TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Sujeet Mate

Automatic Mobile Video Remixing and Collaborative Watching Systems

Julkaisu 1454 • Publication 1454

Tampere 2017

Tampereen teknillinen yliopisto. Julkaisu 1454
Tampere University of Technology. Publication 1454

Sujeet Mate

Automatic Mobile Video Remixing and Collaborative Watching Systems

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 17th of February 2017, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2017

Thesis Supervisor: **Prof. Moncef Gabbouj**
Signal Processing Laboratory
Faculty of Computing and Electrical Engineering
Tampere University of Technology
Tampere, Finland

Pre-Examiners: **Prof. Chaabane Djeraba**
LIFL UMR CNRS, University of Lille Nord Europe,
50, Avenue Halley, B.P. 70478
59658 Villeneuve d'Ascq, France

Prof. Oskar Juhlin
Department of Computer Science and Systems
Stockholm University
SE-16407 Kista, Sweden

Opponents: **Prof. Dr.-Ing. Jörg Ott**
Chair of Connected Mobility
Faculty of Informatics
Technical University of Munich
Munich, Germany

ISBN 978-952-15-3901-5 (printed)
ISBN 978-952-15-3902-2 (PDF)
ISSN 1459-2045

Abstract

In the thesis, the implications of combining collaboration with *automation for remix creation* are analyzed. We first present a sensor-enhanced Automatic Video Remixing System (AVRS), which intelligently processes mobile videos in combination with mobile device sensor information. The sensor-enhanced AVRS system involves certain architectural choices, which meet the key system requirements (leverage user generated content, use sensor information, reduce end user burden), and user experience requirements. Architecture adaptations are required to improve certain key performance parameters. In addition, certain operating parameters need to be constrained, for real world deployment feasibility. Subsequently, sensor-less cloud based AVRS and low footprint sensorless AVRS approaches are presented. The three approaches exemplify the importance of operating parameter tradeoffs for system design. The approaches cover a wide spectrum, ranging from a multimodal multi-user client-server system (sensor-enhanced AVRS) to a mobile application which can automatically generate a multi-camera remix experience from a single video. Next, we present the findings from the four user studies involving 77 users related to automatic mobile video remixing. The goal was to validate selected system design goals, provide insights for additional features and identify the challenges and bottlenecks. Topics studied include the role of automation, the value of a video remix as an event memorabilia, the requirements for different types of events and the perceived user value from creating multi-camera remix from a single video. System design implications derived from the user studies are presented. Subsequently, sport summarization, which is a specific form of remix creation is analyzed. In particular, the role of content capture method is analyzed with two complementary approaches. The first approach performs saliency detection in casually captured mobile videos; in contrast, the second one creates multi-camera summaries from role based captured content. Furthermore, a method for interactive customization of summary is presented. Next, the discussion is extended to include the role of users' situational context and the consumed content in facilitating *collaborative watching experience*. Mobile based collaborative watching architectures are described, which facilitate a common shared context between the participants. The concept of movable multimedia is introduced to highlight the multidevice environment of current day users. The thesis presents results which have been derived from end-to-end system prototypes tested in real world conditions and corroborated with extensive user impact evaluation.

.

Preface

The research presented in this paper has been carried out at Nokia Labs (earlier Nokia Research Center), Tampere. The research work was done in different research teams, but with a focus on creating innovative multimedia technologies.

I express my deepest gratitude to my supervisor Prof. Moncef Gabbouj for the trust, encouragement and guidance throughout the duration of my thesis. I would like to thank the pre-examiners of this thesis, Prof. Chaabane Djeraba and Prof. Oskar Juhlin for the insightful comments and providing them in a compressed schedule. Thanks to Prof. Jorg Ott, for agreeing to be my opponent.

The thesis would not be possible without the help and support from Nokia and my managers (both past and present). I would like to thank my manager Dr. Arto Lehtiniemi and Dr. Ville-Veikko Mattila for enabling the final push for writing the thesis. I would like to thank all my previous managers Dr. Ville-Veikko Mattila, Dr. Igor Curcio, Dr. Miska Hannuksela for their encouragement and support.

A special thanks to Igor Curcio for the long discussions and for his contributions as a co-author in all my publications in the thesis. I would like to sincerely thank all my co-authors for their contributions. I take this opportunity to thank my colleagues Francesco Cricri, Juha Ojanpera, Antti Eronen, Jussi Leppanen, Yu You, Kai Willner, Raphael Stefanini , Kostadin Dabov, Markus Pulkkinen, for the great work environment. I wish to acknowledge and thank Igor Curcio, Miska Hannuksela and Umesh Chandra for offering me research topics which formed the eventual basis of this thesis. I acknowledge and thank Ville-Veikko Mattila and Jyri Huopaniemi for facilitating the research projects on these topics. Thanks to Fehmi Chebil, for giving me the first opportunity to work with Nokia.

Last but not the least, I would like to thank my family, who mean the world to me. This thesis is dedicated to them.

Tampere 2017 Sujeet Shyamsundar Mate

||Shri||

Contents

List of Figures

List of Tables

List of Abbreviations

3.5G	Enhanced Third Generation
AVRS	Automatic Video Remixing System
CASV	Customized Automatic Smart View
DSK	Domain Specific Knowledge
EPG	Electronic Program Guide
FASV	Fully Automatic Smart View
FW	Firewall
GPS	Global Positioning System
H.264	Video Coding standard
HSDPA	High-Speed Downlink Packet Access
HSPA	High-Speed Packet Access
HTTP	Hyper Text Transfer Protocol
IDR	Instantaneous Decoder Refresh
JSON	Java Script Object Notation
LBP	Local Binary Patterns
LTE	Long Term Evolution
MIST	Mobile and Interactive Social Television
MTSV	Multi-Track Smart View
NAT	Network Address Translation
OR	Operating Requirement
PC	Personal Computer

RTSP	Real Time Streaming Protocol
SDP	Session Description Protocol
SE-AVRS	Sensor Enhanced Automatic Video Remixing System
SIP	Session Initiation Protocol
SL-AVRS	Sensor-less Automatic Video Remixing System
SLP	Service Location Protocol, version 2
SNS	Social Networking Service
SMP	Social Media Portal
SV	Smart View
TV	Television
UGC	User Generated Content
UPnP	Universal Plug and Play
URI	Universal Resource Identifier
URL	Universal Resource Locator
VOD	Video On Demand
VSS	Virtual Shared Space
WLAN	Wireless Local Area Network
XML	Extensible Mark-up Language

List of Publications

[P1] S. Mate, I.D.D. Curcio, “Automatic Video Remixing Systems”, IEEE Communications Magazine, Jan. 2017, Vol. 55, No. 1, pp. 180-187, doi: 10.1109/MCOM.2017.1500493CM.

[P2] S. Mate, I.D.D. Curcio, A. Eronen, A. Lehtiniemi, “Automatic Multi-Camera Remix from Single Video”, Proc. 30th ACM Annual Symposium on Applied Computing (SAC’15), 13-17 Apr. 2015, Salamanca, Spain, pp. 1270-1277, doi:10.1145/2695664.2695881.

[P3] S. Vihavainen, S. Mate, L. Seppälä, I.D.D. Curcio, F. Cricri, " We want more: human-computer collaboration in mobile social video remixing of music concerts", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’11), 5-10 May 2012, Austin, USA, pp.651-654, doi:10.1145/1978942.1978983.

[P4] S. Vihavainen, S. Mate, L. Liikanen, I.D.D. Curcio, "Video as Memorabilia: User Needs for Collaborative Automatic Mobile Video Production", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’12), 7-12 May 2012, Vancouver, Canada, pp.287-296, doi:10.1145/2207676.2207768.

[P5] J. Ojala, S. Mate, I.D.D. Curcio, A. Lehtiniemi, K. Väänänen-Vainio-Mattila, ” Automated creation of mobile video remixes: user trial in three event contexts”, Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia (MUM’14), 25-28 Nov., 2014, Melbourne, Australia, pp. 170-179, doi:10.1145/2677972.2677975.

[P6] F. Cricri, S. Mate, I.D.D. Curcio, M. Gabbouj, “Salient Event Detection in Basketball Mobile Videos”, IEEE International Symposium on Multimedia (ISM’14), 10-12 December, 2014, pp. 63-70, doi:10.1109/ISM.2014.67.

[P7] S. Mate, I.D.D. Curcio, R. Shetty, F. Cricri, “Mobiles Devices and Professional Equipment Synergies for Sport Summary Production”, submitted to ACM International Conference on Interactive Experiences for Television and Online Video (TVX’17), Hilversum, The Netherlands, 14-16 June., 2017.

[P8] S. Mate, I.D.D. Curcio, “Mobile and interactive social television”, IEEE Communications Magazine, Vol. 47, No. 12, Dec. 2009, pp. 116-122.

[P9] S. Mate, I.D.D. Curcio, “Consumer Experience Study of Mobile and Interactive Social Television”, Proc. 10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM’09), 15-19 Jun. 2009, Kos, Greece, doi:10.1109/WOWMOM.2009.5282415.

[P10] S. Mate, U. Chandra, I.D.D. Curcio, “Movable-Multimedia: Session Mobility in Ubiquitous Computing Ecosystem”, Proc. ACM 5th International Conference on Mobile and Ubiquitous Multimedia (MUM‘06), 4-6 Dec. 2006, Stanford, CA, U.S.A., doi:10.1145/1186655.1186663.

1 Introduction

We have all been to events, where we have ourselves recorded videos and have seen other people do the same with their mobile devices. It is usually the case that not every person is in a good position to record videos. The recorded content is diverse in terms of the recording position, the direction of recording and the media quality. The recorder who is close to the stage may record a better close-up view of the performers while the recorder who is far behind could find it difficult to do the same, but may have a good wide angle view of the event. Similarly, some people may be recording with a steady hand while some others may be jumping to the music beats while recording. Furthermore, there can be diversity in terms of the recording direction depending on their own subjective interests. While one person may be recording the performers on the stage, the other may be recording the crowd. Thus, the same event is captured with varied viewpoints. However, this content often remains unused on each recorders' device.

The opportunity loss arising with the sub-optimal or disuse of the recorded content is twofold. Firstly, the recorded content often remains unused at an individual level. The raw content which often needs some post-processing (trimming, stabilization, etc.) to make it more usable, is rarely performed. This can be attributed to the users' inability in using the right tools or paucity of time. Secondly, the recorded content from all the user can be utilized together for creating a superior representation of the event than content from a single user. Thus, it can be seen that collaboration can add value with the synergies provided by the content recorded by multiple persons. However, the challenge in leveraging the synergies is due to the lack of quality assurance (in terms of objective as well as subjective quality parameters) of the individual videos and redundancy in the captured content. A manual approach to find the best parts of the clip in terms of viewing value as well as objective media quality is too laborious and complicated for a large demography. Consequently, creating manual edits with multi-angle views is a niche activity.

Automation provides the possibility of leveraging the synergies in the content recorded by the multiple persons in an event, with negligible manual effort. Today's mobile devices can record high quality videos. In addition, they have multiple sensors (accelerometer, magnetic compass, gyroscope, GPS, etc.). These sensors provide additional situational context information about the recorded content. The situational context information may consist of camera motion (e.g., camera tilting and panning) [21][24] and the type of event [22][25][72][76]. The high quality user recorded videos and sensor data recorded by multiple users in an event; provides an opportunity to create a *rich relive experience.* Realizing the adage,*"The whole is greater than the sum of its* *parts" (Aristotle)*. Automation in combining this content can significantly lower the threshold for involving a large demography in *extracting more value from their recorded content*.

The advances in capability of camera enabled devices and high speed Internet have given a fillip to user generated content creation, where the social media portals (like OneDrive [82], Dropbox [35] and YouTube [52]) and social networking services (like Facebook [42], and Twitter [129]) form the hubs and spokes of the social media ecosystem. The increase in the size as well as the resolution of the display of mobile devices, have pushed the popularity of mobile based content consumption of social media overtake the hitherto leader, the personal computer [84]. The popularity of Internet driven content consumption has meant that it is no longer limited to user generated content. There has been a plethora of services offering Internet driven professional content, providing movies, sports, TV broadcast content and even Internet-only professional content.

Tools with rich audio visual presence like Skype based video calls, Face Time and others have become commonplace in consumer domain. There is a drive towards fusing social media activities with Internet driven content consumption, even in news and broadcast content. For example, Twitter and other social network feeds are a channel for providing a barometer of reactions from the audience at large, even as a live telecast of an event is in progress. In spite of these advances in Internet driven services, the paradigm of *watching content together* is still in its early days. *Collaborative watching of content* with people of interest has the potential to enhance the content consumption experience.

1.1 Scope and objective of thesis

The scope of this thesis is to analyze the systems aspect of automatic co-creation of multi-camera edits from mobile videos and mobile based collaborative watching of content. Thus, novel systems and their implications for content creation and consumption will be discussed, with more emphasis on the former. Figure 1 illustrates a simplified framework for automatic co-creation and collaborative watching systems. The thesis covers the end to end aspects of the proposed systems, represented as four steps consisting of capture, transport, create and consume, for simplicity. The analysis will be from the perspective of implications of system design choices on the various performance parameters as well as the impact on the user experience. This involves comparison between different architectural approaches, in terms of parameters such as number of users required, computational resource requirements and multimedia ecosystem support.

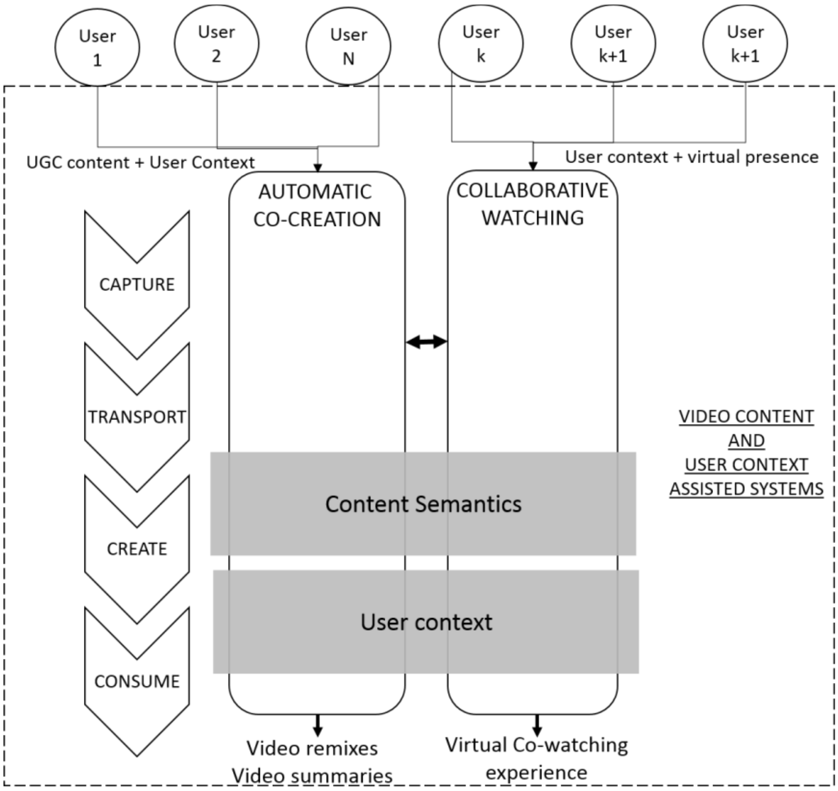


Figure 1. Automatic Co-Creation and Collaborative Watching Systems

While the automatic co-creation system provides value by delivering video remixes and video summaries of events, the collaborative watching system provides virtual co-watching experience as the value to the user. However, what both of these two systems have in common is the use of video content and the use of (recording or consuming users') situation context to generate the respective deliverables. The situational context required by both systems are however different. For example, information, such as event type, recording users' camera motion and their intended subject of interest, is relevant for automatic co-creation. In case of collaborative watching, collaborating users' instantaneous reactions (expressed with facial and body language) and, interactions with other participants are the key information to create a common shared context between users. The handling of collaborative content creation and watching in specific situation are done separately in this thesis, even though, for some type of implementations, interworking between them is possible.

The source content for the automatic co-creation research is recorded by amateur users in a casual manner with their mobile devices, unless specified differently. For example, sport content summarization includes approaches for casually recorded mobile videos as well as role based capture from mobile devices and professional cameras. The collaborative watching is primarily focused on mobile based collaborative watching. Due to the mobile device centered research, the network connectivity in this thesis is wireless.

