

# A Real-time Video Summarizing Service for Community-contributed Contents of Real-life Events

Samantha Vu, Owen Noel Newton Fernando, Mikko Rissanen, Natalie Pang, Schubert Foo

*Nanyang Technological University, Singapore*

*14 Nanyang Drive, Singapore 637332*

*{sgtvu, ofernando, mjrissanen, nlspong, sfoo}@ntu.edu.sg*

**Abstract**—Using the example of a group of spectators watching a live street dance, this paper presents a work-in-progress concept of real-time video summarizing service by 'jumping' in different point of views. This innovative service takes up the challenges from various aspects of multimedia: mobile multimedia, authoring, real time interactive multimedia applications.

**Keywords**—multimedia entertainment; community-contributed content; real-time; real-life events

## I. INTRODUCTION

The increasing popularity of mobile live event capturing is undeniable. More and more people are using their phones for recording video clips and for editing activities. Video recording quality of mobile phones has reached a level whereby the phones are becoming a serious tool for on-the-move content creators. To share these self-edited videos is to share experience which is a social norm in modern communities. However, according to Juhlin et al. [1], people utilizing mobile broadcasting are still struggling with the technology despite all advances in modern camera-equipped smartphones and mobile live video broadcasting services.

In this paper, we focus on two problems in the current state of art in multimedia editing. The first problem is lack of motivation for editing. It is known that people recording the videos rarely watch them because of the amount of work for exploring, and annotating; especially since mobile videos are rarely edited after recording [2]. Kirk et al. [2] take a holistic user centric view on people's practices around home videos. Their results reveal useful information about people's motivations and practices for editing home videos. One of their results is that people do not find any reason to do editing of the short video clips they had taken. To overcome this, our proposed service (which will be explained in details in part II) would make producers more motivated to create the summary video by offering ability to participate in the live event and switch from one point of view to another in real time.

The second problem is a lack of good quality summary videos or remixing videos of real-life events. For instance, a live concert usually attracts a good number of video recordings by members of the audience but there are few compact and coherent videos that can capture the highlights of the concert. Similarly, if a search is done on Youtube about a certain event, for example, Obama's victory speech, there is usually an extremely long list of videos found. Even

though the list is sorted corresponding to relevance, this is not a proper results returned for such a question. We argue that a preferred way is to get a compact presentation of a predefined length, which gives a summary, composed from the views of many people that have witnessed the event. This is not necessarily the best view, but the view that can be created based on the information people provided when uploading the content. However, this "summary" video hardly exists in most cases. Thus, with the motivation to supply such a video, in this paper, we will present the concept of a ubiquitous service for summarizing live event.

Making available a large pool of snapshot digital videos taken by the audience in the same concert can result in higher value material than individual video clips. The individual digital video clips can be remixed into compilations that potentially enhance the perceived value of the event, are useful for various stakeholders such as the artists, and the fans of the artists.

According to Vihavainen et al. [3], remixing can also give the fans the possibility to become creators and not just receivers, and enhance the community feeling between the performers and the fans. Engström et al. [4] analyzed how video jockeys in dance clubs work and suggested that mobile video could enhance the interaction between the club visitors and VJs (Video Jockeys). Engström et al. [5] continued their studies and presented the SwarmCam prototype for video capture and live transmission of mobile video. Club visitors film their dancing on the dance floor and stream the video live to the VJ, who possibly broadcasts it to a big display screen. From our paper's point of view, the study is interesting, given that it concentrates on the interaction. In our usage scenario, the interaction between audiences and artists and distance audiences (remote participants) is emphasized as well. This is also why we explain the proposed service using an example of a street performance.

Our concept emphasized the element of real time editing which allows the producer to create summaries but also participate in the event live. The participation factor has never been discussed before in the area of video summarization of events.

The various aspects of multimedia discussed above (automatic summarizing, video editing and authoring, multi-camera video production, live events, user point of view, real time participation) have been researched in several well-established studies but never have all been integrated in a single service. This is what the proposed service in this paper aims to do.

