Journal of Mobile Multimedia, Vol. 8, No.1 (2012) 025-033

LIVE SHARING WITH MULTIMODAL MODES IN MOBILE NETWORK

XIAO ZENG[1, 2]
[1]Beijing University of Posts and Telecommunications
[2]Nokia Research Center
zengxiao29@gmail.com

KONGQIAO WANG
Nokia Research Center
Kongqiao.Wang@nokia.com

DA HUO
Nokia Research Center
huoda@buaa.edu.cn

Received September 28, 2011
Revised January 29, 2012

Live video sharing is a newly generated and interesting service, with which users can broadcast and view the videos being recorded by mobile phones. However, mobile network usually blocks users to enjoy that service since video transmitting is still a nontrivial task with poor bandwidth. In order to make live sharing easier in mobile environment, a novel service with multimodal modes is proposed in this paper, which could save a lot of bandwidth for sharing and is more adaptive in mobile network. To save bandwidth and introduce differentiated user experience, real-time extracted key-frames, audio or hybrid information can be transmitted instead of original video stream. Both publishers and receivers can select suitable mode according to their preference or network condition. Thanks to the key frame mode of the proposed service, detailed tagging of video content and live cooperation with other SNS can be implemented. Experimental results and user study demonstrate that the proposed multimodal live sharing service is of high adaption of mobile network and introduces direct and interesting user experience.

Key words: live sharing, video, key frame, multimodal, mobile network
Communicated by: K. Liu & A. Hakim

1 Introduction

Nowadays, information sharing becomes a habit of many people. This is one reason of why social network is so popular currently, which allows users share their daily experience with friends and family members.

Because of the rich quantity of information contained in video, it is an important information type that people would like to share on the social network. Video sharing service commonly allows users upload their recorded video files to servers, and set titles, tags or descriptions to declare the video content. The most famous website of this type is YouTube [1]. Recently, a new type of social medium, which allows users to broadcast live video from mobile devices to websites on the internet, is becoming increasingly popular [2]. In this type of service, videos are shared while they are being recorded. Bambuser [3] is a typical live video sharing website, through which users are capable of broadcasting and viewing videos which are under recording meanwhile, with almost no delay. Besides Bambuser, there are several similar services [4-7]. Live video sharing creates a better and more direct user experience than conventional video sharing that has an inherent delay such as those that require downloading after the video is recorded and uploaded. Live video sharing may be beneficial in some cases, for example, when a user is travelling, at a social event, or for a live meeting. Live sharing is useful, vivid, and interesting, and is a growing trend in social networking services (SNS) domain.

However, it is still hard for many mobile users to enjoy live video sharing well because real-time video transmitting is almost only supported by 3G or WIFI network due to the heavy network load of video stream. But currently, neither 3G mobile network nor WIFI is supported by all users' device. And though 3G or WIFI is supported by some users' device, the network is not available anywhere. Moreover, some users just want to save bandwidth (network charges) when they use live video sharing. Therefore, sharing live video data in a mobile environment is still a nontrivial task.

As discussed above, though live sharing is desirable to users, many mobile users are unable to enjoy this service because of poor networks. To minimize these limitations and boost this valuable service, a multimodal live sharing is proposed in this paper, which is capable of saving a lot of bandwidth during live sharing and is adaptive in various network conditions. Besides the basic function of live video broadcasting, extracted multimodal information can be transmitted instead of original video stream to fit users' network bandwidth and calculation resource. This solution is also with high configurability.

The rest of this paper is organized as follows. In Section 2, the proposed multimodal live sharing service is overviewed. The method of real-time extraction for the most important key frame mode is discussed in Section 3. And in Section 4, two additional features of this service, automatic tagging and cooperation with other SNS, are discussed. Experimental results and user study are given and evaluated in Section 5. Finally, in Section 6, conclusions are drawn.

2 A New Manner of Live Sharing

2.1 Multimodal Sharing Modes

As aforementioned, mobile network is the dominant limitation for video sharing due to the narrow bandwidth and high flow price. To facilitate mobile phone users to generate, publish and enjoy personal videos instantly, and give consideration to the adaptability in mobile network, we propose a new live video sharing service in this paper. In our solution, the video publisher and viewer can select proper mode for them to achieve sharing. Supported multimedia modes include:

a) Key frame mode: Key frames of the video under recording can be employed to represent the main visual content. Transmitting key frames instead of raw video stream can obviously save the network bandwidth.

b) Audio mode: In a case that auditory content is more important than visual content, such as during a meeting or presentation, the visual data is unnecessary to transmit and can be removed from the sharing session to save bandwidth.

c) Hybrid mode: Combination of above two modes, which contains both visual and audial information and is still a reduction of bandwidth usage from the original video stream.

d) Video of scalable resolution mode: If the network bandwidth is sufficient, users can transfer entire video stream. In this mode, the proposed service also provides different resolution to receivers in different network conditions.

