MoVi: Mobile Phone based Video Highlights via Collaborative Sensing

Xuan Bao
Department of ECE
Duke University
xuan.bao@duke.edu

Romit Roy Choudhury
Department of ECE
Duke University
romit@ee.duke.edu

ABSTRACT

Sensor networks have been conventionally defined as a network of sensor motes that collaboratively detect events and report them to a remote monitoring station. This paper makes an attempt to extend this notion to the social context by using mobile phones as a replacement for motes. We envision a social application where mobile phones collaboratively sense their ambience and recognize socially "interesting" events. The phone with a good view of the event triggers a video recording, and later, the video-clips from different phones are "stitched" into a video highlights of the occasion. We observe that such a video highlights is akin to the notion of event coverage in conventional sensor networks, only the notion of "event" has changed from physical to social. We have built a Mobile Phone based Video Highlights system (MoVi) using Nokia phones and iPod Nanos, and have experimented in real-life social gatherings. Results show that MoVi-generated video highlights (created offline) are quite similar to those created manually, (i.e., by painstakingly editing the entire video of the occasion). In that sense, MoVi can be viewed as a collaborative information distillation tool capable of filtering events of social relevance.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; C.2.4 [**Computer-Communication Networks**]: Distributed Systems; H.5.5 [**Information Interfaces and Presentations**]: Sound and Music Computing

General Terms

Design, Experimentation, Performance, Algorithms

Keywords

Video Highlights, Mobile Phones, Collaborative Sensing, Context, Fingerprinting

MobiSys'10, June 15–18, 2010, San Francisco, California, USA.

1. INTRODUCTION

The inclusion of multiple sensors on a mobile phone is changing its role from a simple communication device to a life-centric sensor. Similar trends are influencing other personal gadgets such as the iPods, palm-tops, flip-cameras, and wearable devices. Together, these sensors are beginning to "absorb" a high-resolution view of the events unfolding around us. For example, users are frequently taking geo-tagged pictures and videos [1, 2], measuring their carbon footprint [3], monitoring diets [4], creating audio journals and tracking road traffic [5, 6]. With time, these devices are anticipated to funnel in an explosive amount of information, resulting in what has been called as an information overload. Distilling the relevant content from this overload of information, and summarizing it to the end user, will be a prominent challenge of the future. While this challenge calls for a long-term research effort, as a first step, we narrow its scope to a specific application with a clearly defined goal. We ask, assuming that people in a social gathering are carrying smart phones, *can the phones be harnessed to collaboratively create a video highlights of the occasion.* An automatic video highlights could be viewed as a distilled representation of the social occasion, useful to answer questions like "what happened at the party?" The ability to answer such a question may have applications in travel blogging, journalism, emergency response, and distributed surveillance.

This paper makes an attempt to design a Mobile Phone based Video Highlights system (MoVi). Spatially nearby phones collaboratively sense their ambience, looking for event-triggers that suggest a potentially "interesting" moment. For example, an outburst of laughter could be an acoustic trigger. Many people turning towards the wedding speech – detected from the correlated compass orientations of nearby phones – can be another example. Among phones that detect a trigger, the one with the "best quality" view of the event is shortlisted. At the end of the party, the individual recordings from different phones are correlated over time, and "stitched" into a single video highlights of the occasion. If done well, such a system could reduce the burden of manually editing a full-length video. Moreover, some events are often unrecorded in a social occasion, perhaps because no one remembered to take a video, or the designated videographer was not present at that instant. MoVi could be an assistive

solution for improved social event coverage[1].

A natural concern is: *phones are often inside pockets and may not be useful for recording events.* While this is certainly the case today, a variety of wearable mobile devices are already entering the commercial market [7]. Phone sensors may blend into clothing and jewelry (necklaces, wrist watches, shirt buttons), exposing the camera and microphones to the surroundings. Further, smart homes of the future may allow for sensor-assisted cameras on walls, and on other objects in a room. A variety of urban sensing applications is already beginning to exploit these possibilities [8, 9]. MoVi can leverage them too.

Translating this vision into a practical system entails a range of challenges. Phones need to be grouped by social contexts before they can collaboratively sense the ambience. The multi-sensory data from the ambience needs to be scavenged for potential triggers; some of the triggers need to be correlated among multiple phones in the same group. Once a recordable event (and the phones located around it) is identified, the phone with the best view should ideally be chosen.

While addressing all these challenges is non-trivial, the availability of multiple sensing dimensions offers fresh opportunities. Moreover, high-bandwidth wireless access to nearby clouds/servers permits the outsourcing of CPU-intensive tasks [10]. MoVi attempts to make use of these resources to realize the end-goal of collaborative video recording. Although some simplifying assumptions are made along the way, the overall system achieves its goal reasonably well. In our experiments in real social gatherings, 5 users were instrumented with iPod Nanos (taped to their shirt pockets) and Nokia N95 mobile phones clipped to their belts. The iPods video-recorded the events continuously, while the phones sensed the ambience through the available sensors. The videos and sensed data from each user were transmitted offline to the central MoVi server.

The server is used to mine the sensed data, correlate them across different users, select the best views, and extract the duration over which a logical event is likely to have happened. Capturing the logical start and end of the event is desirable, otherwise, the video-clip may only capture a laugh and not the (previous) joke that may have induced it. Once all the video-clips have been shortlisted, they are sorted in time, and "stitched" into an automatic video highlights of the occasion. For a baseline comparison, we used a manually-created video highlights; multiple users were asked to view the full length iPod videos, and mark out events that they believe are worth highlighting. The union of all events (marked by different users) were also stitched into a highlights. We observe considerable temporal overlap in the manual and MoVi-created highlights (the highlights are $15\ minutes$ while the full length videos are around $1.5\ hours$). Moreover, end users responded positively about the results, suggesting the need (and value) for further research in this direction of automatic event coverage and information distillation.

The rest of the paper is organized as follows. The overall system architecture is proposed in Section 2, and the individual design components are presented in Section 3. Section 4 evaluates the system across multiple real-life and mock social settings, followed by user-surveys and exit-interviews. Section 5 discusses the cross-disciplinary related work for MoVi. We discuss the limitations of the proposed system and future work in Section 6. The paper ends with a conclusion in Section 7.

[1]This bears similarity to spatial coverage in sensor networks, except that physical space is now replaced by a space of social-events, that must be covered by multiple sensing dimensions.

2. SYSTEM OVERVIEW

Figure 1 shows the envisioned client/server architecture for MoVi. We briefly describe the general, high level operations and present the details in the next sections. The descriptions are deliberately anchored to a specific scenario – a social party – only to provide a realistic context to the technical discussions. We believe that the core system can be tailored to other scenarios as well.

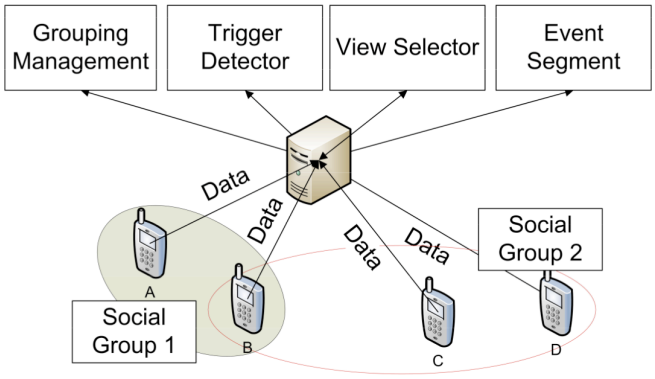

Figure 1: The MoVi architecture.

In general, MoVi assumes that people are wearing a camera and are carrying sensor-equipped mobile devices such as smart phones. The camera can be a separate device attached on a shirt-button or spectacles, or could even be part of the wearable phone (like a pocket-pen, necklace, or wrist watch [11]). In our case, an iPod Nano is taped onto the shirt pocket, and the phone is clipped to a belt or held in the hand. Continuous video from the iPod and sensor data from the phone are sent to the MoVi server offline.

At the MoVi server, a *Group Management* module analyzes the sensed data to compute social groupings among phones. The idea of grouping facilitates collaborativeinferring of social events; only members of the same social group should collaborate for event identification. If real time operations were feasible, the Group Management module could also load-balance among the phones to save energy. Each phone could turn off some sensors and be triggered by the server only when certain events are underway. We are unable to support this sophistication in this paper – optimizing energy consumption and duty-cycling is part of our future work. A *Trigger Detection* module scans the sensed data from different social groups to recognize potentially interesting events. Once an event is suspected, the data is correlated with the data from other phones in that same group.