The next part of the paper will describe the usage scenario and technical architecture of the proposed service in details. Discussions on advantages and disadvantages of the service will be covered afterwards and followed by conclusion.

## II. THE CONCEPT

### A. Usage scenario

Fig. 1 shows the proposed real-time video summarizing service designed for three types of users: spectators, performers, and producers. The spectators record the event using their camera-equipped smartphones with the intention of documenting the performance as well as their experiences during it. Pointing a smartphone to record the event is a natural action for them. The performers are equipped with wearable cameras and optional dataglasses. The remote participants become either normal passive spectators or they act as producers who are in charge of producing a complete documentation of the event for the community. For each event, the community can utilize one or more producers. They are located away from the local site and equipped with computers more powerful than smartphones, i.e. laptop or desktop computers. The producers are able to 'jump' into either the spectators or the performers to share live audio connection and viewpoints through the cameras. The producers are presented as augmented reality avatars connected to the person they are sharing viewpoint with. This is done to emphasize their presence to the spectators, performers and to themselves, which subsequently enhance their participation in the event. User interfaces are demonstrated in a practical scenario shown in Fig. 2.

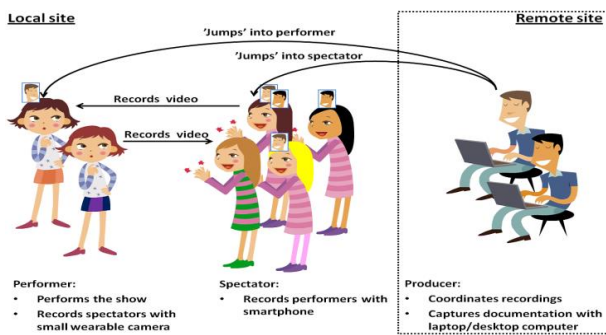


Figure 1. Participation in live events by 'jumping' into others.

Time	Sam the Spectator	Pete the Performer	Stacy the Spectator
1	Ryan jumps to record		Rita jumps to record
2	(Ryan keeps recording)	Ryan jumps to record	(Rita keeps recording)
3	(Ryan keeps recording)	(Ryan keeps recording)	Ryan jumps to record (Rita keeps recording)

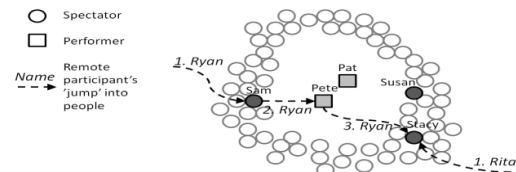


Figure 2. Setup of a live street dance scenario with 7 community members. Ryan captures 3 viewpoints and Rita 1 viewpoint of video recordings.

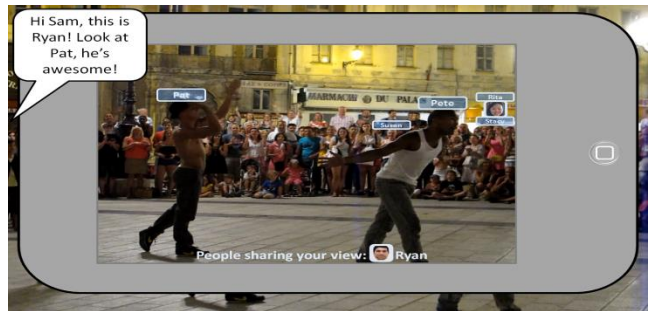


Figure 3. Ryan recording through Sam's smartphone.



Figure 4. Ryan recording through Pete's dataglasses.

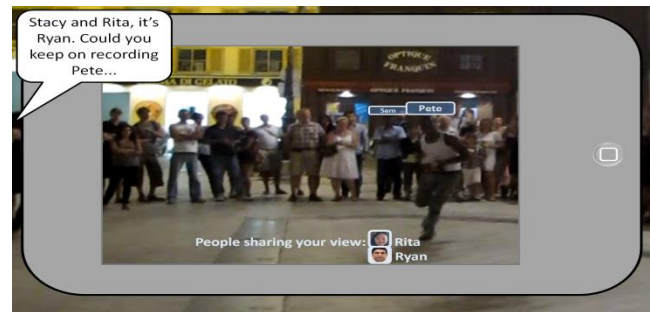











Figure 5. Ryan and Rita recording through Stacy's smartphone.