Though there are several optional modes in a sharing session, available modes in one sharing session primarily depend on the publisher’s selection. Necessary video process (like key frame extraction) may be applied on the server or publisher’s device, also depending on publisher’s mode selection. For example, if the publisher chooses video mode for sharing, all the process tasks can be applied on the shared video stream and the extracted information are available; in this case the receivers can choose any mode to review the shared content, and the video process can be executed on the server to gain a high performance. But, if the publisher chooses other multimodal modes for sharing, the modes available for reviewers are limited, and corresponding calculations have to be taken on his own device.

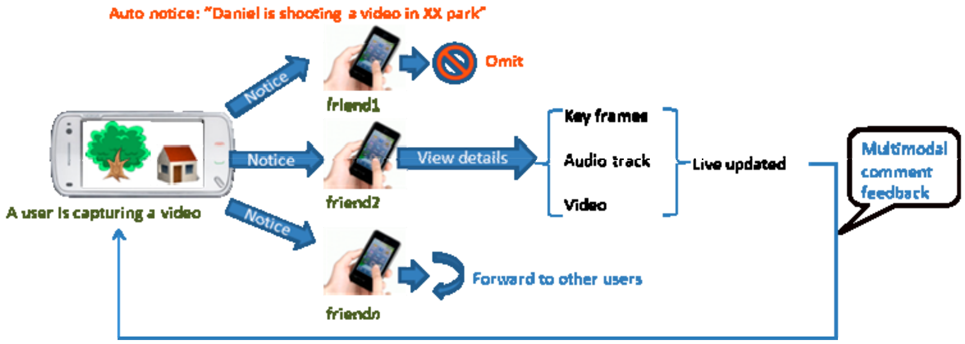


Figure. 1. Illustration of a live sharing session with multimodal modes

To make it clear how the proposed service provides multimodal mode sharing, a typical sharing session is illustrated in figure 1. In this case, a user (publisher) is recording video and sharing it; when the recording starts, his friends will receive an automatically generated message including the information about who (ID of the publisher), where (location from GPS), what (content explored by automatic analysis, like face detection); When a user (receiver) receives the message, he can choose to dive into the details, forward it or just omit it; If a user dives into his friend’s capturing content, he can choose different modes according to his desire or bandwidth limitations, like key frame mode, which are live and dynamically increased/updated while the video content is being collected; Besides receiving the shared content, a viewer can give his comment instantly to other users. The publisher can adjust his recording according to the instant feedbacks, which makes the sharing session more interactive.

Besides live sharing, the contents that are shared may be stored on a server side. Users can access those stored contents anytime, anywhere. Even when the contents are stored on the server, the multimodal modes access are also supported. To make the concept of our solution more clear, a comparison of three types of video sharing service is listed as Table 1.

Table 1. Comparison of different video sharing service

Service	**Conventional Video Sharing (YouTube)**	**Live Video Sharing (Bambuser, Qik, and etc.)**	**Proposed**
Video file upload	√	√	√
Video file browse	√	√	√
Live video upload	×	√	√
Live video browse	×	√	√
Format of shared content	Video stream	Video stream	Multimodal
Applicative mobile network	3G/Wifi	3G/Wifi	2G/3G/Wifi
Cooperate lively with other SNS	×	×	√

2.2 Framework

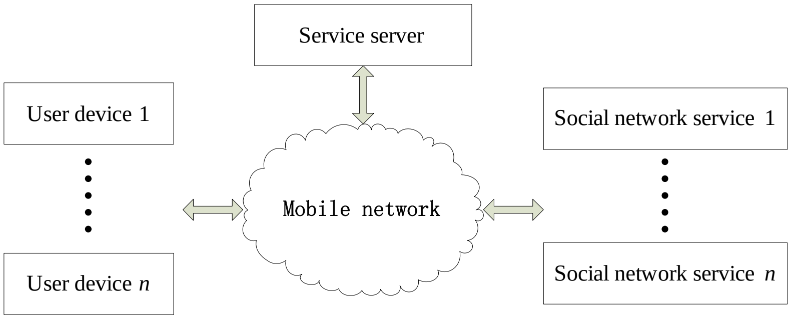


Figure. 2. Framework of the proposed service

Framework of the proposed service is illustrated in figure 2. Users' devices are the end devices for publishing and receiving, whereas server computers provide services including data transmission, storage, analysis, and other processing. If a device is camera embedded, it is capable to be as a publishing device. Captured content is optionally uploaded to the server according to publisher's desire or network condition. A device is also able to act as a receiver terminal, with which a user can view his friend's recording. Different users may choose different modes during once sharing.

