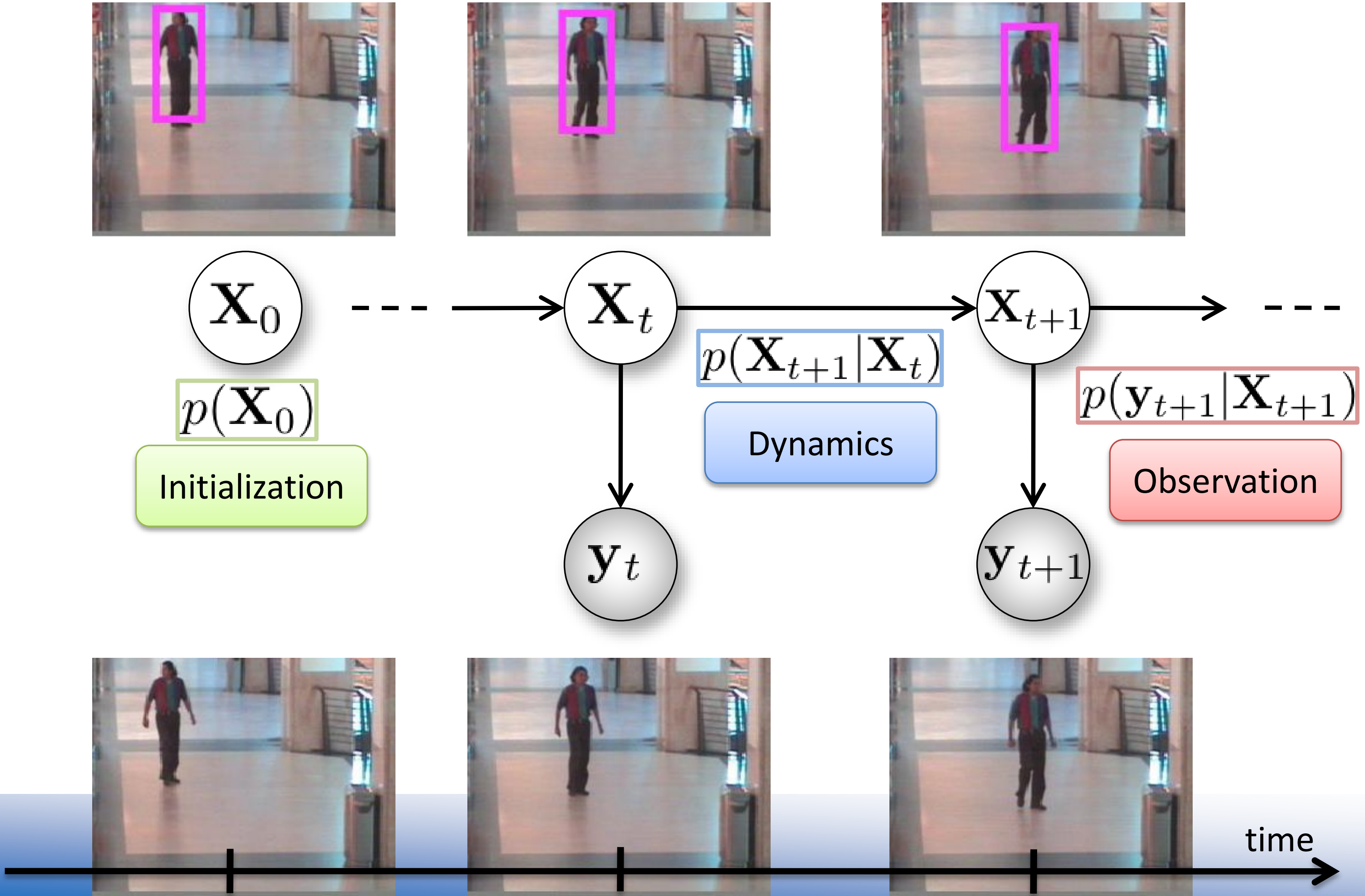
# Group Tracking

# Social behavior analysis

- **Goal**: model human interactions to better understand their **social behavior** and dynamics
- Focus on **group** modeling and tracking
- Why it is hard:
  - Highly non-linear dynamics
  - Non-atomic entities: split and merge
  - Appearance changes quickly
- Modeling **jointly** the tracking of individuals and groups

L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. *CVPR* 2012.
M. Zanotto, L. Bazzani, M. Cristani, and V. Murino. Online bayesian non-parametrics for social group detection. *BMVC* 2012.

# Tracking



$\mathbf{X}_0$
$\mathbf{X}_t$
$\mathbf{X}_{t+1}$
$p(\mathbf{X}_0)$
Initialization
$\mathbf{y}_t$
$p(\mathbf{X}_{t+1}|\mathbf{X}_t)$
Dynamics
$p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1})$
Observation
$\mathbf{y}_{t+1}$
time

