Stabilization Technique for Region-of-Interest Trajectories Made from Video Watching Manipulations

Daisuke OCHI†(a), Nonmember, Hideaki KIMATA†, Member, Yoshinori KUSACHI†, Kosuke TAKAHASHI†, Nonmembers, and Akira KOJIMA†, Member

SUMMARY Due to the recent progress made in camera and network environments, on-line video services enable people around the world to watch or share high-quality HD videos that can record a wider angle without losing objects’ details in each image. As a result, users of these services can watch videos in different ways with different ROIs (Regions of Interest), especially when there are multiple objects in a scene, and thus there are few common ways for them to transfer their impressions for each scene directly. Posting messages is currently the usual way but it does not sufficiently enable all users to transfer their impressions. To transfer a user’s impressions directly and provide users with a richer video watching experience, we propose a system that enables them to extract their favorite parts of videos as ROI trajectories through simple and intuitive manipulation of their tablet device. It also enables them to share a recorded trajectory with others after stabilizing it in a manner that should be satisfactory to every user. Using statistical analysis of user manipulations, we have demonstrated an approach to trajectory stabilization that can eliminate undesirable or uncomfortable elements due to tablet-specific manipulations. The system’s validity has been confirmed by subjective evaluations.

key words: camera trajectory, ROI, region of Interest, video stabilization, tablet manipulation

1. Introduction

The rapid progress and spread of smart phones, tablet devices, and broadband networks is currently expanding on-line video services that enable users to choose and watch their favorite video contents. This has made it possible for users around the world to watch and experience video contents regardless of location and conditions. These services also include a video sharing function that has become a popular way for users to express or transfer their impressions, e.g., tastes or opinions, via messages. Video sharing services have become one of the major social tools to express or communicate people’s impressions to others.

However, although the exchanging of user messages has become a popular means of communication in these services, it does not sufficiently enable users to directly transfer/communicate their real impressions because they all watch video content in different ways especially when there are multiple objects in a scene. Currently, the best way for them to transfer their impressions directly is to extract from videos what we call ROIs (Regions of Interest), i.e., the parts or objects that are of interest to them, rebuild or edit them, then upload them to the services again. Unfortunately, this method lacks practicality since the editing process is not only rather difficult but also tiresome, as indicated in a report [1] that says only about 20% of users have ever uploaded video content to the video sharing services. We can provide users with a richer video sharing experience if we achieve a system with a simple, uncomplicated process that enables them to extract video parts that interest them and communicate and share their impressions with others.

Our idea is to achieve a system that enables users to watch and record their favorite parts of a video through simple manipulations of a tablet device. In this paper we have proposed a system that can extract users’ ROIs from a high resolution panoramic video produced by multiple cameras and stabilize their trajectories in a manner that should be satisfactory to every user. Our proposed system can handle not only panoramic videos but also ordinary videos shot by a single camera. However, we use the former since they provide finer resolution images when users zoom in on their ROIs. To ensure visually plausible trajectory stabilization, we have extracted a number of problems from a preliminary user study and created a model by means of a statistical approach. In the model, we focus on eliminating undesirable or uncomfortable ROI movements that are caused by tablet-specific manipulations, e.g., multi-flickings and multi-pinchings, from the trajectories. To show the validity of our approach, we subject it to subjective evaluations and report the results in this paper.

The following section reviews related studies that have been made in this field. In Sect. 3 we describe the proposed system. Section 4 describes how we extract problems in users’ ROI movement by observing their recorded trajectories. The approach we take to address the problems is explained in Sect. 5. In Sect. 6 we show subjective evaluations of our system to ascertain its validity. Section 7 concludes the paper with a summary of important points.

2. Related Work

Today, people are very familiar with video sharing services that enable them to upload their personal videos and allow others to post messages commenting on them. In some services or studies [2], [3], users can not only post messages but also specify the timing to insert them, and the messages appear with the video content at the specified time. Some stud-
ies also focused on message postings. Lanagan et al. [4] and Shamma et al. [5] used messages on tweeting service [6] to analyze the timing in which the community is interested in a live video. As a result, with these technologies, users who watched the content can not only know but also experience the moment that the community watched it and focused on it most according to the number of posted messages. This approach is considered and has been described as one of “impression transfer with timing.”