3 Real-time Key Frame Extraction

Key frame mode is the most important mode in our solution, which can represent the main visual content of shared video well and save a lot of bandwidth. Since key frame is very representative of video content, the viewer is able to spend less time for grasping the substance of video. Another advantage of key frame is that the viewer can gain a full view of the video at any time without keeping looking at the screen.

Key frame extraction is a hotspot research domain in recent years. But conventional key frame studies merely deal with stored video files; that is, in those studies key frames extraction can only be applied after recording is finished. In order to get key frames in a real-time manner just during the recording, as required in the proposed service, it is necessary to develop a new method.

In the proposed key frame extraction method, there are three issues need be taken into account. First, in various methods, key frame is always extracted based on shot segmentation [8]. But for user recorded video, there is no exact shot boundary generated by video edit. Camera motion is the most important item for key frame extraction in such video type [9] since camera movement often stands for interest change or scene transformation. Therefore, in our solution, camera motion is detected and employed as the criteria when a key frame should be extracted. Second, in our real-time extraction method, at the moment when key frame extraction judgment is made, only the information of already recorded frames is available whereas future frames are unknowable absolutely. In this situation, it is hard to determine when the optimal key frame will appear. Last, if the publisher choose key frame mode, the extraction must be applied on his device. Considering that computation capability of mobile phone is always limited contrast to PC, the extraction algorithm runs on them cannot be of high complexity to meet real-time requirement.

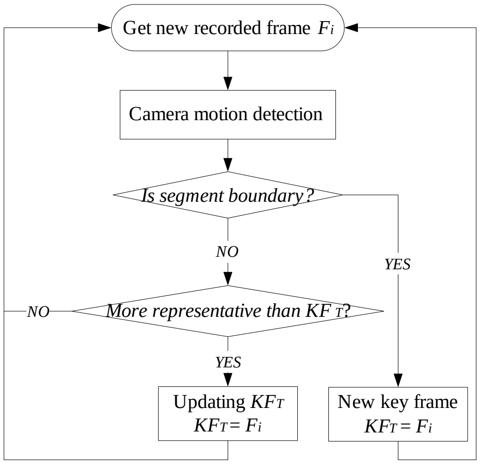


Figure. 3. Process of real-time key frame extraction

Considering above issues, the proposed real-time key frame extraction process is developed and shown in figure 3. The method proposed by Lan et al. [10] which is fast and robust for dominant visual movement judgment is employed for camera motion detection. To further accelerate the method, diamond search strategy [11] is adopted to displace the original one. Motion vector of each frame is accumulated until no motion is detected. If the total amplitude exceeds a threshold, which indicates that the camera motion is obvious enough, a new key frame should be extracted. As descripted above, the key frame may be not optimal, so following recorded frames are keeping compared with the key frame by edges, luminance, colour richness, entropy and other features. If subsequent frame is better than current key frame, the key frame should be updated. This strategy can be called as key frame evolution. In our solution, initial key frame and updated ones are all transmitted instantly once they are extracted but only the latest one will be reserved on receiver's device.

4 Additional Features

4.1. Automatic Tagging of Sharing Content

In existing video sharing, the content of video is always described by a title or some descriptive words. Those texts usually summarize the whole video and are not able to emerge the details of the content; in another aspect, those texts are always typed manually. In the proposed service, the video content can be described in details thanks to fine grit key frame. We can tag each key frame to make it more meaningful by presenting the location, direct, object, and etc. Since the publisher keeps recording in a live sharing session, so the tagging should be automatic to not disturb the publisher.

To implement automatic tagging on key frames, context collection and analysis is integrated into our service. Currently, many mobile devices carry various sensors to collect data of GPS, gravity, and compass. This data is easy to get by recalling the corresponding API, which is useful to make the sharing more understandable. This type of tagging should be done on the publisher's device directly.

Moreover, multimedia content analysis is also helpful for automatic tagging. Many multimedia analysis technologies can be introduced into our service to explore the semantic details in the captured content. For example, human detection [12] and face recognition [13] are helpful for distinguishing the people involved in the key frames. Since multimedia content analysis is time consuming, it can be processed on the server side.