There are a) 2 street dance performers called Pat and Pete, b) 3 spectators called Sam, Susan and Stacy, and c) 2 remote participants Ryan and Rita who act as the producers. Fig. 3, 4, and 5 illustrate how Ryan 'jumps' first into Sam, then Pete and finally to Stacy who is already visited by Rita. The recordings Ryan captures from these 3 viewpoints (shown in Table 1) show others and himself appearing in the video. Default scene selections are done automatically based on the viewpoints Ryan has collected by 'jumping' into others. These can be dropped out or included in the summary video. Yet, Ryan has the option of mixing content from Rita's view or adding them later if post-processing is desired.

TABLE I. VIDEO STREAMS RECORDED BY PRODUCERS (PR) THROUGH SPECTATOR OR PERFORMERS (SP). STACY AND SAM RECORDED RYAN'S 'JUMPS' WHICH RYAN CAN VIEW AS INSTANT REPLAY.

Pr	Sp	Timepoint 1	Timepoint 2	Timepoint 3
Ryan	Sam			
	Pete			
	Stacy			
Rita	Stacy			

B. Architecture

Fig. 6 represents the overall architecture of the ubiquitous service of live events for remote participants. This architecture seeks to address the problem of sharing experience using mobile devices [6] with remote participants in a real world context with practical and technical challenges and social implications on design. This includes: the accessibility of hardware devices, wireless connectivity, current low-tech approaches, cost, and the habits and preferences of the end-users. Live video and audio from smartphones or other portable devices are streamed over a 3G or wireless connection while face recognition algorithm on smartphones processes the data stream [7]. GPS coordinates and orientation data from each spectator's and performer's location are used to decide which image frame is needed for the face recognition and tracking algorithms in order to reduce computational cost. Such tracking system, which combines vision-based and inertial trackers, provides the speed, precision, and stability to support outdoor mobile augmented reality systems [8]. In a crowded festival or indoors, mere GPS location and viewpoint orientation may not be adequate for recognizing the exact position of a community member. Advanced face detection and tracking is incorporated in the processing layer of the service to triangulate the exact position of a specific person appearing in a video stream, which will subsequently help to produce accurate augmented reality avatars with name labels.

Recognition results are then returned in real-time to the mobile devices and augmented with text and images. The mobile device is mostly responsible for sending live video and audio streams to the remote participants for recognition processing, which in turn provides the computed results. Both the spectator and the performer will be able to see augmented video on smartphones or data glasses, and the remote participants can see augmented video while experiencing spatial sound capabilities. The service delivers real-time face recognition information, text and image based information, and it enables functionalities of augmented reality.

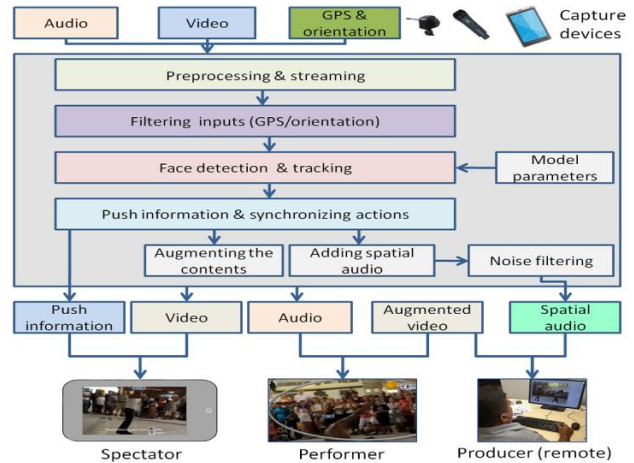


Figure 6. Service architecture.