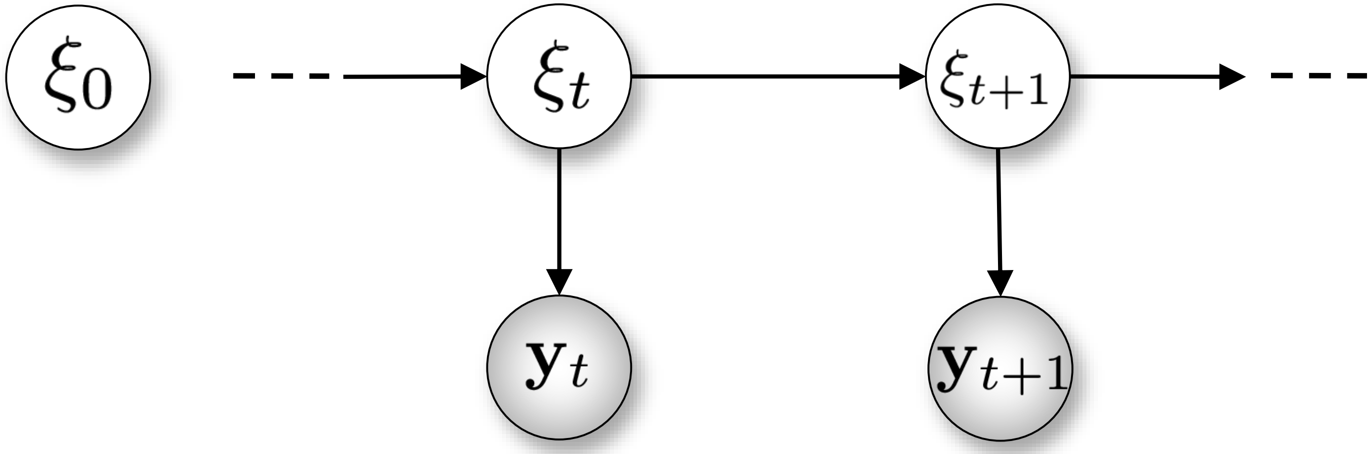
# Joint Individual-Group Tracking

Joint state

$$\xi_t = [\boldsymbol{\Theta}_t, \mathbf{X}_t]$$

Non-linear discrete-time systems

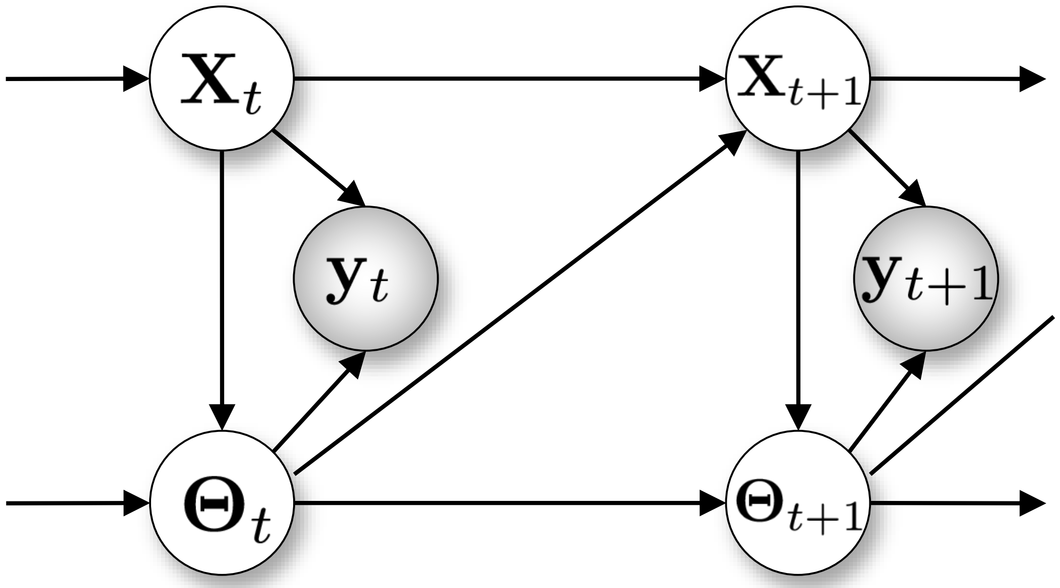
$$\xi_{t+1} = f_t(\xi_t, \eta_t^{\xi}),$$
$$\mathbf{y}_t = h_t(\xi_t, \eta_t^{y})$$



# The Proposed Model for Joint Individual-Group Tracking

$$\mathbf{X}_{t+1} = f_t^X(\mathbf{X}_t, \quad \eta_t^X),$$

$$\mathbf{y}_t = h_t(\mathbf{X}_t, \quad \eta_t^y).$$




L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. *CVPR* 2012.

# Group Modeling

- Group modeling is seen as a problem of **mixture model** fitting
- Mixture model
  - Each group corresponds to a **component** of the mixture
  - Each individual is an **observation** drawn from the mixture
- Gaussian mixture model?
  - No, fixed number of components
- **Dirichlet process** mixture model
  - Potentially **infinite** number of components
- # groups not fixed and may change over time
- Allow probabilistic *soft* assignments (of individuals to groups)

# Qualitative Results

Joint Individual-Group Modeling for Tracking.mp4

# CROWD
## detecting abnormal behaviors

# Abnormality Detection

Examples of abnormalities in crowd


**Walking against crowd**


**Panic**

**Violence**

**Going faster**

- Issues:
  - Heavy occlusion, view points, background clutter, low quality video, etc.
  - Ambiguous definition of abnormal behaviours (context dependent)
  - Lack of adequate abnormal samples (e.g., riots) for model training

# Existing Approaches

- Object-based approaches: detecting and tracking objects and individuals to model motions and interactions
  - Object segmentation and shape estimation [Rittscher et al, cvpr 2005]
  - Counting crowded moving objects [Rabaud et al, CVPR2006]
  - Trajectory-based anomalous event detection [Piciarelli et al, TCSVT2008]
  - Pedestrian agents [Zhou et al., CVPR 2012]
  - …
- Holistic approaches: no object/individual detection and tracking, extracting global motions from the entire scene
  - Optical flow histograms [Krausz et al, ICCV 2011]
  - Social Force Models [Mehran et al CVPR, 2009]
  - Spatial-Temporal Grids [Kratz et al, CVPR 2010]
  - Crowd collectiveness [Zhou et al., CVPR 2013]
  - …

