MoVi: Mobile Phone based Video Highlights via Collaborative Sensing

Xuan Bao, Romit Roy Choudhury

Michael Molignano
mikem@wpi.edu
CS 525w – 3/29/2011

Overview

- Extend notion of sensor motes to social context
- Define “interesting” social event
- Built an automated video highlight system using mobile phones and devices
- Test system both controlled and real-life scenarios

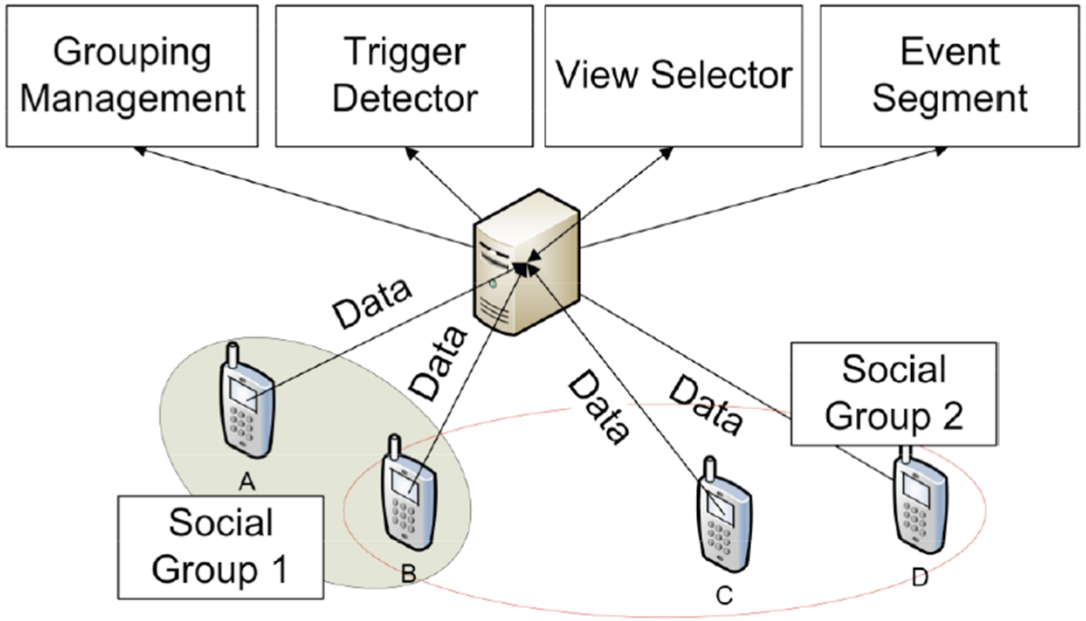
Motivation

- **Sensors on phones and devices everywhere**
  - **Move beyond simple communication**
- **Information gathered from devices is exponentially increasing**
  - **Need to distill and present relevant info.**
- **Want to create automatic video representation of social events**

MoVi Overview

- Spatially nearby devices look for "interesting" event triggers
  - Ex: laughter, people turning same way, etc.
- Device with best view records event
- Individual recordings "stitched" together
- Creates video highlight of event

System Overview



- **Group Management: creates social groups among devices**
- **Trigger Detection: recognizes potentially interesting events**
- **View Selector: picks the "best" device to record event**
- **Event Segment: extracts appropriate segment of video that fully captures the event**

Challenges

- **Group Management**
  - **Attaching each device to at least one zone**
  - **These zones are not necessarily spatial**
- **Event Detection**
  - **“Interesting” events are subjective**
  - **Need clues of when events are occurring**
- **View Selection**
  - **“Best View” is subjective**
  - **Need heuristics to eliminate bad views**
- **Event Segmentation**
  - **Each event has unique start/end to event**

