VUPoints: Collaborative Sensing and Video Recording through Mobile Phones

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

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

ABSTRACT

Mobile phones are becoming a convergent platform for sensing, computation, and communication. This paper envisions *VUPoints*, a collaborative sensing and video-recording system that takes advantage of this convergence. Ideally, when multiple phones in a social gathering run VUPoints, the output is expected to be a short video-highlights of the occasion, created without human intervention. To achieve this, mobile phones must sense their surroundings and collaboratively detect events that qualify for recording. Short video-clips from different phones can be combined to produce the highlights of the occasion. This paper reports exploratory work towards this longer term project. We present a feasibility study, and show how social events can be sensed through mobile phones and used as triggers for video-recording. While false positives cause inclusion of some uninteresting videos, we believe that further research can significantly improve the efficacy of the system.

Categories and Subject Descriptors

C.2.4 [**Computer Communication Networks**]: Distributed Systems – *Distributed Applications*; H.4.3 [**Information Systems Applications**]: Communications Applications – *Information Browsers*; H.5.3 [**Information Interfaces and Presentation**]: Groups and Organization Interfaces – *Collaborative Computing*

General Terms

Design, Experimentation, Measurement, Performance, Human Factors

Keywords

Mobile phones, video recording, participatory sensing, activity recognition, social networks, collaborative ambience sensing, wearable devices, image processing

MobiHeld'09, August 17, 2009, Barcelona, Spain.

1. INTRODUCTION

The inclusion of cameras in mobile phones has enabled people to take spontaneous pictures and short video clips. The pictures/videos are typically taken when a person decides that a certain event is of interest and explicitly points-and-clicks the camera to capture it. However, in social gatherings for instance, people are involved in various activities and often forget to record interesting moments – they realize this in retrospect. Even if one remembers to record, studies show a degree of unwillingness to do so. This is because the person recording must become a passive observer of the event, as opposed to an active participant. Even if a person is willing to passively observe and video-record, parallel events in different parts of the party may be difficult to cover. To that end, even multiple videographers may be inadequate.

We postulate that mobile phones can be harnessed to collaboratively record events of interest. Spatially nearby phones may collaboratively sense their ambience [1,2] and infer event-triggers that suggest an "exciting" moment. For example, an outburst of laughter in a party can be an acoustic trigger for video-recording. Many people turning towards the wedding speech – detected from the correlated compass orientations of the phones – can be another example trigger. Based on such triggers, the phone with the best view can be automatically activated to record the event for a short duration. At the end of the party, the individual recordings from different phones can be correlated over time, and "stitched" into a single video-highlights of the party. Creating automatic video-highlights through mobile phones can enable a variety of new applications in mobile social computing. This paper describes early research aimed at translating this idea into a publicly usable system. We call this system *VUPoints*.

A natural concern is: *phones are often inside pockets and may not be useful for recording events.* While this is certainly the current trend, a variety of wearable mobile phones are already entering the commercial market [3]. Phone sensors may blend into clothing and jewelry (necklaces, wrist watches, shirt buttons), exposing the camera and microphones to the surroundings. A variety of urban sensing applications is already beginning to exploit them [1,4]. *VUPoints* can leverage them too.

Even if phones are mostly in pockets, it may still be useful if *VUPoints* can offer users with cues to record an interesting moment. If the phones collaboratively identify an event, and also locate the person in the best view of this event, that per-

