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

Vittorio Murino
Analysing groups and crowds...
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)
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

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

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

<table>
<thead>
<tr>
<th>Small gathering (2 to 6)</th>
<th>Medium gathering (7 to 12/30)</th>
<th>Large gathering (13/31 to N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Occurring in private, semi-public and public places</em></td>
<td><em>Occurring in private but mostly semi-public/public places</em></td>
<td><em>Occurring in semi-public but mostly in public places</em></td>
</tr>
</tbody>
</table>

- **private places**: home, private garden, car
- **semi-public places**: classroom, office, club, party area
- **public places**: open plaza, transportation, station, walkway, park, street
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...)
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)
Medium gathering (7 to 12-30)

Line at the post office (unfocused)

Classroom group, touring group at the museum (common-focused)

Meeting group, extended family commensal (jointly-focused)

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

In these cases, small gatherings of other typologies of gathering may be present: difficult to catch/model but important to individuate.
Large gathering (13-31 to N)

- line at the airport check in, walking in a street (unfocused)

- sport/theatre/cinema spectators (common-focused)

- mob/riot/sit-in/march participants (common and jointly-focused)

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)*

<table>
<thead>
<tr>
<th></th>
<th>Unfocused</th>
<th>Common focused</th>
<th>Jointly focused</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Static</strong></td>
<td><img src="image1" alt="Unfocused" /></td>
<td><img src="image2" alt="Common focused" /></td>
<td><img src="image3" alt="Jointly focused" /></td>
</tr>
<tr>
<td><strong>Dynamic</strong></td>
<td><img src="image4" alt="Unfocused" /></td>
<td><img src="image5" alt="Common focused" /></td>
<td><img src="image6" alt="Jointly focused" /></td>
</tr>
</tbody>
</table>

*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-focused)

Spectator crowd
• flash-mob, Mass, sport supporters (jointly-focused - Expressive crowd)
• mob/riot/sit-in/march (common&jointly-focused - Protest/Acting crowd)

stronger social relations
harder to model
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

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Easy frames</th>
<th>Hard frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDIAP Poster</td>
<td><img src="image1" alt="IDIAP Poster" /></td>
<td><img src="image2" alt="IDIAP Poster" /></td>
</tr>
<tr>
<td>Cocktail Party</td>
<td><img src="image3" alt="Cocktail Party" /></td>
<td><img src="image4" alt="Cocktail Party" /></td>
</tr>
<tr>
<td>Coffee Break</td>
<td><img src="image5" alt="Coffee Break" /></td>
<td><img src="image6" alt="Coffee Break" /></td>
</tr>
<tr>
<td>GDet</td>
<td><img src="image7" alt="GDet" /></td>
<td><img src="image8" alt="GDet" /></td>
</tr>
</tbody>
</table>
Challenges

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

Slide credit: M. Cristani
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
  o unfocused
  o common focused
  o 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
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]
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]
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
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 \( \left\lceil \frac{2}{3} \cdot |G| \right\rceil \) of their individuals have been selected.
  – Compute **precision** and **recall**

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

\[
\begin{pmatrix}
G \cdot \frac{3}{2} \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
0 & 0 \\
+1 & 0 \\
0 & 0 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
i & j \\
i & j \\
\end{pmatrix}
\]
• 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Mantel test</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRPM</td>
<td>0.55</td>
<td>0.19</td>
<td>0.67</td>
</tr>
<tr>
<td>Our approach</td>
<td>0.85</td>
<td>0.76</td>
<td>0.76</td>
</tr>
</tbody>
</table>
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.

---

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
The method – Step 1

**Frustum**

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

\[(x, y, \theta)\]

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

\(l\) is the length of the Frustum.

All Samples

Valid Samples

VFoV: 120º

Other senses: 20º for each side

Normalized 2D histogram of the samples.

20x20 grid
Our method - Frustum

- A frustum implicitly embeds:
  - Spatial position of each person
  - Biological area in which interactions may occur
  - Each histogram's cell represents the probability of having a conversation in that location
The method – Step 1
Frustum

- A frustum implicitly embeds:
  - Spatial position of each person
  - Biological area in which interactions may occur
  - 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:

  
  \[
  KL(P \parallel Q) = \sum_{i=1}^{n} \left( \log(p_i) \frac{p_i}{q_i} \right)
  \]

  
  \[
  JS(P, Q) = \frac{KL(P \parallel M) + KL(Q \parallel M)}{2}
  \]

  
  \[
  M = \frac{1}{2} (P + Q)
  \]

- 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 $\sigma$ 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.