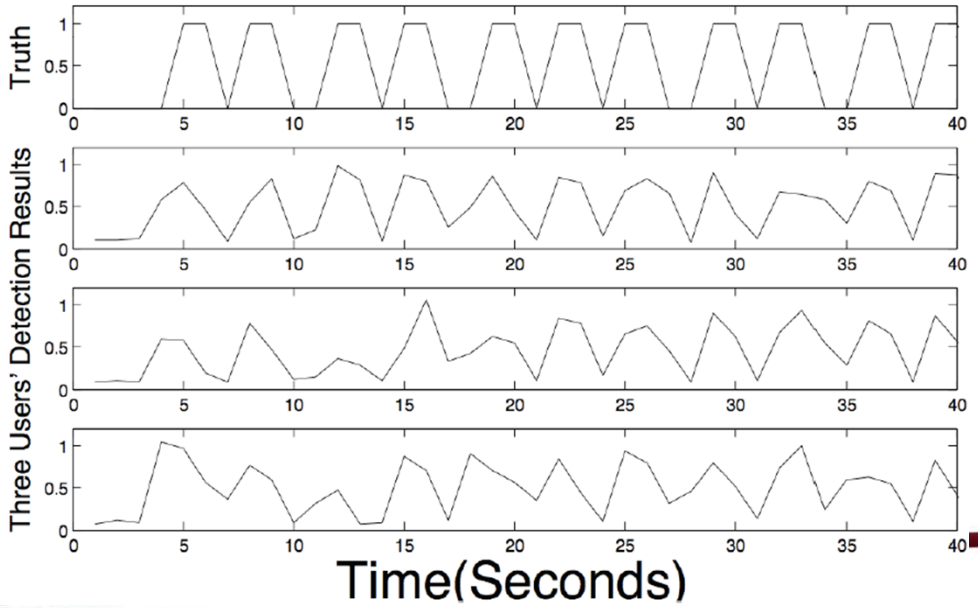
SYSTEM DESIGN

Social Group Identification

- Physical co-location may not be enough
- Uses both visual and acoustic ambience of phones
- Acoustic Grouping
  - Through Ringtone
  - Through Ambient Sound
- Visual Grouping
  - Through Light Intensity
  - Through View Similarity

Grouping Through Ringtone

- **Helps to give approximate grouping**
- **Random phone plays short high-frequency ringtone periodically**
  - **Phones listen for ringtone**



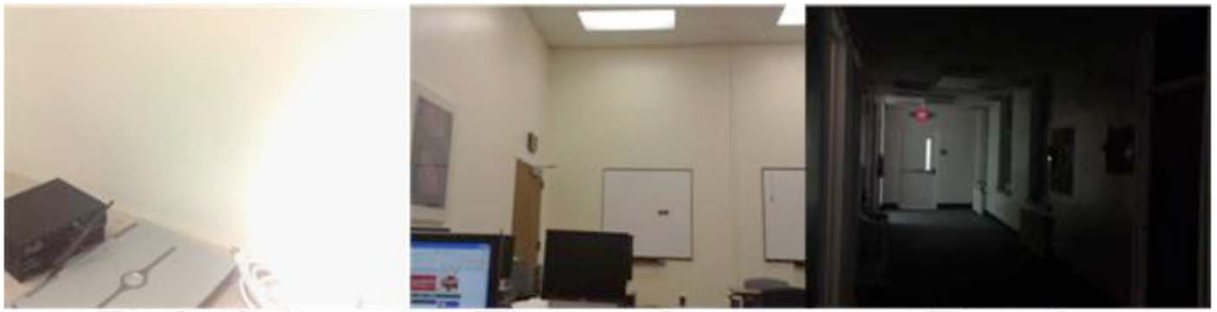
Grouping Through Ambient Sound

- Ringtones not always detectable
- Look at similarity of phones' ambient sounds
- Music, human conversation, and noise
- Use Mel-Frequency Cepstral Coefficients (MFCC) to group phones that "hear" similar classes of sound

Classification Type	Accuracy
Music, Conversation, Noise	98.4535%
Speaker Gender	76.319%
Music Genre	40.3452%