Confirmed of an event, the *View Selector* module surveys the viewing quality of different phones in that group, and re-

cruits the one that is “best". Finally, given the best video view, the *Event Segmentation* module is responsible for extracting the appropriate segment of the video, that fully captures the event. The short, time-stamped video segments are finally correlated over time, and stitched into the video highlights.

Challenges

The different modules in MoVi entail distinct research challenges. We briefly state them here and visit them individually in the next section.

(1) The Group Management module needs to partition the set of mobile devices based on the social context they are associated to. A social zone could be a gathering around an ice-cream corner, a group of children playing a video game, or people on the dance floor. The primary challenges are in identifying these zones, mapping phones to at least one zone, and updating these groups in response to human movement. Importantly, these social groups are not necessarily spatial – two persons in physical proximity may be engaged in different conversations in adjacent dinner tables.

(2) The Event Detection module faces the challenge of recognizing events that are socially “interesting", and hence, worth video recording. This is difficult not only because the notion of “interesting" is subjective, but also because the space of events is large. To be detected, interesting events need to provide explicit clues detectable by the sensors. Therefore, our goal is to develop a rule-book with which (multi-modal) sensor measurements can be classified as “interesting". As the first step towards developing a rule book, we intend to choose rules shared by different events. Our proposed heuristics aim to capture a set of intuitive events (such as laughter, people watching TV, people turning towards a speaker, etc.) that one may believe to be socially interesting. Details about event detection will be discussed in Section 3.2.

(3) The View Selection module chooses the phone that presents the best view of the event. The notion of “best view" is again subjective, however, some of the obviously poor views need to be eliminated. The challenge lies in designing heuristics that can achieve reliable elimination (such as ones with less light, vibration, or camera obstructions), and choose a good candidate from the ones remaining. Details regarding our heuristics will be provided in Section 3.3.

(4) The Event Segmentation module receives a time-stamped event-trigger, and scans through the sensor measurements to identify the logical start and end of that event. Each social event is likely to have an unique/complex projection over the different sensing dimensions. Identifying or learning this projection pattern is a challenge.

MoVi attempts to address these individual challenges by drawing from existing ideas, and combining them with some new opportunities. The challenges are certainly complex, and this system is by no means a mature solution to generating automated highlights. Instead it may be viewed as an early effort to explore the increasingly relevant research space. The overall design and implementation captures some of the inherent opportunities in collaborative, multi-modal sensing, but also exposes unanticipated pitfalls. The evaluation results are limited to a few social occasions, and our ongoing work is focused on far greater testing and refinement. Nevertheless, the reported experiments are real and the results adequately promising to justify the larger effort. In this spirit, we describe the system design and implementation next, followed by evaluation results in Section 4.

3. SYSTEM DESIGN AND BASIC RESULTS

This section discusses the four main modules in MoVi. Where suitable, the design choices are accompanied with measurements and basic results. The measurements/results are drawn from three different testing environments. (1) A set of students gathering in the university lab on a weekend to watch movies, play video games, and perform other fun activities. (2) A research group visiting the Duke SmartHome for a guided-tour. The SmartHome is a residence-laboratory showcasing a variety of research prototypes and latest consumer electronics. (3) A Thanksgiving dinner party at a faculty’s house, attended by the research group members and their friends.

3.1 Social Group Identification

Inferring social events requires collaboration among phones that belong to the same social context. To this end, the scattered phones in a party need to be grouped socially. Interestingly, physical collocation may not be the perfect solution. Two people in adjacent dinner tables (with their backs turned to each other) may be in physical proximity, but still belong to different social conversations (this scenario can be generalized to people engaging in different activities in the same social gathering). Thus people should not video-record just based on spatial interpretation of a social event. In reality, a complex notion of “same social context” unites these phones into a group – MoVi tries to roughly capture this by exploiting multiple dimensions of sensing. For instance, people seated around a table may be facing the same object in the center of the table (e.g., a flower vase), while people near the TV may have a similar acoustic ambience. The group management module correlates both the visual and acoustic ambience of phones to deduce social groups. We begin with the description of the acoustic methods.

(1) Acoustic Grouping

Two sub-techniques are used for acoustic grouping, namely, ringtone and ambient-sound grouping.

Grouping through Ringtone. To begin with an approximate grouping, the *MoVi* server chooses a random phone to play a short high-frequency ring-tone (similar to a wireless beacon) periodically. The ring-tone should ideally be outside the audible frequency range, so that it is not interfered by human voices and also not annoying to people. With Nokia N95 phones, we were able to play narrow-bandwidth tones at the edge of the audible range and use it with almost-inaudible amplitude [2]. The single-sided amplitude spectrum of the ring-tone is shown in Figure 2. The target is to make the ringtone exist only on $3500Hz$. This frequency is high enough to avoid being interfered by indoor noises.

[2]Audible range differs for different individuals. Our choice of frequency, $3500Hz$, was limited by hardware. However, with new devices such as the iPhone, it is now possible to generate and play sounds at much higher frequencies.

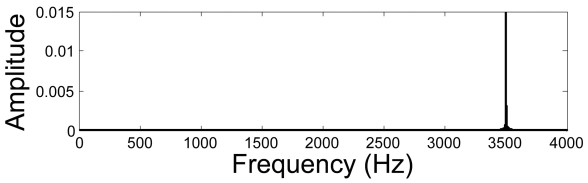

Figure 2: Single-sided amplitude spectrum of the ringtone

Phones in the same acoustic vicinity are expected to hear the ringtone[3]. To detect which phones overheard this ringtone, the MoVi server generates a frequency-domain representation of the sounds reported at each phone (a vector, $\overrightarrow{S}$, with 4000 dimensions), and computes the *similarity* of these vectors with the vector generated from the known ringtone ($\overrightarrow{R}$). The similarity function, expressed below, is essentially a weighted intensity ratio after subtracting white noise (Doppler shifts are explicitly addressed by computing similarity over a wider frequency range).

$$Similarity = \frac{Max\{\overrightarrow{S}(i)|3450 =< i <= 3550\}}{Max\{\overrightarrow{R}(i)|3450 =< i <= 3550\}}$$

Therefore, high similarities are detected when devices are in the vicinity of the ringtone transmitter. The overhearing range of a ringtone defines the auditory space around the transmitter.

Figure 3 shows the similarity values over time at three different phones placed near a ring-tone transmitter. The first curve is the known transmitted ringtone and other three curves are the ones received. As shown in Figure 3, the overheard ringtones are in broad agreement with the true ringtone. All phones that exhibit more than a threshold similarity are assigned to the same acoustic group. A phone may be assigned to multiple acoustic groups. At the end of this operation, the party is said to be "acoustically covered".

Grouping through Ambient Sound. Ringtones may not be always detectable, for example, when there is music in the background, or other electro-mechanical hum from machines/devices on the ringtone's frequency band. An alternative approach is to compute similarities between phones' ambient sounds, and group them accordingly. Authors in [12] address a similar problem – they use high-end, time-synchronized devices to record ambient sound, and compare them directly for signal matching. However, we observed that mobile phones are weakly time-synchronized (in the order of seconds), and hence, direct comparison results will yield errors. Therefore, we classify ambient sound in stable classes using an SVM (Support Vector Machine) on MFCC (Mel-Frequency Cepstral Coefficients), and group phones that "hear" the same classes of sound. We describe the process next.

[3]We avoid bluetooth based grouping because the acoustic signals are better tailored to demarcate the context of human conversations while bluetooth range may not reflect the social partition among people. However, in certain extremely noisy places, bluetooth can be used to simplify the implementation.

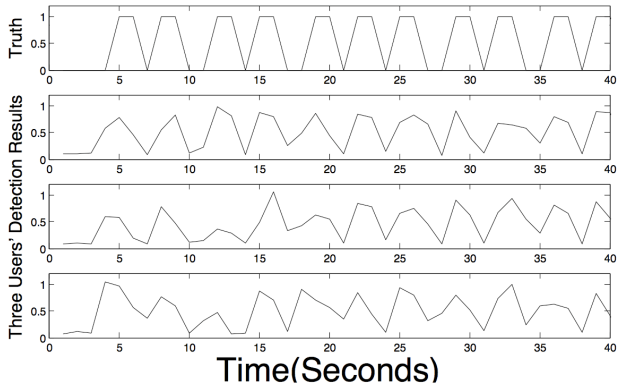

Figure 3: Ringtone detection at phones within the acoustic zone of the transmitter.