Although the current services can transfer not only messages but also the moments when they were posted, they cannot transfer the ROI themselves that users watched or in which they interested. Previous studies reported techniques that enabled users to select and watch their favorite parts of video content interactively. By making use of these technologies, it is possible to achieve to transfer user’s ROIs technically by recording and keeping the ROI information. MAVlankar et al. [7] developed a system that can deliver user-selected parts to smaller devices or GUI windows with correct resolution by using P slices of the H.264/AVC standard. By using this concept, Halawa et al. [8] and MAVlankar et al. [9] achieved a means of delivering lectures to tablet devices; it enables users to move their ROIs by its simple flicking or pinching manipulations. They also evaluated their systems’ image quality and user interfaces, and both were shown to be satisfactory. Kimata et al. [10], [11] developed a similar delivery system using the H.264 Annex H (MVC) standard bonding multiple resolutions with many views that were made by dividing the video into many tile-shaped regions with each view overlapping to some degree. To maintain image clarity when the user zooms the video, they used high-resolution panoramic video about 5000 pixels in width produced by their own technology. They also evaluated the system’s QoE (Quality of Experience) and user interfaces in a field test and obtained positive results. With their system, users can also select parts or scenes interesting to them through intuitive manipulations of their tablet devices.

Techniques such as these, therefore, make it possible to record and maintain ROI information. However, the recorded ROIs are made by performing simple tablet manipulations rather than careful, editing ones, and consequently it will be a challenging task to provide them with visual plausibility that most or all users will find satisfactory. Much study has gone into making or refurbishing camera trajectories from a recorded video; this is known as video retargeting or stabilization. One major retargeting process that has been reported was developed by Liu et al. [12] and Gleicher et al. [13]. They partially cropped scenes from large images without camera shake and made new scenes with new camera trajectories with content plausibility. As for video stabilization, Liu et al. [14], [15] reported another study. They used features extracted from source video images, estimated their position between frames, and moved them to the proper position where the images became stabilized. By distorting each image in a frame, they were ultimately able to refurbish visually plausible video. Their method does not sacrifice any pixel data that was contained in the original video as many previous works do. Another novel study of stabilization is by making use of a free-viewpoint technique that produces interpolate images between cameras. Ballan et al. [16] made a stable and smooth ROI trajectory by using this technique from casually filmed events.

We draw upon some of these studies in developing our approach to achieve our goal, which is to enrich the video sharing experience of users by enabling them to: 1) record their favorite video parts as ROI trajectories through simple and intuitive manipulations of their tablet devices, 2) stabilize the parts in a manner that all or most users will find satisfactory, and 3) share them with others.

3. Interactive Video Watching

We use the system reported by Kimata et al. [10], [11] as our base system and enhance it with a scenario making/sharing function that enables users to not only record their ROIs with simple tablet device manipulations but also share them with others. These are necessary functions to achieve our goal that was described in the previous section.

3.1 Base System

The base system mainly consists of two technologies, one for video rendering and one for encoding/delivering.

In the video rendering (Fig. 1A) we use technology of Isogai et al. [17] inspired by the concept of Kunita et al. [18]. Multiple HD videos from multiple cameras placed in an event site are obtained with all images overlapping to some degree. After color correction, the images are combined by rendering them with color brightness according to the presence probability of an object’s surface at each pixel on a layered plane. This provides a panorama video having a very wide range without any readily discernible joints.

In video encoding/delivering (Fig. 1B), we use the concept of Kimata et al. [10], [11] by encoding every overlapped tile-shaped region with horizontal and vertical shifting, \( W_s \) and \( H_s \) pixels, respectively (Fig. 2). This makes it possible to provide not only high quality images of any size at any part of a video (not necessarily a panorama video) by selecting them from the video, but also smooth ROI movement without any image holes by sending lower-resolution videos.
image at the same time. Furthermore, this process consumes a very small amount of network bandwidth because only two types of images, user-selected high-resolution and low-resolution images, are sent to the client tablet device.