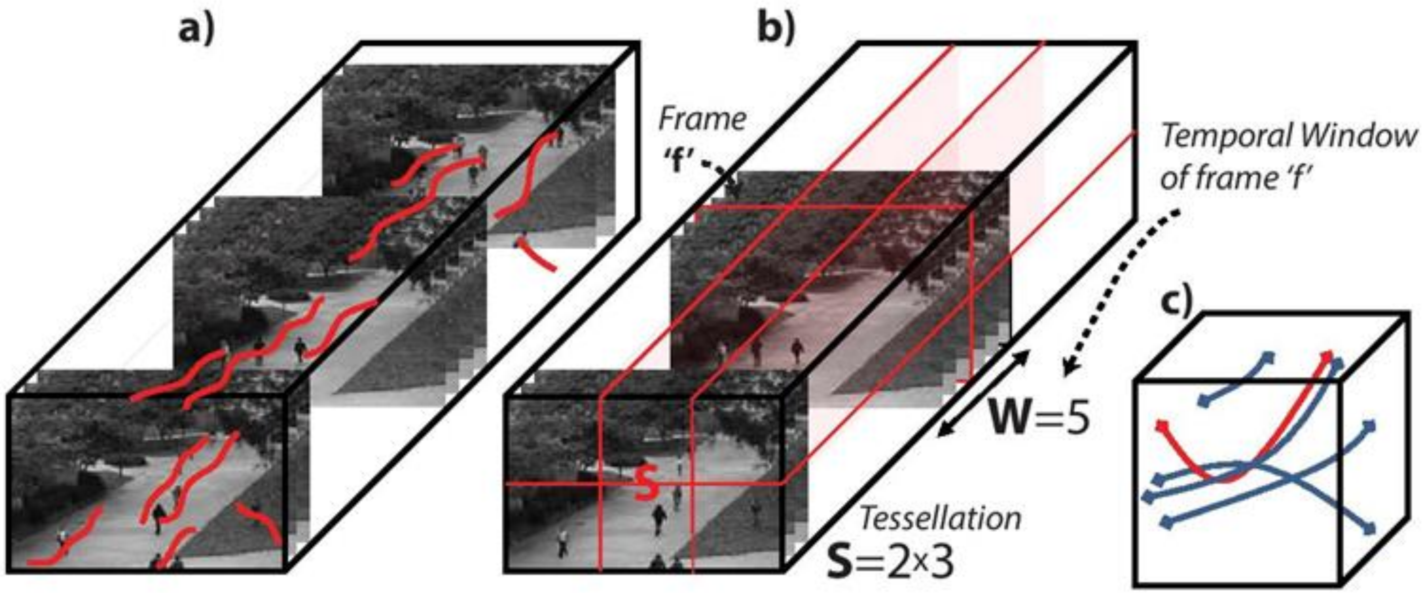
# Our proposed approaches

1. Histogram of Oriented Tracklets, HOT [WACV 2015]
2. Improved HOTs, iHOT [ICIAP 2015]
3. Commotion measure [ICIP 2015]


1. H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, V. Murino, “Analyzing Tracklets for the Detection of Abnormal Crowd Behavior”, *IEEE Winter Conference on Applications of Computer Vision WACV 2015*.
2. H. Mousavi, M. Nabi, H.K. Galoogahi, A. Perina, V. Murino, “Abnormality detection with improved histogram of oriented tracklets”, *18th Int’l Conf. on Image Analysis and Processing ICIAP 2015*.
3. H. Mousavi, M. Nabi, H.K. Galoogahi, A. Perina, V. Murino, “Crowd Motion Monitoring Using Tracklet-based Commotion Measure”, *Int’l Conf. on Image Processing ICIP 2015*.

# Histogram of Oriented Tracklets

a) Tracking interest points over **T** frames to compute tracklets
b) Subdividing the video in spatio-temporal cuboids
c) Computing motion statistics of all trajectories passing through each 3D cuboid