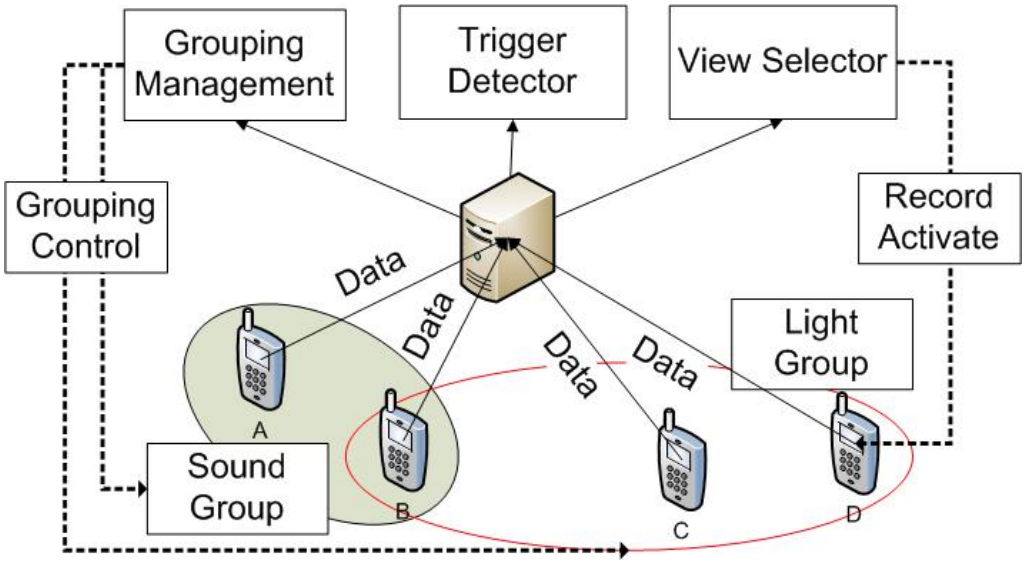

Figure 1: The VUPoints architecture: Phones are clustered according to zones and a zone-monitor reports sensed data to the VUPoints server. The server scans this data for potential triggers that suggest a socially interesting event. Once an event is suspected, the server prescribes a few phones at vantage locations to activate video-recording. The individually recorded video-clips are later "stitched" together by the VUPoints server, making a video highlights of the social occasion.

son's phone can ring. Knowing this to be a cue, the person can record the ongoing events for a suitable duration. Involving multiple people for short durations is a good load-balancing act, precluding any single person from becoming the designated videographer. As an alternative to requiring user participation, one may scatter small cameras in the surrounding, or even utilize installed surveillance cameras. These cameras may record the entire occasion, while collaborative event-triggers from phones can be used to select the portions of interest [5]. This may preclude the need for video-recording with phones and any type of human participation.

Thus, assuming cameras exposed to the surroundings, automatic event coverage via mobile phones may be a useful application. To this end, we identify 3 research components:

(1) In view of energy constraints and load balancing, not all phones need to continuously monitor their surroundings for interesting events. Ideally, a few phones from each *social zone* can be turned on. Social zones can be defined as the group of users/phones involved in the same social activity. People conversing at a dinner table may be considered in the same social zone; other zones may be friends watching TV together, or guests gathered around a birthday cake. Clearly, the social zones may change over time as people mingle with others. The first challenge, therefore, is to dynamically identify social zones, and designate a few phones as zone monitors.

(2) Based on streaming data from different zone monitors, the VUPoints server must detect exciting events that call for a video-recording. Identifying events entails aggregation of sensed data from multiple sensing dimensions, including sound, image, accelerometer, and compasses. To improve confidence, sensed data may need to be further correlated across multiple zone monitors. Once an event is identified, video-recording needs to be triggered.

(3) Not every phone in a social zone may have the same view of the exciting event. The third challenge, therefore, is to pick a few phones that have good views of the event from different angles. This is hard because the best view may be subjective, hence difficult to deduce automatically. The difficulty may be partly overcome by conservatively turning on multiple recordings and later choosing the best views manually. Of course, this comes as a tradeoff with higher battery consumption, memory usage, and (some) manual intervention.

This paper presents early explorations towards building such a collaborative sensing and video-recording system using mobile phones. We explore design considerations, basic approaches, and preliminary (offline) evaluations using Nokia N95 and N6210 phones. A social occasion is fully video-recorded through a dedicated phone camera, and the collaborative triggers (from users' phones) are used offline to extract out short video-clips. We employ similarity in ambient sounds and light-intensities to form the *social groups*. Between members of the same social group, we use view-similarity, laughter recognition, compass orientation, and combinations thereof, to identify an event-trigger. A short video-clip is extracted around the time of each event; each of the clips are concatenated to form a single video-highlights of the occasion. This video-highlights is the output of VUPoints.

2. SYSTEM ARCHITECTURE

Fig. 1 shows the envisioned client/server architecture of VUpoints. We briefly describe the high level operations first, and present details in the next sections. Periodically, all phones upload sensed data to the VUPoints server. The group management module, running at the server, analyzes the sensed data to compute social zones (also called social groups). Phones are notified of the social group they belong to, and zone monitors are designated. The monitors continue to stream sensed data to the server, while a trigger detection module scans the

data to detect events of interest. Once an event is detected, the view selector module selects a few phones likely to have good views of the event. Video-recording at these phones are then activated. As a starting point, we have developed an offline version of the system. Encouraged by the results, our ongoing work is focussed on a fuller implementation.