For classification, we build a data benchmark with labeled music, human conversation, and noise. The music data is a widely used benchmark from Dortmund University [13], composed of 9 types of music. Each sample is ten seconds long and the total volume is for around 2.5 hours. The conversation data set is built by ourselves, and consists of 2 hours of conversation data from different male and female speakers. Samples from each speaker is around ten minutes long. The noise data set is harder to build because it may vary entirely based on the user's background (i.e., the test may arrive from a different distribution than the training set). However, given that MoVi is mostly restricted to indoor usage, we have incorporated samples of A/C noises, microwave hums, and the noise of phone grazing against trousers and table-tops. Each sample is short in length but we have replicated the samples to make their size equal to other acoustic data.

MFCC (Mel-Frequency Cepstral Coefficients) [14, 15] are used as features extracted from sound samples. In sound processing, Mel-frequency cepstrum is a representation of the short-term power spectrum of a sound. MFCC are commonly used as features in speech recognition and music information retrieval. The process of computing MFCC involves four steps: (1) We divide the audio stream into overlapping frames with 25ms frame width and 10ms forward shifts. The overlapping frames better capture the subtle changes in sound (leading to improved performance), but at the expense of higher computing power. (2) Then, for each frame, we perform an FFT to obtain the amplitude spectrum. However, since each frame has a strict cut-off boundary, the FFT causes leakage. We employ the Hann window technique to reduce spectral leakage. Briefly, Hann window is a raised cosine window that essentially acts as a weighting function. The weighing function is applied to the data to reduce the sharp discontinuity at the boundary of frames. This is achieved by matching multiple orders of derivatives, and setting the value of the derivatives to zero [16]. (3) We then take the logarithm on the spectrum,and convert the log spectrum to Mel (perception-based) spectrum. Using Mel scaled units [14] is expected to produce better results than linear units because Mel scale units better approximate human perception of sound. (4) We finally take the Discrete Cosine Transform (DCT) on the Mel spectrum. In [14], the author proves that this step approximates principal components analysis (PCA), the mathematically standard way to decorrelate the components of the feature vectors, in

the context of speech recognition and music retrieval.

After feature extraction, classification is performed using a two-step decision, using support vector machines (SVM), a machine learning method for classification [17]. Coarse classification tries to distinguish music, conversation, and ambient noise. Finer classification is done for classes within conversation and music [18]. Classes for conversation include segregating between male and female voices, which is useful to discriminate between, say, two social groups, one of males, another of females. Similarly, music is classified into multiple genres. The overall cross validation accuracy is shown in Table 1. The reported accuracy is tested on the benchmarks described before. Based on such classification, Figure 4 shows the grouping among two pairs of phones – <A,B> and <A,C> – during the Thanksgiving party. Users of phones A and C are close friends and were often together in the party, while user of phone B joined A during some events as in. Accordingly, A and C are more often grouped as in Figure 4(b) while user A and B are usually separated (Figure 4(a)).

Table 1: Cross Validation Accuracy on Sound Benchmarks

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

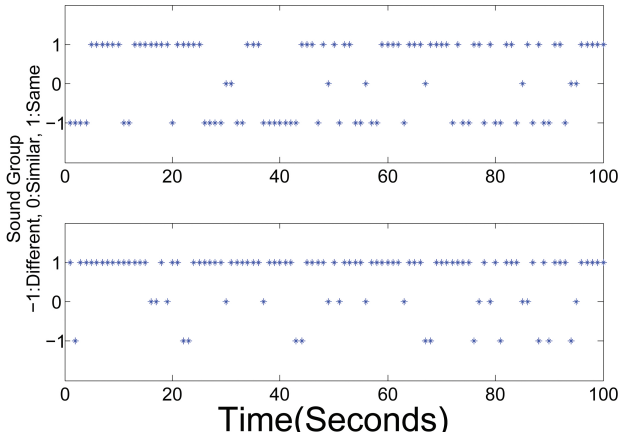

Figure 4: Grouping based on acoustic ambience: (a) users A and B's acoustic ambiences' similarity. (b) users A and C's acoustic ambiences' similarity.

(2) Visual Grouping

As mentioned earlier, acoustic ambience alone is not a reliable indicator of social groups. Similarity in visual ambience, including light intensity, surrounding color, and objects, can offer greater confidence on the phone's context [19]. We describe our visual grouping schemes here.

Grouping through Light Intensity. In some cases, light intensities vary across different areas in a social setting. Some people may be in an outdoor porch, others in a well-lit indoor kitchen, and still others in a darker living room, watching TV. We implemented light-based grouping using analogous *similarity* functions as used with sound. However, we found that the light intensity is often sensitive to the user's orientation, nearby shadows, and obstructions in front of the camera. To achieve robustness, we conservatively classified light intensity into three classes, namely, bright, regular, and dark. Most phones were associated to any one of these classes; some phones with fluctuating light readings, were not visually-grouped at all. Figure 5 illustrates samples from three light classes from the social gathering at the university.

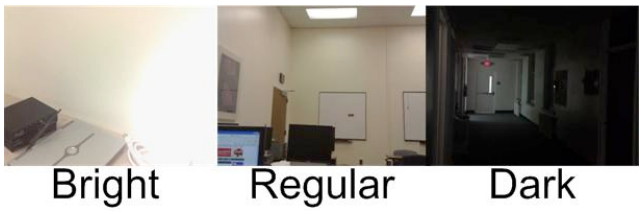

Figure 5: Grouping based on light intensity – samples from 3 intensity classes.

Grouping through View Similarity. A second way of visual grouping pertains to similarity in the images from different phone cameras. Multiple people may simultaneously look at the person making a wedding toast, or towards an entering celebrity, or just towards the center of a table with a birthday cake on it. MoVi intends to exploit this opportunity of common view. To this end, we use an image generalization technique called spatiogram [20]. Spatiograms are essentially color histograms encoded with spatial information. Briefly, through such a representation, pictures with similar spatial organization of colors and edges exhibit high similarity. The second order of spatiogram can be represented as:

$$h_I(b) = \langle n_b, \mu_b, \sigma_b \rangle, b = 1, 2, 3 \cdots B$$

where n_b is the number of pixels whose values are in the b^{th} bin (each bin is a range in color space), and μ_b and σ_b are the mean vector and covariance matrices, respectively, of the coordinates of those pixels. B is the number of bins. Figures 6(a) and (b) show the view from two phones while their owners are playing a multi-player video-game on a projector screen. Both cameras capture the screen as the major part of the picture. Importantly, the views are from different instants and angles, yet, the spatiogram similarities are high. Comparing to the top two pictures, the views in Figure 6(c) and (d) are not facing the screen, therefore exhibiting a much lower view similarity.

The MoVi server mines through the acoustic and visual information (offline), and combines them to form a single audio-visual group. View similarity is assigned highest priority, while audio and light intensity are weighed with a lower, equal priority. This final group is later used for collaboratively inferring the occurrence of events. Towards this goal, we proceed to the discussion of event-triggers.

3.2 Trigger Detection

From the (recorded) multi-sensory information, the MoVi server must identify patterns that suggest events of potential social interest. This is challenging because of two factors. First, the notion of interesting is subjective; second, the space

Figure 6: Grouping based on view similarity – top two phones (a, b) are in the same group watching video games, while the bottom two (c, d) are in the same room but not watching the games..

of social events (defined by human cognition) is significantly larger than what today's sensing/inferring technology may be able to discern. We admittedly lower our targets, and try to identify some opportunities to detect event-triggers. We design three categories, namely (1) Specific Event Signature, (2) Group Behavior Pattern, and (3) Neighbor Assistance.

(1) Specific Event Signatures

These signatures pertain to specific sensory triggers derived from human activities that, in general, are considered worth recording. Examples of interest include, laughter, clapping, shouting, whistling, singing, etc. Since we cannot enumerate all possible events, we intend to take advantage of collaboration using triggers related to group behavior instead of relying heavily on specific event signatures. Therefore, as a starting point, we have designed specific acoustic signatures only for laughter [21] using MFCC. Validation across a sample of 10 to 15 minutes of laughter, from 4 different students, offered evidence that our laughter-signature is robust to independent individuals. Negative samples are human conversation and background noise. Figure 7 shows the distribution of self-similarity between laughter-samples and cross similarity between laughter and other negative samples. In other words, the laughter samples and negative samples form different clusters in the 12 dimensional space. We achieved a cross-validation accuracy of 76% on our benchmark.