Consequently, in this base system, by using the panoramic video as an input to the encoding/delivering technique, users can select and watch any part from an area wider than that of normal videos as their ROIs simply by manipulating their tablet device, e.g., by flicking or pinching it, as if they were actually looking around the site.

### 3.2 Scenario Making/Sharing

At the same time that users perform their manipulations while watching a video using the base system, in our system they can record their favorite ROIs (Fig. 1C, Fig. 3). A trajectory (which we call a “scenario”) that consists of these ROIs with the information about its 2-D location and angle values makes it possible to reproduce user’s manipulations. The scenario is first stored in a user’s tablet device and then he/she can upload it in the delivery server. Since scenarios from many users are stored in one server, anyone wishing to watch along with one of them can download it and reproduce the ROI trajectory (if the maker has given permission for others to watch it).

With this function, users can not only make/keep many ways to watch a video but also share their favorite ROI trajectory as a scenario with others.

### 4. Preliminary User Study

To evaluate our interactive video watching system, we have conducted a preliminary user study. We asked ten subjects to watch a five-minute-long live performance (Fig. 8 (a), the same content we use later) with the system, then reproduce and observe their manipulations with recorded scenarios. All of them had had previous experience in using smart phones or tablet devices (the same applies to the following all subjective evaluations). System environments were the same as our subjective evaluation later (Sect. 6.1). As a result, we have obtained two undesirable or uncomfortable elements in their scenarios:

1. Users make many instantaneous pauses during long panning or deep zooming.
2. Users make slight changes in ROI position or size during/after panning and zooming.

No. 1 is caused by tablet-specific manipulations, e.g., multi-flicking and multi-pinching, when subjects move outside the display size of the tablet device and zoomed outside of the distance between their thumb and index finger (Fig. 4). No. 2 is caused by manipulations the subjects made in order to make the ROI positions meet their intentions. Some subjects passed over their target object and retargeted it soon afterward since they had no idea of the direction in which the object had gone. Sometimes they tended to flick to fix the ROI position, especially during zoom operations. Users perform these operations not only because they do not have enough distance between their pinch fingers but also because they want to confirm the ROI position after/during the zoom operation.

It is difficult for users to be aware that they are performing these actions when they are doing watching manipulations, but they stand out once they and others reproduce their recorded scenarios without any manipulations as discontinuous or non-smooth camera trajectories. These undesirable or uncomfortable elements lower the quality of recorded scenario so that should be eliminated.

---

To evaluate our interactive video watching system, we have conducted a preliminary user study. We asked ten subjects to watch a five-minute-long live performance (Fig. 8 (a), the same content we use later) with the system, then reproduce and observe their manipulations with recorded scenarios. All of them had had previous experience in using smart phones or tablet devices (the same applies to the following all subjective evaluations). System environments were the same as our subjective evaluation later (Sect. 6.1). As a result, we have obtained two undesirable or uncomfortable elements in their scenarios:

1. Users make many instantaneous pauses during long panning or deep zooming.
2. Users make slight changes in ROI position or size during/after panning and zooming.

No. 1 is caused by tablet-specific manipulations, e.g., multi-flicking and multi-pinching, when subjects move outside the display size of the tablet device and zoomed outside of the distance between their thumb and index finger (Fig. 4). No. 2 is caused by manipulations the subjects made in order to make the ROI positions meet their intentions. Some subjects passed over their target object and retargeted it soon afterward since they had no idea of the direction in which the object had gone. Sometimes they tended to flick to fix the ROI position, especially during zoom operations. Users perform these operations not only because they do not have enough distance between their pinch fingers but also because they want to confirm the ROI position after/during the zoom operation.