With the combination of above two types of tagging method, the shared key frames are annotated automatically so that the information of publisher's recording is represented more clearly to the receivers.

4.2. Real-time Cooperation with Other SNS Service

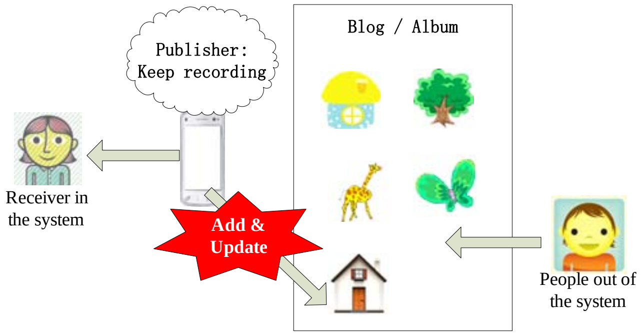
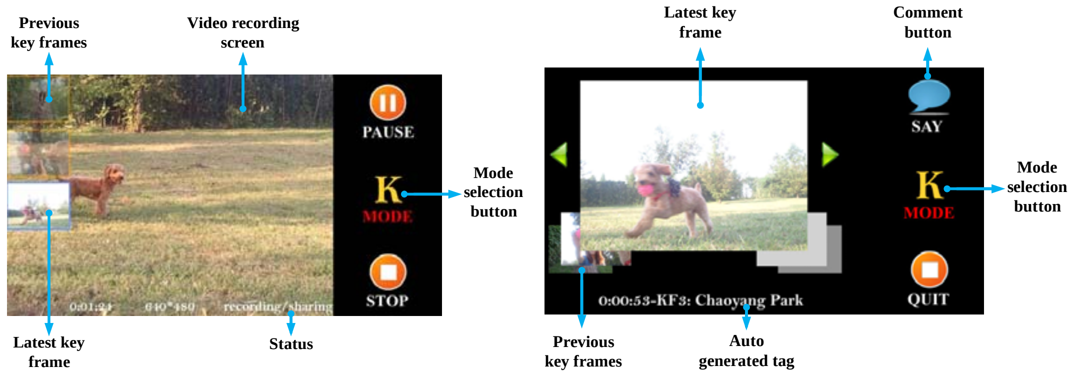


Figure. 4. Illustration of cooperation with other SNS service

In other video sharing services, the video content only can be shared to other SNS when the recording finishes [3, 6]. Hence it is impossible for any user, who does not register in the video sharing system, to enjoy the live sharing service. This drawback can be avoided in the proposed system. In our service, thanks to the instant key frame generation ability, it is easily to cooperate with other popular SNS in live manner, such as micro blog and online albums (flicker, picasa). While the publisher keeps recording video, the real-time extracted key-frames can be transmitted to those SNS to publish & update the content instantly. Therefore, the proposed service enables more people to enjoy live sharing no matter whether they register in our system or what device and network they have. Figure 4 is an illustration of the cooperation with other SNS. The live content can be shared to receivers in our system directly, or shared by way of a SNS so that other people may view the content that is being shared either lively, or at a later time.

5 Experiment and User Study

We have implemented the proposed service application both on device side and server side. On device side, the terminal program is developed based on Nokia N8 Smartphone with Symbian 3 platform; and on server side, the service program is developed based on Windows XP platform using Visual C++.



(a) Publisher's device

(b) Receiver's device

Figure. 5. Screenshot of publisher and receiver's devices during a sharing session of key frame mode

Figure 5 illustrates publisher and receiver's device screen during a sharing session, wherein key frame mode is selected. As shown in figure 5 (a), publisher's device is keep recording and sharing as the "status" text shows. Mode selection button is on the right column, and the yellow letter "K" denotes that the publisher chooses key frame mode. Real-time extracted key frames are listed at the left of the screen. Previous key frames are semi-transparent to not cover the recording content. The screenshot of receiver is captured simultaneously, as shown in figure 5 (b). The biggest image on the screen is the latest key frame which may be updated by evolution strategy, whereas the previous ones are at the left and of smaller size. The receiver can swipe the screen to browse different key frames. Every key frame is automatically tagged on the publisher's device and the tag text is displayed at the bottom of the receiver's screen. Besides viewing, the receiver also can press the comment button to input words and send to the publisher immediately, which will display at the top of the publisher's screen.

