Synchronized Dissemination Framework for Supporting High-Quality Tele-Immersive Shared Activity

Zixia Huang
Department of Computer Science
University of Illinois at Urbana-Champaign
zhuang21@illinois.edu

ABSTRACT

My Ph.D. thesis aims to develop a synchronized dissemination framework for supporting high-quality *tele-immersive shared activity* (TISA). Synchronization in TISA is complicated by (a) the multi-modal, bandwidth-savvy and timing-dependent media streaming over shared network resources, and (b) the delay-sensitive and interaction-critical nature of TISA in the real two-site and multi-site applications. This paper provides an outline of my thesis which includes four major contributions: (1) proposing a generalized layered framework for multi-stream synchronization sourced at one or multiple sites, (2) presenting an adaptive media packet scheduling scheme based on multi-stream timing correlations under Internet dynamics, (3) proposing a synchronized multicast topology based on diverse user interests, and (4) studying human subjective satisfactions to guide the system adaptation. This study is expected to output research results significant for next-generation systems, where timing-dependent media multi-modality is critical.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms

Design, Performance, Experimentation

Keywords

Tele-Immersive Shared Activity, Synchronization

1. INTRODUCTION

A tele-immersive (TI) system can offer a joint virtual space for distributed users to conduct meaningful shared activities. To achieve a realistic user experience, each TI site is configured with media devices with different functionalities, including multiple 3D cameras, microphones and haptic devices outputting timing-dependent 3D video, audio and haptic sensory streams respectively. A gateway at each site aggregates these local media streams and communicates with the remote sites who are then responsible for media rendering at the output devices (Fig. 1).

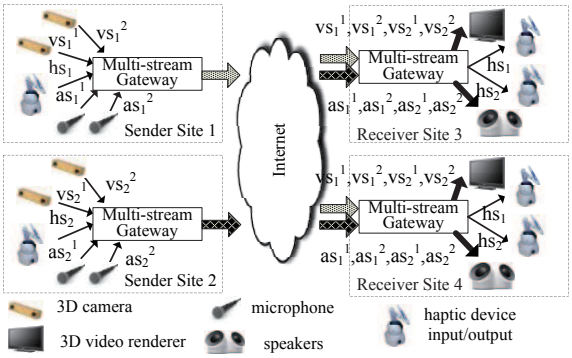

Figure 1: TI multi-stream configurations. vs: 3D video stream, as: audio stream, hs: haptic stream.

To provide seamless collaboration similar to a face-to-face interaction, an ideal TI system should disseminate the media signals with minimal latency, while preserving their rendering quality and timing synchronization. However, current best-effort network can delay the media delivery and introduce packet jitters and losses which will inevitably downgrade the user subjective satisfactions. This motivates me to design a synchronized dissemination and adaptation framework for supporting high-quality *tele-immersive shared activity* (TISA) over the Internet.

In the two-site TI system, three factors complicate the synchronized dissemination of the timing-dependent media streams. (a) *Shared network resources.* Because multiple bandwidth-savvy TI media streams share and compete the same network resources, packets can be dropped when bandwidth is not abundant. This can degrade the media rendering, and may lose the synchronization information which is important at the receiver. (b) *Heterogeneous characteristics.* Media with different traffic characteristics usually demand diverse streaming QoS, so they can employ their own transport protocols and adaptation algorithms in response to the Internet dynamics. This can lead to the end-to-end delay (EED) heterogeneity of different media streams between the two sites, and create synchronization skews across the media (*inter-media*) and among the streams within each *media bundle* (*inter-stream*). (c) *Interactivity.* While the receiver buffer can be used to smooth packet jitters (for rendering quality and *inter-frame synchronization*) and remedy the

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

EED differences (for *inter-stream* and *inter-media synchronization*), the tradeoff is that the buffer size can be added to the EED which will degrade the TISA interactivity.