It is difficult for users to be aware that they are performing these actions when they are doing watching manipulations, but they stand out once they and others reproduce their recorded scenarios without any manipulations as discontinuous or non-smooth camera trajectories. These undesirable or uncomfortable elements lower the quality of recorded scenario so that should be eliminated.
5. Approach

To address the problem of the above-mentioned elements, a model to eliminate these elements and stabilize the trajectories has been made.

In observing the subjects’ recorded scenarios, we have noticed that almost all of them eventually stopped their manipulations and watched their favorite parts or objects for a while without changing ROI position or size. We focus on the fact that all the observed undesirable or uncomfortable elements were made in transition states between moving from one ROI to another. It is possible for us to refurbish scenarios with natural ROI trajectories if we can determine users’ manipulation start/end points of the transitions and replace the original transitions with smooth ones.

5.1 Transition Term Decision Algorithm

The start/end point searches have been made for all values \((x, y, r)\) independently (Fig. 5), where \(x\) and \(y\) are for the ROI’s 2-D positions with respect to panning and \(r\) is for its angle with respect to zooming. The value \(r\) is the ratio of the diagonal pixel size of the user’s ROI to that of the tablet’s display. The ratio of width to height for the angle value is determined in accordance with the resolution of the watching device that the user uses.

5.1.1 Start Point Search

The start point search begins only when no other start point has been elected.

1. Condition for panning
   \(S_p\). A point whose absolute speed is above a threshold \(v_s\) pixel/sec.

2. Condition for zooming
   \(S_z\). A point when the user’s pinching manipulation starts.

where the speed is defined as the user’s manipulation speed measured with the pixels in the original source video.

As for panning, since the users’ intended ROI movements are expected to occur at a slow speed, \(v_s\) pixel/sec, movements above the speed are regarded as transitions between ROIs and those points where it exceeds the threshold are regarded as start points \((S_p)\). As for zooming, the beginning of pinching manipulations is simply considered to be the start point for zooming \((S_z)\).

5.1.2 End Point Search

Once the start point is elected, the end point search begins. The end point search is a little more complicated since it is difficult to tell intended manipulations from unintended ones when some manipulation gaps (no value change) that are caused by multi-flicking/pinching or mistakes the user made exist in the recorded scenario.

1. Conditions for panning
   \(E_{p-1}\). A point where the user releases his/her fingers except for the one that he/she re-touches and moves to the same direction again within \(t_1\) seconds.
   \(E_{p-2}\). A point where the moving directions are reversed except for any that are reversed again within \(t_2\) seconds.
   \(E_{p-3}\). A point where the absolute speed is below a threshold \(v_e\) pixel/sec.
   \(E_{p-4}\). A point where the user stops watching the scene/content.

2. Conditions for zooming
   \(E_{z-1}\). A point where the user releases his/her fingers except for the one that he/she re-touches and zooms to the same direction again within \(t_3\) seconds.
   \(E_{z-2}\). A point where the user’s pinching manipulations finish except for any that are performed again within \(t_4\) seconds.

There should be gaps of \(t_1\) and \(t_3\) seconds for panning and zooming respectively (Fig. 6 (a)) in the scenario when the user does multi-flicking/pinching manipulations because his/her finger(s) is(are) off the touch screen. Since these gaps are ones unintended by the user, we do not elect them as the end points and consider that they are in the transition process \((E_{p-1} \text{ and } E_{z-1})\). Under these conditions, the undesirable or uncomfortable elements no. 1 described in Sect. 4 can be eliminated from users’ recorded scenarios.