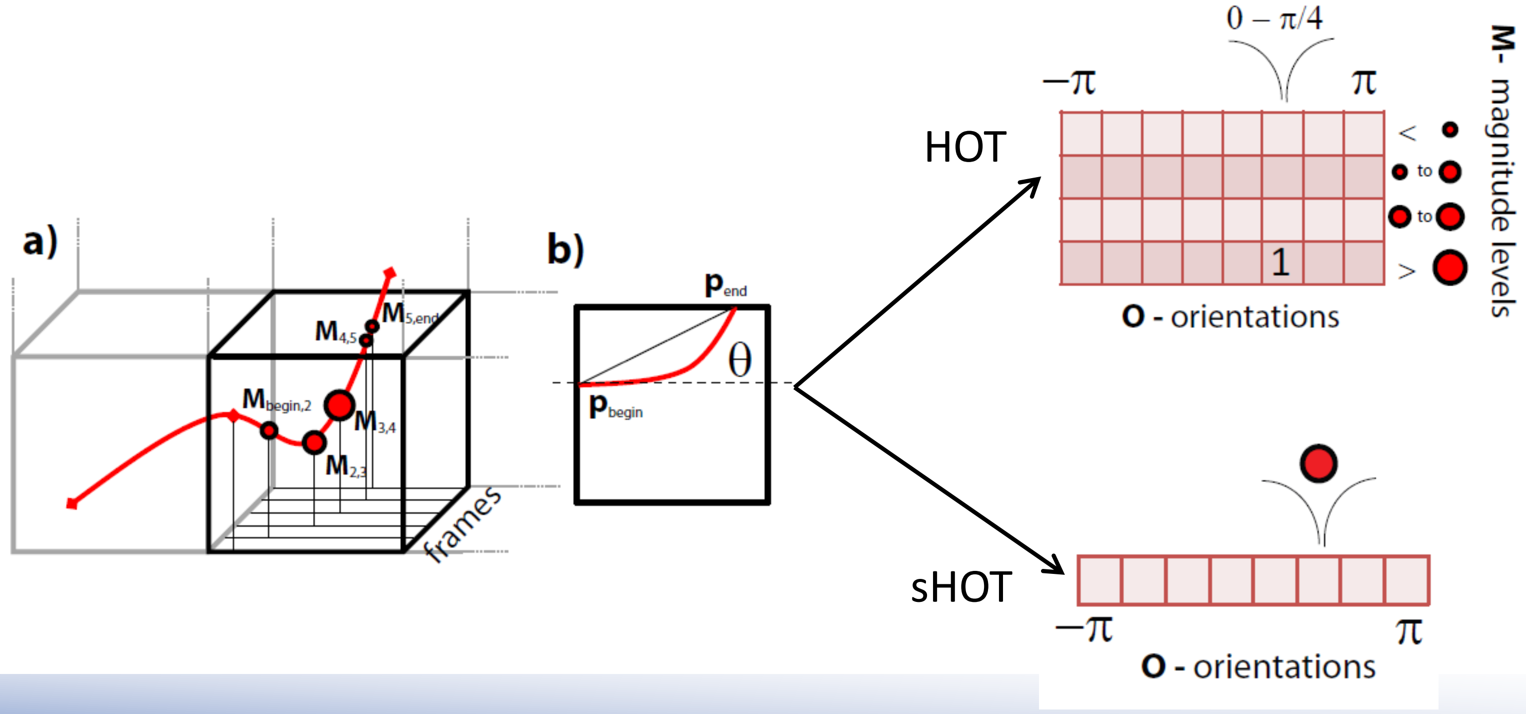



H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, V. Murino, "Analyzing Tracklets for the Detection of Abnormal Crowd Behavior", *IEEE Winter Conference on Applications of Computer Vision WACV 2015*.

# Statistics of motion

For each 3D cuboid:

1. Compute magnitude and orientation of each tracklet passing through the cuboid
2. Quantize all magnitudes and orientations of tracklets across the cuboid to form a 2D or 1D (simplified) **H**istogram of **O**riented **T**racklets (HOT).



# Detection strategies

- Learning by **generative (LDA)** or **discriminative (SVM) models:** training and test phases accordingly
- **Full bag of words – BW :**

  HOT descriptors are summed across sectors (patches)

  $$D^f = \sum_s H_{o,m}^{s,f} \qquad and \qquad D = \{D^f\}_{f=1}^F$$

- **Per-frame, Per-sector – FS :**

  HOTs from all the different sectors are concatenated in a single descriptor

  $$D^f = \left\{H_{o,m}^{1,f} \middle| H_{o,m}^{2,f} \middle| \ldots \middle| H_{o,m}^{s,f}\right\} \qquad and \qquad D = \{D^f\}_{f=1}^F$$

- **Per-frame, Per-independent-sector – FiS :**

  Learn an independent Latent Dirichlet Allocation (LDA) model per-sector

  $$D^f = H_{o,m}^{s,f}$$

iit PAVIS

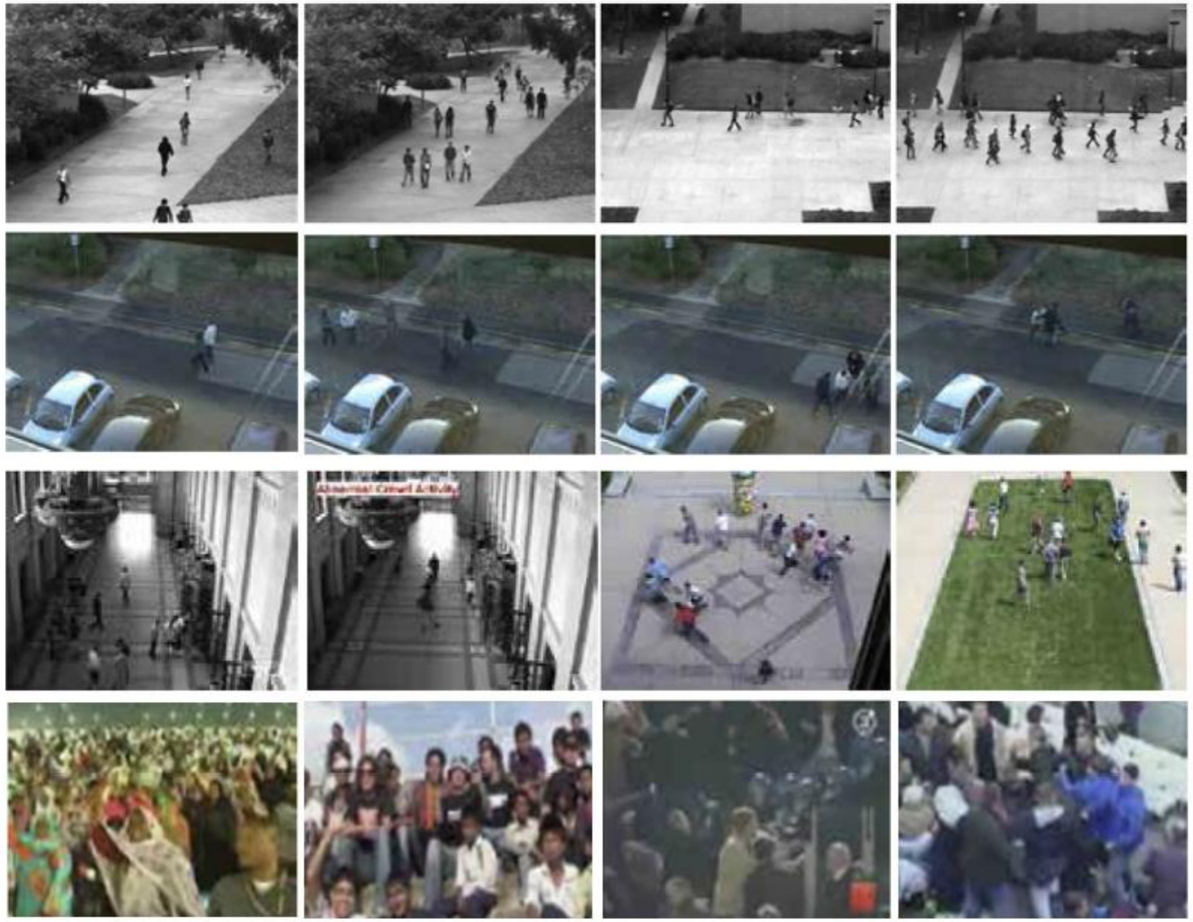
# Experiments: datasets

UCSD

Behave

UMN

Violence



Only normal situations in training

Only normal situations in training

# Experimental results

- Learning by generative (LDA) or discriminative (SVM) models, when possible
- Evaluating on semi-crowded (UCSD) and dense crowded datasets (Violence In Crowd)
- Comparing with the social force model (Mehran et al, CVPR'09) and the other state of the art methods