Additionally, in terms of participation, we integrate spatial audio capabilities together with noise filtering to bring a natural experience of 'being there' with the spectators and performers. The architecture as a whole requires intensive research and development both in the communication and in the processing layers. As of present, the project is still in its initial stage of concept. However, we foresee some potential technical implementation issues e.g network connectivity will dictate the real-timeliness and quality of service. The service is conceptualised with the assumption of stable 3G and 4G networks in Singapore context. In case of connectivity failure, an automatic video recording will be activated to record the happenings from different viewpoints. This acts as a backup which the producer (remote) can still refer to after the real-life event.

III. DISCUSSION

There are previous studies on automatic summarizing systems or applications for community-contributed contents of a real-life event. Kenney et al. [9] proposed an automatic video mixing of concerts using audio fingerprints. Vihavainen et al. [3] presented an automatic remixing system for community contributed content from music concerts whereby users record and upload videos during live events and the shared content is then synchronized based on the creation timestamps. The single audio tracks of the synchronized videos are compiled to produce a master audio track. Through automatically detecting regions of interest in this master track, video remixes of a concert are automatically produced. However, as Gunes et al. [10] pointed out, for any event, elements like human emotions are still very difficult to be recognized automatically. Furthermore, the fully automatic video authoring methods are not in existence yet. Recent studies like Fabro et al. [11] focus on how to crawl community-contributed multimedia content from Youtube, Flickr to make video summaries of social events (using the Royal wedding of William and Kate as case study). The summaries produced are results of an algorithm which combines the multimedia content based on three criteria: quality, diversity and coverage. It is basically

data crawling without human acting as the producer. In our proposed service, we believe that authoring is best left to the human producers rather than algorithms. Furthermore, the previously developed systems enable only asynchronized summarizing whilst our service provides real-time synchronized summarization through the ability to jump across the different viewpoints.

Similar to our proposed service in terms of real-time editing, Engstrom et al [5] presented a live video broadcast video system whereby the camera operators interact and decide filming angles to collaboratively create a coherent narrative of an event. In this system, the different camera operators are the producers who collectively produce one summary video. On the other hand, in our concept, any single producer does not have to negotiate with another producer about angles, storyline, etc since the producer(s) can be the participant(s) anytime by jumping from one view point to another view point, including another producer's viewpoint. Thus, each producer creates their own summarized video which can be more coherent than that produced collaboratively by producers in Engstrom et al. [12].

There are a few areas that would eventually complete our service in the future. Current video broadcasting services with mobile wearable cameras do not provide any virtual embodiment for the remote participants (producers) yet as there is no display with augmented reality capabilities. Current smartphones support augmented reality using applications such as Layar but lack the functionality of producing audio and video stream (e.g. FaceTime on iPhone) while recording. Nor it is usually possible to use both of the smartphone cameras (one pointing forward and the other at the user's face) simultaneously, which would enable the producer to get further indications about what the spectators and performers are feeling. Discussion on hardware components is beyond the scope of this project.

Recognition of joyous moments e.g smiling and laughter could be a valuable support for selecting views where interesting things might be happening. According to Jacucci et al. [13], many of the most interesting moments that people wish to document happen spontaneously. Therefore, it would be also essential to provide a flexible and highly automated instant replay function that is already segmented to appropriate size for the producer to view it and clip it in the documentary while watching the primary view of the person 'jumped' into. However, this replay function is subjected to the capabilities of the devices used to record. Devices like Samsung Galaxy Note already has the instant replay function while other lower end devices like Aakash tablet does not.

#### IV. CONCLUSION AND FUTURE WORK

We presented a real-time video summary service that enables producers, who can be remote participants, to switch from one viewpoint to another viewpoint to create summary videos of live events. This service provides added value to the current literature of video summarizing applications by proposing the idea of viewpoint 'jumping' and how it can be

incorporated in accordance with other elements in a basic video summarizing and broadcasting system.