A multi-site TI system can further complicate the inter-stream and inter-media synchronization, because the media streams from a same sender to a same receiver can follow multiple paths by relaying through different intermediate sites. In addition, two new problems arise in the multi-site support. (a) *Inter-site synchronization.* Because of the diverse network conditions, different streams from multiple senders can experience heterogeneous EED to a same receiver (*inter-sender skew*). Streams from a sender can also arrive at multiple receivers at different time (*inter-receiver skew*). Depending on activities, a TI system can demand different constraints on inter-site synchronization skews according to the user satisfactions. (b) *User interest diversity.* Multiple receivers may request different subsets of the multi-stream from each sender because of their diverse interests. For example, not all 3D video streams from a sender has visual contributions to the user view at a receiver, and thus only those useful streams need to be delivered.

Synchronization, interactivity and media rendering are three integral attributes combinedly impacting the user experience in TISA. Existing adaptation algorithms, however, usually focus on optimizing objective metrics which are either irrelevant or partially relevant to the real human perceptions [7, 14]. To realize high-quality TISA, subjective evaluations are needed to find the user preferences. The drawback is that subjective tests are expensive and can only be conducted off-line. Therefore, a good design should be able to generalize the offline subjective findings to guide the online system adaptations.

2. RESEARCH CONTRIBUTIONS

Based on the motivation and problems stated above, I present an outline of my research contributions in this section. The goal of my study is to design a dissemination, adaptation and synchronization framework which can maximize the user perceptions in both two-site and multi-site TISA.

1. My study first answers what are the *synchronization quality* and synchronization skew for multi-modal media multi-stream bundle sourced at one or multiple sites in the interactive multimedia. The definitions are missing in the previous related literature where the prevalent studies are on the audio-visual synchronization of 2D videos. In my thesis, a generalized layered architecture is proposed to describe and simplify the complications. Specifically, the four *sync layers* from bottom to top are: inter-frame, inter-stream, inter-media and inter-site synchronization. My study formulates the synchronization skew at each layer, and model the interrelations of synchronization quality across different layers. The *dominance* of each layer is also prescribed for synchronization references.

2. An adaptive dissemination scheme is proposed for in-time delivery of media streams under Internet dynamics. Specifically, because 3D video streams demand a much larger bandwidth overhead compared to audios and haptic sensory streams, my study presents a cooperative frame rate allocation scheme at the sender for the videos based on online bandwidth estimation. This can either be sender-driven (by source localization of 3D spatial audios [11]) or receiver-driven (based on the user interests of the 3D videos [14]). The TI system also improves the rendering and synchronization quality by adapting the media packet scheduling with receiver buffering control according to network conditions.

3. In a multi-site TISA scenario, I extend the previous ViewCast study [14], and build a multicast topology based upon the diversity of user interests and the prioritization of dissemination path options. The new scheme is able to (a) satisfy the inter-site synchronization demand, (b) minimize the EED of dissemination paths between senders and receivers, and (c) maximize the network resource utilization and reduce the number of undelivered *victim* streams because of bandwidth inadequacy.

4. User-observable objective and subjective metrics affecting the human perceptions are identified for both two-site and multi-site TISA. The objective metrics are able to capture the effects of synchronization, interactivity and media rendering quality, while the subjective counterparts can describe the overall user satisfactions. My research investigates the interrelations and trade-offs among these metrics at different TI activities. It also finds a mapping from objective metrics to subjective findings, and generalizes the offline subjective results. All these efforts can contribute to the online perception-driven system adaptation for achieving high-quality TISA under Internet dynamics.

5. The prerequisite of the above studies is that we are able to obtain accurate synchronized timestamps across multiple distributed sites, and among different input/output media devices within a single site. For practical considerations, I study the clock synchronization among/across different devices/sites, as an integral part of the synchronized dissemination framework. My thesis is intended to answer the following two questions. (a) *How often do we need to synchronize clocks* ? (b) *How do we minimize the impact of clock skews in our framework* ? Because [15] has discussed the problems of delay predictions errors on several network applications, my plan is to extend their approaches for this TI synchronization study.

3. RELATED WORK

Previous studies are unable to model the timing correlations of media multi-modality, and to identify the impact of real human subjective preferences in TISA. Therefore, there work cannot be directly applied to our TI system design.