In the description and presentation of the research and results, the focus of the thesis is to develop the end to end system as a whole, the description and analysis of individual semantic analysis algorithms is not the focus of the thesis, hence it will mainly be referenced. Selected algorithms will be presented to clearly establish the link between systemic change and performance improvement. The thesis also explores user experience impact and presents findings, in order to validate selected system design goals. Furthermore, the user experience impact studies provide insights into the need for additional features as well as the challenges and bottlenecks experienced by the key stakeholders of the system. The analysis of user experience impact emphasize the practical impact rather than theoretical models, which will only be referenced where applicable.

The thesis presents the impact of architectural choices while designing end to end systems for automatic video remixing, summarization and collaborative watching. The different architectural approaches improve particular performance parameters for certain operating scenarios while reducing the compromise on other parameters.

The research approach is both top-down and bottom-up, depending on the research question to be answered. Figure 2 gives an overview of the research flow in the thesis. Iterated versions of the parts of the sensor enhanced remixing system in section 3.2 described in publication [P1] is used as the basis system to perform the user impact studies in publications [P3], [P4] and [P5]. The lessons learnt from these studies and key stakeholder requirements were used as the input for the research work in publications [P2], [P6] and [P7]. For the collaborative watching systems, publication [P10] explored concepts for aiding multi-device concepts. The top down implementation is explored for user experience impact in publication [P9], and presented in a consolidated manner in publication [P8].

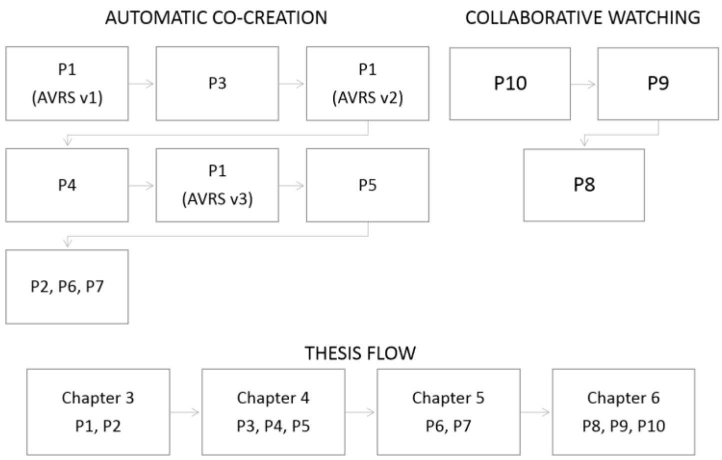


Figure 2. Research flow in the Publications and the thesis

1.2 Outline of thesis

The thesis is organized as follows. Chapter 2 introduces terms and concepts which are important for understanding the subsequent discussions in the thesis. Chapter 3 presents novel architectures for automatic video remixing systems. System architecture for sensor-enhanced remixing, sensor-less remixing and low footprint remixing are presented. The different architectures exemplify the need for system adaptation to comply with operating parameter constrains for real world deployment feasibility. Chapter 4 discusses the user experience impact of automatic collaborative video remix creation. The motivations, methods and key findings from the user studies are presented. Four user studies, covering the role of automation in remixing, the use of automatic remixes as a memorabilia, the event specific requirements and the multi-camera remix creation from a single video, are presented. Chapter 5 presents summarization approaches for sports events with two different capture techniques. The first approach is the unconstrained capture of mobile videos by amateur users. The second approach is a novel role based capture technique which uses a mix of professional cameras and mobile devices to capture content. The chapter presents the saliency detection technique for basketball sport events using both the approaches. Subsequently, a tunable multi-camera summary creation approach which leverages the earlier user experience findings is presented. Chapter 6 presents the concept, realization and user experience requirements of mobile based collaborative watching systems. Furthermore, the chapter presents the concept of movable multimedia sessions, its benefits and the current state of support.

1.3 Publications and author’s contribution

The research work presented in this dissertation consists of 10 publications [P1-P10], all of which are done in a team environment, thus more than one person contributed to the work. The main contributing person is identified as the first author of these publications. For publications where the author is not the first author, the author’s contribution has been essential as detailed in the sequel.

Publication [P1] presents the different approaches for realizing the automatic mobile video remixing systems. The author is the main contributor of the paper. He is the co-inventor of the sensor-enhanced automatic video remixing, cloud based remixing and low footprint remixing approaches. The author contributed with the main ideas behind the work, supervised the implementation of the end to end systems and did most of the writing for the paper.

A method for automatic creation of multi-camera remix experience from a single video is presented in Publication [P2]. The author is again the main contributor to the publication and wrote most of the paper. He is the co-inventor of the main idea in the paper. He also supervised the implementation of the prototype system. He planned, designed and implemented the user study.

Publication [P3] presents the user study investigating the role of automation in video remixing. The author contributed to the technical aspects of the trial and in delivering the automatic video remixes for the user study. He also contributed to the writing of the paper.

Publication [P4] is to understand the utility of automatic video remixes as a memorabilia. The author contributed to the planning, design and implementation of the user study. The author also contributed to the technical aspects of the trial and delivering the automatic video remixes as well as one of the manual remixes for the user study. He contributed to conducting the data collection trial and writing of the paper.

Requirements imposed by different types of events for automatic remixing systems is analyzed in Publication [P5]. The Author contributed to the planning, design and implementation of the user study. The Author was responsible for delivering the automatic remixes for the user study. He also contributed to the writing of the paper.

Publication [P6] presents a saliency detection method for basketball game videos recorded by end users without any constraints. The author supervised the work and the research path during the research project; and he contributed also to the paper writing. The author also contributed to the planning and execution of the data collection for this research.

Role based capture for basketball sport saliency detection and user defined summary duration is presented in Publication [P7]. The author is the main contributor and a co-inventor for the idea behind the role based capture technique used in the paper. He is also a co-inventor of the motion based saliency detection method used in this paper. He supervised the implementation of the prototype system and did most of the writing for this paper.

Publication [P8] presents mobile based collaborative watching systems approaches and user experience needs. The Author is the main author of the paper. He contributed by providing the main ideas behind the system design and supervising the implementation. He wrote most of the paper.

Consumer study of collaborative watching with mobile devices is presented in Publication [P9]. The Author is the main author of the paper. He contributed by planning, designing and implementing the user study. He wrote most of the paper.

Publication [P10] is regarding the movable multimedia sessions. The Author is the main author of this paper. He contributed by proposing majority of the ideas behind the paper and wrote most of the paper.

2 Video Creation and Consumption

This chapter establishes the background terms and concepts, as used in the thesis. The introduced topics are related to automatic creation of mobile video remixes, social media and collaborative content consumption.

2.1 Video remixing concepts

A video remix is typically a video clip, and is used in this thesis as *“a variant of the originally captured one or more video clips, from one or more cameras, by one or more users”.* The originally captured content is also referred to as source videos. A remix may consist of only multimedia rendering metadata with references to the source videos, as discussed in section 3.4. We will introduce the various approaches for creating a video remix.

2.1.1 Remixing approaches

Manual Remix

The most common method of creating content to suite a specific purpose, consists of editing done by a human for the originally captured video clips using manual video editing tools. This approach gives full creative freedom and control to the editor. The biggest drawback of this approach is that it is laborious and time consuming [P3]. A manual approach becomes untenable with increase in the number of source videos from multiple cameras.

Automatic Remix

An automatic remix is generated with the aid of information derived by semantic analysis of the source videos to understand the content. Some examples of this approach are [5][7][111][126].Typically, the derived semantic information is used in combination with heuristics or cinematic rules, to mimic a real director. There have been works which further model the editing rules with camera switching regime trained using professionally edited concert videos [105]. Thus, automatic approach is best suited for users who want to create value added content from user generated content with minimal effort. Automation enables leveraging a large amount of source video content from multiple users. The opportunities and challenges associated with this approach will be discussed in more detail, in chapters 3, 4 and 5 of the thesis.

Semi-Automatic Remix

As the name suggests, semi-automatic approach uses both manual work and automation for producing a video remix. This approach replaces parts of the manual editing work with automation but at the same time includes human input in the work flow for the other tasks. This approach can be used to design a remix creation work flow which addresses the challenges of heavy user effort and lack of creative freedom (in manual and automatic approaches). Involvement of the user in fine tuning the remixes have shown improved user acceptance [P2].

2.1.2 Multimedia analysis techniques

Content analysis

This approach involves using the recorded audio and video content from the source video clips to derive semantic information from the content. This is the most dominant method for extracting semantic information from audio-visual content. This method provides greater flexibility in defining a concept to be detected, compared to the other methods. Concepts that contain motion as well as without any movements can be detected with this method. Due to the large amount of data, especially the visual content, this method is computationally demanding. The work in [69] surveys articles on content-based multimedia information retrieval. Survey of visual content based video indexing is presented in [57] whereas [10] surveys content analysis based methods for automatic video classification. An example of audio content based music tempo estimation can be seen in [41].

Sensor analysis

Sensor data based semantic analysis has gained increased interest in the last years. This has been driven by the availability of in-built sensors such as accelerometer, magnetic compass, gyroscope, positioning sensor. The sensors provide motion and position information in a compact form. For example, to understand camera movement information, analysis of a full HD video at 30 fps would require the analysis of 62 million pixels per second. On the other hand, with magnetic compass, 10 samples per second need to be analyzed [21] [24]. In [25], sensor data is analyzed to generate semantic information to assist in mobile video remixing. Each sensor captures only a specific abstraction of the scene, hence there is less flexibility in defining a concept of detection.

Multimodal Analysis

Data belonging to different modalities capture and represent information from the recorded scene differently. This diversity afforded by analyzing data from multiple modalities (e.g., audio, video, magnetic compass, accelerometer, etc.) is a useful tool to improve the robustness of content understanding. Combining analysis information from multiple modalities has demonstrated improvement in accuracy of content indexing, according to [16]. A multi-user multimodal approach for audio-visual content captured by users with their mobile devices in an unconstrained manner is used to determine the sport type automatically [22]. The multimodal approach in [24] uses sensor data in combination with audio content for determination of semantic information from videos recorded with mobile devices.

Multi-User or Single-User

Source videos for generating a remix can be from one or more cameras. Single camera source content is inherently non-overlapping whereas multi-camera source content can have temporal overlaps. This provides an opportunity in the form of diversity of source content, which can be exploited for semantic analysis. The challenge with using such content is the additional complexity for time alignment of such source videos. This has been solved using various techniques, for example, by using the camera flashes in [125] and audio based time alignment in [64][94][95][124]. Determination of direction of interest in an event in [23] and robust sport type classification in [22] utilize data from multiple users to determine semantic information which may not be meaningful or sufficiently robust if analyzed for a single user's data. Thus, multi-user analysis provides an advantage in terms of robustness facilitated by multiple sources at the cost of increase in computational resource requirements. This has an impact on the type of analysis which can be performed in resource constrained conditions, such as a mobile device. Detailed analysis about multi-user and multimodal analysis of mobile device videos is presented in [20].

2.2 Social media

Widespread use of mobile devices with high quality audio-visual content capture capability has led to an increase in UGC. Reliable and high speed Internet connectivity has enabled sharing and consumption of UGC at a massive scale. As discussed earlier, the social media portals (SMPs) and social networking services (SNSs) are the hubs and spokes of the ecosystem for users to share, consume and collaborate UGC. Some well-known examples of SMPs are YouTube, Facebook, OneDrive; among many more. The

SMPs not only provide the means for users to consume content directly from dedicated applications (both mobile based and PC based) or webpages, they also provide APIs for other applications and services to view and upload content. In this section we will discuss the concept of “events“, as applicable in the thesis. This is followed by a brief introduction to some terms related to social media creation and consumption.

2.2.1 Event

An *“event”* is defined *as a social occasion or activity*. Events can be of different types. A typical event can be defined as something that happens in a single place or area, during a specific interval of time, typically ranging from few hours (e.g., a rock concert or a football match) to multiple days (a festival i.e. Roskilde in Denmark) [72]. This definition makes some events difficult to describe, e.g. New Year celebrations that take place almost all over the world, but nearly simultaneously. The focus of the thesis will be primarily on music dominated events such as concerts, parties, social celebrations and sport events.

2.2.2 User generated content

User generated content in the context of this thesis refers to videos recorded by users with their mobile devices. The mobile devices are assumed to be hand held and the user is assumed to be recording without any specific constrains (unless specified otherwise). In this thesis, we will be mainly dealing with mobile videos recorded in an unconstrained environment. This introduces, both intentional and unintentional motion in the recorded content, further complicating content analysis. From the perspective of objective media quality, the video segment during the panning is likely to be stable or blurry, depending on the speed of panning [23]. This is in contrast with constrained recording where the camera may be mounted on a fixed or swiveling tripod. The work in [53] describes the characteristics of UGC being unstructured, more diverse and also unedited.

2.2.3 Crowd sourcing

A large number of users attending an event, if they collaborate to co-create a video remix with their recorded content, such a method is referred to as *crowdsourced video remix creation*. *Crowdsourcing*, a modern business term coined in 2006, is defined by [81] as *the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, especially an online community, rather than from employees or suppliers*. The content recorded by the collaborating crowd is the user generated content and their contribution is crowdsourced contribution. Crowd sourced contribution may be spread out over a time period. Consequently, the source videos will be available incrementally.

If the actual process of generating a remix or summary is automatic, after receiving the source videos with crowdsourced contribution, the process is referred to as *automatic co-creation or automatic remixing*.

2.2.4 Value added content

Content captured during an event is seldom perfectly matching the intended use. The value addition occurs by modifying the raw content for the intended end use. In its simplest form, for videos, it can be trimming i.e. removing unnecessary temporal segments after manual perusing of the video. Such value added content can take on many different forms. The focus in the thesis will be on creating multi-camera video *remixes* and *video summaries* using raw videos captured with one or more cameras (see Figure 3).

- In case of single camera content, since content is linear, the key challenge is determination of salient segments for summarization. Consequently, the value added content can take the form of a short summary which includes the best parts of an event [85]. A summary can either be a temporal summary consisting of the selected time segments or a spatiotemporal summary consisting of different spatial regions corresponding to the selected time segments. In case of multi-camera summaries, a salient event may be rendered using one or more (sequential or overlapping) camera views. A multi-camera summary, can show the salient temporal segments from different viewpoints to give a better grasp of the event. For example, scoring attempts or successful scores in a sport game with different zoom levels or perspectives. In Figure 3B, S_i represents salient segments which can be rendered with one or more viewing angles (V_i). Sport summarization techniques are discussed in more detail in chapter 4.
- A multi-camera remix usually follows a linear timeline (depending on the type of content), and may consist of one or more views from different cameras, to give a multi-angle continuous viewing experience of an event. For example, a multi-camera music video of a song performed in a concert is an example of such a remix video. In case of multi-camera video remix creation, determination of appropriate switching instance, switching interval and view selection are the key challenges [79][111][126]. In Figure 3A, C_i represents the video clips recorded by different users; V_i and A_i represent the video and audio components selected from the different video clips to generate a video remix.

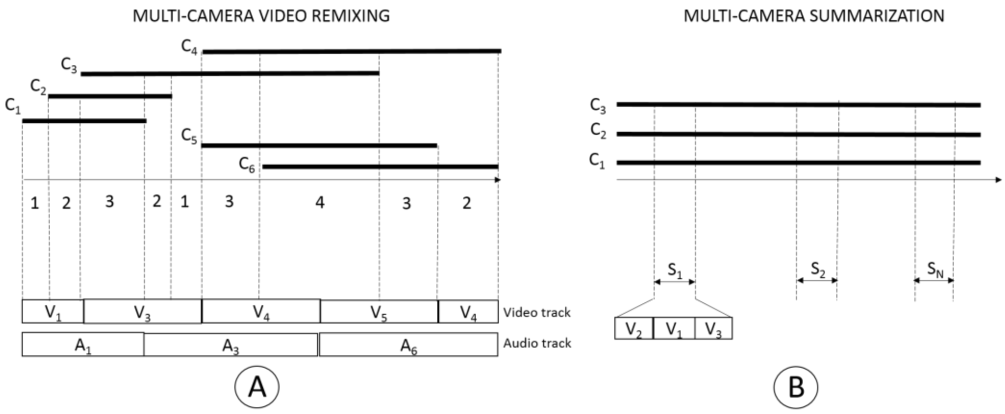


Figure 3. Multi-camera video remixing (A) and summarization (B)

2.3 Collaborative Watching

This refers to the idea of users situated in different locations, watch content mediated with such a system that creates a feeling of watching together. The feeling of watching together is created by leveraging rich interaction and presence sharing tools, which help in creating a common context. Collaborative watching is also referred to as co-watching, in the thesis. Co-watching systems are of mainly three types. The first type is optimized for living room scenarios [8][46]. Some other systems addressed the mobility aspect of collaborative consumption [116]. In the thesis, we will focus on mobile based collaborative watching aspect, which will be discussed in chapter 6. With the advent of mobile based VR [114], VR based system indicate a future of collaborative content consumption with high immersion.