3. SYSTEM DESIGN

This section discusses the main components of VUPoints, namely, Social Group Identification, Trigger Detection, and Video Activation.

3.1 Social Group Identification

To detect interesting events in a gathering, one approach is to require all phones to become ambience monitors. However, in view of energy and bandwidth constraints, we propose to partition the occasion into social groups, and designate a few monitors in each of them. Of course, these social groups are not well defined – they are not always spatial because two people in proximity may be engaged in different conversations in adjacent dinner tables. In fact, these groups have a social nature, meaning that the social ambience each group perceives may be similar. For instance, people seated around a table may be facing the same object in the center of the table, while people near the TV may have a similar acoustic ambience. We plan to capture this similarity in ambience to approximately group mobile phones. To this end, ambient sound and light are of interest.

Acoustic Grouping

To begin with an approximate grouping, the *VUPoints* server chooses a random phone to play a short high-frequency ring-tone (similar to a wireless beacon). The ring-tone should ideally be outside the audible frequency range, such that it is not interfered by human voices (with Nokia N95 phones, we were able to generate narrow-bandwidth tones at the edge of the audible range). Once the tone is sent, the server requests all phones to report back their overheard sounds. It then generates a frequency-domain representation of the sounds reported by each phone (a vector, $\overrightarrow{S}$, with 4000 dimensions), and computes the *similarity* of this vectors with the vector generated from the known ring-tone ($\overrightarrow{R}$). The similarity function, expressed below, is essentially a weighted intensity ratio (Doppler shifts are explicitly addressed by computing similarity over a wider frequency range).

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

Fig. 2 shows the similarity values over time at two different phones placed near a ring-tone transmitter. The similarity spike is around the same time, indicating that they may have overheard the same ring-tone. All phones that exhibit more than a threshold similarity are grouped by the server – called an *acoustic group*. Among these phones, a few monitors are randomly selected and tasked to keep their microphones on. The server separates out phones that do not belong to this acoustic group and instructs a random one from them to send a ring-tone. The acoustic grouping process continues, until all phones have been assigned to at least one acoustic group. At this point, the party is said to be “acoustically covered" because each acoustic zone is monitored by at least one phone.

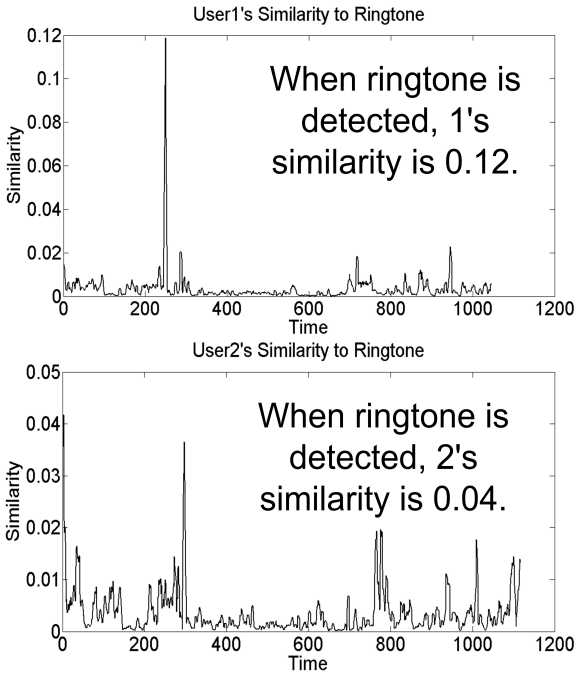

Figure 2: Frequency domain similarity between two users' sensed sound and the known ring-tone.

Periodic high-frequency ring-tones may be annoying. Further, it may sometimes not accurately reflect the socio-acoustic groups. An alternate approach is to compute similarities between phones' ambient sounds, and cluster them accordingly. Thus, the VUPoints server also collects ambient sound samples from each phone, computes the pair-wise similarity, and uses them to cluster phones in socio-acoustic groups [6–8]. The pair-wise similarity is again computed in the frequency domain (50 to 1600 Hz), using the definition of cosine similarity as follows:

$$Similarity = \frac{\overrightarrow{F(A)} \cdot \overrightarrow{F(B)}}{\|\overrightarrow{F(A)}\| \cdot \|\overrightarrow{F(B)}\|}$$

where F(A) and F(B) are the frequency vector of sounds from phones A and B. The similarity values are then clustered, each cluster representing a social group. VUPoints adopts both the ring-tone and ambience-sound approaches.