Ten persons with various networks (5 of 2G, 3 of 3G and 2 of WIFI) are invited to test and evaluate the proposed service. In the testing, the participants totally start 58 sharing sessions with each other and the bit rate of each mode is recorded for comparison, as listed in Table 2. It is worth noting that the bit rate of key frame mode is a statistic since different key frames are not transmitted successively. The comparison results show that key frame mode needs the lowest bit rate, and audio mode and hybrid mode also cost only 16.6% and 20.7% bit rate of the video mode. Therefore, the proposed multimodal sharing modes can efficiently save the bandwidth and have stronger adaptability in various network conditions.

Table 2. Relationship among available modes

MODE	Key frame	Audio
Bitrate (kbps)	3.2 (.jpg, 176×144, average 4 frame/min)	12.8
MODE	Hybrid	Video
Bitrate (kbps)	16	77.2 (.3gp, 176×144, 15fps)

Table 3. Statistical results of user study

Index	**Item**	**Yes**	**No**
1	Are you used to video sharing service on your mobile phone, like Youtube or Bambuser, or some other else?	30%	70%
2	If “no” for question 1, do you think the poor mobile network condition is the main reason?	60%	10%
3	If “yes” for question 1, do you think it is more interesting and direct to share or watch a video when it is just under recording?	30%	0%
4	Do you think the multimodal modes of the proposed service make live sharing easier in mobile network?	80%	20%
5	Do you think the key frame mode of the proposed service represents the visual content well?	70%	30%
6	Do you like to integrate this service into your mobile phone?	80%	20%
7	Will you introduce this service to others?	40%	60%

The proposed service aims to provide a new way of living sharing to mobile users. In order to evaluate whether the user experience can be improved, some user study is carried out. After the ten participants finished the test, they are asked to fill a questionnaire. The statistical results are listed in Table 3. From the first two questions it can be found that most people are not used to enjoy video sharing on mobile phones due to the poor network. Answer of Question 3 shows that users consider that live sharing is more interesting and direct than conventional ones. 80% participants think that the proposed multimodal modes is helpful in live sharing and 70% participants feel that key frame mode can express the video content well enough. Furthermore, 80% users show their interest for integrating our service into their phones so that they can use it in the future; and 40% users also express that they will introduce this service to other users.

6 Conclusions

In this paper, a novel live multimedia sharing service with multimodal modes is proposed to make live sharing service available in various mobile networks. The key concept of the proposed service is to support three other modes (key frame, audio and hybrid) besides conventional video stream sharing to save network bandwidth and enhance the adaptability of live sharing. To achieve live multimodal sharing, real-time information extraction and sensor data collection are integrated, wherein the live key frame extraction is a new developed method. Thanks to the key frame mode of the proposed service, the detailed tagging of video content and live cooperation with other SNS can be implemented. From the experimental results, it is found that the proposed service indeed saves a lot of bit rate in network transmitting, which benefits network adaption and charge saving. Furthermore, the new user experience introduced by the proposed service is also highly accepted and preferred by the users.

It is notable that the proposed service is well configurable and extendable. For example, the key frame extraction method on device side and server side can be different for most adaption of calculation complexity and performance. And many pattern recognition methods besides the referred ones can be integrated into our service for automatic tagging purpose.

References

1. http://www.youtube.com
2. Oskar Juhlin, Arvid Engström, and Erika Reponen. Mobile Broadcasting – The Whats and Hows of Live Video as a Social Medium. MobileHCI '10 Proceedings of the 12th international conference on Human computer interaction with mobile devices and services, 2010, pp. 35-43
3. http://bambuser.com
4. http://www.sourcebits.com/iphone/knockinglivevideo
5. http://techcrunch.com/2009/12/09/iphone-live-streaming-ustream/#
6. http://qik.com
7. http://www.livecast.com
8. Hanjalic, A., and Zhang, H. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. IEEE Transactions on Circuits and Systems for Video Technology, 9(8), pp. 1280–1289, Aug. 1999.
9. Jiebo L., Papin C., and Costello, K. Towards Extracting Semantically Meaningful Key Frames From Personal Video Clips: From Humans to Computers. IEEE Transactions on Circuits and Systems for Video Technology, 19(2), pp. 289-301, Feb. 2009
10. DongJun L., YuFei M., and HongJiang Z.. A novel motion-based representation for video mining. IEEE International Conference on Multimedia and Expo, 3, pp. 469-472, 2003
11. Zhu S., and Ma K. A new diamond search algorithm for fast block-matching motion estimation. IEEE Transacions on Image Processing, 9(2), pp. 287-290. 2000
12. Navneet D., and Bill T. Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, pp. 886-893, 2005.
13. Imran N., Roberto T., and Mohammed B. Linear Regression for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2106-2112, 2010