2.4 System design concepts

Operating environment and infrastructure based constraints informs the choices while choosing the appropriate architecture configuration for a particular application. CAFCR (Customer objectives, Application, Functional, Conceptual and Realization) framework is an example of a process for system architecting [87]. The framework is an iterative process, which is repeated with the help of modeling, simulations and prototyping. The process ensures clear linkages between key user requirements and the resultant system implementation (see Figure 4). The CAFCR framework operates as a cyclic process, from left to right with motivations and requirements as the driver, and from right to left takes into account constraints and capabilities. There are other methods described in literature [75] and [102]. CAFCR is used as an example to illustrate the process. The CAFCR model is introduced (although not used in the thesis) to provide a system design perspective.

In the thesis, research goals, user experience requirements and piloting scenario constraints formed the “what” aspect of system design. The research goals, technical enablers, and real world operating constraints derived from piloting scenarios and piloting formed the “how” aspect of system design. As can be seen in the subsequent chapters, real world operating parameters and user experience requirements have a direct impact on the system design and operation.

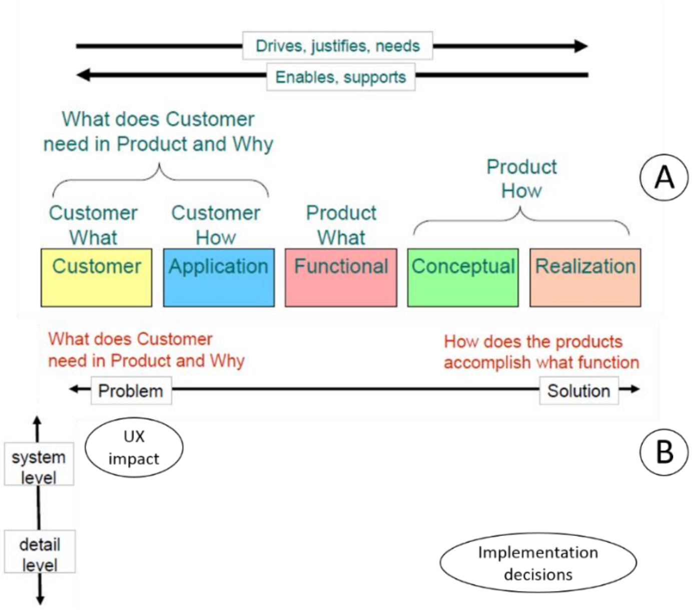


Figure 4. CAFCR framework (A), implementation method (B). Adopted from [87]

The impact of real world operating parameters may result in moving a particular functionality from the server-side to the client-side if the network latency or bandwidth is the bottleneck. In contrast, a constraint on computational resources or battery usage, may require moving certain media processing task to the server-side. We will discuss three types of systems, which broadly cover the range of options available while designing client-server systems. A tripartite pattern has been observed as part of research prototyping of end to end systems in the thesis. It should be noted, that each of these types share a few elements with the other type(s). The purpose of introducing these types is to assist in making design choices of different functionalities rather than for classification.

2.4.1 Client centric systems

This approach emphasizes the use of client side resources as much as possible. This approach is well suited for environments where network connectivity is unavailable, unreliable or too costly. The deployment cost of such systems is much less compared to the former approach, since there is no need for server side infrastructure development and maintenance. This is a popular approach in the mobile application ecosystem [6][51], since it offloads the cost of operation to the user's device. The drawback of this approach is that the application functionality is limited by the client device computational resource availability (memory, CPU, battery, etc.).

2.4.2 Server centric systems

With this approach, the goal is to transfer resource intensive functionality to the server side with the goal of making the client side resource requirements as low as feasible. This approach is also referred to as thin client approach. A VT100 terminal is a typical, but an extreme example of this approach. The client is expected to support only the functionality necessary to enable user input and output interaction with the system. The biggest advantage of this approach is low resource footprint in the client device. Depending on the application, the latency and bandwidth requirements may vary, but network connectivity is an essential requirement. Another requirement is operational maintenance of cloud infrastructure to host the server-side functionality, which may result in additional costs.

2.4.3 Hybrid systems

As the name suggests, the approach here is to leverage both the server and the client resources to implement the necessary functionality. With increasing use of cloud based infrastructure, resource availability in mobile devices (CPU, display resolution, memory, battery, etc.) and Internet connectivity, this approach is more feasible than ever before. The drawback of this approach is increased complexity and cost of such a system.

2.4.4 Limitations

The constraints which drive the choice of the system architecture is informed by the use case and the operating environment parameters. Some key limitations are presented which are often encountered while designing mobile centric multimedia applications and services. In Table 1, the first column represents the limiting parameter and its criticality for the three client-server models described above.

TABLE 1. Parameter constraints for different systems

	Client-centric constraint	Server-centric constraint	Hybrid constraint
Battery	High	Low	Moderate
CPU	High	Low	Moderate
Memory	High	Low	Moderate
Storage	Moderate	Low	Low
Network connectivity	Low	High	High
Sharing	High	Low	Low
Sensors	High	High	High

3 Automatic Mobile Video Remixing Systems

In this chapter, we describe an automated system which leverages the high quality content capture from multiple users in combination with sensor data. This approach has the following key benefits. Firstly, it reduces the effort in creating value added content such as video remixes with their own content or that from multiple users. Secondly, the use of in-built sensors in mobile devices can help produce a high quality remix with a higher efficiency in terms of computational resource usage. This chapter covers content from publications [P1] and [P2].

The next section discusses the prior work related to the publication [P1]. After the related work, we present the sensor enhanced automatic video remixing system (SE-AVRS) and the corresponding system requirements. Subsequently, the sensor-less AVRS (SL-AVRS) adaptations which are optimized for different operating scenarios and key performance parameters are described. Furthermore, the implications of system design choices in terms of benefits and compromises for the sensor-enhanced as well as sensor-less AVRS systems are discussed, to conclude the chapter. The user experience aspects of the sensor-enhanced and sensor-less AVRS systems will be discussed in Chapter 4.

3.1 Related work

In this section we present related work in the area of automatic video remixing, which uses user generated content. In [126], the proposed system utilizes audio-visual content analysis in combination with pre-defined criteria as a measure for interestingness for generating the mash-up. This approach does not leverage the sensor information to determine semantic information. The system proposed in [111] utilizes video quality, tilt of the camera, diversity of views and learning from professional edits. In comparison, our system utilizes multimodal analysis involving sensor and content data where higher level semantic information is used in combination with cinematic rules to drive the switching instance and view selection. The work [7] presents a collaborative sensor and content based approach for detecting interesting events from an occasion like a birthday party. The system consists of grouping related content, followed by determining which view might be interesting and finally the interesting segment of that view. Our approach takes the sensor analysis as well as content analysis into account to generate semantically more significant information from the recorded sensor data (region of interest) as well as video data (audio quality, audio rhythm, etc.). The approach in [5] uses the concept of focus of multiple users to determine the value of a particular part of the event. The focus is determined by estimating camera pose of the devices using content analysis. This approach also utilizes cinematic rules as well as the 180-degree rule for content editing. Compared to this approach, ours is significantly less computationally intensive, since we utilize audio-based alignment of content and also sensor-based semantic information. A narrative description based approach for generating video edits is presented in [137]. This approach utilizes end user programming for generating remixes corresponding to different scenarios.

Most of the previous research delves on the aspect of using different approaches using audio-visual data and sensor data. We will address issues related to the effect of architectural choices on performance parameters for certain operating scenarios. The underlying goal of this research is to achieve systems which improve the chosen performance parameter while minimizing the adverse impact on other parameters.

3.2 Sensor-enhanced Automatic Video Remixing System

3.2.1 End-to-End system overview

The sensor-enhanced AVRS has been implemented as a client-server system, with HTTP [43] based APIs with JSON [61] based information exchange, to enable user interaction with the system, either using a mobile application or a web browser (Figure 5).

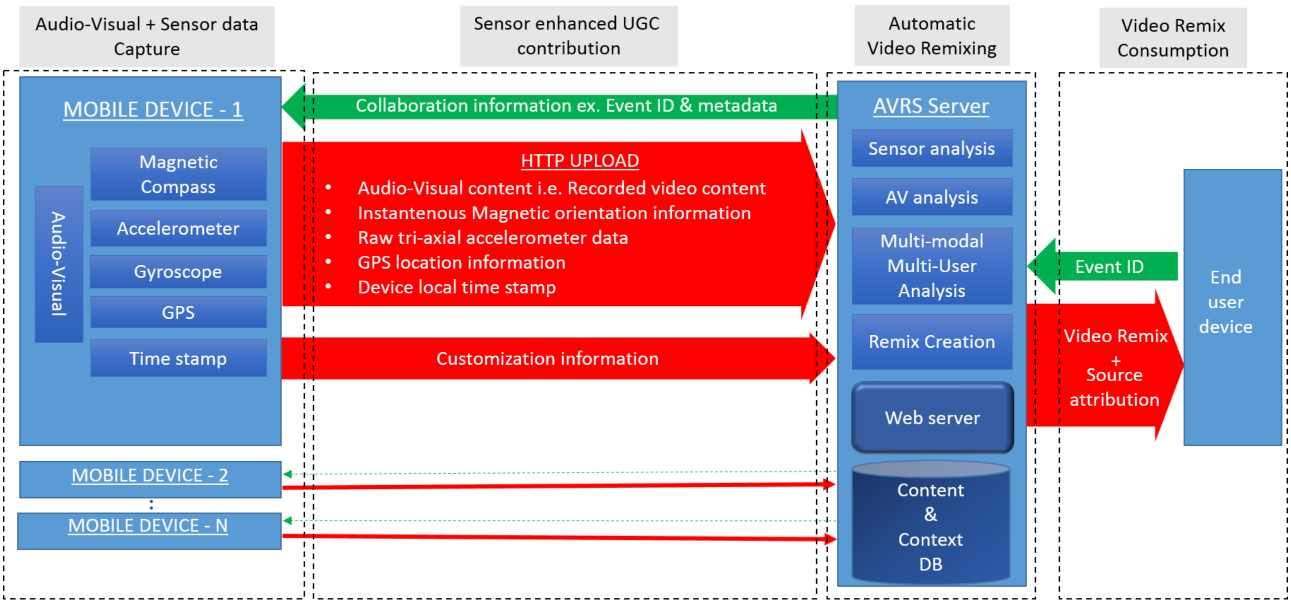


Figure 5. Sensor-enhanced AVRS E2E overview

The sensor-enhanced AVRS (SE-AVRS) functioning can be broadly divided into four main steps.

The first step consists of capturing media and associated time-aligned sensor information from the user's recording device, which includes data from magnetic compass, accelerometer, GPS, etc. The sensor data provide motion and location information of the recording device. The sensor data is encrypted and stored in the same file container as the video file.

The second step involves having an Internet based service set up which facilitates collaboration between multiple users attending an event, to effectively co-create a video remix. A logical hub or a nodal point for this collaboration and source media contribution is the virtual "event". This event placeholder is created in the system by one of the participants of the event itself or the organizers of the event. Based on the user's selection, media items (along with the associated sensor data) are uploaded to the server. In order to ensure robustness over an unreliable network, upload with small chunks of data over HTTP is used.

The third step starts with processing of all the contributed source media, which consists of sensor data in addition to the audio-visual data. This is performed to extract semantic information and determine the objective media quality of the received media from multiple users. The sensor data from heterogeneous devices is normalized to a common baseline and utilize vendor specific sensor data quality parameters to filter data. The SE-AVRS is expected to use crowd contributed UGC as source media, all of which is not received at the same time. This necessitates support for iterative and incremental remix creation. Successive remixes can include portions from the newly contributed content if they offer new and better views compared to the previous version of remix. The method in [80] proposes a criteria based sampling approach for identifying the right time for having a remix which is meaningful for end user consumption.

The fourth and the final step involves storing the video remix as a video file. The remix video file also includes metadata to acknowledge the contributing users for transparency and due accreditation. The user attribution is done by overlaying the contributing user's information when her contributed source segment is rendered.

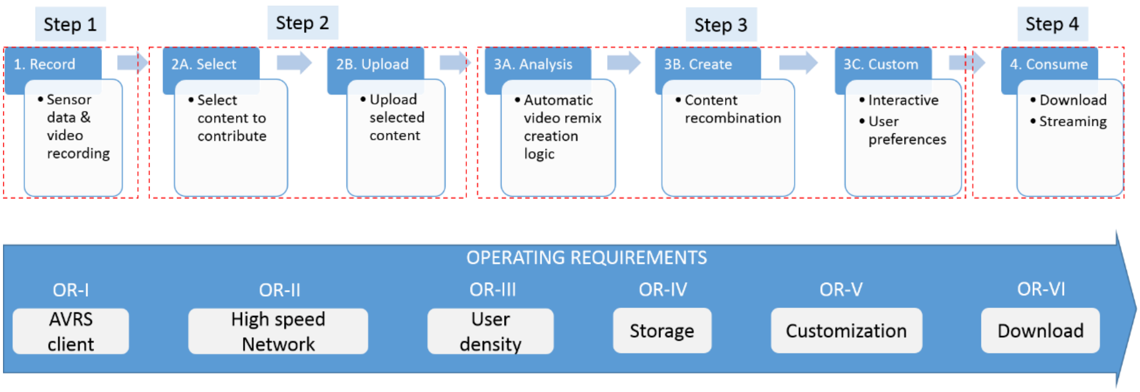


Figure 6. Sensor-enhanced AVRS functional overview.

The functional steps and the resultant operating requirements are shown are in Figure 6. In the next section, we will discuss the details of video remix creation methodology.

3.2.2 Video remixing methodology

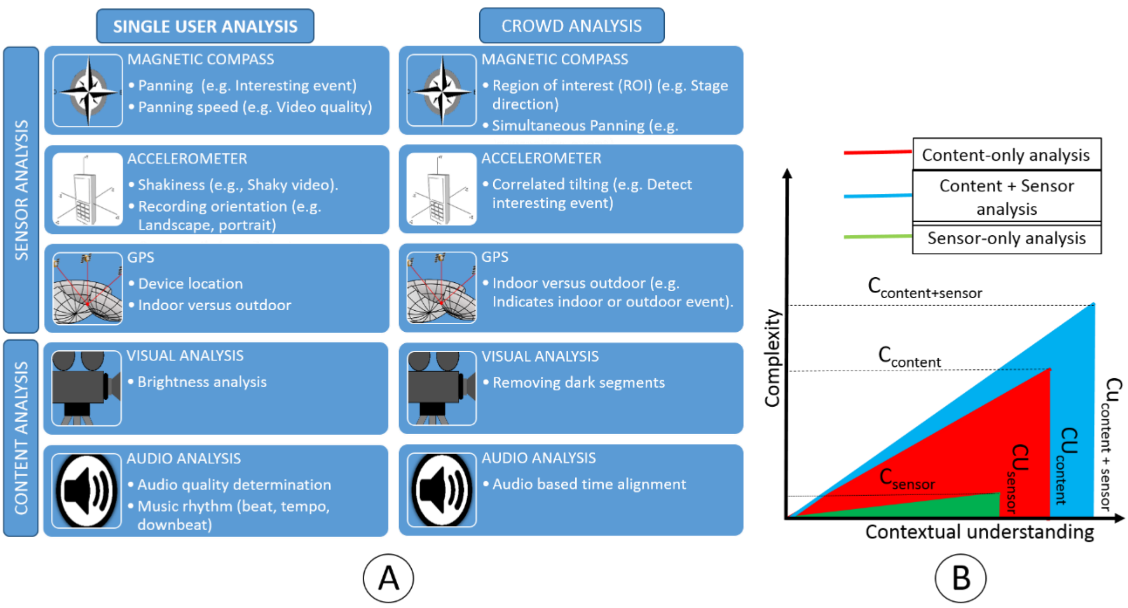


Figure 7. Sensor and content analysis methods (A) and their comparison (B), Adopted from publication [P1], Figure 2

The SE-AVRS analysis process consists of four main steps, *bad content removal, crowd-sourced media analysis, content understanding, and master switching logic*. The use of sensor data, in addition to the traditional content analysis only approach, provides significant advantages. Figure 7A presents in brief the sensor and content analysis methods utilized in this system. Figure 7B indicates high efficiency for contextual understanding can be achieved by using sensor data, whereas better contextual understanding can be obtained by combining sensor and content analysis [21][22][23][24]. Thus, sensors can play a significant role in improving efficiency as well as expanding the envelope of semantic understanding. We will next discuss the remixing steps.