Grouping through Light Intensity

In some cases, light intensities vary across different zones. 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. Light intensity can be considered as another dimension of partitioning the social gathering; monitors can be designated for each light group. Upon receiving a light-based trigger from one of the monitors, multiple members in that light group can be activated for video-recording. 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 and nearby shadows. To ensure robustness, we used conservative approaches for classification – we defined only 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 associated at all. Figure 3 illustrates the 3 light classes as experienced in our experiments.

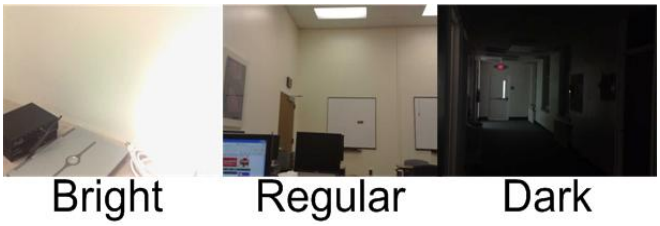

Figure 3: Light intensity classification

3.2 Trigger Detection

Sifting through a steady flow of sensed information from different phone monitors, the VUPoints server must identify patterns indicative of an interesting event. This is hard because the notion of "interesting" is subjective. Thus, in an attempt to approximate the notion of "interesting", we scan for unusual homogeneity or diversity in the sensor readings. Changes in the ambience, sudden user activity, and combinations thereof are suspected as triggers to initiate video-recording. We present some possibilities.

Detecting View Similarity

When phone cameras are found to be viewing the same object from different angles, it could be an event of interest (EoI). The birthday cake on a table, a wedding toast, a celebrity's arrival, are some examples. In such scenarios, it might be useful to video-record from multiple viewpoints, and reconstruct the scene from all angles. Of course, we need the trigger that detects when a social group is viewing the same object. For this, we use an image generalization technique called spatiogram [9, 10]. Spatiograms are essentially color histograms encoded with spatial information. With such a representation, pictures of the same object from different viewing angles can be shown to have 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, 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. Fig. 4(a) and (b) show the view from two phones while their owners are playing a multi-player video-game on a projector screen. Observe that the views are not of the same instant. Yet, the similarity proves to be 0.75, much higher than the similarity observed when one of the cameras faced away from the screen. Hence, when VUPoints observes such a high similarity among the views of the zone monitors, it immediately triggers video recording among all the phones in that zone. Of course, the view similarity can be combined with other triggers to improve the confidence. Multi-sensor triggers is a part of our ongoing work.

(a) User1 Left (b) User2 Right

Figure 4: Views with different similarities

Detecting Acoustic Signatures

Human reactions like laughters, screaming, clapping, whistling, can be viewed as acoustic responses to interesting events. It may be feasible to recognize these sound signatures. As a starting point, we are able to design a fingerprint for laughter. Validation across a sample size of 100 laughters, from 4 different students, offered evidence that our laughter-signature is independent of the individual. Hence, we used this signature at the server, and computed the similarity with ambient-sound measurements arriving from different zone monitors. Since more than half of human-voice energy is typically concentrated on frequencies below 2000 Hz, we used a weighted cosine similarity (with higher weights in this band). Whenever we detected a similarity greater than a empirically tuned threshold, we activated video-recording at all the phones in that social group.

Detecting 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, VUPoints 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 reasonable with compasses). The compass-based orientation triggers can be further combined with accelerometer triggers, indicating that people have turned and moved together. The confidence in the trigger can then be higher.

Detecting Ambience Fluctuations

In addition to specific specific signatures, the general ambience of a place may fluctuate as a whole. Lights may be turned off for 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 detectible across multiple users, they can result in a high-confidence trigger. VUPoints employs such kind of collaborative schemes on the photo-acoustic ambience. Different thresholds on fluctuations are empirically set – the thresholds are higher for individual sensors, and relatively lower for joint sensing. Whenever any of the sensors (or combined) exceed the corresponding threshold, all the cameras of phones are triggered for video-recording. Fig. 5 shows an example of the sound fluctuation in time domain. The dark lines specify the time-points when

the average of one-second time windows exceed a threshold. These are accepted as triggers, and all phones are instructed to video-record. The individual clips are “stitched" offline to generate the desired video-highlights.