LDA

| UCSD | EER-ped1 | EER-ped2 |
|---|---|---|
| Dynamic-texture | 22.9 % | 27.9 % |
| Social Force Model | 36.5 % | 35.0 % |
| **Hist. Orient. Tracklets** | 20.49% | 21.20 % |

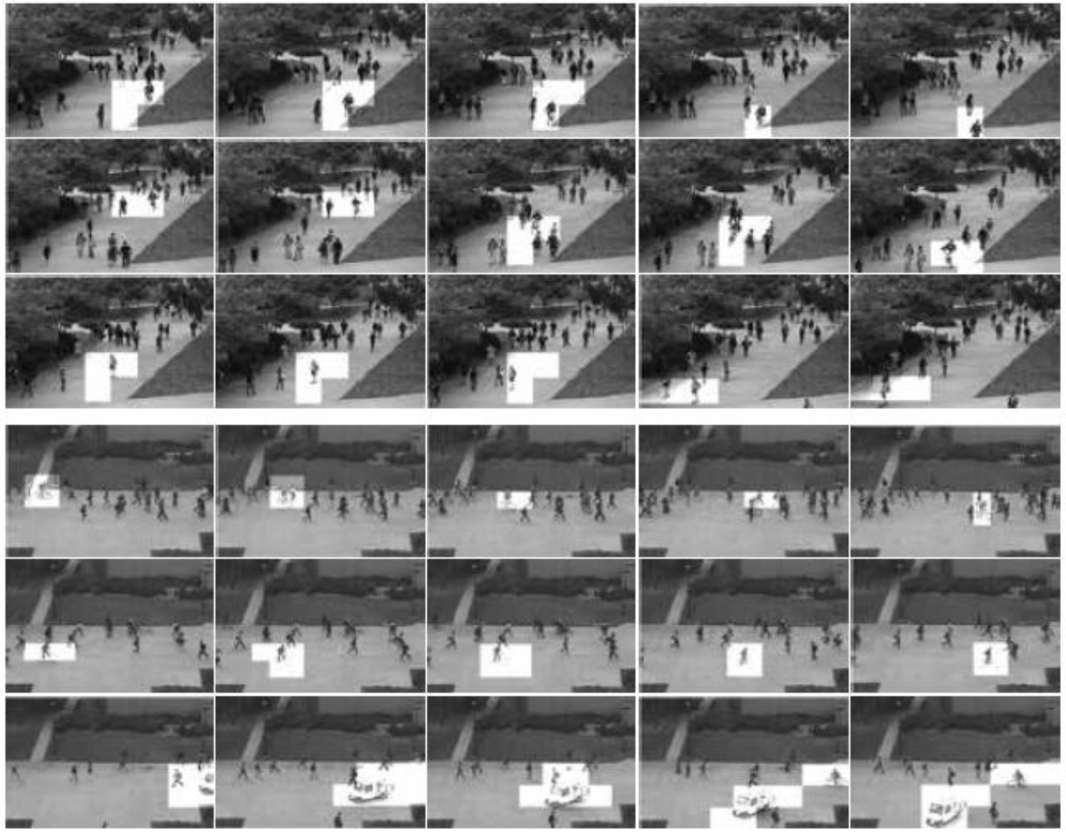

5-fold cross-validation SVM

| Violence in Crowds | Accuracy |
|---|---|
| Violent Flows | 81.30 % |
| Social Force Model | 80.45 % |
| **Hist. Orient. Tracklets** | 82.30 % |

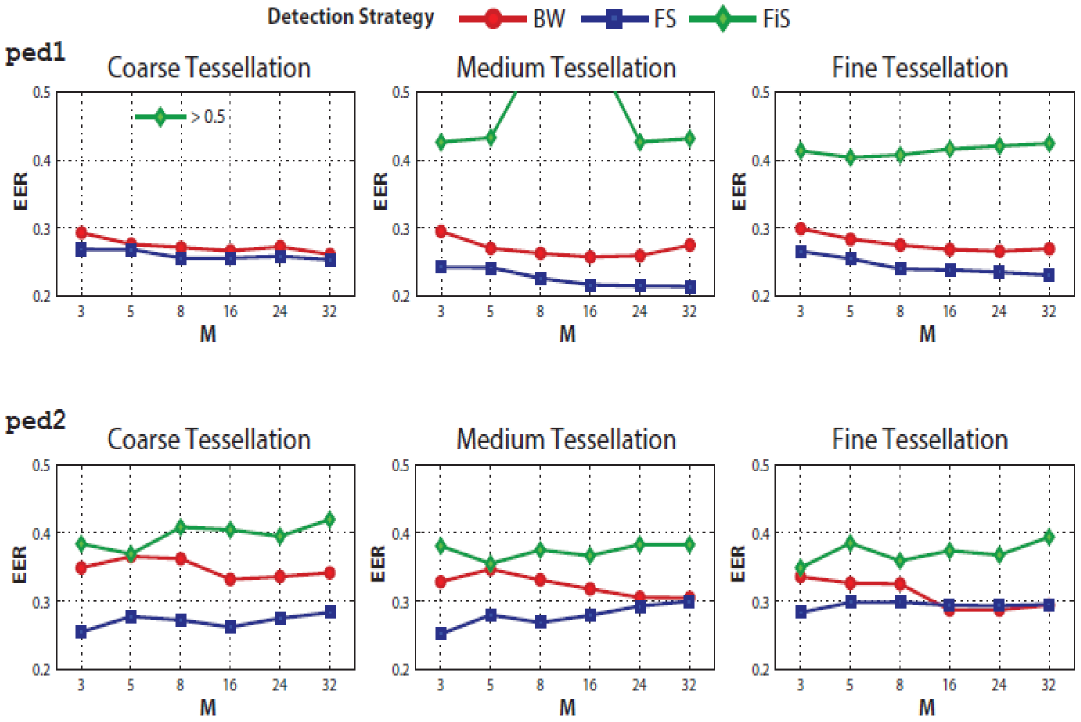
W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. TPAMI 2014.
R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. CVPR 2009.