Bad content removal, primarily involves removing content segments with poor objective quality. Sensor-based methods (using accelerometer and magnetic compass data) can be applied on each video file to remove shaky or blurred video segments, segments recorded with incorrect orientation (portrait versus landscape), and also those which may be recording irrelevant parts, such as feet. Dark segments are removed with content analysis [29]. Compared to the traditional content-analysis only approach, use of content analysis and motion sensor data analysis is more efficient [21].

Crowd-sourced media analysis, consists of analyzing source media and the corresponding sensor data contribution by multiple users in the event. The information, which may be insignificant for one user, when combined with the same information from multiple users in the same event, can provide valuable semantic information about the salient features of the event. For example, using magnetic compass data from all the contributing users, we can determine the significant direction of interest (e.g., a stage) in the event. Simultaneous pannings/tiltings can indicate occurrence of an interesting event [23][24]. Some methods to understand the semantic information and event type with the help of multimodal analysis have been described in [21][22][23][25]. Precise time alignment of all the contributed videos is done by analyzing the source media audio content envelope [94][95]. This is an essential requirement for seamless recombination of different source videos. The power of the crowd and the sensor information add significant value without requiring heavy computational requirements.

Content understanding, starts with determining the characteristics of the source media. Sensor data corresponding to each source media item can efficiently provide orientation (w.r.t. the magnetic North as well as the horizontal plane), fast or slow panning/tilting information about the recorded content [24][79]. Other information consists of beat, tempo, downbeat information in case of music [37][41][98][99], face information from videos [58], which is determined with content analysis. This data is used to find the appropriate instance for changing a view, and for selection of the appropriate view segment from the multiple available views.

Master switching logic, embodies the use of all the information generated in the previous steps in combination with cinematic rules to create a multi-camera video remix. The master switching logic determines the appropriate switching times of views for a multi-camera experience, and uses a method for ranking the views based on the interestingness derived from the previous steps. Bad quality content is penalized. A seamless audio experience is obtained by selecting the best quality audio track from the source content and switching to a different track, only when the currently selected track ends. These features were derived as lessons learnt in publication [P3][P4][P5]. The video remix can be personalized by providing user specific preferences to the master switching logic parameters: for example, users can indicate whether they prefer more frequent view switches or they would like to have more of their own content as part of the video remix.

The video remixing methodology is analogous to method illustrated in Figure 3A of section 2.2.4, and it is optimized for music dominated ambience. Sport content summarization will be discussed in chapter 5.

3.2.3 Operating requirements

The operating requirements for SE-AVRS are custom AVRS recording client, high speed Internet, user density, storage, customization, downloading (see Figure 6). In summary, the operating scenario generally expects the capability of sensor data capture in parallel with video recording on the participating users' mobile device and the capability of the service side infrastructure to process the sensor data together with the audio-visual data. In addition, there is a need for high-speed upload capability and minimum critical density of sensor data enriched video contributors. Overall, the above choices aim for high quality user experience without constraints on resource requirements. The implication details of operating requirements are discussed in section 3 of publication [P1]. An approach for reduced upload (operating requirement II) has been proposed in [28] which leverages sensor data, but it entails increase in system complexity (increased signaling) between the mobile device and the AVRS server.

Next we will present the sensor-less AVRS adaptation which leverages cloud based media from social media portals to address pain points experienced in the SE-AVRS system.

3.3 Sensor-less Cloud based AVRS system

Real world deployment scenarios inhibit the support for requirements needed for SE-AVRS. For example, there is limited support for devices with sensor data annotated capture of videos, as well as support for handling sensor data in the mainstream social media portals. These limitations affect directly the achieving of minimum critical density of users who can participate. This consequently affects the business model, as such a system would require proprietary support for end-to-end system realization. To overcome these limitations, a sensor-less architecture adaptation of the SE-AVRS is required, which is optimized for a different set of operating scenario parameters. In the following, a sensor-less AVRS (SL-AVRS) architecture adaptation is presented.

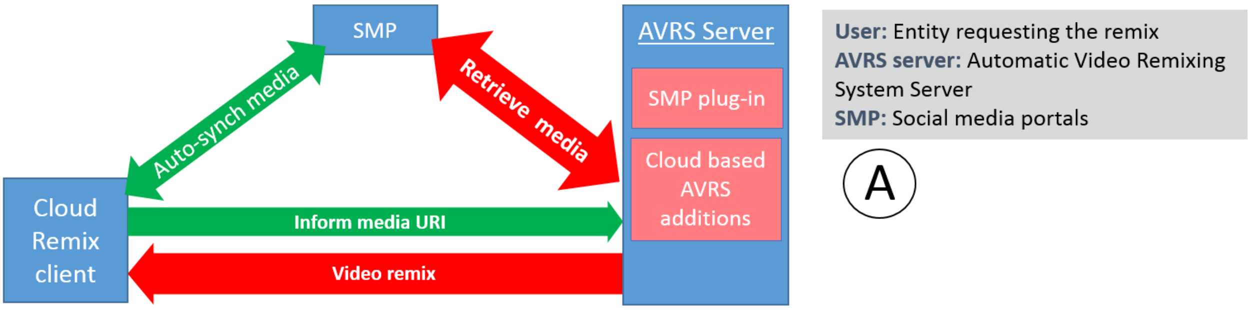
3.3.1 Motivation

From the sensor-based AVRS described above, it was found that a custom video capture client (OR-I in Figure 6), needs wide availability of devices equipped with a non-standard video recording client. Thus, devices that do not have such client would not be able to contribute. Consequently, the user density (OR-III in Figure 6), for user density, might also be compromised. In addition, the need for high speed Internet (OR-II in Figure 6), would be difficult to fulfill for users in regions having low network bandwidth, unreliable connectivity or high data usage costs. The problem is more pronounced in terms of user experience impact when a user explicitly uploads videos to get a video remix, because she has limited patience to wait for seeing any result. Based on our trials and pilot experience, contributing content to the sensor-enhanced AVRS by uploading videos was identified as a pain point by the users, during testing and user trials. Consequently, this architecture adaptation of the video remixing system envisages removing the need for uploading videos with the sole purpose of generating a video remix.

3.3.2 System Overview

The cloud remixing system envisages, retrieving source media directly from social media portals (e.g., YouTube [52]). This approach leverages the content uploaded by other users from the same event. In addition, this approach enables the users to leverage the uploaded content for sharing it with friends, in addition to creating remixes. Generally, all content available in the social media portals can be used for video remix creation. In practice, the content retrieval directly from the cloud can be done in two ways.

The first method (see Figure 3 from publication [P1]) consists of the user querying one or more Social Media Portals (SMPs) for content of interest using the search parameters supported by the respective SMPs (Step 1). The SMPs return the results based on the search parameters (Step 2). The user previews the media and selects the source media to be used for generating the video remix (Step 3). Preview and selection of optimal source content plays an important part in influencing in the quality of the video remix [77]. The selected media URLs are signaled to the AVRS sever (Step 4). The AVRS server retrieves the source media using the signaled URLs directly from the SMPs (Step 5 and 6). The automatic Video Remix video is generated in the AVRS server (Step 7). Finally, the video remix URL is signaled to the user (Step 8). The video remix file is stored on the AVRS server for a limited period, during which the user is notified to view and download the video.



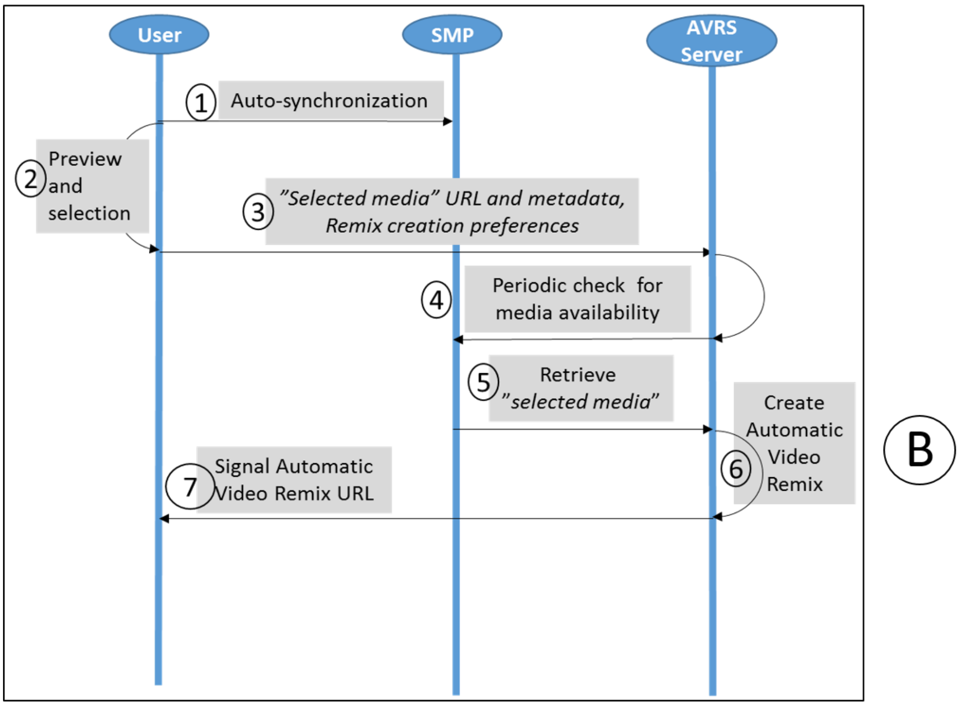


Figure 8. Cloud based SL-AVRS with Auto-Synch overview (A) and sequence (B)

In the second method (see Figure 8), the cloud remix system leverages the auto-synchronization of media on the device and the cloud (e.g., Dropbox [35], Microsoft OneDrive [82], YouTube [52], Google Drive [50], Facebook [42], etc.), which is available on increasing number of mobile devices. This feature can be used by the cloud remixing client on the users' mobile device to contribute their content to the AVRS server, and it mitigates significantly the perceived delay in the upload, even though the content selection is explicit, the upload is implicit. The contributed source media URLs or media identifiers are signaled from the cloud remix client to the AVRS server. The AVRS server periodically checks for the availability of the contributed source media on the user's SMP. When the source media is available on the user's SMP, the AVRS server retrieves the content directly from the SMP. The AVRS server creates the video remix, and subsequently stores it for a limited duration (as described in the above paragraph).

3.4 Low footprint sensor-less AVRS system

This section is derived from publication [P2] and presents an architecture adaptation of SL-AVRS system that can work completely on a mobile device, without the need for any network connectivity for generating the video remix. In addition, it is envisaged that this architecture adaptation of the video remixing system should enable creation of a multi-camera remix experience from as few *as a single user recording a single video clip* from an event. Consequently, the operating parameters are clearly different from the sensor-based AVRS and sensor-less cloud based AVRS. This requires a different architecture compared to systems discussed earlier in this chapter, while retaining the essential aspects of the video remixing methodology. This implies that the core cinematic rules, content understanding aspects and low footprint are essential for such a system.

3.4.1 Related work

We will discuss the work related for a low footprint sensor-less AVRS system. A “Zoomable videos” concept was presented in [10] and [89] as a way to interact with videos to zoom or pan a video for better clarity of certain spatial regions on the video. The viewports in [89] are interactively chosen by the users viewing a video based on his/her needs to focus on certain portions of the video. Zoomable video presents a method for creating media suitable for region of interest based streaming, to improve bandwidth efficiency when playing a high resolution video with zoom functionality [89]. The work in [10], provides an interaction overlay for interactively viewing the content. Our work, on the other hand, creates an automatic multi-camera viewing experience by utilizing semantic information in the content. In the previous sections, systems are described which utilize crowd sourced content from multiple cameras to generate a single video remix. In the low footprint adaptation, a contrasting approach that creates a multi-camera viewing experience from a single video in a music dominated environment. Carlier et al. present a crowd sourced zoom and pan detection method to create a retargeted video [11][12]. There is no dependency on initial crowd training data for our proposed system, since such data may not be available for videos that are not viewed by a large audience or the video content is for consumption in small private groups. The SmartPlayer [17], adjusts the temporal playback speed based on content identification, with the primary goal of skipping uninteresting parts in a video. In addition, the user preferences are also taken into account to tune the viewing experience, such that it matches the viewer's preferences. Our work also employs the modification of content playback to deliver the desired viewing experience. Differently from the prior art, the modification is done by understanding the relevant portions to be presented at the right time in synch with the content rhythm, for creating a multi-camera viewing experience. Cropping as an operator has been presented in [109], even though many new retargeting methods have been proposed, which acquires significance even in videos for selective zooming of certain spatial regions. Our work, on the other hand, focuses on generating the desired narrative based on fusion of multimodal analysis features and cinematic rules. The main goal is to generate a pleasing overall viewing experience rather than focus on maintaining maximum similarity with the source content. El-Alfy et al. present a method for cropping a video for surveillance application [36]. The work in [70] proposes a method for video retargeting of edited videos by understanding the visual aspects of the content. Compared to [36] and [70], our system can work with user generated content, which does not always have clean scene cuts. The low footprint system utilizes audio characteristics in addition to the visual features to make the remixing decisions. Another instance of cropping based retargeting is the commercially available application, Smart Resize [83]. This application tries to understand the content in a still image and crops it in such a way that important subjects remain intact. This approach enables adaptation to different sizes and aspect ratios. Our work extends the adaptation to videos. A lot of work has been done in interactive content retargeting by utilizing various methods. For example, in [136], manual zoom and pan are used to browse content that is much larger than the screen size. In [130], gaze tracking is used to gather information about the salient aspects of the content in the viewed scene. This can then be employed for tracking the object of interest as it moves along the video timeline. A study of user interactions presented in [13] indicates the high frequency of interaction as well as preference for watching content of interest with a zoom-in by the users in order to view the video. In contrast, our work employs automatic analysis for making the zoom-in choices.

3.4.2 Motivation

The motivation driving low footprint architecture is to remove the need for high speed network, user density and storage, as defined in section 3.3 in publication [P1]. Zooming in to different spatial regions of interests (spatial ROI) of a video for different temporal intervals can be used to create a video narrative, such that it optimally utilizes the content for a particular display resolution. We utilize the paradigm of time-dependent spatial sub-region zooming to create the desired viewing experience. In this paper, we present an automatic system that uses this paradigm to create a *multi-camera video remix viewing experience from a single video*, see Figure 1 from publication [P2]. The low footprint system is referred to as “SmartView” (SV). The details are presented in publication [P2]. The Multi-Track SmartView (MTSV) extends the SV concept to incorporate multiple videos. The MTSV creation involves analyzing the multiple videos to generate rendering metadata in a similar fashion to SV, which is used by a metadata-aware player.

3.4.3 System Overview

The video remix creation is initiated (see Figure 9) using the one or more selected videos (Step 1). The SV Application (SVA) extracts the one or more audio tracks and time aligns the multiple videos using their audio track information (Step 2). In step 3, audio characteristics like music tempo and downbeat information is determined to derive semantically coherent switching points, for rendering different views. This information is used to analyze the video frames corresponding to the switching instances. This analysis can consist of detecting faces in the video frames from one (SV) or more source videos (MTSV) to rank the inclusion of different views for each temporal segment (Step 4). Such information is used in combination with cinematic rules for generating rendering metadata (Step 5). The rendering metadata consists of source media identifier(s) for audio and visual track rendering for each temporal segment (Step 6). The spatio-temporal rendering coordinate information is stored as SV or MTSV rendering metadata. A SmartView rendering is performed with the help of a player application on the same device which is able to scale the video rendering and/or render the different source videos to deliver the desired multi-camera remix experience (Step 7). Details of the low footprint implementation can be found from publication [P2].