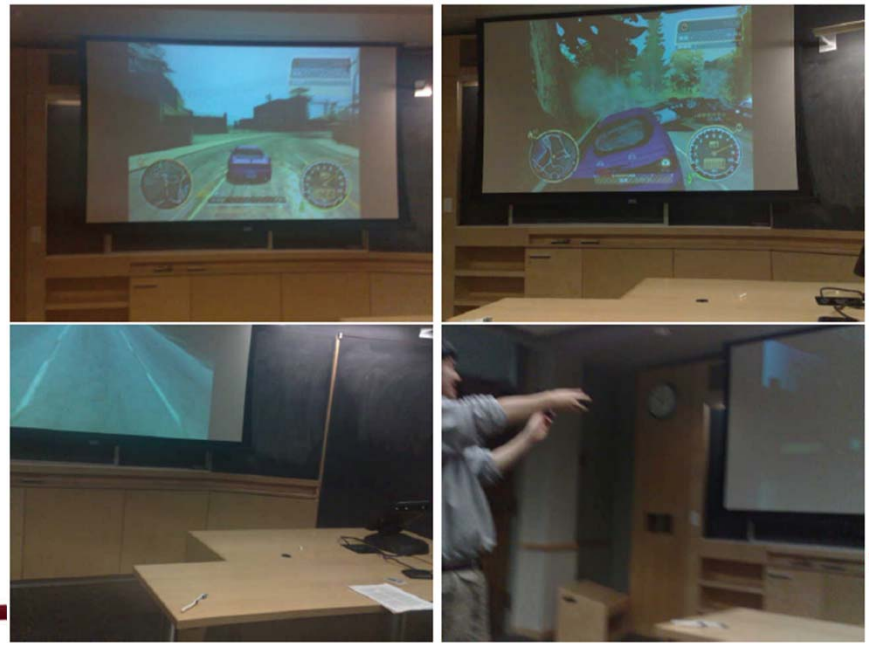
Grouping Through Light Intensity

- Light intensities vary in different areas of same social setting
- Found that light often sensitive to orientation of device
- Used three classes of light



Grouping Through View Similarity

- Look at similarities from different cameras
- Use image technique called spatiogram
  - Pictures with similar spatial organization of colors and edges have high similarity



Trigger Detection

- MoVi must identify patterns that represent socially interesting events
- Interesting events is subjective
- Devices limited in sensing/inferring

- Use three categories to identify
  - Specific Event Signatures
  - Group Behavior Pattern
  - Neighbor Assistance

Specific Event Signatures

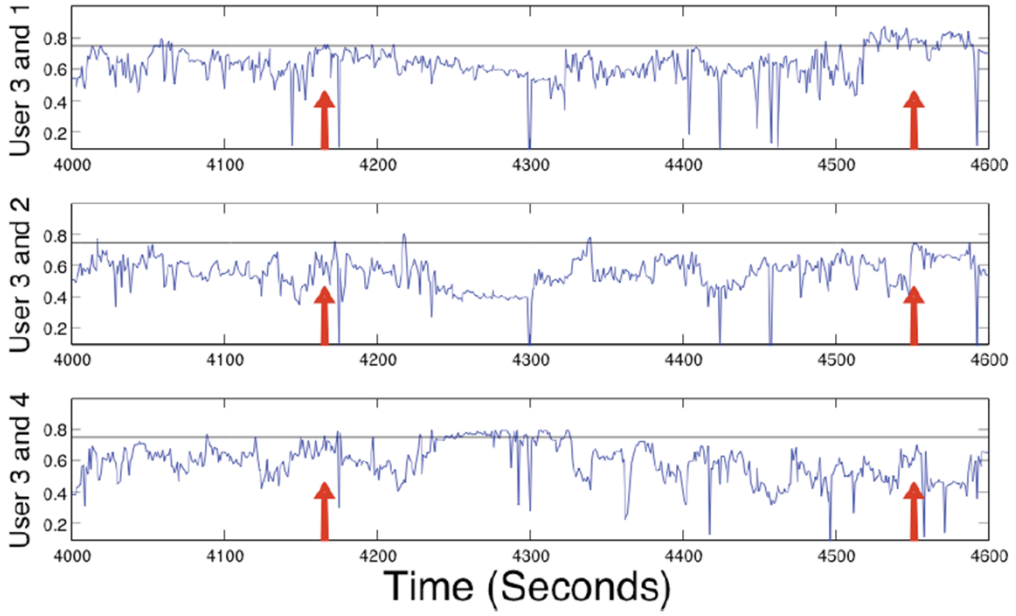
- Pertain to specific sensory triggers
  - Laughing, clapping, shouting, whistling, etc.
- They started with only laughter
- Use samples of laughter 10-15 minutes of 4 students
- Achieved an accuracy of 76%

Group Behavior Pattern

- **Look at similarity in sensory fluctuations of a group**
- **Broken into three triggers**
  - **Unusual view similarity**
  - **Group rotation**
  - **Ambience fluctuation**

Unusual View Similarity

- Similar to the technique used in grouping
- However this must last for extended period of time