(2) Group Behavior Pattern

The second event-trigger category exploits similarity in sensory fluctuations across users in a group. When we observe most members of a group behaving similarly, or experiencing similar variances in ambience, we infer that a potentially interesting event may be underway. Example triggers in this category are view similarity detection, group rotation, and acoustic-ambience fluctuation.

Unusual View Similarity. When phone cameras are found viewing the same object from different angles, it could be an event of interest (EoI). As mentioned earlier, some ex-

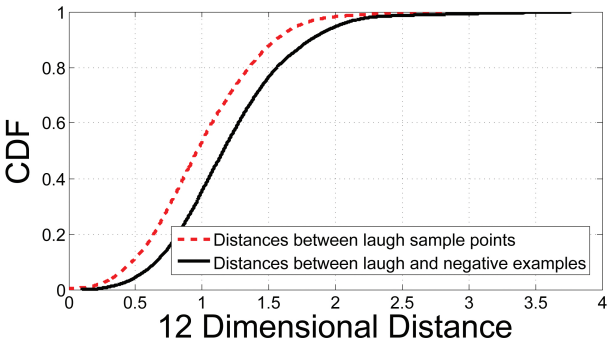

Figure 7: The CDFs show the distances between pairs of laugh samples, and distances between laugh and other sound samples.

amples are people watching the birthday cake on a table, paying attention to a wedding toast, or everyone attracted by a celebrity's arrival. Recall that view-similarity was also used as a grouping mechanism. However, to qualify as a trigger, the view must remain similar for more than a threshold duration. Thus, augmenting the same notion of spatiogram with a minimum temporal correlation factor, we find good event-triggers. In Figure 8, each curve shows a pairwise similarity between two views in a group. The arrows show the two time-points at which three (2 pairs) out of four users are watching the same objects, that is i.e. their views show higher similarity than the threshold (empirically set as 0.75). Those three users are all triggered at the two time-points. Of course, such a trigger may be further correlated with other similarities in the acoustic and motion domains. Multi-sensor triggers is a part of our ongoing work.

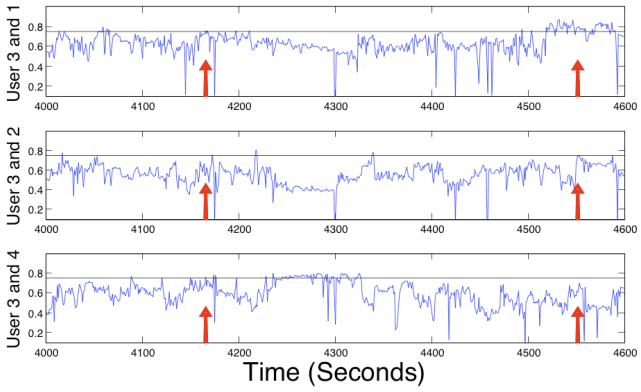

Figure 8: Pair-wise view similarity, at least among 3 phones, qualifies as a video trigger. Users 3, 1, and 4 are all above the threshold at around 4100 seconds; users 3, 1, and 2 see a trigger at around 4500 seconds.

Group Rotation. An interesting event may prompt a large number of people to rotate towards the event (a birthday cake arrives on the table). Such "group rotation" – captured through the compasses in several modern phones – can be used as a trigger. If more than a threshold fraction of the people turn within a reasonably small time window, MoVi considers this a trigger for an interesting event. For this, the compasses of the phones are always turned on (we measured

that the battery consumption is negligible). The compass-based orientation triggers are further combined with accelerometer triggers, indicating that people have turned and moved together. The confidence in the trigger can then be higher. Such a situation often happens, e.g., when a break-out session ends in a conference, and everyone turns towards the next speaker/performer.

Ambience Fluctuation. The general ambience of a social group may fluctuate as a whole. Lights may be turned off on a dance floor, music may be turned on, or even the whole gathering may lapse into silence in anticipation of an event. If such fluctuations are detectable across multiple users, they may be interpreted as a good trigger. MoVi attempts to make use of such collaborative sensor information. Different thresholds on fluctuations are empirically set – with higher thresholds for individual sensors, and relatively lower for joint sensing. The current goal is to satisfy a specific trigger density, no more than two triggers for each five minutes. Of course, this parameter can also be tuned for specific needs. Whenever any of the sensor's reading (or combined) exceed the corresponding threshold, all the videos from the cameras become candidates for inclusion in the highlights. Figure 9 shows an example of the sound fluctuation in time domain, taken from the SmartHome visit. The dark lines specify the time-points when the average of one-second time windows exceed a threshold. These are accepted as triggers. The video-clips around these time-points are eventually "stitched" into the video highlights.

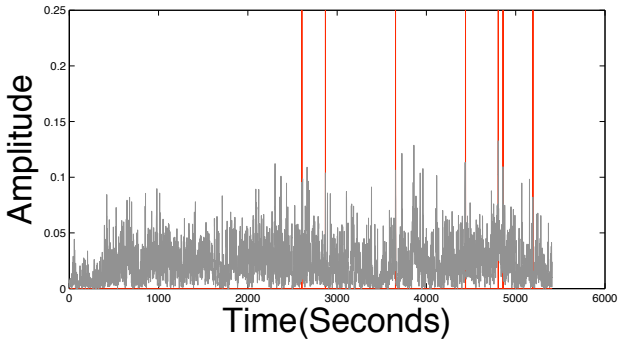

Figure 9: The fluctuations in the acoustic ambience are interpreted as triggers (time-points shown in black lines).

(3) Neighbor Assistance

The last category of event-trigger opportunistically uses human participation. Whenever a user explicitly takes a picture from the phone camera, the phone is programmed to send an acoustic signal, along with the phone's compass orientation. Other cameras in the vicinity overhear this signal, and if they are also oriented in a similar direction, the videos from the camera are recruited as candidates for highlights. The intuition is that humans are likely to take a picture of an interesting event, and including that situation in the highlights may be worthwhile. In this sense, MoVi brings the human into the loop.

3.3 View Selection

The view selection module is tasked to select videos that have a good view. Given that cameras are wearable (taped on shirt pockets in our case), the views are also blocked by objects, or pointed towards uninteresting directions. Yet, many of the views are often interesting because they are more personal, and captures the perspectives of a person. For this, we again rely on multi-dimensional sensing.

Four heuristics are jointly considered to converge on the "best view" among all the iPods that recorded that event. (1) Face count: views with more human faces are given the highest priority. This is because human interests are often focused on people. Moreover, faces ensure that camera is facing a reasonable height, not to the ceiling or the floor. (2) Accelerometer reading ranking: to pick a stable view, the cameras with the least accelerometer variance are assigned proportionally higher points. More stable cameras are chosen to minimize the possibility of motion blurs in the video. (3) Light intensity: to ensure clarity and visibility, we ranked the views in the "regular" light class higher, and significantly de-prioritize the darker pictures. This is used only to rule out extremely dark pictures, which mostly are caused by blocking. (4) Human in the loop: finally, if a view is triggered by "neighbor assistance", the score for that view is increased.

Figure 10 shows two rows corresponding to two examples of view selection; pictures were drawn from different iPod videos during the Thanksgiving party. The first view in each instance is selected and seems to be more interesting than the rest of views. Figure 11 illustrates the same over time. At each time-point, the blue circle tags the human selected view while the red cross tags the MoVi select one. When two symbols overlap, the view selection achieves right result. The most common reason that view selection fails is that all four views exhibit limited quality. Therefore, even for human selection, the chosen one is only marginally better.

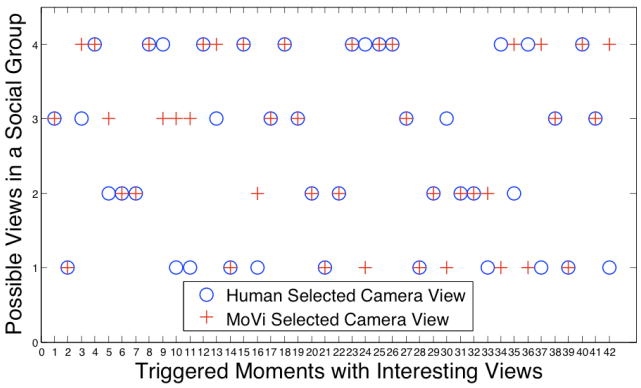

Figure 11: MoVi selects views that are similar to human selected ones.