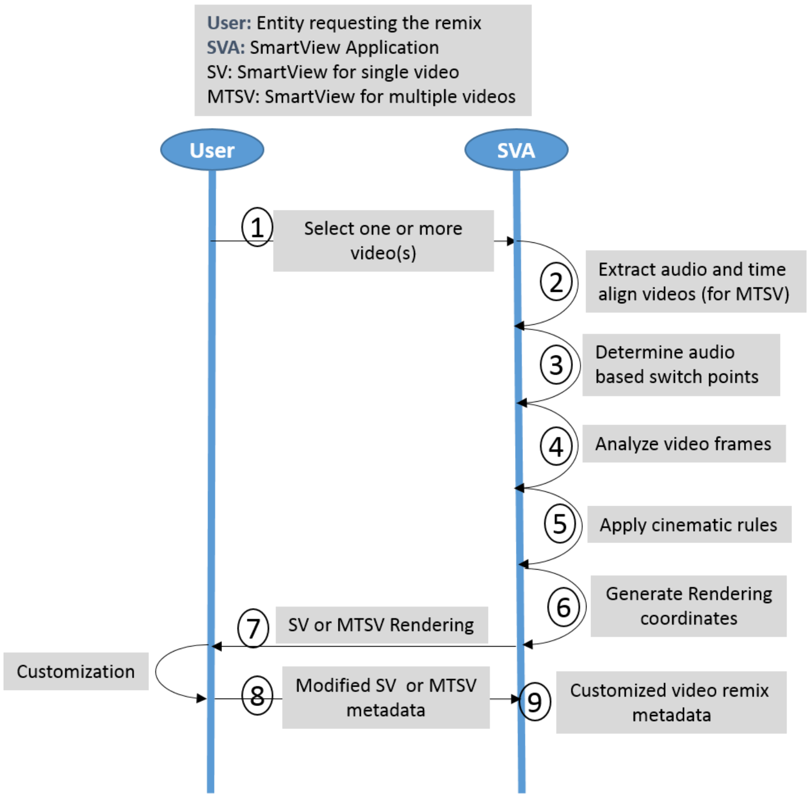


Figure 9. Low footprint sensor-less AVRS

The remix creation is limited to generating metadata and does not involve video editing or re-encoding. Consequently, the overall footprint of such a system is minimized to enable video remix creation, completely on the device. This approach also enables instantaneous interactive customization [78] of the video remix by the user without involving any media processing (Step 8). The modified SV metadata is stored within the original video file in a suitable format in case of a single source video input. For multiple source videos, the MTSV metadata is either stored in the source videos or stored separately (Step 9).

MTSV can use side loading to obtain multiple temporally overlapping videos for creating a multi-camera video remix viewing experience on a device. Such a setup can operate without the need of network connectivity and remove any dependency on the cloud. The remix creation process for multiple videos scenario is similar to the single video scenario, except for the addition step of time alignment of the multiple source videos. In case of multiple source videos, step 4 can either be repeated to rank different source videos or analyze objective visual quality to avoid bad quality views (Step 4). For multiple source videos, the rendering coordinates consist of a source video identifier for video and audio track for each temporal segment [27].

3.5 Comparison and advantages of the solutions

Architectural changes in the system to meet the application usage requirements has an impact on different operating parameters. In this section we will discuss the implications on different parameters (see Table 2).

TABLE 2. Comparison of video remixing systems.

	Low footprint SL-AVRS	Cloud based SL-AVRS	Sensor-enhanced AVRS
Min. # of source videos	1	>1	>1
Min. # of people	1	1	>1
Source videos	Locally captured; downloaded	YouTube or other portals (no capture required), CE devices, mobile platforms	AVRS client, iPhone, Android, CE devices
Explicit Upload required	No	No (or autosync services)	Yes
Final output downloading required	No	Optional	Optional (streaming is preferred)
Manual customization capability	Yes	No (some customization is possible)	No (some customization is possible)

The *sensor-enhanced AVRS* utilizes sensor augmented source media from a large number of users. This enables the video remixing process to have a higher amount of information to generate a high quality video remix. The practical aspects related to wide penetration of sensor equipped multimedia capture clients adversely affects the user density requirement. The lack of inherent support for sensor data enriched UGC media from popular SMPs, inhibits widespread use due to increased system complexity and infrastructure requirements. Furthermore, a proprietary end-to-end set up requires single purpose content upload for video remixing for the end user and higher costs for the operator.

The *cloud based sensor-less AVRS*, since it relies on the content from SMPs, may or may not have sensor augmented source media. This reduces the amount of semantic information available for choosing the views in the remix (e.g., device landscape/portrait orientation during recording). However, this approach not only removes the need for users to upload content for a single specific purpose of video remix creation and but also allows use of various SMPs. The user density requirement is down to one person, since it allows leveraging the content available on various SMPs. Consequently, it is of great advantage in terms of managing costs and reducing system complexity.

The *low footprint sensor-less AVRS* architecture is the leanest since there are no dedicated infrastructure requirements. It achieves good user experience in focused operating scenarios (e.g., music dominated situations). It is ideal for a single or a small group of users, since the user density requirement threshold is just one.

A comparison of remix quality and overall complexity for the sensor-enhanced and sensor-less approach is presented in Figure 5A, in publication [P1]. Figure 5B in publication [P1] illustrates the comparison between the user density requirements and the upload effort as well as the storage server requirement.

4 Automatic Mobile Video Remixing – UX aspects

In this chapter the key findings from four user studies are presented. The user studies involve 77 users and consider different aspects of collaborative and automatic video remixing. This chapter is derived from publications [P2], [P3], [P4]and [P5]. The word "*automatic*" means a machine or device having controls that allow something to work or happen *without being directly controlled by a person* [81]. *Video remixing*, on the other hand, is an artistic endeavor of the editor or the director. Hence at first glance, the two may seem immiscible and impossible to co-exist. However, the user study results, suggest, it is not exactly true. Diverse topics related to the impact on user experience were investigated in these user studies (see Figure 10).

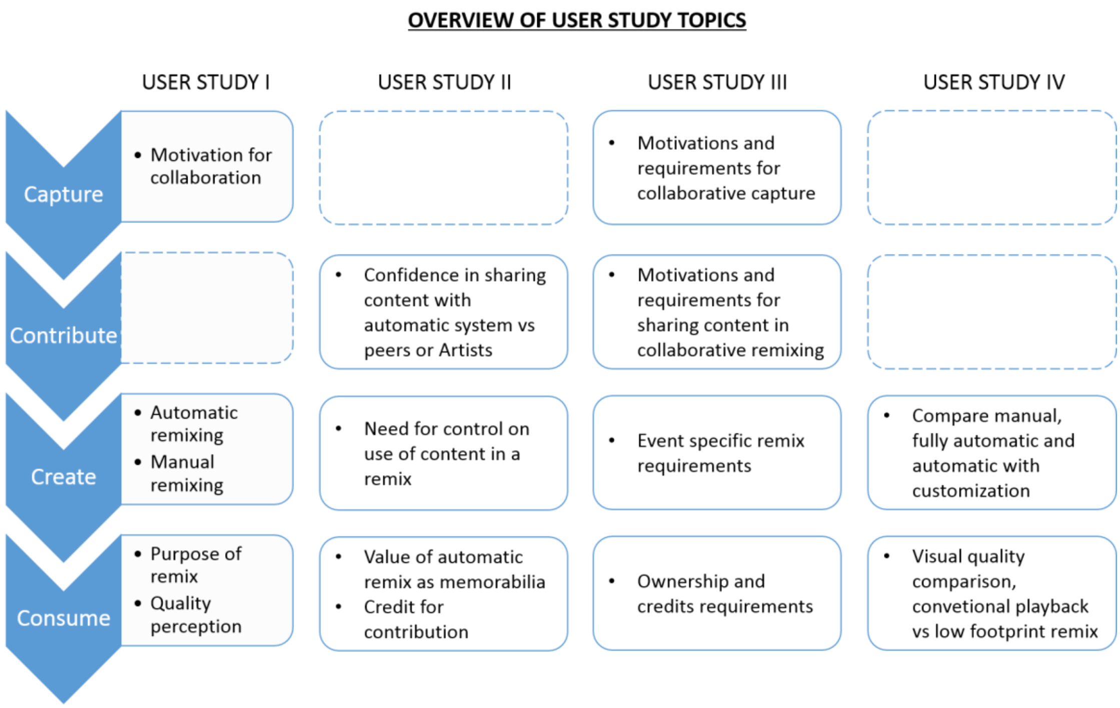


Figure 10. Overview of the topics covered in each user study.

The rest of the chapter is structured as follows. We start with presenting the motivations in brief for conducting the user studies. Subsequently, we discuss the background work related to the user studies. This is followed by presenting the experimental findings from the user studies. We conclude the chapter with design implications derived from the user studies. User study details like research questions, procedure and detailed findings can be referred from the corresponding publications.

4.1 Motivation

The user studies were conducted to understand the user impact of the video remixing systems (described in chapter 3). In addition, the objective was to identify new features which can help in improving the user experience as well as identify the pain points for the stakeholders of the system. The key motivations of each of the user study are described in the following.

The first user study was performed to understand the *role of automation in automatic and collaborative video remix creation*. First version of the sensor-enhanced AVRS method (see section 3.2.2) was used to generate the automatic remixes. The user study corresponds to publication [P3].

The second user study was to determine the value of *mobile video remix as an event memorabilia*. This user study (published as [P4]) explores the use of concert video recordings and video remixes as a memorabilia. A second iterated version of the sensor-enhanced AVRS was used for this study. In addition, the detailed dynamics of collaboration for video remixes was studied. In particular, the aspect of control on content contributions and acknowledgment for the use of source content in a remix video, was studied.

The third user study explored the *requirements imposed by different types of events* on video remix creation (as described in publication [P5]). The three events chosen for this study were, Ice Hockey game (a sport event), Doctoral dissertation (a private event) and a music concert. A third iterated version of the sensor-enhanced AVRS was used for this study.

The fourth user study objective was to understand the effectiveness of a low footprint SL-AVRS system for creating a multi-camera remix experience from a single video, recorded by a single camera (as explained in section 3.4) and in publication [P2].

4.2 Related work

In this section, the background work related to the user studies is presented.

Role of automation in video remix creation:

Video remix creation for dance club scenario was studied in [38], which suggests mobile videos recorded by club patrons can enhance interactions between the club visitors and

VJs. Live remixing system proposed in [39], was implemented to prototype a scenario which involved club visitors providing the mobile recorded live video stream which can be subsequently switched for the desired output stream. From the perspective of this user study, this provides a multi-stake-holder interaction in a music dominated dark environment. This is in variance with a post-event production approach in our study. In [44], automatic analysis of audio and video track of the recorded video is done to create a music video. Additionally, from the perspective of this paper, the semi-automatic approach where the user manually selects the video clip and the clips are automatically synchronized with the audio track. In [64] a system for synchronization and organization of user contributed videos is presented. Their work provides useful cues regarding the representation of overlapping user generated content and practical approach for implementing such a representation. In [47], the semi-automatic approach creates customized videos using home videos captured with basic home video cameras. The work focuses on home videos rather than music dominated videos. It is interesting from the perspective of different levels of automation. In [44] both automatic and semi-automatic approaches are presented which uses significant audio changes and matching temporal video segments. In [48], a user study on the semi-automatic system referred to in [44] and [47] is presented. The results of the user study suggest a useful balance between automation and user control. The work in [65], presents a holistic study of practices around home videos. The work also suggests that short videos are not considered worth editing. Multimedia research should shift from semantics to pragmatics by designing systems, is proposed in [122], such that, it can utilize the specific context in which the media is being used. This is particularly relevant for designing automatic remixing systems, which may have to deal with different types of events having similar type of audio-visual concepts present in the scene. In addition to video remix creation applications, [60] and [115] is related research on mobile content creation and sharing.

Video editing, multi-camera video production, music videos, automation, live contexts have been studied in previous studies. However, none of the previous studies have combined all the aspects in the same study. Furthermore, the differentiated views of the key stakeholders, Artists and Fans in a concert scenario.

Mobile video remix as a memorabilia:

The work in [14] and [68] indicate the importance of visual content in relive experience for an event. The work in [59] shows the prominent role of user captured content in reconstruction of a shared event experience in case of large scale events. The context of the user study in a similar large festival provides good grounds for conducting research regarding memorabilia creation with user generated content. Automation and user control need to be balanced to ensure a favorable user experience, which is supported by

[123]. Excessive automation can have adverse impact on user participation in such a service, especially if the automation is not matching user needs and the user cannot exercise control [P3]. User rights management form an important aspect of the remix creation ecosystem and culture. The works [33][74][86] provide useful insights about the authorship issues in remixing culture, visual content re-use aspects and effectiveness of attribution in online remixing.

Video remix requirements for different types of events:

The habit of amateur mobile video creation is a growing phenomenon [62][63]. Social media portals and social networking services are the media storage and sharing hubs. The UGC is the content driving the ecosystem. In a study by Lehmuskallio et al. [68], editing these snapshot videos is a prominent problem that the users face. Automatic collaborative remixing provides a low threshold barrier for a large demography of users. The work in [131] presents the drivers and obstacles for social experience with a focus on web services. The goal of the current study is to understand the user habits and derive requirements for automatic remixing in a collaborative scenario. There is large body of work that explores content recording and sharing [92][96][132], as well as collaborative video creation [7][32][38][40]. However, there is a need to further research the event context specific requirements for automatic collaborative remixing. Collaboration in video creation requires learning, which is addressed in the work by Weilenmann et al [135]. The learning can happen playfully by imitating the professionals, as the work by Juhlin et al [62] suggests. The prior art studies [62] and [135] consider systems which require collaboration during the capture phase. In the contrary, the system in current study corresponds to create phase from the collectively captured and shared videos. Interaction with the system in the moment of capturing is to be kept minimal.

Publication [P4] studied use of automatic collaborative remixing in the context of a large-scale festival. In that study, the users posited trust in an automatic remixing service, even though they stated that they did not want public acknowledgment by default, if their content ended up in the remix. Monroy-Hernandez et al. [86] divide acknowledgement in the content to “attribution” (automatic and computer generated) and “credit” (by other users). The interestingness of the content to a user depends on how closely the user can identify oneself with the content, and this feeling of closeness influences the need for attribution. A similar study is warranted for automatic collaborative remixing for UGC captured in different types of events.

Multi-camera remix from a single video:

The related work has been discussed in section 3.4.1 and publication [P2] (for details).

4.3 Experimental findings

In this section we will present the key findings from the user studies. They are presented as four groups, corresponding to each of the user studies.

4.3.1 Role of automation in video remix creation

Motivations for collaborative remixing

Different stakeholders (Fans and Artists) had different motivations.

- The Fans' main motivation was to use the concert recording as a *memorabilia*.
- The Artists saw the video remixes as a method to *promote the band image* as well as use them as *promotion material* for the venue owners. The Artists saw the remixes of live concerts of great value to demonstrate the interaction between the band and the crowd, especially for those who did not attend the concert.
- Significantly, both the Artists and the Fans, saw the *video remixes as a method to expand the timeline of the concert*. Furthermore, collaborative remixes promote interaction between Fans.

Reactions towards manual remixing

Manual remix creation was taken as a personal challenge and users were open to publishing it, if there was an appropriate opportunity.

- Manual remixing created personal involvement and a sense of accomplishment. However, the effort was seen to be daunting for most users.
- The lone user who created manual remix, decided to concentrate on only one song, since it was *too difficult to keep track of multiple camera views*. The users had difficulty in finding good scenes from others' content so ended up using her own content most of the time. It becomes clear that the amount of *multi-camera content becomes quickly overwhelming* for fully manual editing.
- The manual remixes (one reference and one Fan made) were both recognized by 5 out of 6 participants, as made by a human. The reference mix was well received for the *continuous audio track* and the *synchronous camera view*

change with the music. The Fan made mix received negative feedback due to a discontinuous audio track.

Reactions towards automatic remixing

In the focus group discussion, many participants expressed curiosity, interest and skepticism for automatic remix creation. This is not surprising given that it is not a commonly available system.

- The Fans found it bothersome to record content without knowing how the automatic system would use it. The editing phase was on the minds of the Fans during the video recording. Consequently, the users tried to facilitate the automatic remixing, even though they had no knowledge about how it worked.
- Five out of six participants recognized the automatic video remix clip as made by a machine. The automatic remix was not liked as much as the manual remix. However, when the Fans came to know that they were made by a machine, their characterization of the automatic remix video clips became more positive. Also, the Artists did not find the automatic compilations suitable for publishing.
- The suggestions included incorporation of *music synchronized switching* between cameras, accurate *audio-visual synchronization* and *removal of dark segments*. Furthermore, the *need for human intervention* was emphasized by the Artists.

4.3.2 Mobile video remix as a memorabilia

The key findings are derived from the responses to the web questionnaire requested from 43 trial participants via email, out of which 19 participants responded (10 males and 9 females).

Automatic remixes as memorabilia

- The best manual remixes are rated better than the automatic remixes for overall quality. However, *automatic remixes perform as good as the best manual remixes* for their value as a memorabilia.
- Significantly, for a memorabilia, the users are *more accommodative of an off-beat switch or a shaky video segment* included in the remix.
- According to some users, the switching pace and an occasional shaky video segment in the *video remix seemed to portray the concert ambience well*.

The need for control of clips

This is an important aspect, since it gives the contributing users a sense of security. A user who is insecure about how her contributed content will be used is less likely to contribute.