Another unintended manipulation can be included. Users stop panning when the object suddenly stops against their will while they are panning, and then the ROI passes over the object and the users try to move in the opposite direction to fix their ROI position/size. These kinds of short
The function we use is an “ease-in the smooth interpolation functions are replaced in the term.

be described as follows:

Parameters, \( v_s, v_e, a, \) and \( t_{en} \) (\( n = 1, 2, 3, 4 \)), should be determined for our approach. Both \( v_s \) and \( v_e \) are the velocities of user’s careful manipulations, so they are treated as \( v_s = v_e \). We set 0.005 pixels/msec = 5 pixels/sec, which is a sufficiently small value. For simplicity, \( a = 0.05 \) is selected as a panning and zooming factor and no dynamic fitting that can determine the value \( a \) to fit the original trajectory described in [12] is implemented this time. \( t_{en} \) (\( n = 1, 2, 3, 4 \)) are the values of multi-flicking/pinch and short time ROI fixing intervals, which are the key values to eliminate undesirable or uncomfortable elements.

\[ f_K(t) = \begin{cases} 1 - d \left( \frac{t - t_1}{t_2 - t_1} \right) f_K(t_1) + d \left( \frac{t - t_1}{t_2 - t_1} \right) f_K(t_2) & (1) \\
\end{cases} \]

where \( t_1 \leq t \leq t_2 \) and the ease-in/ease-out function \( d(u) \) can be described as follows:

\[ d(u) = \begin{cases} \frac{u^2}{2a - 2a^2}, & u < a \\
\frac{u}{2a} - \frac{u^2}{2a^2}, & a \leq u \leq 1 - a \\
\frac{1 - u}{2a - 2a^2}, & u > 1 - a \end{cases} \]

where \( a \) is a constant value (0 < \( a \leq 0.5 \)).

5.2 Transition Interpolation

Once the start/end points of a transition have been obtained, the smooth interpolation functions are replaced in the term. The function we use is an “ease-in/ease-out” function [12].

Assume that the user’s recorded value is \( f_K(t) \) where \( K \) denotes \( x, y, \) and \( r \), the type of values we deal with and \( t \) is content time or frame position. If we let \( f_K'(t) \) be the smooth interpolation function between the start and end points, \( t_1 \) and \( t_2 \) respectively, it can be described as follows (Fig. 7):

\[ f_K'(t) = \begin{cases} 1 - d \left( \frac{t - t_1}{t_2 - t_1} \right) f_K(t_1) + d \left( \frac{t - t_1}{t_2 - t_1} \right) f_K(t_2) & (1) \\
\end{cases} \]

where \( t_1 \leq t \leq t_2 \) and the ease-in/ease-out function \( d(u) \) can be described as follows:

\[ d(u) = \begin{cases} \frac{u^2}{2a - 2a^2}, & u < a \\
\frac{u}{2a} - \frac{u^2}{2a^2}, & a \leq u \leq 1 - a \\
\frac{1 - u}{2a - 2a^2}, & u > 1 - a \end{cases} \]

where \( a \) is a constant value (0 < \( a \leq 0.5 \)).

5.3 Parameter Settings

Parameters, \( v_s, v_e, a, \) and \( t_{en} \) (\( n = 1, 2, 3, 4 \)), should be determined for our approach. Both \( v_s \) and \( v_e \) are the velocities of user’s careful manipulations, so they are treated as \( v_s = v_e \). We set 0.005 pixels/msec = 5 pixels/sec, which is a sufficiently small value. For simplicity, \( a = 0.05 \) is selected as a panning and zooming factor and no dynamic fitting that can determine the value \( a \) to fit the original trajectory described in [12] is implemented this time. \( t_{en} \) (\( n = 1, 2, 3, 4 \)) are the values of multi-flicking/pinch and short time ROI fixing intervals, which are the key values to eliminate undesirable or uncomfortable elements.
tervals from two subject's (ranging in age from their fortiess to their fifties) ball-following tasks for a 5-minute soccer video (Fig. 8(b)). Sudden ROI changes within 2.5 seconds were made 165 times and their frequencies are shown in Fig. 10. The reason we chose this type of video content is that a user's unintentional short time ROI fixings can often occur in content such as sports because the players move fast and sometimes stop or change direction suddenly. Confidence intervals of 95% for short time ROI fixing intervals assuming exponential distribution have been calculated for the population means \( \mu_s \) and the results are \( 462 \text{ msec} \leq \mu_s \leq 505 \text{ msec} \).