# Experimental results

## Localization at cuboid level



## Approach robust to quantization and tesselation



# In summary …

- Robust to quantization levels
- Robust to tessellation size
- Localization possible at cuboid level
- Can be used with generative (Latent Dirichlet Allocation, LDA) or discriminative (SVM) models
- Robust to LDA number of topics

# CROWD behavior

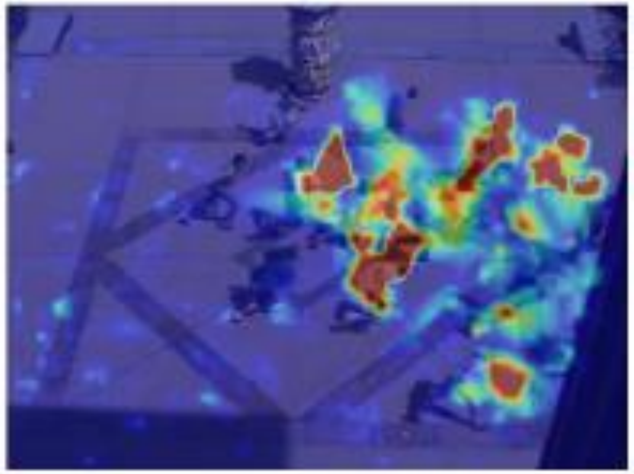
# Violence Detection using Substantial Derivative

# Abnormality/Violence Detection

# A popular approach

- **Physics based approach**
  (e.g., [R.Mehran et al., CVPR09])

Video Frame



Social Force Model

✓ Easy for simulating crowd behavior

✗ Too simple to reveal wide range of crowd dynamics in a real scenarios

R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. CVPR 2009.

# Motivations

- Physics-inspired approaches such as Social Force Model (SFM) have been succesfully employed to detect abnormality in crowd scenarios **[Mehran et al, CVPR09]**
- As major drawback these methods are not able to capture the whole range of abnormal patterns
- Actually, phyics-based approaches have considered *temporal* information as a main source of information
- However, sociological studies show that *structure of motion* has a significant effect on pedestrain behaviors in crowded scenes **[W. Chao, and T. Li , ICCCI11]**

R. Mehran, A. Oyama, M.Shah, "Abnormal crowd behavior detection using social force model" CVPR 2009
W.Chao, and T.Li "Simulating Riot for Virtual Crowds with a social communication Model" ICCCI 2011

# Substantial Derivative

Consider a velocity vector $\boldsymbol{U} = U(\boldsymbol{P},t)$ at a location $\boldsymbol{P} = (x, y)$ and time $t$, the acceleration of objects moving through a velocity field can be described as:

$$\frac{\mathbf{DU}}{\mathrm{D}t} = \frac{\partial \mathbf{U}}{\partial t} + \left(u\frac{\partial \mathbf{U}}{\partial x} + v\frac{\partial \mathbf{U}}{\partial y}\right) = \mathbf{U}_t + (\nabla \mathbf{U})\mathbf{U} \;\rightarrow \mathbf{F}^{\mathrm{L}} + \mathbf{F}^{\mathrm{Cv}}$$

where

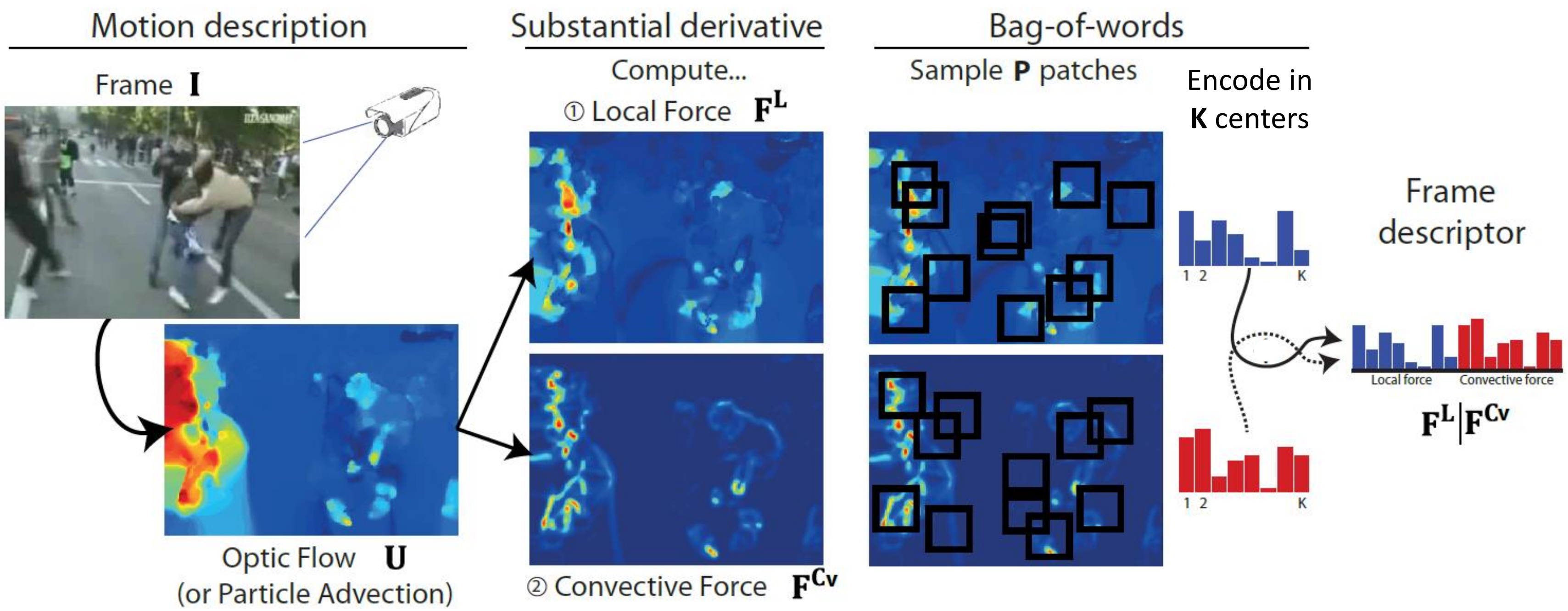
- $\frac{\mathbf{DU}}{\mathrm{D}t}$ : ***substantial derivative*** or *total acceleration* of certain particle in fluid
- $\mathbf{U}_t$: ***local acceleration*** rate of change $\mathbf{U}$ at the temporal domain
- $(\nabla \boldsymbol{U})\boldsymbol{U}$ : ***convective acceleration*** explains spatial variation of velocity field