- The users desired *more control* on the clips when contributing or sharing content with an *entity they trusted less*.
- The users' *need for control is the highest when contributing or sharing content with an unknown peer*. On the other hand, the need for control is less when the contributed entity is Artists and *least for an automatic remixing service*.
- Trust and risk factors are crucial in a multi-agency system like an automatic collaborative mobile video remixing system [71]. *Deterministic behavior is an important factor*, as the user expects such a system as less likely to violate her impression management goals.

Attitudes for public acknowledgment

- The users do not want public acknowledgment for their content contributions if the remix is generated by an automatic remix creation system.
- Users are keen to have acknowledgment if the remix is created by the Artists. This is because the users want to be associated with the Artists and it contributes positively to impression management goals.
- Most users expressed desire to review the final outcome before providing their consent for being acknowledged in the video remix. The final video remix quality and reputation of the publication forum inform the users' preference for acknowledgment.

4.3.3 Video remix requirements for different types of events

These are divided into three broad categories.

Motivations for capturing and sharing videos

- An important aim of the automatic collaborative remixing system is to add reciprocity to the video capturing and contribution. When a user contributes to a collaborative video remix system, the user gets others' content in return. This *motivates the users* by getting other viewpoints and temporal segments which were not captured by themselves. This experience also adds a *feeling of connectedness with other capturers* [92]. Furthermore, others' material can enhance their own captured material. This adds social dimension, which also *encourages users to capture more content* [73][92].

- *Ease of creating the remixes* was stated to be the main benefit of using the automatic remixing service (see section 3.2.2). In absence of such an easy way to create remixes, many users felt that many videos would be left unused on their devices. *Automatic remixing service provides a channel for unused unedited content*.
- Motivation to be *creative and express themselves* was inspired by the knowledge about the presence of other capturers for collaborative remixing. In addition, presence of many recorders of the event, gives a *sense of flexibility* for recording the unexpected and interesting views in the event.
- The aspect of *sharing of content with non-attendees* was considered most important in case of large public events like Ice Hockey games, on the other hand, the *relive aspect* was most important for *concert attendees*. For a relatively *small and private event* like doctoral dissertation, there was higher interest in knowing the *identity of the person who recorded the clips*.

Requirements for different event contexts

- For sports events, the *180-degree line* was considered a key criteria to avoid the alternation between the left and right sides of the venue. *Smooth narrative*, without frequent switch in the camera views was considered important. In addition, a *continuous audio track* in the remix was seen as desirable, even if there would be a switch between cameras.
- For music concerts, in addition to the findings from the previous user studies in section 4.3.1 and 4.3.2, a key finding from this study was that users do not have very high expectations from videos recorded with mobile devices. This may be due to the poor illumination conditions in many concerts (e.g., if they are indoors and dominated with strobe lights). Although this may change in future with improvements in mobile device capabilities. Users were *interested to see viewpoints of other recording users* in the remix. Coverage of the *concert ambience* was considered important, which includes the *audience and the band*. An abrupt cut would *break an in-progress narrative* resulting in poor user experience.
- Formal events could have significantly different requirements depending on the specific type of event. For example, for a dissertation presentation which is speech dominated, clear legibility of what is being spoken was considered to be important. Furthermore, *video capturing in formal events needs to be discrete*. Additionally, it was important for the *key persons to be presented in the remix according to their roles*.

Collaboration and Ownership

- Collaborating users *preferred to have a layered approach*, *for sharing* their recorded content and remixes. The users also wanted the ability to *utilize content in a layered approach*, especially for large events for collaborative remix creation. For example, create a remix with viewpoints of friends or such closed group of people.
- Uses of video remixes included being a *gift* to friends and relatives, or as a bigger group *memorabilia*.
- User accreditation was seen to be important, this can be gauged from a strong interest in knowing which remixes their contributions are used. *Acknowledgment in the remixes was preferred by some users. This finding was in variance with the previous study*. This suggests providing controls to users for managing the acknowledgment is important.
- The type of event affected the need for small groups in collaborative remixing. Users from concerts and dissertation event wanted this more than the users from Ice Hockey game. Collaborative remixing service which also makes the source content available, provides a channel for content discovery to the users [92][93].
- The fundamental *idea of co-ownership of the remix* by all the contributors was supported by all the users, even if their individual clips did not end up in the final remix.

4.3.4 Multi-camera remix from a single video

The findings are grouped into three main categories. The first category presents findings related to value addition from creating a multi-camera remix experience from a single video. The second category presents findings on the impact on visual quality due to the zooming of a selected region of interest in the original video. The third category discussed the user feedback regarding sharing and ownership of such type of remixes.

Before presenting the findings, the terms used in describing the findings are explained. The process of creating content analysis derived rendering of a video clip is termed as Smart View (SV) in publication [P2]. The term FASV (Fully Automatic Smart View) playback refers to a multi-camera remix playback experience generated from a single video. The term CV (Conventional Video) playback refers to the conventional playback, consisting of full video frame resizing to match the native video resolution with the display resolution. The term CASV (Customized Automatic Smart View) refers to an end user customized version of FASV.

Value addition over conventional content playback

- The study suggests that 5 out of 9 users liked the FASV playback experience over the CV playback. The synch between the virtual view switches with the music and close-up shots of the persons in the video was liked by those who preferred FASV over CV playback experience. This is supported by a key finding, already discovered in the first user study findings 4.3.1.
- In contrast, for those who preferred the CV playback more, felt the switches were too often and distracting the attention from the main subject (violin music) in the video. This suggests that even though the video switches may be in synch with the video, the switching regime may not suite a particular user's taste.
- The need for *customization and user control* was indicated even before the users were made aware of the possibility to customize the FASV playback. After customizing the FASV playback, the customized playback was liked by all the 9 users. The customization of an automatic remix introduces user control in an otherwise black-box process. Thus *customization significantly enhances user involvement and a sense of own creation*, which significantly enhances its acceptability for the user. Overall, 8 out of 9 indicated that they see the value in the cascaded use of automation and customization. This is suggested by findings from the first user study in 4.3.1. Interestingly, one user suggested use of only the interactive customization for creating the multi-camera experience.

Visual quality perception

- This was a novel situation involving subjective visual quality feedback compared to the conventional visual quality tests which are often standalone. In this test, a playback experience with different size of virtual view areas (region of interest selected from the complete video frame) requiring different levels of zooming to fit the native display resolution, is the scenario.
- The visual quality perception is not adversely affected by SV playback. Significantly, CASV playback gets better visual quality rating compared to FASV playback. Similar trend but with lower change (and statistically not very significant) is seen for those who prefer FASV over CV and vice versa.
- Interestingly, these results suggest that overall visual quality perception is informed by the view switching experience.

Sharing and ownership

- For most users in the study, while the users are open to sharing of videos with FASV/CASV playback possibility, they are not so open to the idea of allowing others to create an SV playback of their own video.
- The users in the study perceived a greater risk of somebody creating a narrative which may violate the users' impression management goals, there is less openness towards this possibility. This is supported by the findings in the second user study in section 4.3.2, which suggest a stronger need for control when the editing agency behavior may be less deterministic.

4.4 Design Recommendations

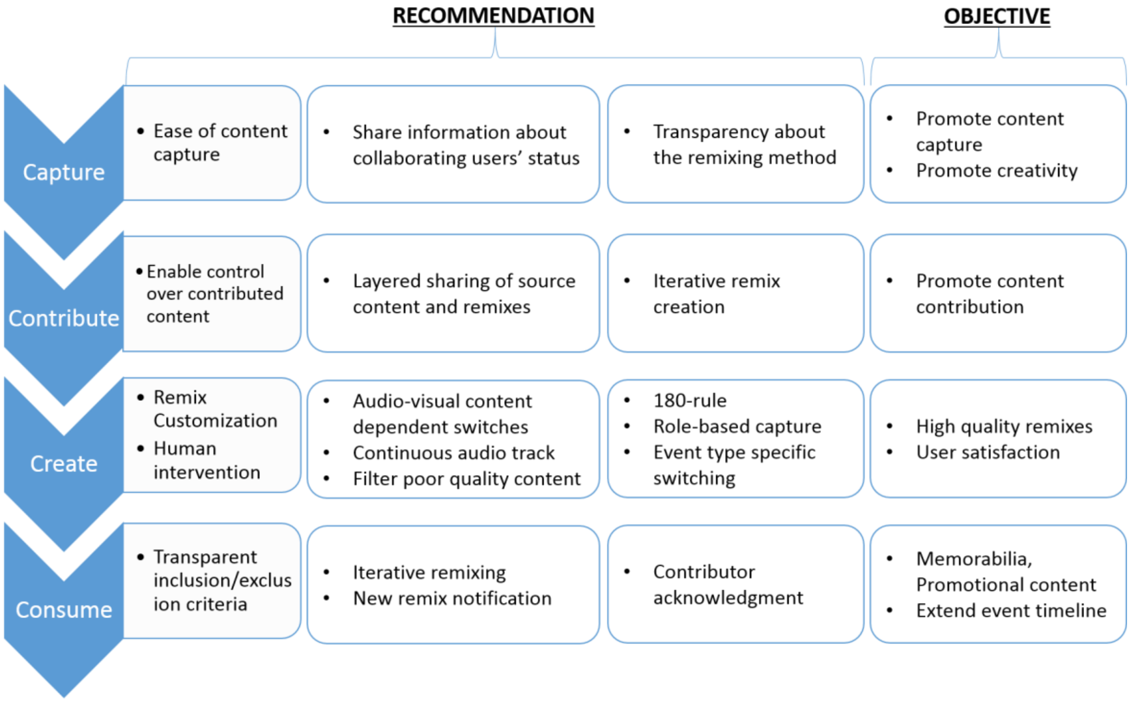


Figure 11. Overview of design recommendations.

In this section design recommendations and expected outcomes are presented (see Figure 11). These are derived from the four user studies presented in the previous sections. The recommendations cover all the stages for an automatic collaborative mobile video remixing system, discussed in chapter 3.

4.4.1 Capture

- *Transparency* about the automatic remixing method, can help the interested users, to adapt their recording styles to assist in creating remixes. There is less concern regarding unexpected use of the recorded content.
- *Easy content capture* is important to avoid diverting user's attention from enjoying the event.
- *Information about collaborating recorders'* presence and activities in the event provides feeling of connectedness with other collaborating capturers. The social dimension in the collaborative scenario encourages users to capture more content and to be more creative.

4.4.2 Contribute

- *Need for control* in the remix creation is essential to facilitate the users' content contribution. It should be easy for the users to withdraw the contributed content.
- Enable *layered sharing for the contributed content*, depending on the contributing user's needs, facilitates collaboration and content contribution. This allows for controlled sharing of content with a small group, a wider audience or make it public.
- Enable *layered sharing for the remixes* created is important. In case of collaborative scenario, there should be a clearly defined authority (one of the users in the group or based on majority voting) for changing the status of the video remix to public or to withdraw the video clip.
- Enabling *iterative remix creation* achieves two objectives. First, it extends timeline of the event. Second, incorporating crowdsourced contributions arriving after the latest version of the video remix. On the other hand, for small groups, contributions can be monitored more closely, hence iterations can be minimized to reduce the remixing infrastructure usage.

4.4.3 Create

- *Continuous audio track* while switching viewpoints provides a cohesive experience for the viewer. This requires accurate time alignment of the contributed content.
- *View switches in synch with the music characteristics* and avoidance of bad quality content, gives a semblance of content understanding for the automatic remixing system. This is important for good viewing experience and acceptance of the automatic remix by the user.
- *Filtering* non-processed and casually captured content *for poor objective media quality* (like shakiness, dark segments, and poor quality audio) is important.

- *180-degree rule* compliance is essential for *sport content*. Important person identification and role based capture in private or formal events can help in creating a meaningful narrative in the video remix. For concerts, covering the entire ambience, including the stage, the audience and the surroundings is important.
- Possibility of *customization* with human intervention in modifying the automatic remix is essential to give the user *a sense of control and involvement* in the remixing process.
- Ideally, customization capability should also allow a user to *create a remix completely manually* (even if with limited capabilities), for situations not supported by the automatic remixing methodology. This is important, especially for personal use, which covers a wide range of situations.

4.4.4 Consume

- *Public acknowledgment* in the remix video, provides a channel to credit the content contributing users. The public acknowledgment should be a configurable option for the user. If situation permits, the user acknowledgment decision should be after reviewing the final remix.
- Promotional material, memorabilia, extension of timeline of the concert are key motivations for creating remixes. *Iterative remixing* allows incorporating additional views of better subjective or objective quality.
- In iterative remixing paradigm, while a user's contributed content is included in the Nth iteration but may be excluded in the (N+1) iteration. Such scenario is not commonly experienced by users. Hence, the *final remix ownership criteria* should be transparent and made clear, to avoid disappointment to content contributors.
- Iterative remixing also requires *notifying the user* when an updated version is available. This requires integration of AVRS system with a suitable push notification service or a polling based mechanism to know about new remix versions.

The next chapter leverages the lessons learnt from this chapter for creating high quality multi-camera summaries for sport events.

5 Automatic Mobile Video Sport Summarization

This chapter discusses summarization of sport events. Content from publications [P6] and [P7] form the basis of this chapter. Creation as well as presentation of a sport summary has different requirements compared to a multi-camera music videos, which was discussed in the previous chapter. A continuous timeline is required to experience a song in a concert or a dance performance in a folk festival. On the other hand, a long duration content is more convenient to consume, when broken into bite-sized pieces. A video summary of a particular sport event is a step towards that goal. The first step in creating summaries is identifying the salient instances of the sport event. The second step is extracting the appropriate content segments from one or more cameras for presentation (see section 2.2.4, Figure 3B). Salient events are usually defined with domain specific knowledge (DSK), for example, a successful basket or a goal in a sport event correspond to a highlight.

In this chapter, salient event detection methods for basketball which utilize two complementary methods for capture of source content, are discussed. The first method detects salient events from unconstrained UGC captured by amateur users with mobile devices. On the other hand, a role based capture set-up which leverages the synergies between professional equipment and mobile devices, is used in the second method. Subsequently, we will present a method for creating multi-camera tunable summaries, where the end user defines the duration of the final multi-camera summary video (this work leverages the design recommendations derived from the work in chapters 3 and 4).

5.1 Related work

We will now discuss background related to publications [P6] and [P7]. The focus in this chapter is on making sport video summaries with content captured by amateurs in a casual setting without any constraints and content captured with assigned roles.

Mobile devices equipped with sensors such as magnetometers and accelerometers have already been used for deriving semantic information from unconstrained UGC. In [22], sport classification is done by using multimodal analysis. In [24] for salient event detection in concert videos. However the same approach was not utilized for summary creation but switching views for creating a multi-camera remix video from the concert.

We will now look at the methods which use multiple camera recorded content. In [3], multiple cameras are used to detect and track multiple players. The approach does not address the salient event detection aspect to extract salient temporal segments. The work in [15] is a similar approach to our work in terms of using a salient object for determining the value of the content. However, this approach does not cover the aspects of efficient content capture, as well as the tuning of the content summary duration. The work described in [134], uses raw camera feeds from professional cameras and uses multimodal analysis for performing automatic camera selection and view switching. The work is similar to our approach as it leverages a structured format for soccer games. This work is not targeted towards creating summaries but rather create broadcast stream. Furthermore, this work does not use role based capture or leverage a mix of professional equipment and mobile devices. The work in [98] performs multimodal analysis for detecting events in broadcast sport videos. First they extract low-level, mid-level and high-level features from the audio and visual content (mainly from color and motion information). Some of the detected high-level features are ”Audience”, ”Field”, ”Goal”, ”Close-up of player”, ”Tennis point” and ”Soccer goal”. Subsequently, summary segments are detected by discovering certain temporal patterns (co-occurrence, sequence) of high-level features.

There have been other works which use a single video, usually a broadcast content to derive semantic information for summarization, for example [104]. The method proposed in [53] identifies specific human actions, which are detected as salient events. In spite of the fact that *basketball shooting* action is one of the considered actions, the test dataset is temporally segmented. This is different from continuously captured videos as the raw content, which was used in our scenario. The method in [103] uses a combination of salient object detection and salient human action to determine a salient event. The approach in this thesis leverages the concept of using a combination of salient aspects, in our case spatial ROI and temporal segment to determine a salient event. In contrast to the prior work, our goal is to create summaries from content which may not be available via traditional broadcast feeds, and may contain content which does not conform to typical broadcast content. The paper [45] presents a multimodal approach for sport highlight recognition, focusing on American football. This approach uses a cascade approach to first determine the banner, followed by the game clock. This is a similar approach to the one we have used. Such approach focuses on the detection of salient segments only, whereas our approach also incorporates a capture framework for role based capture to combine the benefits of saliency detection and high quality summary video. Furthermore, there are many monitoring and surveillance based solutions which provide region of interest based motion detection; for example [88]. The prior art solution uses a simple motion based highlight detection method which provides significantly high number of false positives, compared to the method presented for role based capture based saliency detection. The difference in our approach with respect to such solutions, is that simple motion-based highlight detection provides significantly high number of false positives compared to a system which incorporates additional validation steps to ascertain a salient event. Background modeling methods are very beneficial for sport summarization by transforming the deployment set-up similar to a surveillance scenario (fixed static camera) [9][127].