3.4 Event Segmentation

The Event Segmentation module is designed to identify the logical start and end of an event. A clap after the "happy birthday" song could be the acoustic trigger for video inclusion. However, the event segmentation module should ideally include the song as well, as a part of the highlights. The same applies to a laughter trigger; MoVi should be able to capture the joke that perhaps prompted it. In general, the challenge is to scan through the sensor data received before and after the trigger, and detect the logical start and end that may associate with the trigger.

Figure 10: View selection based on a multiple sensing dimensions. The first view is chosen for inclusion in the highlights because of its better lighting quality, more number of distinct human faces, and less acceleration.

For event segmentation, we use the sound state-transition, computed during the sound classification/grouping phase, time as clues [6]. For example, when laughter is detected during conversation, we rewind on the video, and try to identify the start of a conversation. Gender based voice classification offers a finer ability to segment the video – if multiple people were talking, and a women's voice prompted the joke, MoVi may be able to identify that voice, and segment the video from where that voice started. Figure 12 shows our key idea for event segmentation.

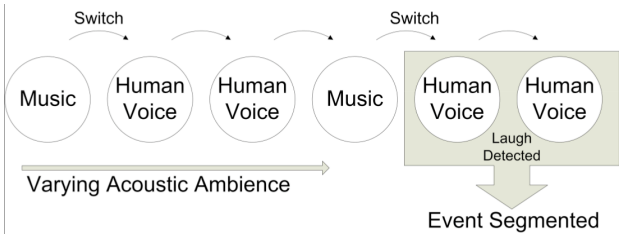

Figure 12: The scheme for segmenting events.

4. EVALUATION

This section attempts to asses MoVi's overall efficacy in creating a video highlight. Due to the subjective/social nature of this work, we choose to evaluate our work by combining users' assessment with metrics from information retrieval research. We describe our experimental set-up and evaluation metrics next, followed by the actual results.

4.1 Experiment Set-up

Our experiments have been performed in one controlled setting and two natural social occasions. In each of these scenarios, 5 volunteers wore the iPod video cameras on their shirts, and clipped the Nokia N95 phones on their belts. Figure 13 shows an example of students taped with iPod Nanos near their shirt pockets. The iPods recorded continuous video for around 1.5 hours (5400 seconds), while the phones logged data from the accelerometer, compass, and microphone. In two of the three occasions, a few phone cameras were strategically positioned on a table or cabinet, to record the activities from a static perspective. All videos and sensor measurements were downloaded to the (MATLAB-based) MoVi server. Each video was organized into a sequence of 1 second clips. Together, the video clips from the volunteers form a 5×5400 matrix, with an unique iPod-device number for each row, and time (in seconds) indexed on each column. The sensor readings from the phones are similarly indexed into this matrix. MoVi's target may now be defined as the efficacy to pick the "socially interesting" elements from this large matrix.

Figure 13: Users wearing iPods and Nokia phones.

The MoVi server analyzes the $<$ device, time $>$-indexed sensor readings to first form the social groups. During a particular time-window, matrix rows 1, 2, and 5 may be in the first group, and rows 3 and 4 in the second. Figure 14(2) shows an example grouping over time using two colors. Then, for every second (i.e., along each column of the matrix), MoVi scans through the readings of each phone to identify event triggers. Detecting a possible trigger in an element of the matrix, the server correlates it to other members of its group. If correlation results meet the desired threshold, MoVi performs view selection across members of that group. It is certainly possible that at time t_i, phone 2's sensor readings match the trigger, but phone 5's view is the best for recording this event(Figure 14(3)). MoVi selects this element $<5, t_i>$, and advances to perform event segmentation. For this, the system checks for the elements along the 5^{th} row, and around column t_i. From these elements, the logical event segment is picked

based on observed state-transitions. The segment could be the elements <5, t_{i-1} > to <5, t_{i+1} >, a 3 second video clip (Figure 14(4)). Many such video clips get generated after MoVi completes a scan over the entire matrix. These video clips are sorted in time, and "stitched" into a "movie". Temporal overlaps between clips are possible, and they are pruned by selecting the better view.

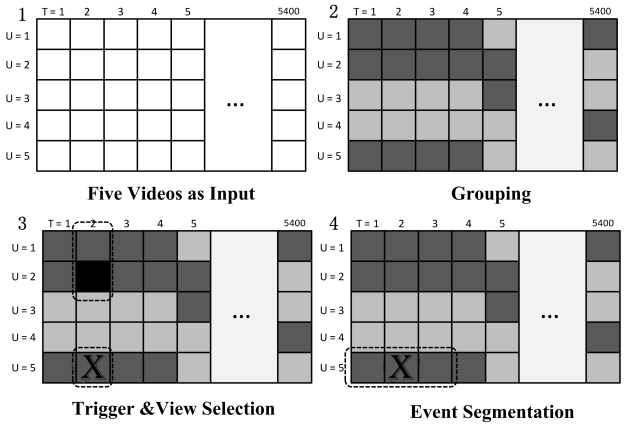

Figure 14: MoVi operations illustrated via a matrix.

4.2 Evaluation Metrics

We use the metrics of *Precision*, *Recall*, and *Fall-out* for the two uncontrolled experiments. These are standard metrics in the area of information retrieval.

$$Precision = \frac{|\{\text{Human Selected} \cap \text{MoVi Selected}\}|}{|\{\text{MoVi Selected}\}|} \quad (1)$$

$$Recall = \frac{|\{\text{Human Selected} \cap \text{MoVi Selected}\}|}{|\{\text{Human Selected}\}|} \quad (2)$$

$$Fall-out = \frac{|\{\text{Non-Relevant} \cap \text{MoVi Selected}\}|}{|\{\text{Non-Relevant}\}|} \quad (3)$$

The "Human Selected" parts are obtained by requesting a person to look through the videos and demarcate time windows that they believe are worth including in a highlights. To avoid bias from a specific person, we have obtained time-windows from multiple humans and also combined them (i.e., a union operation) into a single highlight[4]. We will report results for both cases. "Non-Relevant" moments refer to those not selected by humans. The "MoVi Selected" moments are self evident.

4.3 Performance Results
(1) Controlled Experiment

The aim of the controlled experiment is to verify whether all the components of MoVi can operate in conjunction. To this end, a weekend gathering is planned with pre-planned activities, including watching a movie, playing video-games, chatting over newspaper articles, etc. This experiment is assessed rather qualitatively, ensuring that the expected known exciting events are captured well. Table 2 shows event-detection results. The first two columns show the designed events and their occurrence times; the next two columns show the type of triggers that detected them and the corresponding detection times. Evidently, at least one of the triggers were able to capture the events, suggesting that MoVi achieves a reasonably good event coverage. However, it also included a number of events that were not worthy of recording (false positives). We note that the human-selected portions of the video summed up to 1.5 minutes (while the original video was for 5 minutes). The MoVi highlights covered the full human-selected video with good accuracy (Table 3), and selected an additional one minute of false positives. Clearly, this is not a fair evaluation, and will be drastically different in real occasions. However, it is a sanity check that MoVi can achieve what it absolutely should.

[4]For each experiment, one human reviewer has watched one full video from one camera, which lasts for more than an hour. All video sources from all cameras are covered.

Table 2: Per-Trigger results in single experiment (false positives not reported)

Event Truth	Time	Trigger	Det. Time
Ringtone	25:56	RT, SF	25:56
All watch a game	26:46	IMG	27:09
Game sound	26:58	SF	27:22
2 users see board	28:07	IMG	28:33
2 users see demo	28:58	SF	29:00
Demo ends	31:18	missed	
Laughing	34:53	LH, SF	34:55
Screaming	36:12	SF	36:17
Going outside	36:42	IMG, LI	37:18
RT:ringtone SF:sound fluctuation LI:light intensity			
IMG:image similarity LH:fingerprint			

Table 3: Average Trigger Accuracy and Event Detection latency (including false positives)

Triggers	Coverage	Latency	False Positive.
RT	100%	1 second	10%
IMG	80%	30 seconds	33%
LH	75%	3 seconds	33%
LI	80%	30 seconds	0%
SF	75%	5 second	20%

(2) Field Experiment: Thanksgiving Party

The two field experiments were performed to understand MoVi's ability to create a highlights in real social occasions. This is significantly more challenging in view of a far larger event space, potentially shaking cameras from real excitement, greater mobility within the party, background music, other noise in the surroundings, etc. The first experiment was at a Thanksgiving party, attended by 14 people. Five attendants were instrumented with iPods and phones. After the party, videos from the five cameras were distributed to five different people for evaluation. Manually selecting the highlights from the full-length video was unanimously agreed to be a difficult and annoying task (often done as a professional service). However, with help from friends, we were able to obtain the *Human Selected* moments. The MoVi generated highlights were also generated, and compared against the manual version.