# Substantial Derivative

**Properties:**

- *Local acceleration*
  - Occurs when the flow is unsteady
  - Useful to capture instant velocity changes in crowd
- *Convective acceleration*
  - Occurs when the flow is non-uniform
  - Useful to capture structural motion change in crowd

# Overview of the method proposed



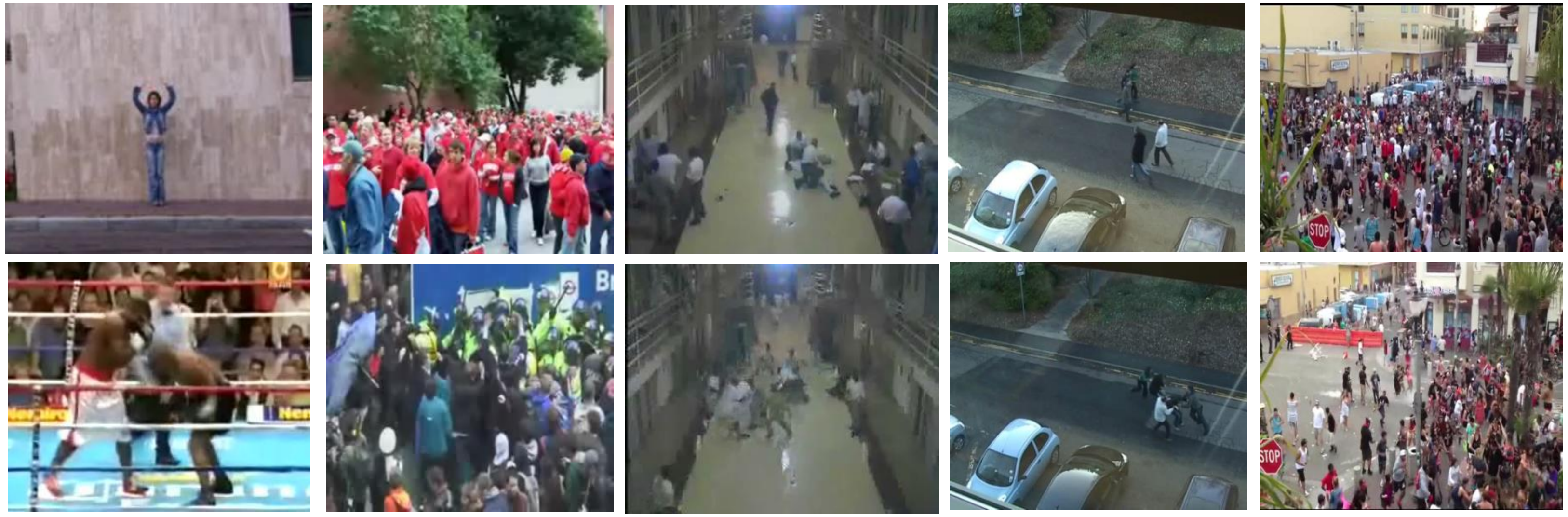
# Experimental Results: Datasets

Five different datasets are selected for evaluation purpose

Violence in Movies | Violence in Crowd | Riot In Prison | Behave | Panic

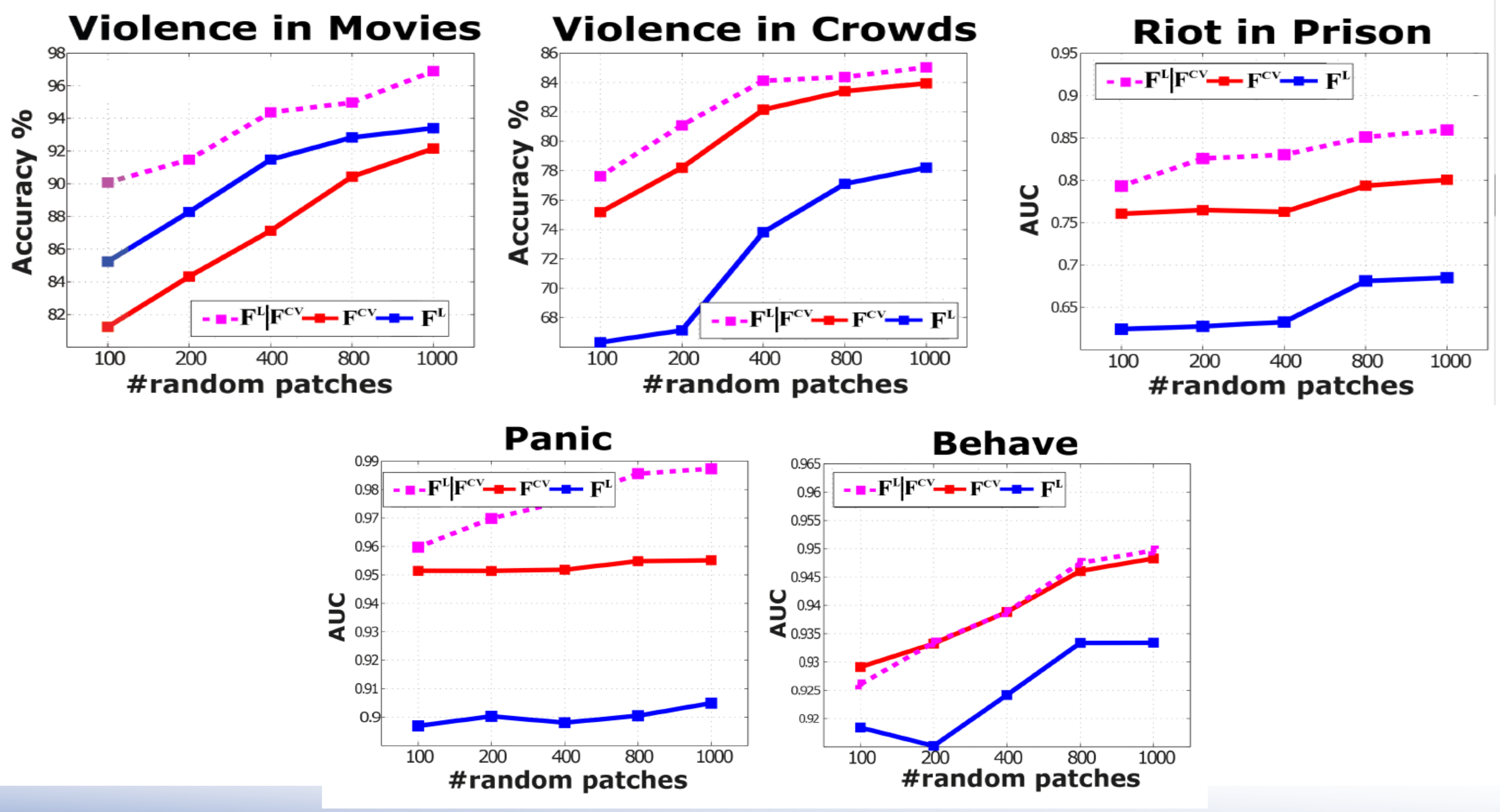
Normal

Abnormal



# Effect of Number of Random Sample Patches

**Number of random patches varies in the range of** $P \in \{100, 200, 400, 800, 1000\}$



# Comparison with State-of-the-Art methods:
# Violence in Movies

**95% confidence interval using SVM**
**With 5-fold cross validation**

| Method | Accuracy |
|---|---|
| STIP(HOF) | 50.5% |
| MoSIFT | 89.5% |
| Optical Flow | 91.31±1.06% |
| Interaction Force | 95.51±0.79% |
| Jerk | 95.02±0.56% |
| Local Force $F^L$ | 93.4±1.24% |
| Convective Force $F^{Cv}$ | 92.16±1.13% |
| $F^L \| F^{Cv}$ | **96.89±0.21%** |

**Normal** **Fight**

# Comparison with State-of-the-Art methods:
# Violence in crowd

Average accuracy with 95% confidence interval using SVM, with 5-fold cross validation

| Method | Accuracy |
|---|---|
| HOT | 82.3% |
| LTP | 71.53±0.15% |
| Optical Flow | 79.38±0.14% |
| Interaction Force | 81.30±0.18% |
| Jerk | 74.05±0.65% |
| Local Force $F^L$ | 78.14±0.92% |
| Convective Force $F^{Cv}$ | 84.03±1.34% |
| $F^L \| F^{Cv}$ | **85.43±0.21%** |

**Normal** **Violence**

# Comparison with State-of-the-Art methods:
## Riots in prison

AUC with 95% confidence interval using LDA

| Method | Riot in Prison | Panic |
|---|---|---|
| Optical Flow | 0.76±0.052 | 0.89±0.0136 |