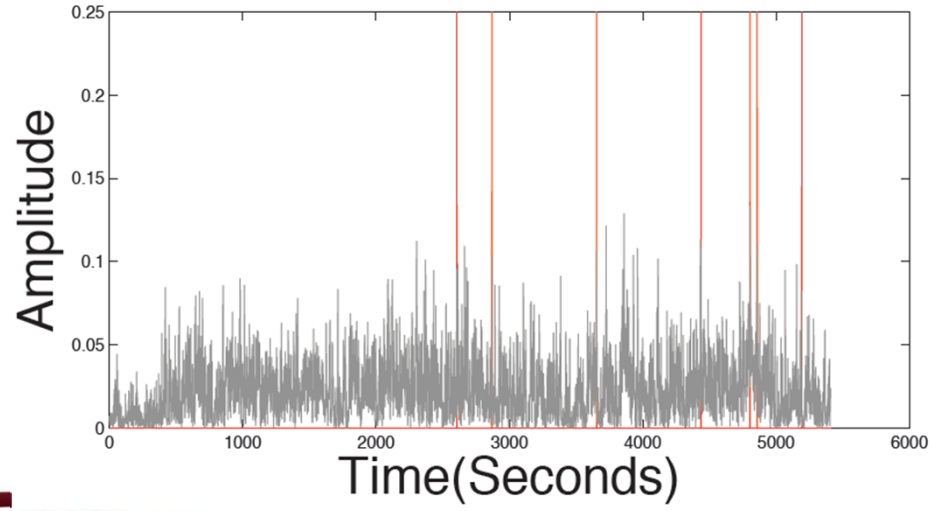


Group Rotation

- Event may prompt large group to all rotate towards same direction
- Must occur within small time window
- Can be captured through compass readings
- Examples
  - Everyone turning towards speaker
  - Everyone turning towards entering celebrity

Ambience Fluctuation

- Ambience of a group may change
- Different threshold set of lighting or sound
- Examples
  - Lights turning on/off
  - Music turning on/off



Neighbor Assistance

- Uses human participation
- When a user takes a picture
  - Send acoustic signal and compass position
  - Other cameras record event

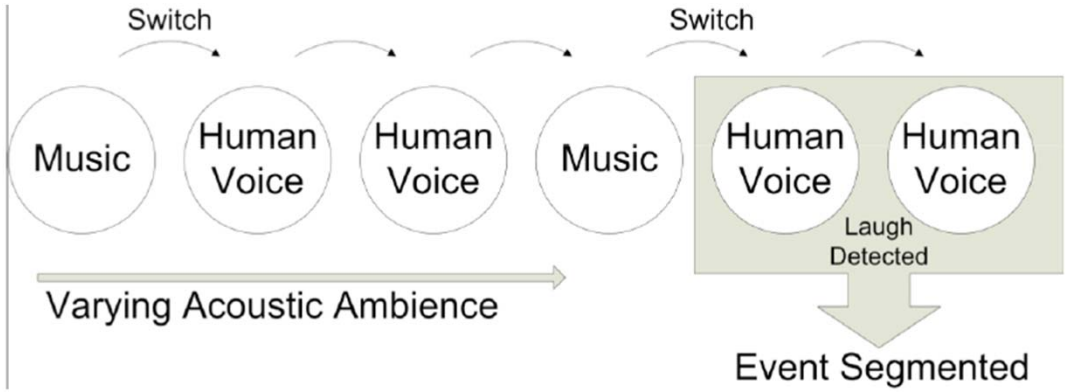
- Intuition is humans likely to take picture of interesting events

View Selection

- **Select phone available with best view**
- **Four heuristics used:**
  - **Face count: more human faces is better**
  - **Accelerometer reading: want stable cameras**
  - **Light intensity: help rule out dark views**
  - **Human in the loop: if triggered by "neighborhood assistance", that view is higher**

Event Segmentation

- **Last step in creating video of event**
- **Finds the logical start and end of event**
- **Use sound state-transition as clues**
  - **Find when conversation started before laughter is heard, etc.**