Figure 15 shows the comparative results *at the granularity of one second.* The X-axis models the passage of time, and the

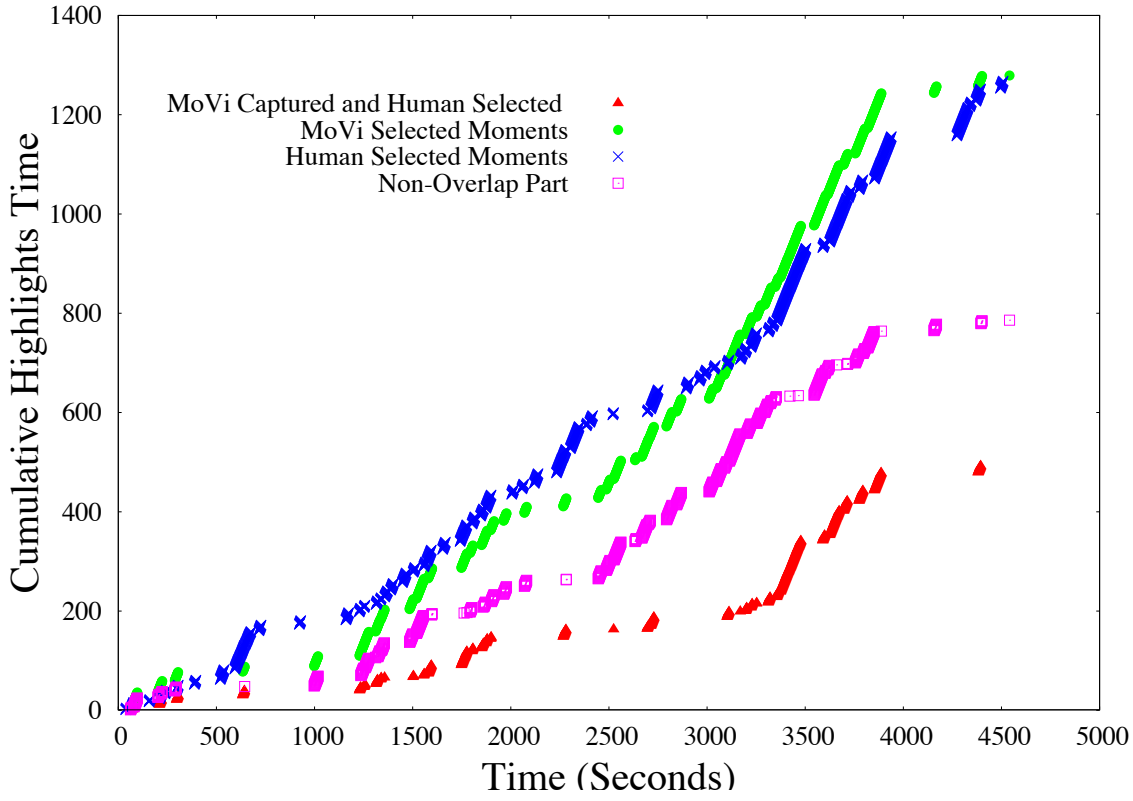

Figure 15: Comparison between MoVi and human identified event list (Thanksgiving)

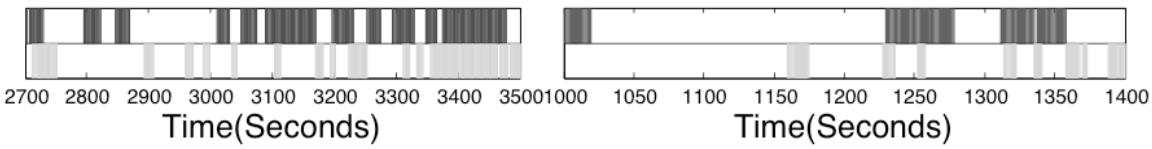

Figure 16: Zoom in view for two parts of Figure 15. Dark gray: MoVi, light gray: human selected

Y-axis counts the *cumulative* highlights duration selected until a given time. For instance, Y-axis = 100 (at X-axis = 1200) implies that 100 seconds of highlights were selected from the first 1200 seconds of the party. Figure 16 presents a zoom-in view for the time windows 2700-3500 and 1000-1400 seconds. We observe that the MoVi highlights reasonably tracks the Human Selected (HS) highlights. The curve (composed of triangles) shows the time-points that *both* MoVi and HS identified as interesting. The non-overlapping parts (i.e., MoVi selects that time, but HS does not) reflect the false positives (curve composed of squares).

Based on this evaluation, we computed the overall *Precision* to be 0.3852, *Recall* to be 0.3885, and *Fall-out* to be 0.2109. Notice that the overall precision is computed by using the union of all human selected video as the retrieval target. Therefore, if a moment is labeled as interesting by one user, it is considered interesting. We also compared the improvement over a random selection of clips (i.e., percentage of MoVi's overlap with human (MoH) minus percentage of Random's overlap with Human (RoH), divided by RoH). MoVi's improvement is 101% on average.

In general, the false positives mainly arise due to two reasons: (1) Falsely detected triggers: since the sensor-based event detection method cannot achieve 100% accuracy, false positives can occur. Since we assign more weight to infrequently happening triggers such as laughter, we trade off some precision for better recall. (2) Subjective choice: the user reviewing the video may declare some of the events (even with triggers) as not interesting. Since this is a subjective judgment, false positive will occur.

Table 4 shows the per-user performance when the MoVi highlights is compared with individual user's selections. Since each user only selects a very small portion of the entire video, according to equation 1, the computed precision is expected to be low. As a result, Recall and performance gains over the Random scheme are more important metrics in this case. The average improvement proves to be 101%.

The results are clearly not perfect, however, we believe, are quite reasonable. To elaborate on this, we make three obser-

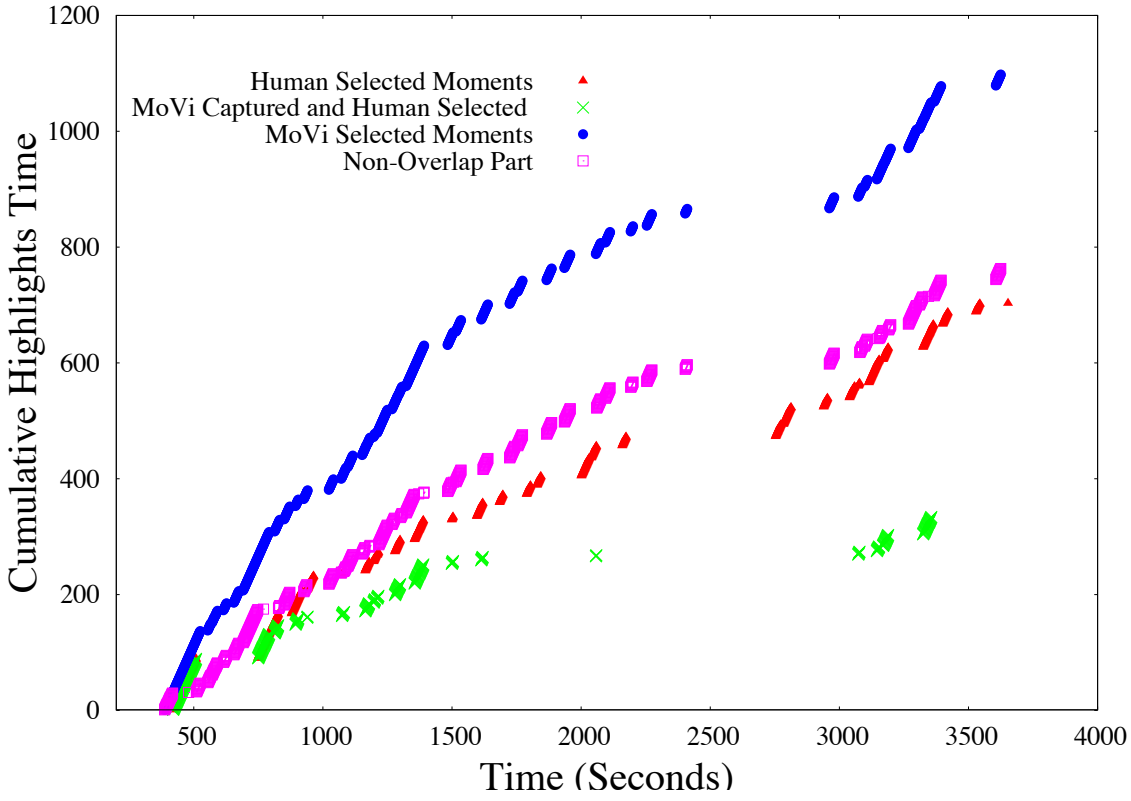

Figure 17: Comparison between MoVi and human identified event list (SmartHome)