Media synchronization. Although approaches and techniques have been heavily studied in the past for synchronization in one or multiple sync layers [1, 6, 8], no literature has systematically modeled the synchronization problem of all layers and investigated the interrelations across the layers. In addition, none of existing work manages to study the impact of network resource inadequacy over the synchronization quality, and to formulate the resulting synchronization skews at the presence of multi-source multi-modal media streams, which are expected to be prevalent in the next-generation media system. A new generalized layered synchronization architecture is, thus, imperative in the design.

Synchronized multicast topology. Existing studies on multicast topology for multi-site interactive multimedia aim to minimize the delay and delay variation with or without bandwidth constraints [12, 13, 16]. They oversimplify the problem by assuming the homogeneity of the media data requested at different receiver sites. The ViewCast algorithm [14], proposed specifically for multi-site TI system, can disseminate multiple video streams and allow cross-tree

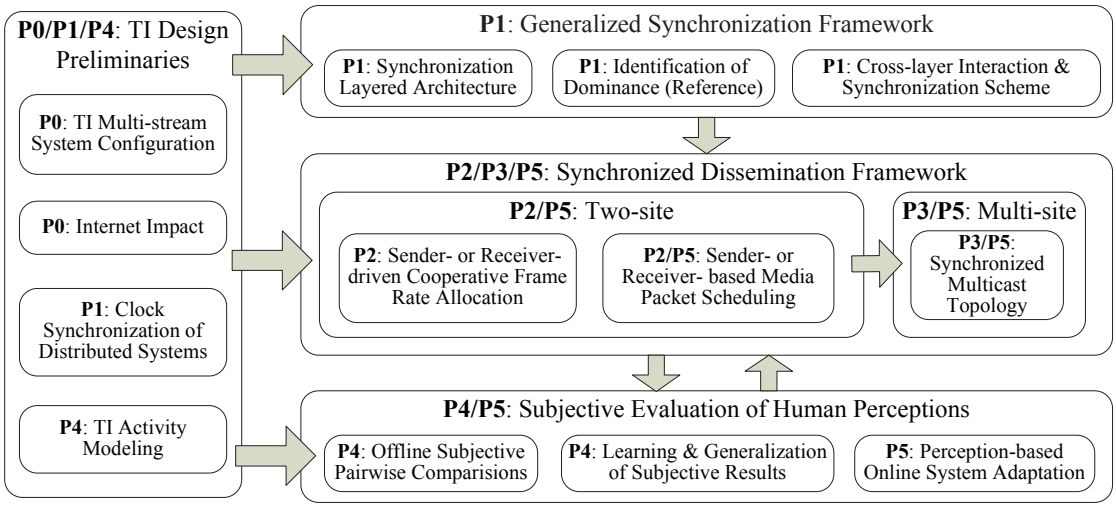

Figure 2: An overview diagram of research study

adjustment under bandwidth constraints and user interest heterogeneity. But ViewCast does not take into account the interactivity and synchronization aspects of the dissemination problem. In addition, the algorithm only studies the video streams, and does not identify the impact of audio or haptic sensory streams. A new multicast algorithm is needed to address these issues.

Impact of human perceptions. ITU G.1070 [5] estimates the synchronization, interactivity and media rendering quality in a closed multiplicative form for 2D video conferencing. But its assumption is that the three quality attributes can be computed independently. In addition, the standard only studies the conversational task and ignores the impact of TI activity heterogeneity on human perceptions. [10] concludes that people are unable to notice a maximum audio-visual skew of 80 ms, but only for best-quality on-demand videos. In our previous study, we have developed a VoIP algorithm that is able to adapt the audio buffer based upon the subjective perceptions of the audio quality and the conversational interactivity [2, 9]. But in TISA, the huge bandwidth overhead and the multi-modal stream correlations can complicate the design.

4. RESEARCH APPROACH DETAILS

In this section, I will show the detailed six-stage research approaches solving the whole problem. Fig. 2 presents an overview diagram of my study.

P0: I configure the TI system with multiple 3D video, audio and haptic streams. To evaluate its performance in different network conditions, I emulate a real TISA setting in PlanetLab, prerecord these sensory streams, and replay them using UDP. Media streams are sent based on their real frame size and arrival time at the sender gateway. To study the sole Internet impact, I do not impose any sender- or receiver-based system adaptation. Bandwidth is not limited at both sides. I study the TI streaming behaviors, and classify Internet traces into different categories based on the network delay, jitter and loss statistics. Repeatable experiments can be conducted and analyzed in each category.