User studies will be conducted to test the proof of concept. A mockup dancing performance with a group of minimum five people including performers, spectators and remote audience will be conducted to evaluate the switching point-of-view interaction, level of participation felt by the producer, how the summaries are perceived by people other than the producer who have or have not attended the live events, and whether task allocation does result in the intended quality of experience.

Future work can explore how to perform this type of real time video summaries on longer events such as Olympic Games with many parallel sub-events.

#### ACKNOWLEDGMENT

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

#### REFERENCES

- [1] O. Juhlin, A. Engström, and E. Reponen, "Mobile broadcasting – The whats and hows of live video as a social medium," Proc. MobileHCI'10. The 12th international conference on Human computer interaction with mobile devices and services, ACM Press, Sept. 2010, pp. 35-44, doi: 10.1145/1851600.1851610.
- [2] D. Kirk, A. Sellen, R. J. Harper, and K. Wood, "Understanding videowork, " Proc. CHI 2007, ACM Press, 2007, pp. 61-70, doi: 10.1145/1240624.1240634.
- [3] S. Vihavainen, S. Mate, L. Seppälä, F. Cricri, and I. D. Curcio, "We want more: human-computer collaboration in mobile social video remixing of music concerts," Proc. CHI'11. The 2011 annual conference on Human factors in computing systems, ACM Press, May 2011, pp. 287-296, doi: 10.1145/1978942.1978983.
- [4] A. Engström, M. Esbjörnsson, and O. Juhlin, "Nighttime visual media production in club environments," Night and darkness: interaction after dark. Workshop at CHI 2008.
- [5] A. Engström, M. Esbjörnsson, and O. Juhlin, "Mobile collaborative live video mixing," Proc. MobileHCI 2008. The 10th international conference on Human computer interaction with mobile devices and services, ACM Press, Sept. 2008, pp. 157-166, doi: 10.1145/1409240.1409258.
- [6] S. Järvinen, J. Peltola, J. Lahti, and A. Sachinopoulou, "Multimedia service creation platform for mobile experience sharing," Proc. MUM'09. The 8th International Conference on Mobile and Ubiquitous Multimedia, ACM Press, Nov. 2009, pp. 1-9, doi: 10.1145/1658550.1658556.
- [7] B. Chen, J. Shen, and H. Sun, "A fast face recognition system on mobile phone," Proc. IEEE. International Conference on Systems and Informatics (ICSAI 12), IEEE Press, May 2012, pp. 1783-1786, doi: 10.1109/ICSAI.2012.6223389.
- [8] K. Satoh, M. Anabuki, H. Yamamoto, and H. Tamura, H, "A hybrid registration method for outdoor augmented reality," Proc. IEEE and ACM Symp. International Symposium on Augmented Reality (ISAR'01), IEEE Press, Oct. 2001, pp. 67-76, doi: 10.1109/ISAR.2001.970516.

- [9] L. Kennedy, and M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos," Proc. WWW 2009. The 18th international conference on World wide web (WWW 2009), ACM Press, Apr. 2009, pp. 311-320, doi: 10.1145/1526709.1526752.
- [10] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," Proc. FG'11. IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG'11), IEEE Press, Mar. 2011, pp. 827-834, doi: 10.1109/FG.2011.5771357.
- [11] M. Fabro, A. Sobe, and L. Boszormenyi, "Summarization of Real-life events based on community-contributed content," Proc. MMEDIA 2012. The 4th international conference on Advances in Multimedia (MMEDIA 12), IARIA, May 2012, pp. 119-126.
- [12] A. Engström, M. Perry, and O. Juhlin, "Amateur vision and recreational orientation: Creating live video together," Proc. CSCW. The ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12), ACM Press, Feb. 2012, pp. 651-660, doi: 10.1145/2145204.2145304.
- [13] G. Jacucci, A. Oulasvirta, and A. Salovaara, "Active construction of experience through mobile media: a field study with implications for recording and sharing," Personal and Ubiquitous Computing, vol. 11(4), Apr. 2007, pp. 215-234, doi: 10.1007/s00779-006-0084-5.