vations. (1) We chose a strict metric wherein MoVi-selected clips are not rewarded even if they are very close (in time) to the Human Selected clips. In reality, social events are not bounded by drastic separations, and are likely to "fade away" slowly over time. We observed that MoVi was often close to the human selected segments; but was not rewarded for it. (2) We believe that our human selected videos are partly biased – all users enthusiastically picked more clips towards the beginning, and became conservative/impatient over time. On the other hand, MoVi continued to automatically pick videos based on pure event-triggers. This partly reduced performance. (3) Finally, we emphasize that "human interest" is a sophisticated notion and may not always project into the sensing domains we are exploring. In particular, we observed that humans identified a lot of videos based on the topics of conversation, based on views that included food and decorative objects, etc. Automatically detecting such interests will perhaps require sophisticated speech recognition and image processing. In light of Google's recent launch of the Google Goggles, an image search technology, we are considering its application to MoVi. If MoVi searches its camera pictures through Google Goggles, and retrieves that the view is of a wedding dress, say, it could be a prospective trigger. Our current triggers are unable to achieve such levels of distinction. Yet, the MoVi-generated highlights was still interesting. Several viewers showed excitement at the prospect that it was generated without human intervention.

Table 4: Per-user performance (Thanksgiving party)

User	Precision	Recall	Fall-out	Over Random
1	21%	39%	23%	51%
2	5%	33%	12%	162%
3	9%	37%	25%	46%
4	18%	74%	20%	222%
5	4%	22%	17%	26%

Field Experiment: SmartHome Tour

The Duke SmartHome is a live-in laboratory dedicated to innovation and demonstration of future residential technology. Eleven members of our research group attended a guided tour into the SmartHome. Five users wore the iPods and carried the N95 phones. Figure 17 shows the results.

In this experiment, the human highlights creator did not find too many interesting events. This was due to the academic nature of the tour with mostly discussions and references to what is planned for future. The human selected moments proved to be very sparse, making it difficult to capture them precisely. MoVi's *Precision* still is 0.3048, *Recall* is 0.4759, and *Fall-out* is 0.2318. Put differently, MoVi captured most of the human selected moments but also selected many other moments (false positives). Compared to *Random* (discussed earlier), the performance gain is 102% on average. Table 5 shows the performance when manual highlights was created from the union of multiple user-selections.

In summary, we find that inferring human interest (espe-

Table 5: Per-user performance (SmartHome)

User	Precision	Recall	Fall-out	Over Random
1	21%	62%	23%	124%
2	19%	45%	25%	67%
3	6%	50%	22%	116%

cially semantically defined ones) is hard. Although this is a current limitation, MoVi's trigger mechanism can capture most events that have an explicit sensor clue. The highlighted video is of reasonably good quality in terms of camera-angle, lighting, and content. Although not a human-videographer replacement, we believe that MoVi can serve as an additional tool to complement today's methods of video-recording and manual editing.

5. RELATED WORK

The ideas, algorithms, and the design of MoVi is drawn from a number of fields in computer science and electrical engineering. Due to limited space, it is difficult to discuss the entire body of related work in each of these areas. We discuss some of the relevant papers from each field, followed by works that synthesize them on the mobile computing platform.

Wearable Computing and SenseCam. Recent advances in wearable devices are beginning to influence mobile computing trends. A new genre of sensing devices is beginning to blend into the human clothing, jewelry, and in the social ambience. The Nokia Morph [11], SixthSense camera-projectors [9], LifeWear, Kodak 1881 locket camera [22], and many more are beginning to enter the commercial market. A large number of projects, including MIT GroupMedia, Smart Clothes, AuraNet and Gesture Pendant [23–25] have exploited these devices to build context-aware applications. Microsoft Research has recently developed SenseCam, a wearable camera equipped with multiple sensors. The camera takes a photo whenever the sensor readings meet a specified degree of fluctuations in the environment (e.g., change in light levels, above-average body heat). The photos are later used as a pictorial diary to refresh the user's memory, perhaps after a vacation [7]. MoVi draws from many of these projects to develop a collaborative sensing and event-coverage system on the mobile phone platform.

Computer Vision. Researchers in Computer Vision have studied the possibility of extracting semantic information from pictures and videos. Of particular interest are works that use audio-information to segment video into logical events [26, 27]. Another body of work attempts scene understanding and reconstruction [28, 29] by combining multiple views of the same scene/landmark to a iconic scene graph. On a different direction, authors in [30] have investigated the reason for longer human-attention on certain pictures; the study helps in developing heuristics that are useful to shortlist "good" pictures. For instance, pictures that display greater symmetry, or have a moderate number of faces (identifiable through face recognition), are typically viewed longer [31]. Clearly, MoVi is aligned to take advantage of these findings. We are by no means experts in Computer Vision, and hence, will draw on the existing tools to infer social events and select viewing angles. Additional processing/algorithms will still be necessary over the other dimensions of sensing.

Information Retrieval. Information retrieval (IR) [32] deals with the representation, storage, and organization of (and access to) information items. Mature work in this area, in collaboration with Artificial Intelligence (AI) and Natural Language Processing (NLP), have attempted to interpret the semantics of a query, and answer it by drawing from disparate information sources [33]. Some research on mobile information retrieval [34] have focused on clustering retrieval results to accommodate small display devices. Our objective of extracting the "highlights" can be viewed as a query, and the mobile phone sensors as the disparate sources of information. MoVi is designed to utilize metrics and algorithms from information retrieval.

Sensor Network of Cameras. Recently, distributed camera networks have received significant research attention. Of interest are projects that observe and model sequences of human activity. For example, BehaviorScope [35] builds a home sensor network to monitor and help elders that live home alone. Distributed views are used to infer networked cameras' locations. Smart cameras [36] are deployed to track real time traffic load. These works provide us useful models to organize information from multiple sensors/mobile nodes in a manner that will provide good coverage and correlation.

People-Centric Sensing. In mobile computing, people-centric, participatory sensing through mobile devices are gaining rapid popularity. Example applications include CenseMe [8], which detects the user's activity status through sensor readings and shares this status over online social networks. SoundSense [6] implements audio processing and learning algorithms on the phone to classify ambient sound types – the authors propose an audio journal as an application. YinzCam [37] enables watching sports games through different camera angles on mobile devices. While these systems are individual specific, others correlate information from multiple sources to generate a higher level view of the environment. PEIR, Micro-Blog, Urban Tomography [38, 39], are few examples in this area.

Our work may be considered a mash-up of diverse techniques that together realize a fuller system. Customizing the techniques to the target application often presents new types of research challenges that are imperceptible when viewed in isolation. As an example, deducing human collocation based on ambient acoustics have been a studied problem [40]. Yet, when applied to the social context, two physically nearby individuals may be participating in conversations in two adjacent dinner tables. Segregating them into distinct social groups is non-trivial. MoVi makes an attempt to assimilate the rich information feeds from mobile phones and process them using a combination of existing techniques drawn from vision, data-mining, and signal processing. In that sense, it is a new mash-up of existing ideas. Our novelty comes from the collaboration of devices and the automatic detection of interesting events. Our preliminary ideas have been published in [41].

6. LIMITATIONS AND ONGOING WORK

MoVi is a first step towards a longer term project on collaborative sensing in social settings. The reported work has

limitations, several of which stem from the non-trivial nature of the problem. We discuss these limitations along with avenues to address some of them.