| Interaction Force | 0.66±0.024 | 0.89±0.004 |
| Jerk | 0.65±0.036 | 0.90±0.009 |
| Local force $F^L$ | 0.68±0.027 | 0.90±0.0079 |
| Convective Force $F^{Cv}$ | 0.79±0.014 | 0.95±0.0023 |
| $F^L \mid F^{Cv}$ | **0.85±0.077** | **0.98±0.0055** |

**Riot in Prison** **Panic**

**Normal**

**Abnormal**

# Comparison with State-of-the-Art methods: Behave

AUC with 95% confidence interval using LDA

| Method | AUC |
|---|---|
| Optical flow | 0.901±0.032 |
| Interaction Force | 0.925±0.008 |
| Local force $F^{L}$ | 0.933±0.073 |
| Convective Force $F^{Cv}$ | 0.946±0.032 |
| $F^{L} \| F^{Cv}$ | **0.948±0.054** |

**Normal** **Abnormal**

# Summary

- Novel computational framework based on spatial-temporal characteristics of substantial derivative to detect act of violence in crowd

- Spatial information captured from *convective acceleration* <u>mainly</u> has significant effect to detect violence in crowd scenarios.

- Robustness of the proposed method has been proven in various abnormal situations such as panic

# Conclusions & Take-Home Message

- Groups and crowd behavior analysis cannot be faced by pure CV approaches only
- Heuristics and cognitive approaches needed, psychology and sociology findings must be taken into account
- Need of high-level models but strong necessity of reliable and robust *low-level* algorithms (for detection, tracking, orientations)
- Motion pattern modeling has strong relevance, descriptors should be able to finely capture people movements, locally and globally
- Learning models seem not to have a high relevance to date, but actual capabilities are still to be fully exploited

# Acknowledgments & Credits


- Marco Cristani
- Loris Bazzani
- Sebastiano Vascon
- Sadegh Mohammadi
- Hossein Moussavi
- Hamed Kiani Galoogahi
- Chiara Bassetti