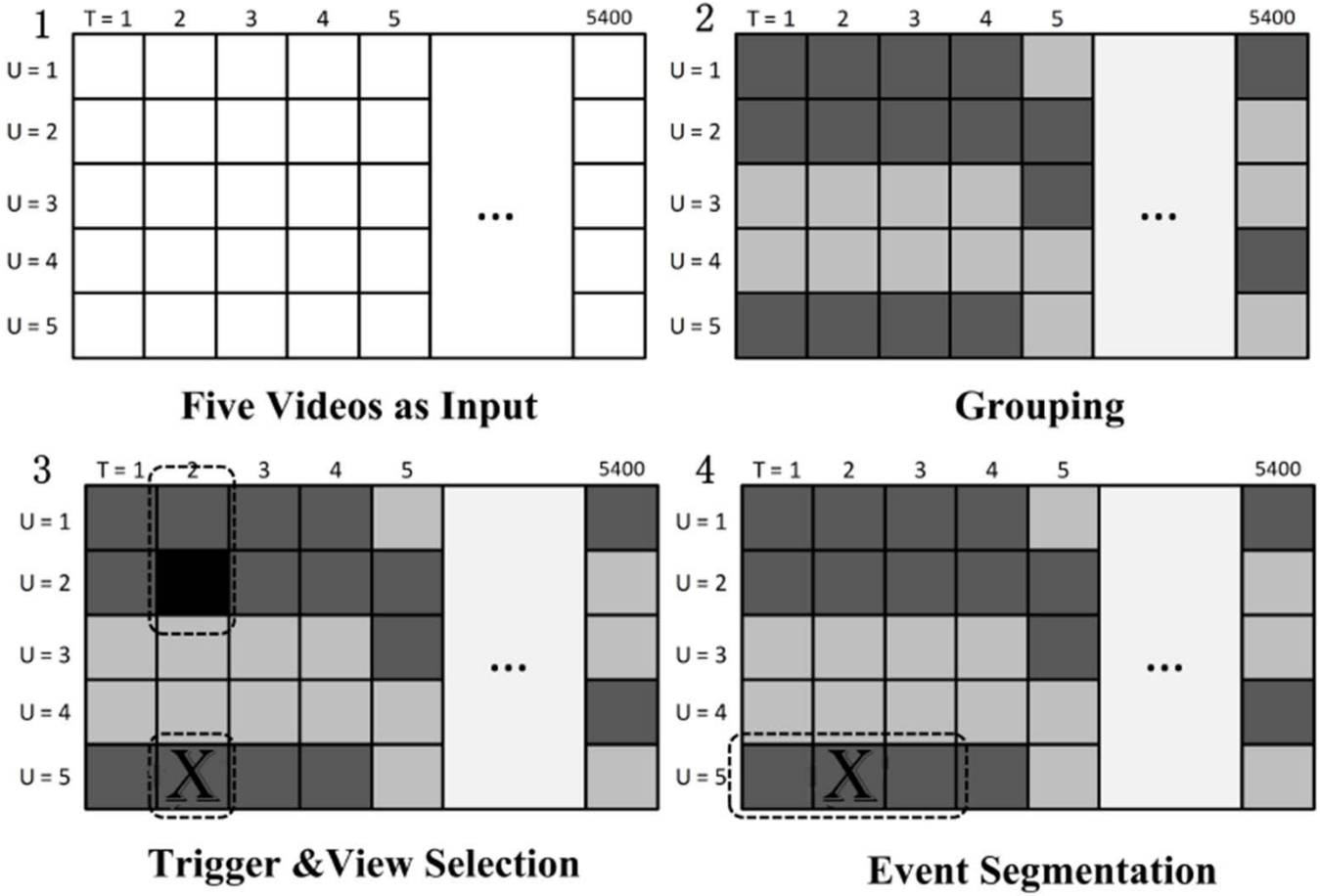


EVALUATION AND RESULTS

Experiments

- Used one controlled and two natural settings
- 5 volunteers
  - iPod video cameras on shirt
  - Nokia N95 phones on belts
- Recorded video for around 1.5 hours (5400 sec)
- Phones used accelerometer, compass, and microphone
- Broke video clips into 5x5400 matrix (1 sec clips)
- Evaluated MoVi's efficacy to pick "socially interesting" elements from matrix