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 $O = \{1 \ldots n\}$ (pure strategies), an $n \times n$ affinity matrix $A_{ij}$ (payoff matrix) the aim is finding the Evolutionary Stable Strategy $x = (x_1 \ldots x_n)^T \in \Delta^n$ that maximize the expected payoff $u(x) = x^T Ax$

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

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

- At convergence of the RD, the support of $x$ correspond to a group.

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

Experiments

- **Evaluation criteria:** A group is correctly detected if at least \( \frac{2}{3} |G| \) of its members matches the ground truth [8]

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

---

### Results

**Single Frame analysis**

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Prec</th>
<th>Rec</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFF</td>
<td>0.82</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>DS</td>
<td>0.68</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>MULTISCALE</td>
<td>0.82</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>Our KL</strong></td>
<td>0.80</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Our JS</strong></td>
<td>0.83</td>
<td>0.89</td>
<td>0.86</td>
</tr>
</tbody>
</table>

- **Parameter search:** $\sigma=[0.1 : 0.9]$, $l=[20 : 150]$
- **Maximum variance for precision and recall** $\sim 0.74\%$
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

Tracking

\[ X_0 \xrightarrow{p(X_0)} X_t \xrightarrow{p(X_{t+1}|X_t)} X_{t+1} \]

Initialization \quad Dynamics \quad Observation

\[ y_t \quad y_{t+1} \]

Time
Joint Individual-Group Tracking

Joint state

$$\xi_t = [\Theta_t, X_t]$$

Non-linear discrete-time systems

$$\xi_{t+1} = f_t(\xi_t, \eta_t^\xi), \quad y_t = h_t(\xi_t, \eta_t^y)$$
The Proposed Model for Joint Individual-Group Tracking

\[ X_{t+1} = f_t^X(X_t, \eta^X_t), \]

\[ y_t = h_t(X_t, \eta^y_t). \]
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, CVPR 2006]
  – Trajectory-based anomalous event detection [Piciarelli et al, TCSVT 2008]
  – 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]
  – 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]

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

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) Histogram of Oriented Tracklets (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)
  \[ D_f = \sum_s H_{0,m}^{s,f} \quad \text{and} \quad D = \{D_f\}_{f=1}^F \]

- **Per-frame, Per-sector – FS:**
  HOTs from all the different sectors are concatenated in a single descriptor
  \[ D_f = \{H_{0,m}^{1,f} | H_{0,m}^{2,f} | ... | H_{0,m}^{s,f}\} \quad \text{and} \quad D = \{D_f\}_{f=1}^F \]

- **Per-frame, Per-independent-sector – FiS:**
  Learn an independent Latent Dirichlet Allocation (LDA) model per-sector
  \[ D_f = H_{0,m}^{s,f} \]
Experiments: datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Images</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCSD</td>
<td><img src="image1.png" alt="UCSD images" /></td>
<td>Only normal situations in training</td>
</tr>
<tr>
<td>Behave</td>
<td><img src="image2.png" alt="Behave images" /></td>
<td></td>
</tr>
<tr>
<td>UMN</td>
<td><img src="image3.png" alt="UMN images" /></td>
<td>Only normal situations in training</td>
</tr>
<tr>
<td>Violence</td>
<td><img src="image4.png" alt="Violence images" /></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>UCSD</th>
<th>EER-ped1</th>
<th>EER-ped2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic-texture</td>
<td>22.9 %</td>
<td>27.9 %</td>
<td></td>
</tr>
<tr>
<td>Social Force Model</td>
<td>36.5 %</td>
<td>35.0 %</td>
<td></td>
</tr>
<tr>
<td>Hist. Orient. Tracklets</td>
<td>20.49%</td>
<td>21.20 %</td>
<td></td>
</tr>
</tbody>
</table>

5-fold cross-validation SVM

<table>
<thead>
<tr>
<th></th>
<th>Violence in Crowds</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent Flows</td>
<td></td>
<td>81.30 %</td>
</tr>
<tr>
<td>Social Force Model</td>
<td></td>
<td>80.45 %</td>
</tr>
<tr>
<td>Hist. Orient. Tracklets</td>
<td></td>
<td>82.30 %</td>
</tr>
</tbody>
</table>

Experimental results

Localization at cuboid level

Approach robust to quantization and tessellation
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

Polish Hooligan Amateur video
Courtesy of TVN, Poland
A popular approach

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

- Easy for simulating crowd behavior
- Too simple to reveal wide range of crowd dynamics in a real scenarios

---

Motivations

- Physics-inspired approaches such as Social Force Model (SFM) have been successfully 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, physics-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 pedestrian behaviors in crowded scenes [W. Chao, and T. Li, ICCCI11]
Substantial Derivative