Retrieval accuracy. The overall precision of our system certainly has room for improvement. Since "human interest" is a semantically sophisticated notion, to achieve perfect accuracy is challenging. However, as an early step towards social event retrieval, the precision of around 43% can be considered encouraging [27, 33, 42].

Unsatisfying camera views. Though view selection is used, cameras in a group may all have unsatisfying views of a specific event. The video highlights for these events exhibit limited quality. This problem can be partly addressed by introducing some static cameras into the system to provide a degree of all-time performance guarantee. The ideas in this paper can be extended to these static wall mounted/wearable cameras equipped with multiple sensors.

Energy consumption. Continuous video-recording on the iPod Nanos persists for less than 2 hours. The mobile phone sensors can last for around 4 hours. Thus, in parallel to improving our event detection algorithms, we are beginning to consider energy as a first class design primitive. One option is to explore peer to peer coordination among phones – few phones may monitor a social zone, allowing other phones to sleep. Lightweight duty cycling, perhaps with periodic help from the server, is a part of our future effort.

Privacy. User privacy is certainly a concern in a system like MoVi. For this paper, we have assumed that attendants in a social party may share mutual trust, and hence, may agree to collaborative video-recording. This may not scale to other social occasions. Certain other applications, such as travel blogging or distributed surveillance may be amenable to MoVi. Even then, the privacy concerns need to be carefully considered.

Greater algorithmic sophistication. We have drawn from preliminary ideas, tools, and algorithms, in data mining, information retrieval, signal processing, and image processing. A problem such as this requires greater sophistication in these algorithms. Our ongoing work is focused towards this direction, with a specific goal of prioritizing among different event triggers. One advantage of prioritizing will permit relative ranking between event-triggers; this may in turn allow for creating MoVi highlights for a user-specified duration. At present, the MoVi highlights are of a fixed duration.

Dissimilar movement between phones and iPods. We often observed that the acceleration in the phone was not necessarily correlated to the vibration in the video-clip. This is a result of the phone being on the belt and the iPod taped to the chest. Sensors on different parts of the body may sense differently, leading to potential false positives. One possibility is to apply image stabilization algorithms on the video itself to gain better view quality.

7. CONCLUSION

This paper explores a new notion of "social activity coverage". Like spatial coverage in sensor networks (where any point in space needs to be within the sensing range of at least one sensor), *social activity coverage* pertains to covering moments of social interest. Moreover, the notion of social activity is subjective, and thus identifying triggers to cover them is challenging. We take a first step through a system called Mobile Phone based Video Highlights (MoVi). MoVi collaboratively senses the ambience through multiple mobile phones and captures social moments worth recording. The short video-clips from different times and viewing angles are stitched offline to form a video highlights of the social occasion. We believe that MoVi is one instantiation of *social activity coverage*; the future is likely to witness a variety of other applications built on this primitive of collaborative sensing and information distillation.

8. ACKNOWLEDGEMENT

We sincerely thank our shepherd Stefan Saroiu, as well as the anonymous reviewers, for their immensely valuable feedback on this paper. We are also grateful to Victor Bahl for his suggestions during the formative stages of MoVi. We also thank Souvik Sen, Sandip Agarwal, Jie Xiong, Martin Azizyan, and Rahul Ghosh for wearing the iPods on their shirts during live experiments. Finally, we thank our all research group members, including Justin Manweiler, Ionut Constandache, and Naveen Santhapuri for the numerous insightful discussions during the research and evaluation phase.

9. REFERENCES

[1] S. Gaonkar, J. Li, R. R. Choudhury, L. Cox, and A. Schmidt, "Micro-Blog: Sharing and querying content through mobile phones and social participation," in *ACM MobiSys*, 2008.

[2] C. Torniai, S. Battle, and S. Cayzer, "Sharing, discovering and browsing geotagged pictures on the web," *Multimedia Integration & Communication Centre, University Firenze, Firenze, Italy, Hewlett-Packard Development Company, LP*, 2007.

[3] A. Dada, F. Graf von Reischach, and T. Staake, "Displaying dynamic carbon footprints of products on mobile phones," *Adjunct Proc. Pervasive 2008*.

[4] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, and M. Hansen, "Image browsing, processing, and clustering for participatory sensing: Lessons from a dietsense prototype," in *ACM EmNets*, 2007.

[5] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *ACM SenSys*, 2008.

[6] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "SoundSense: scalable sound sensing for people-centric applications on mobile phones," in *ACM MobiSys*, 2009.

[7] E. Berry, N. Kapur, L. Williams, S. Hodges, P. Watson, G. Smyth, J. Srinivasan, R. Smith, B. Wilson, and K. Wood, "The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report," *Neuropsychological Rehabilitation*, 2007.

[8] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, "Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of CenceMe Application," in *ACM Sensys*, 2008.

[9] P. Mistry., "The thrilling potential of SixthSense technology," *TED India*, 2009.

[10] E. Cuervo, A. Balasubramanian, D. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making Smartphones Last Longer with Code Offload," in *ACM MobiSys*, 2010.

[11] S. Virpioja, J.J. Vayrynen, M. Creutz, and M. Sadeniemi, "Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner," *Machine Translation Summit XI*, 2007.

[12] T. Nakakura, Y. Sumi, and T. Nishida, "Neary: conversation field detection based on similarity of auditory situation," *ACM HotMobile*, 2009.

[13] H. Homburg, I. Mierswa, B. Moller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering," in *ISMIR*, 2005.

[14] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *ISMIR*, 2000.

[15] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice hall, 1993.

[16] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, 1978.

[17] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[18] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in *ISMIR*, 2003.

[19] M. Azizyan, I. Constandache, and R. Roy Choudhury, "SurroundSense: mobile phone localization via ambience fingerprinting," in *ACM MobiCom*, 2009.

[20] S. T. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," *IEEE CVPR*, 2005.

[21] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST Meeting Recognition Workshop*, 2004.

[22] "Kodak 1881 locket camera," *http://www.redwoodhouse.com/wearable/*.

[23] S. Mann, "Smart clothing: The wearable computer and wearcam," *Personal and Ubiquitous Computing*, 1997.

[24] J. Schneider, G. Kortuem, D. Preuitt, S. Fickas, and Z. Segall, "Auranet: Trust and face-to-face interactions in a wearable community," *Informe técnico WCL-TR*, 2004.

[25] T. Starner, J. Auxier, D. Ashbrook, and M. Gandy, "The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring," in *IEEE ISWC*, 2000.

[26] T. Zhang and C. C. J. Kuo, "Audio-guided audiovisual data segmentation, indexing, and retrieval," in *SPIE*, 1998.

[27] M. Baillie and J. M. Jose, "An audio-based sports video segmentation and event detection algorithm," in *CVPRW*, 2004.

[28] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. M. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *Proc. ECCV*, 2008.

[29] "Microsoft Photosynth," *http://photosynth.net/*.

[30] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *IEEE CVPR*, 2006.

[31] M. Nilsson, J. Nordberg, and I. Claesson, "Face detection using local SMQT features and split up snow classifier," in *IEEE ICASSP*, 2007.

[32] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, Addison-Wesley Reading, MA, 1999.

[33] D.A. Grossman and O. Frieder, *Information retrieval: Algorithms and heuristics*, Kluwer Academic Pub, 2004.

[34] C. Carpineto, S. Mizzaro, G. Romano, and M. Snidero, "Mobile information retrieval with search results clustering: Prototypes and evaluations," *Journal of the ASIST*, 2009.

[35] T. Teixeira and A. Savvides, "Lightweight people counting and localizing in indoor spaces using camera sensor nodes," in *ACM/IEEE ICDSC*, 2007.

[36] M. Bramberger, J. Brunner, B. Rinner, and H. Schwabach, "Real-time video analysis on an embedded smart camera for traffic surveillance," in *RTAS*, 2004.

[37] "YinzCam," *http://www.yinzcam.com/*.

[38] "UrbanTomograph," *http://research.cens.ucla.edu/events/?event_id=178*.

[39] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda, "PEIR, the personal environmental impact report, as a platform for participatory sensing systems research," in *ACM Mobisys*, 2009.

[40] N. Eagle, "Dealing with Distance: Capturing the Details of Collocation with Wearable Computers," in *ICIS*, 2003.

[41] X. Bao and R.R. Choudhury, "VUPoints: collaborative sensing and video recording through mobile phones," in *ACM Mobiheld*, 2009.

[42] M.S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM TOMCCAP*, 2006.