In addition to the basic saliency detection, the thesis also explores the role of mobile devices for video production with professionals or prosumers. The work by Holz et al. [56], analyzes the use of mobile devices while watching TV with primary broadcast content. In literature, studies regarding the use of hybrid production set-ups has been limited, compared to the mobile device role as a companion device or a second screen consumption device. The work in [101] focuses on the use of a smartphone-camera based annotation system for creating rough cuts for “Adobe Premier”. The paper outlines the design, implementation, and example usage of this production and editing assistant, which is aimed at supporting small independent documentary filmmaking teams. Our approach, on the other hand, proposes a hybrid approach using professional and mobile camera which uses automatic saliency detection to obtain customized basketball summaries. In [110], a study about the use of smart phones and small mobile devices that allow audio-visual content capture on the go. The paper includes the design and evaluation of a mobile video capture suite that allows for cooperative ad hoc production. Our work proposes a hybrid approach which aims to selectively use the salient aspect of mobile devices for reducing the drawbacks of professional equipment (cost, physical footprint, among others) and manual workflow (with help of automation). The work in [31] proposes the use of robotic arms and automation for camera switching for improved adaptation to changes in the scene, but does not address the aspect of creating summaries for customized needs. Our approach, in contrast, uses mobile devices and automation for reducing the human effort as well as overall cost of production, and encompasses the full chain from capture to creation of customized summaries. In [66], Kopponen et al. present the use of mobile devices in the field of professional news content production. The key challenges regarding insufficient integration with the existing editorial systems and poor captured content quality. Our approach of utilizing abstracted metadata (e.g., timestamp metadata) reduces the challenges with interworking between professional equipment and mobile devices. The prior work underlines the value of our proposed approach, which leverages the best aspects of professional cameras (content quality) and mobile devices (lower footprint).

5.2 Saliency detection from unconstrained UGC

In this section we will present a method for basketball salient event detection from unconstrained mobile videos. This section presents results from publication [P6]. Unconstrained mobile videos in this context means videos captured using handheld mobile devices and recorded by amateur users, as they would capture without any specific roles or instructions. Section 2.2.2 provides further details about properties of such casually captured UGC. The basketball salient event predefined for detection is a scoring attempt. Based on the DSK, a typical situation (or morphology) for a scoring attempt consists of presence of the basketball ball in a close proximity of the basket. In publication [P6], the key static reference position marker, such as the basket, is referred to as "anchor-object". The "anchor-object" provides a static reference position to determine saliency direction. Consequently, if the user is assumed to be stationary for the duration of a video recording, the relative position of the basket also remains unchanged. In this method, sensor data consisting of magnetic compass (or magnetometer) data is used in combination with the video data. The magnetometer provides horizontal orientation of the mobile device with respect to the magnetic North. The magnetometer data is captured at ten samples per second in parallel with the audio-visual content, using a custom built mobile device application (a variant of the SE-AVRS client discussed in section 3.2.1).

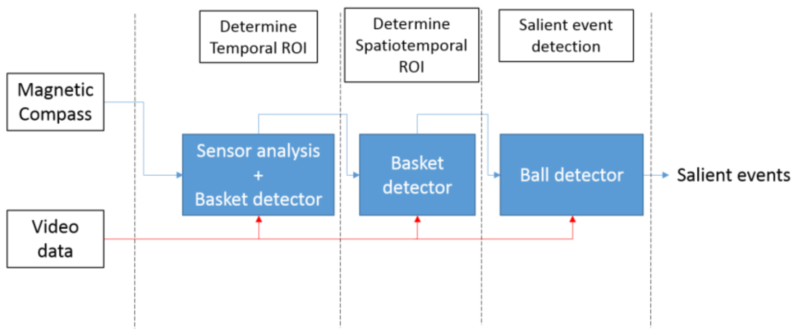


Figure 12. Salient event detection approach for unconstrained UGC.

A simplified view of the framework can be seen in Figure 12. A more detailed view can be seen in Figure 1 from publication [P6], which gives an overview of the proposed framework for salient event detection. The analysis is performed using the magnetometer data and the video data, separately for each video. A salient event is detected with a two-step approach. *The first step* consists of identifying the presence of basket in the frames (temporal aspect) and their position in each of the frames (spatial aspect) in the video. *In the second step*, a salient event is determined when a ball is detected in a predefined bounding box around the spatiotemporal ROIs generated in the first step.

Figure 13 illustrates the process for salient event detection using the content analysis approach and the multimodal analysis approach.

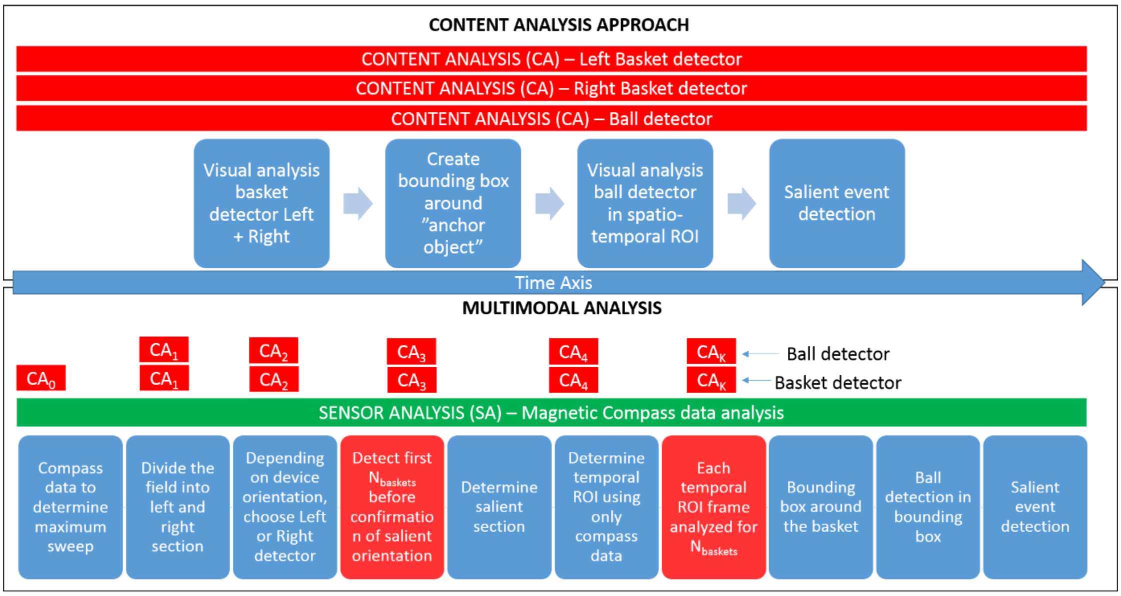


Figure 13. Salient event detection with content-only versus multimodal analysis approach.

5.2.1 Determine spatiotemporal ROI

The first step consists of analyzing magnetic compass (magnetometer) data corresponding to each video, to determine the angular sweep (boundaries of horizontal orientation α_{Right} and α_{Left}). The left and right angular sections correspond to horizontal orientation range intervals $[\alpha_{left}, \alpha_{center})$ and $(\alpha_{center}, \alpha_{right}]$ respectively. This information enables selection of the appropriate visual detector for left or the right basket to determine the horizontal orientation of the anchor-point, which is basket in this case. This makes the detection process more efficient and reduces the risk of false positives due to the use of the incorrect basket visual detector. In order to minimize the chances of false positive detection of the basket, a predefined threshold for consecutive detections of $N_{baskets}$ within a spatial region is used. This corresponds to the red block CA_0 in Figure 13. The basket detectors used in this work are based on cascade classifiers analyzing Local Binary Pattern (LBP) features [100]. The classifiers were trained by using about 2000 training images from basketball matches other than the test match. The magnetic compass orientation for basket detection is the left and right salient angle, corresponding to basket positions.

The magnetic compass orientations which are different by less than a predefined threshold with respect to the left and right salient angles represent temporal segments of interest. The temporal segments obtained by analyzing magnetometer data is classified into left or right section. This information is used to analyze the temporal segments of interest with the correct basket detector (left or right visual detector), to provide spatiotemporal ROIs. The red blocks CA_1, CA_2, etc., correspond to the temporal segments determined with magnetometer data and subsequently analyzed with visual detectors. A similar criteria for $N_{baskets}$ is used for robustness of spatiotemporal ROI detection. In Figure 13, the sensor analysis is represented in green and content analysis in red. The multimodal approach employs content analysis selectively, thereby saving computing resources.

5.2.2 Detect salient event

The spatiotemporal ROIs, once determined, provide the anchor-region for defining the criteria for salient event occurrence. The criteria is the detection of a ball in the spatial ROI, which is identified as a rectangular region surrounding the basket and whose width and height are proportional to the basket size. Using the DSK, the ROI is prolonged towards the right side for the left basket and towards the left side for the right basket – see Figure 3 in publication [P6]. If the ball is detected successfully for at least a predefined threshold number of N_{balls} consecutive frames, the corresponding frames are classified as salient event frames. For detection of the ball, a ball detector, similar to the basket detector, was built by extracting LBP features from about 2000 training images and by using cascade classifiers for training the model. In Figure 13, the CA_i corresponds to the i^{th} temporal ROI or segment of interest where each red box corresponds to content analysis duty cycle, irrespective of whether the resulting spatiotemporal ROI segments includes each. Some temporal ROIs will be dropped if the refinement step does not detect a basket successfully with content analysis.

5.2.3 Results

The above described method was evaluated by comparing the content-only based approach and the multimodal approach. The evaluation content consisted of 104 minutes of videos, the average length of videos was 5.8 minutes, with a minimum length of 11 seconds and a maximum length of about 15 minutes. The experiments were performed on a machine equipped with 92 GB of RAM and an 8-core 2.53 GHz processor; no parallelization was used for obtaining the analysis times. In Table 3, detection of temporal ROIs is presented (P stands for precision, R for recall and F for balanced F-measure). The spatial refinement row, refers to the use of basket detection for spatiotemporal ROIs.

TABLE 3. Comparison of temporal ROI detections.

METHOD	P	R	F	Analysis time (s)
Sensor-based (without spatial refinement)	0.72	0.96	0.82	0.06
Sensor-based (with spatial refinement)	0.82	0.78	0.80	0.77
Content-only	0.56	0.76	0.64	2.37

We can see from the above results that sensor based method outperforms the content-only based approach. The sensor-based method is about 21 times faster and also more accurate than the content-only based approach, which demonstrates the efficiency gains by using sensor data. In addition, the sensor-based method with spatial refinement shows improvement in avoiding false positives but as an undesired side-effect the number of false negatives has also increased. This suggests that even though the mobile device was oriented towards the salient direction, it may not have the basket in its field of view or the visual detector may have failed to detect the basket.

TABLE 4. Comparison of salient event detection.

METHOD	P	R	F
Sensor assisted	0.40	0.39	0.40
Content only	0.30	0.37	0.30

Table 4 shows the experimental results for salient event detection with or without the spatial ROI determination. The sensor-assisted saliency detection performs better than the content-only based approach, primarily due to better temporal ROI detection performance. But overall numbers for either of the methods are not high. Improvement in the visual detectors for presence of ball and basket is required to improve the performance.

An average user recording videos casually cannot always ensure that his/her video includes the visual content necessary to detect salient events. For example, the recorded video may focus on other subjects of interest (e.g., close-up of a player). In addition, if a video contains just one basket or in the worst case, no basket, then detecting a basket scoring event will not succeed. To overcome limitations of unconstrained UGC, the next section presents an approach which incorporates some constraints on content capture. The purpose is to improve saliency detection and obtain high quality video summaries.

5.3 Saliency detection from role based capture

In a shift from the previous section, we will discuss an approach for salient event detection from role based captured content. Furthermore, we explore a new production technique which leverages the synergies between mobile devices professional equipment. The combination can be much more versatile than either mobile device based capture or professional camera capture individually. The work in this section is derived from publication [P7]. The proposed novel capture setup and workflow has three-pronged goals. The first goal is to have a robust salient event detection system. The second goal is to enable creation of high quality multi-camera sport highlights. The third goal is to combine the best aspects of professional equipment (high quality content capture and high zoom-in capability for close-up shots) and mobile devices (lower cost and unobtrusive form factor). This section is organized as follows: first a role-based capture setup is presented; subsequently, a saliency detection method is presented. We conclude this section by introducing a tunable summary creation approach.

5.3.1 Role based recording setup and workflow

The motivation behind the role based capture is presented in the following. Optimal camera position for content viewing is not always the same as optimal camera position for content understanding. Certain camera positions and camera view settings (wide angle shot, mid-shot, close-up-shot) are more suited to allow semantic content analysis. On the other hand, other camera positions and camera view settings are more suited to provide a high quality viewing experience. For example, while a close-up shot, following the player may have high subjective viewing quality, such content may not be suitable to detect a successful basket score, since the basket may not be in the field of view.

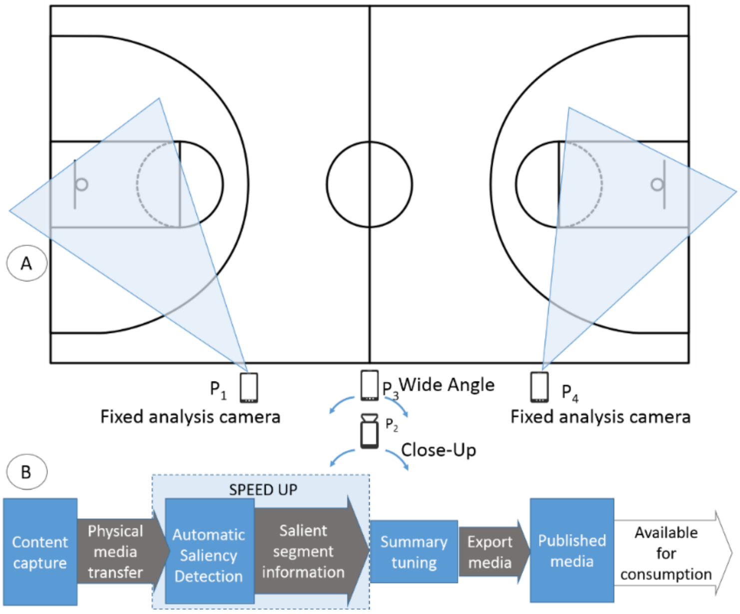


Figure 14. Role based capture set-up and workflow.

The proposed set-up consists of two sets of cameras for content capture, referred to as "fixed analysis camera" and "view cameras" (Figure 14). The cameras labelled "analysis" are situated such that their captured content is optimal for semantic analysis. For example, their field of view overlapping with the intended region of interest (e.g. the baskets in case of basketball, the goal-post in case of football or soccer, etc.). The cameras labelled "view" are situated in such a way that they cover the event from an optimal position for aesthetically pleasing content. The analysis cameras are used to analyze salient events and subsequently extract the relevant content segments from the view cameras. Due to the assignment of roles, this method is referred to as role based recording setup. The alignment between analysis content and the view content can be done with audio based time alignment (which was used in our system) or any suitable method. The specific method of time alignment is not in scope of the thesis.

In Figure 14A, P_1 and P_4 are fixed *analysis cameras* (mounted on a tripod), these need to have sufficient field of view and resolution but need not have a high zoom-in capability. P_2 and P_3 are operated by camera operators (mounted on a swiveling mount) to ensure the right objects and views are always tracked during the game. P_2 needs to have a high zoom capability to ensure professional grade close-up shots. P_3 needs to have a large field of view to give a proper wide angle shot. Consequently, P_1, P_3 and P_4 were chosen to be high-end mobile devices; P_2 was chosen to be a professional camera. The proposed setup requires two persons (with only one professional camera operator) to operate. This is in contrast with conventional setup which consists of four professional cameras operated by four professionals (see Figure 1 in publication [P7]). This reduces costs of equipment as well as personnel needed.