5.4 Implementation

We have implemented the algorithm in Sect. 5.1 and the interpolation function in Sect. 5.2 in our delivery server (Fig. 11). The flow automatically starts immediately after the user uploads the scenario and the start/end point searches and interpolations last repeatedly until the end of the scenario. The largest interval values are used, i.e., \( t_{e1} = 793 \text{ msec} \) for multi-flicking, \( t_{e3} = 1516 \text{ msec} \) for multi-pinching, and \( t_{e2} = 505 \text{ msec} \) for short time ROI fixing. The same value is used for \( t_{e2} \) and \( t_{e4} \) since we assume no difference in the direction change of zooming and panning for a short duration. Two representative results for panning and zooming are shown in Fig. 12.

In the panning results (Fig. 12(a)), a jagged trajectory caused by multi-flicking and short time ROI fixing (the dotted-circled part in the graph) have been eliminated and replaced with a smooth one. Non-uniform zooming speed has been replaced in the zooming results (Fig. 12(b)). Both the panning and zooming results are also compared with simple moving averages of before and after four frames as a general low-pass filter technique, but this technique often misses the user-intended start/end points (the solid-circled part in the graphs). The results we obtained for our proposed technique confirm that with it the proper election of start/end points functions.

6. Subjective Evaluation

Subjective evaluations have been conducted for different content types, i.e., a live performance and a soccer game (Fig. 8), to show the validity of our system. For users to make their watching experience in the video sharing services richer, it is important to make scenarios satisfactory not only for themselves but also for others. The evaluations were made for user-made scenarios from the following points of view:

- Evaluations made by users themselves
  
  Ev.(a) The users’ degree of satisfaction with their own scenarios.

  Ev.(b) The users’ degree of confidence in their own scenarios (How much they want to show it to others).
• Evaluation made by other users
  Ev.(c) Completeness of user-made scenarios for others.

6.1 System

We used a notebook PC (CPU: 4 Cores 2.4GHz, RAM: 4GB, OS: CentOS 5.8, SSD: 128GB) as the content delivery server. The watching client was an NTT docomo Galaxy Tab 10.1 LTE (manufactured by Samsung) with 1280×800 display resolution. They were connected via a wireless LAN router over IEEE802.11a. The panorama content in the delivery server was a live performance, that shown in Fig. 8 (a) and referred to in Sect. 5.3.1. The content was 1 minute long. The server delivered the content to the client over network bandwidth of about 4Mbps (2Mbps for both user-selected and low-resolution views).

6.2 Method

For live performance content, 76 subjects (Japanese, ranging in age from their twenties through their fifties) were asked to watch a favorite object in the panoramic video by manipulating the client tablet device. All of them belonged to a fan community for some of the objects in the content not to manipulate at random. Their watching manipulations were recorded in scenarios after being stabilized.

6.3 Results

The charts in Fig. 13 show the live performance results for Ev.(a) and Ev.(b). Over 80% of the subjects are satisfied with the stabilized scenarios and most of them report that they were able to make the scenarios they had intended with their ROI trajectories. Additionally, 60% of them feels confident about the scenarios they have made. After the soccer content manipulations, both subjects commented that although it was difficult for them to follow the ball or the players because the ball and the players moved very fast, the reproductions along with the stabilized scenarios were not so bad for them. This indicates that they were not confident in but satisfied with their scenarios.

The results for Ev.(c) have confirmed the effectiveness of the scenario stabilization at the 0.01 or 0.05 levels of probability (Tables 1 and 2). They demonstrate that our method can provide stable and natural ROI trajectories in terms of the factors “continuity” and “human warmth”. The “sensitivity” factor results indicate users felt comfortable when they watched stabilized trajectories made by others. They also demonstrate that the value differences between before and after modification for the soccer game content are smaller than those for the live performance content. This is because the soccer content includes many sudden ROI di-
rection changes (often start/end points emergences) so that the benefits our trajectory stabilizing technique provides are less noticeable.