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

$$\frac{D\mathbf{U}}{Dt} = \frac{\partial \mathbf{U}}{\partial t} + \left( u \frac{\partial \mathbf{U}}{\partial x} + v \frac{\partial \mathbf{U}}{\partial y} \right) = \mathbf{U}_t + (\nabla \mathbf{U}) \mathbf{U} \to \mathbf{F}_L + \mathbf{F}_{Cv}$$

where

– $\frac{D\mathbf{U}}{Dt}$: **substantial derivative** or **total acceleration** of certain particle in fluid

– $\mathbf{U}_t$: **local acceleration** rate of change $\mathbf{U}$ at the temporal domain

– $(\nabla \mathbf{U}) \mathbf{U}$: **convective acceleration** explains spatial variation of velocity field
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

Motion description
Frame $I$
Optic Flow $U$ (or Particle Advection)

Substantial derivative
1. Local Force $F_L$
2. Convective Force $F_Cv$

Bag-of-words
Sample $P$ patches
Encode in $K$ centers

Frame descriptor
Local force Convective force $F_L|F_Cv$
Experimental Results: Datasets

Five different datasets are selected for evaluation purpose:

- Normal
  - Violence in Movies
  - Violence in Crowd
  - Riot In Prison
  - Behave
  - Panic

- Abnormal
  - Violence in Movies
  - Violence in Crowd
  - Riot In Prison
  - Behave
  - Panic
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

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIP(HOF)</td>
<td>50.5%</td>
</tr>
<tr>
<td>MoSIFT</td>
<td>89.5%</td>
</tr>
<tr>
<td>Optical Flow</td>
<td>91.31±1.06%</td>
</tr>
<tr>
<td>Interaction Force</td>
<td>95.51±0.79%</td>
</tr>
<tr>
<td>Jerk</td>
<td>95.02±0.56%</td>
</tr>
<tr>
<td>Local Force $F^L$</td>
<td>93.4±1.24%</td>
</tr>
<tr>
<td>Convective Force $F^{Cv}$</td>
<td>92.16±1.13%</td>
</tr>
<tr>
<td>$F^L \mid F^{Cv}$</td>
<td><strong>96.89±0.21%</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOT</td>
<td>82.3%</td>
</tr>
<tr>
<td>LTP</td>
<td>71.53±0.15%</td>
</tr>
<tr>
<td>Optical Flow</td>
<td>79.38±0.14%</td>
</tr>
<tr>
<td>Interaction Force</td>
<td>81.30±0.18%</td>
</tr>
<tr>
<td>Jerk</td>
<td>74.05±0.65%</td>
</tr>
<tr>
<td>Local Force $F_L$</td>
<td>78.14±0.92%</td>
</tr>
<tr>
<td>Convective Force $F_{Cv}$</td>
<td>84.03±1.34%</td>
</tr>
</tbody>
</table>

$F_L | F_{Cv}$  | 85.43±0.21% |
Comparison with State-of-the-Art methods:
Riots in prison

AUC with 95\% confidence interval using LDA

<table>
<thead>
<tr>
<th>Method</th>
<th>Riot in Prison</th>
<th>Panic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Flow</td>
<td>0.76±0.052</td>
<td>0.89±0.0136</td>
</tr>
<tr>
<td>Interaction Force</td>
<td>0.66±0.024</td>
<td>0.89±0.0044</td>
</tr>
<tr>
<td>Jerk</td>
<td>0.65±0.036</td>
<td>0.90±0.0099</td>
</tr>
<tr>
<td>Local force $F^L$</td>
<td>0.68±0.027</td>
<td>0.90±0.0079</td>
</tr>
<tr>
<td>Convective Force $F^{Cv}$</td>
<td>0.79±0.014</td>
<td>0.95±0.0023</td>
</tr>
<tr>
<td>$F^L</td>
<td>F^{Cv}$</td>
<td>0.85±0.077</td>
</tr>
</tbody>
</table>

Riot in Prison

Panic

Normal

Abnormal

AUC with 95\% confidence interval using LDA
Comparison with State-of-the-Art methods: Behave

AUC with 95% confidence interval using LDA

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical flow</td>
<td>0.901±0.032</td>
</tr>
<tr>
<td>Interaction Force</td>
<td>0.925±0.008</td>
</tr>
<tr>
<td>Local force $F_L$</td>
<td>0.933±0.073</td>
</tr>
<tr>
<td>Convective Force $F_{cv}$</td>
<td>0.946±0.032</td>
</tr>
<tr>
<td>$F_L</td>
<td>F_{cv}$</td>
</tr>
</tbody>
</table>
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* mainly 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