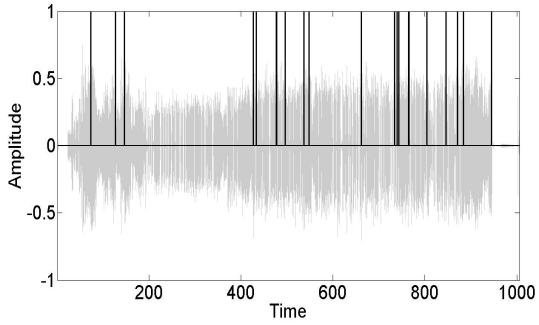

Figure 5: Sound fluctuation in time domain

4. EXPERIMENTS AND EVALUATION

We ran four experiments for testing VUPoints. The experiments involved 3 to 4 users, pretending to be in different types of gatherings. Each user taped a Nokia N95 phone near his shirt pocket. The N95 model has a 5 megapixel camera, and a 3-axis accelerometer. Two of the users also carried a Nokia N6210 in their pockets – the N6210 has a compass that the N95s do not have. The user-carried phones formed social groups and detected triggers throughout the entire occasion. The occasions were also continuously video-recorded by a separate phone. At the end, all sensed and video-recorded data (from all the phones) were downloaded, and processed in MATLAB. The triggers were identified, and using their time-stamps, a 20-second video-clip for each trigger was extracted from the continuous video file. All the clips were then “stitched" in a chronological manner. When two clips overlapped in time, both the clips were included.

We evaluate the system by asking a person (unaware of the entire experiment) to watch the full video, and identify the interesting events that she would recommend for recording. We then compare VUPoints' highlights with the recommended events. Detailed results from one of the experiments is shown in Fig. 6 and Table 1.

Fig. 6 shows the grouping results. The first two rows are the actual (sound and light) groups as people performed social/individual activities; the next two rows are the groups prescribed by VUPoints. The start time – 01:25:56pm – is the time of the first ring-tone transmission. The sound grouping at this time is based on ring-tone comparison. Succeeding sound groupings are all based on ambient-sound similarities (thereby avoiding frequent ring-tone transmissions). Ring-tone groupings are certainly better than ambience-based grouping, particularly because when an individual speaks, her ambient sound is drowned by her own voice. Ring-tones avoid this issue and can create reasonably good groups. Light intensity based grouping also prove to be robust. Evident from the figure, VUPoints always detects the light groups correctly in this experiment. We observe similar trends in other experiments we performed in different lighting environments.

Table 1 shows early results for event detection. The first

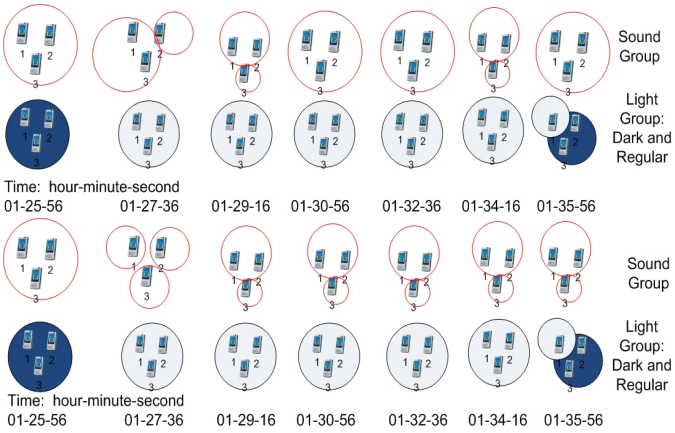

Figure 6: Actual social grouping Vs VUPoints' grouping over different sensing axes.

two columns show the recommended events and their occurrence times, while the next two columns show the type of triggers that detected them, and the corresponding detection times. Evident from the table, at least one of the triggers were able to capture the events, suggesting that VUPoints achieves a good coverage of events. However, it also included a number of events that were not worthy of recording (hence, false positives). We note that the recommended portions of the video summed up to 1.5 minutes (while the original video was for 5 minutes). The VUPoints highlights proved to be for 2.5 minutes and covered all the recommended portions(indicating a false positive of 1 minute).

Table 1: 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 2: Average Trigger Accuracy (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%
ACC, COM	unreliable		unreliable
ACC:accelerometer, COM:group rotation			

To understand the overall performance, we define *event coverage* as $C = E_{rec}/E_{int}$, where E_{rec} is the number of events recorded by VUPoints, and E_{int} is the number of socially interesting events according to a random person. To understand the rate of false positives, we compute $F = E_{unint}/E_{int}$, where E_{unint} is the number of uninteresting events captured by VUPoints. Across 4 experiments, average event coverage proved to be 80%, while average false positive rate proved to be 33%.