6.4 Discussion

Our proposal and approach not only enable scenario makers (senders) to be satisfied with their scenarios personally, but also enable others (receivers) to feel natural and comfortable with them. Good social relations can be established between senders and receivers only if senders can send their scenarios without hesitation. However, even though senders still feel unconfident in scenarios with content that includes fast moving objects, it is easy to persuade them to have greater confidence by having them watch their resulting stabilized scenarios because they are satisfied with the scenarios once they watch them. This implies that our system has been able to enrich both their personal and social watching experiences in using video sharing services. Other subjects’ comments that attracted potential users are that they can a) make many contents of their own with different ROIs and collect some favorite ones and b) show and exchange them with others in a community to enhance inter-community communication. These advantages also seem to have helped our system receive the positive results it gathered.

To make our system more sophisticated and user-friendly, object retargeting for stabilized ROIs by using precise object extracting or tracking can be an effective approach to support the capture of users’ desired objects when the stabilized ROI trajectories missed them. Another enhancement is to add customized stabilization in accordance with each user’s personal habits because users’ manipulation speeds and styles vary considerably as demonstrated in the preliminary study. From the viewpoint of a user’s manipulation style, it is also necessary to examine the effects from the client’s screen size. Screens that are smaller or larger than the current 10.1 inch screen of tablet devices, such as the screens on smartphones or other devices, may affect user manipulation factors such as the zoom ratio since smaller or larger objects are watched. Transformation of scenarios may be an effective approach to take for users watching scenarios on different screen sizes.

7. Conclusion

We have proposed a system that features functions to:

1. allow users to record their favorite parts of videos as scenarios through simple manipulations of their tablet devices
2. plausibly stabilize discontinuous ROI (Region of Interest) trajectories due to tablet-specific manipulations
3. allow users to watch scenarios repeatedly and share them with others

We have confirmed both the validity of our stabilization approach and the user satisfaction it provides. This new approach can provide a richer watching experience and contribute to the growth of existing services and communities that use them. In future work, we will need to refine our approach and system to meet the individual user’s intentions.