MoVi Operation



Evaluation Metrics

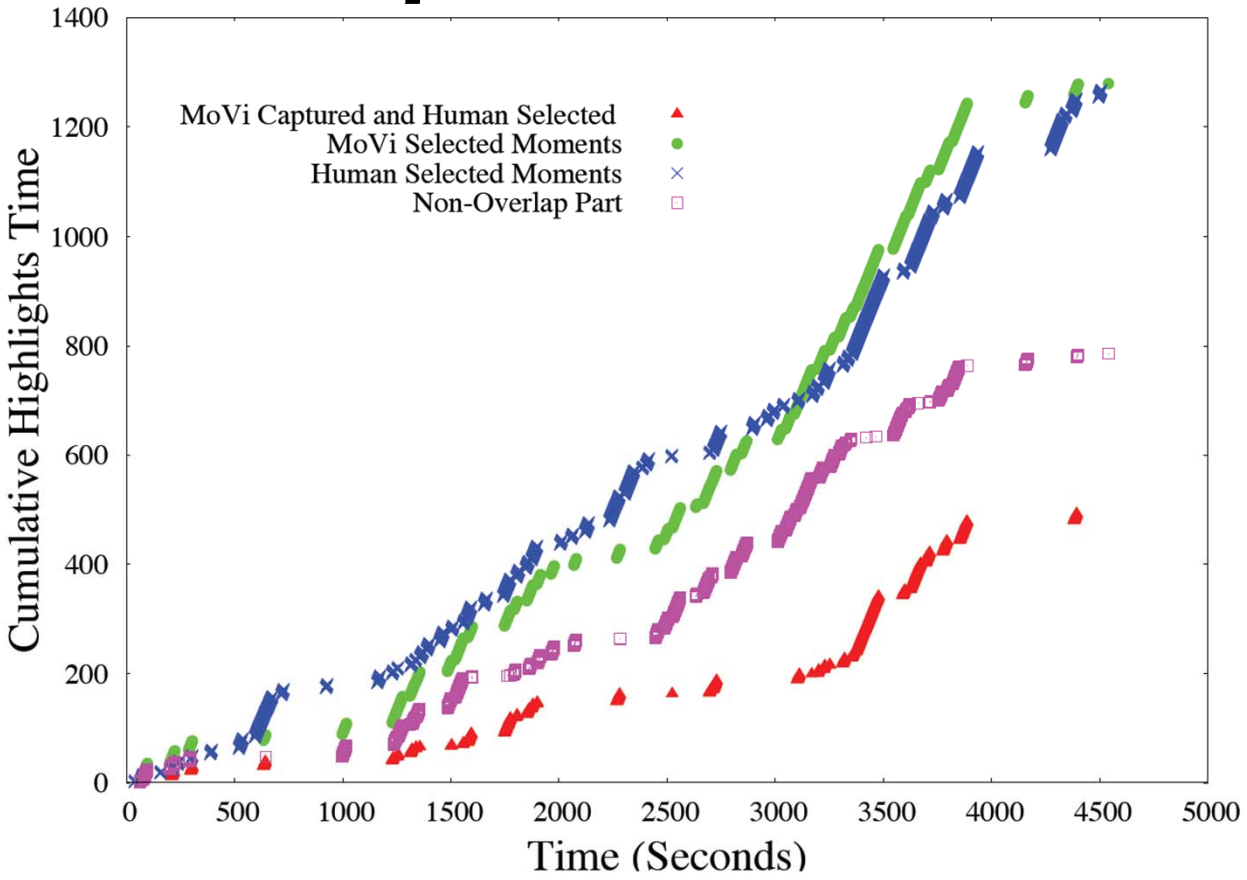
$$Precision = \frac{|\{\text{Human Selected} \cap \text{MoVi Selected}\}|}{|\{\text{MoVi Selected}\}|}$$

$$Recall = \frac{|\{\text{Human Selected} \cap \text{MoVi Selected}\}|}{|\{\text{Human Selected}\}|}$$

$$Fall - out = \frac{|\{\text{Non-Relevant} \cap \text{MoVi Selected}\}|}{|\{\text{Non-Relevant}\}|}$$

- **Human selected parts by multiple humans and combined them**
- **Non-relevant are those not selected by humans**

First Experiment Results



- **Precision = 0.3852, Recall = 0.3885, Fall-out = 0.2109**
- **MoVi's improvement over Random is 101%**

Observations on Results

- Not perfect but reasonable
- They used strict metric for co-selection
  - This caused a lower overlap from MoVi to Human selection, even when partial overlap existed
- Human selected videos is biased
  - Picked lots at beginning, less at end
- Human “interest” is subjective and requires significant research and sensors

RELATED WORK AND CONCLUSIONS

Related Work

- Wearable Computing and SenseCam
- Computer Vision
- Information Retrieval
- Sensor Network of Cameras
- People-Centric Sensing

Conclusions (p1)

- MoVi looks into social event coverage
  - Automated by wearable sensors

- Looks at identifying social groups
- Listening and looking for event triggers
- Finding and recording the events
- Creating a highlight reel of all recorded events of interest

Conclusions (p2)

- They tested MoVi with three events
- Had human selection pick events of interest to compare MoVi to
- MoVi selected a lot of events of interest as well as many not selected by humans
- Overall idea is great start
  - Needs more research into event triggers
  - “Important” events are subjective

Future Work

- Improving accuracy of trigger detection
- Introduce static cameras
  - Help deal with poor views
- Better energy consumption
  - Continuous video recording eats battery life
- Privacy concerns
- Improvements on algorithms
  - Segmentation and triggers