P1: I identify the overall synchronization problem using a four-layer architecture. Each sync layer is self-contained so that the system configuration changes within a layer can create minimal impacts on other layers. I present systematic rules to decide the application- (or activity-) specific dominant frame, stream, media or site based upon the user interests. I propose metrics to compute the synchronization skew from the dominance at each layer, and present methodology to model the interrelations of synchronization quality across layers. To investigate the impact of the clock drift for computers all over the world, I plan to synchronize these computers to the local NTP servers periodically. At a span of 24-hour period, I measure the clock skews of the distributed computers, and examine whether they will affect the real applications. An automated clock skew detection and repair scheme is expected to be part of our framework.

P2: I start with two-site dissemination which includes the design of (a) the cooperative resource allocation and (b) the media packet scheduling. In (a), I allocate the available bandwidth to different 3D video streams (from multiple 3D cameras) based on the ordering of their visual contributions. The corresponding frame rate is then computed by estimating future frame sizes. I use two methods to determine the video stream with the largest contribution. The first method is sender-driven, meaning that it picks the camera that is closest to the sender speaker whose location can be obtained by the microphone array. The second one is receiver-driven similar to [14] in which the algorithm decides the camera whose orientation is closest to the receiver user view. The quality of audio and haptic sensory streams is not changed due to their low bandwidth demand. For (b), I intend to propose an adaptation algorithm which schedules media packet buffering and delivery at both sender and receiver based on the network conditions. The algorithm can be further extended in P5 by taking into account the human preferences.

P3: I extend the research findings in P2 for multi-site support. In addition, a synchronized multicast tree topology is expected for inter-site synchronization. The prolonged EED as a result of the multi-hop dissemination should be minimized for preserving the TISA interactivity. To maximize the bandwidth utilization, I intend to propose rules to prioritize media streams and dissemination path candidates. Note that users can impose different expectations on inter-site synchronization quality at different TI activities. Therefore, the upper bounds of both inter-sender and inter-receiver synchronization skews can only be decided by subjective evaluations (P5).

Table 1: SyncCast [4] vs ViewCast [14]

	SyncCast	ViewCast
Delay	Minimize delay	Delay-bounded
Bandwidth	Bandwidth-bound	Bandwidth-bound
Path Prioritization	Yes	No
Stream Prioritization	Yes	Yes
Inter-stream Sync	Yes	No
Inter-media Sync	Yes	No
Inter-site Sync	Yes	No

P4: TI activities are modeled according to their functionalities. Some activities attach more importance to the uninterrupted audio conversation, while people in other activities may prefer smooth video rendering. Multiple user-observable objective metrics are also identified to describe the human perceptions of synchronization, interactivity and media rendering quality. In each activity, I generate test samples characterizing these metrics at different values. Pairwise comparisons are expected in the subjective test for the purpose of computing the degradation score of a sample to the other. Similar to my previous VoIP study [2], I use regression to generalize the subjective results in this thesis.

P5: By employing the learned regression model at run time, I propose perception-driven algorithm which can dynamically and consistently select the operating point for media packet scheduling (P2) that leads to the best user experience, even under unseen conditions. The upper bounds of the inter-site synchronization skews are also empirically obtained from the subjective tests at different activities, as a critical factor in the multicast tree construction (P3).

5. CURRENT PROGRESS & NEXT STEPS

Current Progress. Most of my accomplished studies up to date are focusing on the dissemination of multiple 3D video streams and one single audio stream from each site. I have evaluated the multi-stream characteristics in PlanetLab (P0). The results demonstrate strong correlations of delay and loss distributions of the two media over the Internet. I have proposed receiver-driven video cooperative frame rate allocation scheme (P2) in [3] based on the dominance of the audio and video streams (P1). I have also presented synchronized multicast algorithm (P3) in [4] which have two new features: synchronization functionality and bandwidth utilization optimization, as compared to the previous ViewCast study [14] (Table 1). I have identified four user-observable objective metrics that respectively describe the subjective 3D video quality, audio quality, interactivity and synchronization (P4). I have also presented the subjective results of the pairwise comparisons for samples characterizing different values of the four metrics (P4). I have shown that G.1070 [5] is unable to describe the complications and tradeoffs of different quality attributes, and to identify the TI activity heterogeneity on user perceptions. I have generalized the offline subjective findings to guide the online media packet scheduling. The study is only limited to two-site scenario at this stage.