References

[1] H.T. Chen, “Understanding content consumers and content creators in the Web 2.0 Era: A case study of YouTube users,” Proc. Comm. Tech. Div. 2008 Conf. Int’l Comm. Association, May 2008.
[2] niwango, inc., “niconico,” http://www.nicovideo.jp/, accessed April 1, 2013.
[3] T. Nakanishi, S. Shimada, A. Kojima, and Y. Fukuhara, “Lecture video and scene-related knowledge sharing common platform design and its prototyping, with nursing skill videos,” Proc. 18th ICCE, pp.270–274, Nov. 2010.
[4] J. Lanagan and A.F. Smeaton, “Using Twitter to detect and tag important events in live sports,” Proc. 5th AAAI Int’l Conf. on Weblogs and Social Media, pp.542–545, July 2011.
[5] D.A. Shamma, L. Kennedy, and E.F. Churchill, “Tweet the debates: Understanding community annotation of uncollected sources,” Proc. 1st ACM SGMM Workshop on Social Media, pp.3–10, Oct. 2009.
[6] Twitter inc., “Twitter,” http://twitter.com/, accessed April 1, 2013.
[7] A. Mavlankar and B. Girod, “Spatial-random-access-enabled video coding for interactive virtual pan/tilt/zoom functionality,” IEEE Trans. Circuits Syst. Video Technol., vol.21, no.5, pp.577–588, May 2011.
[8] S. Halawa, D. Pang, N.M. Cheung, and B. Girod, “ClassX - An open source interactive lecture streaming system,” Proc. ACM MM 2011, pp.719–722, Nov. 2011.
[9] A. Mavlankar, P. Agrawal, D. Pang, S. Halawa, N.M. Cheung, and B. Girod, “An interactive region-of-interest video streaming system for online lecture viewing,” Proc. 2010 IEEE 18th Int’l Packet Video Workshop, pp.64–71, Dec. 2010.
[10] H. Kimata, D. Ochi, A. Kameda, H. Noto, K. Fukazawa, and A. Kojima, “Mobile and multi-device interactive panorama video distribution system,” Proc. IEEE GCC 2012, Oct. 2012.
[11] H. Kimata, S. Shimizu, Y. Kunita, M. Isogai, and Y. Ohtani, “Panorama video coding for user-driven interactive video application,” Proc. IEEE Int’l. Symp. Consumer Electronorics 2009, pp.112–114, May 2009.
[12] F. Liu and M. Gleicher, “Video retargeting: automating pan and scan,” Proc. ACM MM 2006, 241–250, Oct. 2006.
[13] M. Gleicher and F. Liu, “Re-cinematography: Improving the camera dynamics of casual video,” Proc. ACM MM 2007, pp.27–36, Sept. 2007.
[14] F. Liu, M. Gleicher, J. Wang, H. Jin, and A. Agarwala, “Subspace video stabilization,” ACM Trans. Graphics, vol.30, Issue. 1, no.4, Jan. 2011.
[15] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, “Content-preserving warps for 3D video stabilization,” ACM Trans. Graphics, vol.28, Issue 3, no.44, Aug. 2009.
[16] L. Ballan, G.J. Brostow, J. Puwein, and M. Pollefeys, “Unstructured video-based rendering: Interactive exploration of casually captured videos,” ACM Trans. Graphics, vol.29, no.4, Article 87, July 2010.
[17] M. Isogai, H. Noto, and N. Matsuura, “Panoramic video rendering and display system using probability mapping method,” Proc. IEEE Int'l. Symp. Consumer Electronorics 2011, pp.264–267, June 2011.
[18] Y. Kunita, M. Ueno, and K. Tanaka, “Real-time rendering system of 3D images using layered probability maps,” J. ITE, vol.60, no.7, pp.1102–1110, 2006. (In Japanese)
[19] D. Kato, A. Ishikawa, T. Tsuda, H. Fukushima, and M. Yamada, “Analysis of the camera work and the subjective evaluation experiments,” J. ITE, vol.53, no.9, pp.1315–1324, 1999. (In Japanese)
Daisuke Ochi received the B.E. degree in Department of Electrical Engineering and Computer Science, and the M.E. degree in Information Engineering, from Kyushu University, Fukuoka, Japan, in 2001 and 2003, respectively. He is currently a Research Engineer in NTT Media Intelligence Laboratories, NTT Corporation. His research interests include computer-human interactions, human interfaces, and multimedia information systems.

Hideaki Kimata received the B.E. and M.E. degrees in applied physics in 1993 and 1995 and Ph.D. degree in electrical engineering in 2006 respectively from Nagoya University, Nagoya, Japan. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1995 and has been engaged in R&D of video coding algorithms, and visual communication systems. His research interests also include audio-visual processing for realistic communication. He is currently a Senior Research Engineer, Supervisor, in NTT Media Intelligence Laboratories. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE).

Yoshinori Kusachi received the B.E. in 1995 from Kyoto University, and M.E. degrees in 1997 from Nara Institute of Science and Technology (NAIST), and subsequently became a researcher at NTT Laboratories, where he has been engaged in research and development of pattern recognition and computer vision. In 2007, he received Ph.D. in engineering degree from NAIST. He is currently a Senior Research Engineer in NTT Media Intelligence Laboratories.

Kosuke Takahashi received his B.Sc. in Engineering and M.Sc in Informatics from Kyoto University, Kyoto, Japan, in 2010 and 2012 respectively. He currently works at NTT Media Intelligence Laboratories, NTT Corporation. His research interests include computer vision and computer-human interaction.

Akira Kojima is Senior Research Engineer, Supervisor, Visual Media Project, NTT Media Intelligence Laboratories. He received the B.E. and M.E. degree in Mathematical Engineering and Information Physics from the University of Tokyo in 1988 and 1990, respectively. Since joining NTT in 1990, he has been engaged in research and development on video database, digital library, multimedia information retrieval, video surveillance and high-reality visual communication. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Institute of Image Electronics Engineers of Japan (IIEEJ) and ACM.