Table 2 shows the average accuracy on a per-trigger basis. Evidently, ring-tones and light-based triggers perform reasonably well, with a high rate of event detection and reasonably low false positives. Image triggers also achieve good event detection, but incur several false positives. Most of the laugh triggers were captured – a few false positives was incurred when everyone was excited and spoke in a loud voice. The compass trigger did not perform reliably. This is probably because we had only two compass-phones, and they were not adequate to detect correlated turns. We expect that greater number of compass readings may yield better results.

5. LIMITATIONS AND ONGOING WORK

This paper reports exploratory work on a longer term project on collaborative sensing for social-event coverage (akin to spatial coverage in sensor networks). Clearly, there are several limitations of the system in its current form. (1) The number of triggers are limited and may not be sufficient to capture all the socially "interesting" moments that arise. Improved information processing is necessary to identify complex patterns that are together indicative of a prospective event. (2) Even if most events are captured, some moments may be over before VUPoints can trigger video-recording (e.g., people may laugh at a joke, but recording the incident after the laugh will not cover the joke). (3) The energy and privacy concerns with the system are certainly open questions. Continuous sensing on multiple sensors, as well as periodic communication to the VUPoints server, is likely to drain the phone's battery. Video-recording on the phones will further add to the consumption. Our ongoing work is increasing the information processing burden on the phone (to reduce the communication overhead), and load-balancing across multiple zone-monitors (to prevent continuous sensing). (4) The evaluation results reported here are in an artificially generated setting with few students. A realistic social function may pose greater challenges in grouping and trigger detection; however, the collaboration between many more phones may greatly improve the efficacy of zone demarcation and trigger detection.

Extending the VUPoints prototype to a fuller, deployable systems is the focus of our ongoing work. Assuming that such a system can accomplish the desired efficacy, a variety of new applications may emerge on top of such a collaborative ambience-sensing framework. For instance, we are exploring a system called *Location based RSS Feeds*, where a user subscribes to a physical location to learn about specific events at that location. A person may set up an RSS feed on the plaza of a mall, expressing interest in live music at the plaza. People at the mall can collaboratively sense the music in the ambience and notify all those that have subscribed to that feed. Another application we are developing is *Context Aware Personal Home-Page*. The main idea is to automatically generate a personal home-page based on the surrounding context, and share it in social proximity networks. Thus, a professor may expose her research interests and papers while she is at a conference, but share her movie and music tastes while she is at a cultural festival. We expect VUPoints to enable the context-awareness needed to support these kind of applications.

6. 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 every interesting activity by at least one mobile phone. Of course, the notion of social activity is subjective, and thus identifying triggers to cover them is challenging. We take a first step through VUPoints, a system that 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 VUPoints is one instantiation of *social activity coverage*; the future is likely to witness a variety of other applications built on this collaborative sensing primitive.

7. REFERENCES

[1] E. Miluzzo et. al., "Sensing meets mobile social networks: The design, implementation and evaluation of cenceme application," in *ACM Sensys*, 2008.

[2] Martin Azizyan and Romit Roy Choudhury, "Surroundsense: Mobile phone localization using ambient sound and light," *Poster, ACM Mobicom*, 2008.

[3] Emma Berry, Narinder Kapur, and et. al, "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*, vol. 17, pp. 582–601, August 2007.

[4] Roy L. Ashok and Dharma P. Agrawal, "Next-generation wearable networks," *Computer*, vol. 36, no. 11, pp. 31–39, November 2003.

[5] Lenin Ravindranath Venkat Padmanabhan, Piyush Agrawal, "Sixthsense: Rfid-based enterprise intelligence," *ACM Mobisys, Association for Computing Machinery, Inc*, June 2008.

[6] Toshiya Nakakura, Yasuyuki Sumi, and Toyoaki Nishida, "Neary: conversation field detection based on similarity of auditory situation," *Proceedings of the 10th workshop on Mobile Computing Systems and Applications*, , no. 14, 2009.

[7] Engstrom et. al., "Mobile collaborative live video production," *Mobile Multimedia Workshop (with MobileHCI)*, Sep 2008.

[8] H. Susono and et. al., "Creating digital and analog storytelling for collaborative learning," in *TECHNOLOGY AND TEACHER EDUCATION ANNUAL, Vol 2*, 2007.

[9] S. T. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1158–1163, June 2005.

[10] Tingxin Yan, Deepak Ganesan, and R. Manmatha, "Distributed image search in camera sensor networks," *ACM SenSys*, pp. 155–168, Nov 2008.