Next Steps. I will extend the current study to the TI system configuration with the support of multiple multi-source video, audio, and haptic sensory streams. I will propose the generalized synchronization framework and identify the dominance at each sync layer (P1). I will study the microphone array characteristics and investigate how it will help guide the network resource allocation (P2). I will also identify the impact of user subjective perceptions in multi-site TISA (P4) and achieve optimal perception-driven system adaptations (P5). All these work are expected to be done within the next 15 months.

6. IMPACTS ON COMMUNITY

The impact of our study on the research community is expected to be twofold. First, the generalized synchronization and dissemination framework can be applied to next-generation distributed systems or virtual gaming environment where multi-modal sensory streams with strong timing correlations are becoming prevalent. Second, the perception-driven streaming adaptation scheme can also be extended to any two-site or multi-site real-time 2D or 3D (stereoscopic) video conferencing for enhancing the user satisfactions.

7. REFERENCES

[1] F. Boronat, M. Montagud, and J. C. Guerri. Multimedia group synchronization approach for one-way cluster-to-cluster applications. In *Proc. IEEE Conference on Local Computer Networks*, pages 177–184, 2009.

[2] Z. Huang, B. Sat, and B. W. Wah. Automated learning of play-out scheduling algorithms for improving the perceptual conversational quality in multi-party VoIP. In *Proc. IEEE ICME*, pages 493–496, July 2008.

[3] Z. Huang, W. Wu, K. Nahrstedt, A. Arefin, and R. Rivas. Tsync: A new synchronization framework for multi-site 3d tele-immersion. In *Proc. ACM NOSSDAV*, June 2010.

[4] Z. Huang, W. Wu, K. Nahrstedt, R. Rivas, and A. Arefin. Synccast: synchronized dissemination in multi-site interactive 3D tele-immersion. In *ACM MMSys*, 2011.

[5] ITU-G.1070. Opinion model for video-telephony applications, 2007.

[6] K. Nahrstedt and L. Qiao. Stability and adaptation control for lip synchronization skews. Tchnical report, University of Illinois, USA, 1997.

[7] D. Ott and K. Mayer-Patel. Coordinated multi-streaming for 3D tele-immersion. In *ACM MM*, 2004.

[8] T. Peng and K. Du. Requirements and strategy for presentation stage synchronization of multi-object multimedia applications. In *Int'l Conference on Advanced Information Networking and Applications*, 2009.

[9] B. Sat and B. W. Wah. Playout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality. In *ACM MM*, 2007.

[10] R. Steinmetz. Human perception of jitter and media synchronization. *IEEE Journal on Selected Areas in Communications*, 14(1):61–72, 1996.

[11] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau. Robust sound source localization using a microphone array on a mobile robot. In *Proc. IEEE Int'l Conference on Intelligent Robots and Systems*, pages 1228–1233, 2003.

[12] K.-H. Vik, P. Halvorsen, and C. Griwodz. Multicast tree diameter for dynamic distributed interactive applications. In *IEEE INFOCOM*, 2008.

[13] Z. Yang. Multi-stream synchronization for 3D tele-immersive and collaborative environment. In *Proc. of ICST Conference on Immersive Telecommunications*, 2009.

[14] Z. Yang, W. Wu, K. Nahrstedt, G. Kurillo, and R. Bajcsy. Enabling multi-party 3d tele-immersive environments with viewcast. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(2), 2010.

[15] R. Zhang, C. Tang, Y. C. Hu, S. Fahmy, and X. Lin. Impact of the inaccuracy of distance prediction algorithms on internet applications: an analytical and comparative study. In *IEEE INFOCOM*, 2006.

[16] R. Zimmermann and K. Liang. Spatialized audio streaming for networked virtual environments. In *ACM MM*, 2008.