As can be seen from Figure 14B, the workflow consists of role based recording, automatic saliency detection and summary tuning. The details of automatic saliency detection method, the results and the tunable summary creation method will be presented in section 5.3.2, 5.3.3 and 5.3.4, respectively.

5.3.2 Saliency detection for basketball

The approach is outlined in Figure 15. In the *first part*, the *spatial ROI* is determined. In the *second part*, the *temporal ROIs* are determined by detecting the ball in the proximate region surrounding the spatial ROI. The *third part* consists of obtaining a *salient events* from a set of detected salient frames.

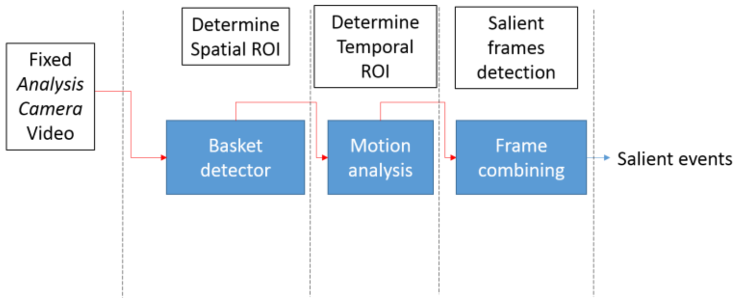


Figure 15. Salient event detection approach for role based capture

Part 1: Spatial ROI detection

Due to the use of a *fixed analysis camera*, it is sufficient to obtain the spatial ROI only once. Since we are considering basketball, the anchor-object is the basket. Spatial ROI detection is done using the visual detector for basket that was used in section 5.2.1. In order to improve the robustness of the spatial ROI determination, a predefined threshold number $N_{baskets}$ is used to confirm the basket detection.

Part 2: Temporal ROI determination

Detection of the ball in the proximity or within the desired region of interest, determines that whether a particular frame is salient or otherwise. Thus ball detection determines the temporal aspect of the spatiotemporal salient even detection. Ball detection was seen to be underperforming for detecting salient events with the unconstrained mobile videos as source content in 5.2. Consequently, sensitive methods which do not result in excessive false positives were explored. A motion based ball detection approach was chosen for detecting temporal ROIs. This method consists of the following steps:

- Calculate frame difference between current and previous frames. Threshold frame difference to get motion contours.
- Apply noise reduction techniques to filter out noise and enhance motion contours.
- Background modeling to reduce false positives. This is done using an adaptive Gaussian mixture modelling technique [127].
- Analyze the shape of the motion contour to determine saliency. The shape verification is implemented using a polygon estimation method as per the Douglas-Pecker algorithm [34], to further reduce the false positives.

Please refer to sub-section Automatic Saliency Detection of publication [P7] for more details. The motion based ball detection is shown in Figure 17.

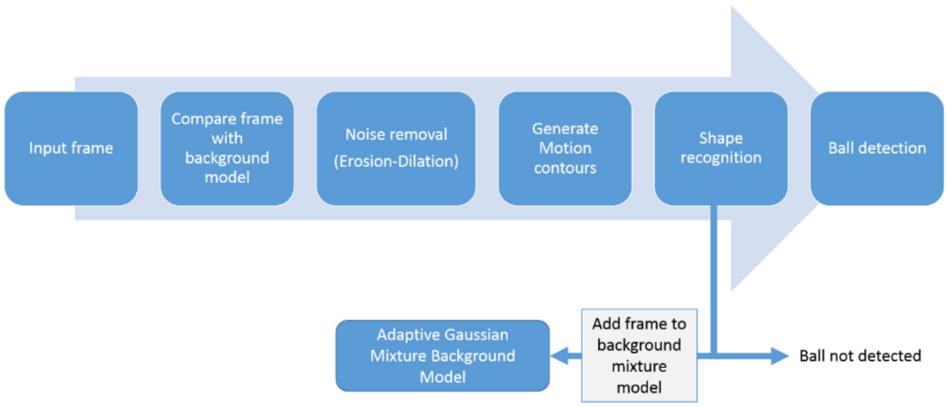


Figure 16. Ball detection process overview.

Part 3: Salient events detection

In this step, salient frames are first identified by the detection of the ball in the spatial ROI. Detection of ball in the proximity of the spatial ROIs for at least two seconds represents a salient event. This is the heuristic hypothesis for a salient event. In addition, the non-causal use of detection information reduced the false positives.

5.3.3 Results

This sections presents the results from a test event captured with a role based recording set-up described in 5.3.1. The saliency detection is performed for video recorded by *fixed analysis camera* P2 in Figure 14. The ground truth consisted of 45 salient events annotated manually in a video of 40 minutes duration. Saliency detection with frame difference followed by noise removal and thresholding resulted in 100% recall rate, although with a significant number of false detection (precision 74%). With the use of background modeling and shape recognition, the precision increased 32% to 97.8%. This suggests strong promise, which *needs to be verified with a larger data set* (see Table 5).

TABLE 5. Salient event detection performance

Method	Precision	Recall	F measure
LBP-based classifiers	0.80	0.17	0.28
Frame difference based motion contours	0.74	1.00	0.85
Background subtraction	0.82	1.00	0.90
Background subtraction + Shape identification	0.98	1.00	0.99
Successful basket detection	0.95	0.84	0.89

5.3.4 Tunable summary creation

Tunable summaries are required to provide users, the control to obtain a right-sized summary, which is optimized by taking into account the end use. For example, different length summaries are needed for showing short clips within a news program versus highlights of the whole game. The summary tuning control is available at two levels.

Prioritized salient event selection

The *first level* controls the number of salient events included for making a summary of a specified duration. This requires selection of one or more salient events from a set S, where $S = \{S_1, S_2, S_3, \ldots\ldots, S_N\}$. The key requirement at this level of tuning is to include the salient event which adds the maximum subjective value to the viewer of the summary. For example, inclusion of successful basket attempts is likely to be more important than an unsuccessful attempt but on the other hand, a false salient event would degrade the viewing experience. Consequently, salient events are ranked with a combination of whether the scoring attempt is successful and the salient events' confidence value. A successful basket detection is ranked above an unsuccessful score attempt (even if the former has a lower algorithmic confidence value). Successful basket detection is done by detecting motion in the "inner ROI", which is the lower middle block formed by dividing the spatial ROI into nine blocks (see Figure 7 from publication [P7]). The successful scoring event classification resulted in 25 instances, out of which 18 were true positives, one false positive and 6 false negatives as a result, achieving 84.21% recall and precision of 94.73%. Further details about successful basket detection can be seen from publication [P7].

Salient event adjustment

The *second level* of control is at the level of tuning the duration of each segment of a single salient event's multi-camera presentation, consisting of three sections[$\{Main\ Section\}, \{Replay\}, \{Additional\ View\}$]. The summary employs the multiple camera angles by using the cinematic rules described in the section *Tunable Summary Production* in publication [P7]. Figure 18, gives a brief overview tunable summary system.

Overall, the tunable summary system consists of three aspects. Firstly, as discussed above is the salient event ranking. Secondly, the use of cinematic rules to present a salient event in a manner, that is both aiding user understanding as well as aesthetically pleasant for viewing. Thirdly, leverage the low footprint method of using metadata based playback control to facilitate instant preview by changing parameters for the two levels described above. This method employs the low footprint approach of metadata based rendering discussed in section 3.4

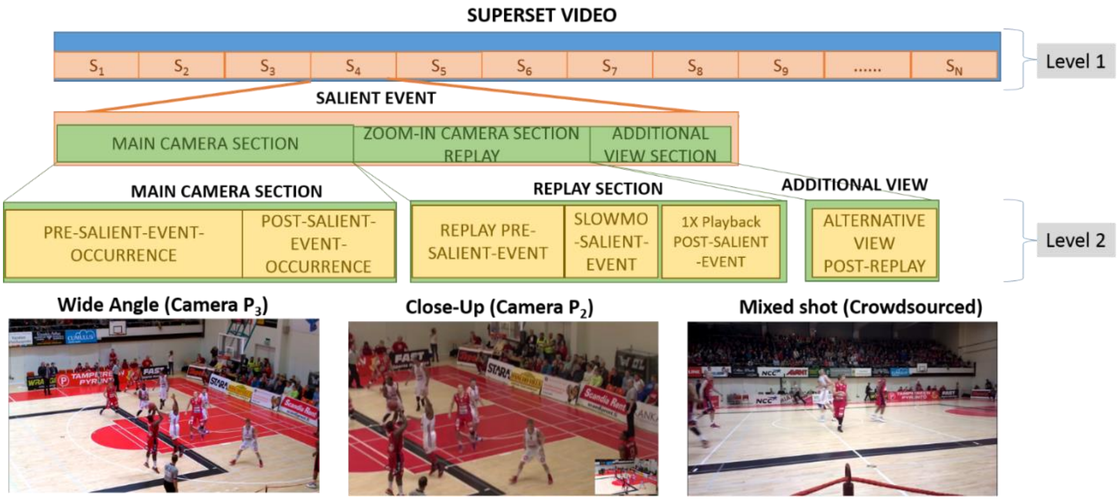


Figure 17. Tunable summary overview.

5.4 Implications of unconstrained and role based capture

The *unconstrained mobile video capture* technique is suitable for *amateur* end users who casually record videos in different types of events. The act of recording usually distracts the user from enjoying the event [P5]. Thus, minimal effort for performing the recording is a key requirement. However, the saliency detection for unconstrained UGC has clear challenges imposed by the field of view constraints and unintentional movements in the mobile device.

The *role based capture* technique is suitable for *professional and prosumer* category of users. The proposed set-up provides a lower cost (compared to conventional professional set-up) alternative which combines the elements of simplicity (e.g., automatic saliency detection) and professional quality (e.g., cinematic rules, close-up shots from long range) to deliver a high quality summary. The *crowdsourced content* can be leveraged with the automatic saliency detection framework to provide much needed variation in the views used in the summary and at the same time benefit the recorder by receiving the salient event indexes. The ability to tune the summary allows a *user to control what she wants and only as much as she wants*.

In the next chapter, we will shift focus from collaborative content creation to collaborative content consumption.

6 Mobile based collaborative watching

In the earlier chapters we have discussed the use of content derived semantics and the recording users' situational context (e.g., camera motion, event information, etc.) for collaborative creation of video remixes and summaries. Now, we will discuss the use of situational context for collaborative consumption of TV or video content. An example of such a collaborative watching method was proposed as the Mobile and Interactive Social Television (MIST) [P8][P9]. This concept envisaged a mobile based *virtually collocated content consumption* experience, between people who may in reality be present in different locations.

In this chapter, we will first present the requirements and the method for a virtual co-watching experience. Thereafter, novel architectural approaches are presented for realizing such a system as well as initial findings about the user experience aspects for such type of systems. This is followed by a discussion regarding the seamless transfer of multimedia consumption between different devices, covering [P10]. We will conclude this chapter by discussing the state of the art in this topic.

6.1 Role of context and content in collaborative watching

Traditionally video consumption has been dominated by broadcast content watched on TV. The widespread availability of mobile devices equipped with high quality video playback capability and high bandwidth network connectivity, mobile based content consumption has become commonplace. However, content consumption experience continues to be a substantially solo activity.

It has often been observed since the birth of TV that people prefer to watch a game or a movie with other people, due to the social experience that it offers [67]. The content often becomes a medium of interaction between people and can sometimes make interaction between users to be more important than the content itself [117]. Similar motivations have been realized in a static context in [1][8][18][46]. An example of collaborative watching with mobile devices is implemented for broadcast video delivery. Although, DVB-H is not widespread any more, the system provides useful insights into the value of audio based interaction for collaborative watching [116][117][118]. The work in [128], presents community streaming with interactive visual overlays, such that a dedicated space is left for interaction content and the other space is left for the consumed content. On the other hand, our method is amenable for overlays to move dynamically so that important objects of interest in collaborative consumption are not occluded by visual overlays.

The motivation for the MIST system was to provide a *watching together* experience, as though the users are collocated in the same location. The “*watching together*” aspect is facilitated by the virtual presence between the participating users. The creation of virtual presence is achieved by capturing and sharing the users’ situation context with the other users. The level of virtual presence is influenced by the richness of user’s situational context. The virtual presence can be shared in the form of facial expressions, sounds, text message, which are referred to as interaction content in publication [P8]. The interaction content may consist of simple text based interaction for sharing views and reactions. A higher degree of virtual presence can be obtained via sharing real time audio based interactions between the collaborating users. A further enhancement of the virtual presence involves audio-visual interaction between the users. In addition, the *watching together* experience is contributed by the commonality in the content being watched by the participating users. *The participating users’ situational context and the watched content as a mediation channel form the basis of context and content mediated collaborative watching*. Figure 1 in publication [P8] gives an illustrative overview of the concept.

In the next section, two architectures are presented for realizing the mobile based collaborative watching system.

6.2 Collaborative watching architectural approaches

As discussed above, sharing of situational context facilitates virtual presence. Consequently, the degree in richness of *virtual presence* influences the cohesion among participants in the collaborative watching session. In our case, the virtual presence is achieved with real-time audio-visual interaction. A common shared context between the participating users is created with the help of the consumed content and the sharing of their situational context. Hence, this is also referred to as *context and content mediated consumption*. The following two media delivery requirements are important for successfully creating a common shared context:

- The interaction responses consisting of the users’ comments (both text and audio) and visual feedback (e.g., facial expressions, gestures) are viewed in synch with the consumed content. Hence, the delivery of such interactions should be with low latency, to maintain their contextual meaning.

- The content consumption should be in synch for all the participating users, in order to maintain a common baseline.

Realization of such a system on a resource constrained mobile device presents many challenges. Physical constraints like display size, computational resource availability, battery and network connectivity need to be considered for defining a suitable architecture (for more details see section 2.4). Equally important are the user experience related requirements from architectural perspective. The proposed architectural approaches are the centralized mixing or a thin client approach (see section 2.4.2) and the end-point mixing or a thick client approach (see section 2.4.1), proposed in publication [P8].

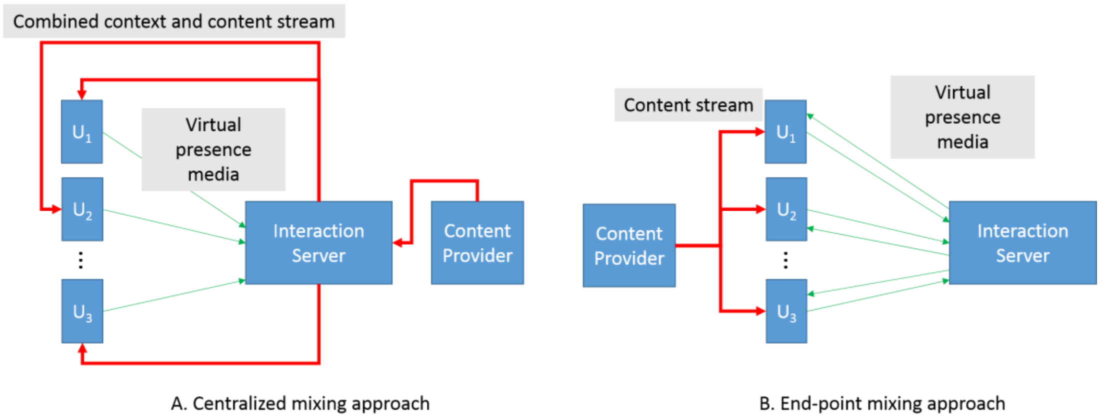


Figure 18. Overview of the centralized (A), end-point mixing (B) approaches.

6.2.1 Centralized mixing architecture

This architectural approach is designed to minimize the computational and other resource requirements for the mobile device participating in the collaborative watching session. The architecture has three main entities, the Content Provider (CP), the Interaction Server (IS) and the mobile clients (hosted by the user's mobile device). Figure 18A gives an overview of such an architecture. The Content Provider delivers the content to be watched collaboratively. The users' situational context is captured by their respective mobile devices and transmitted as *virtual presence* media (as audio, video and text modality) to the Interaction server. The Interaction server mixes the content from the CP with the virtual presence media to generate a combined audio-visual stream as the output. This stream is subsequently delivered to all the participants, comparable with conversational applications like video telephony.

The main advantage of this scheme is the need to decode and playback only one combined stream. The centralized mixing scheme resembles a start topology, with the IS forming the hub while the mobile clients and the CP forming the spokes. The advantage of this topology is that the combined stream can be adapted for each mobile client's video playback capability (e.g., in terms of resolution) as well as network specific bandwidth adaptation to maintain the desired latency.

6.2.2 End-point mixing architecture

In this approach, the *virtual presence* media received from IS and the content received from the CP is mixed in the users' mobile device. The virtual presence media is transmitted from the users' mobile devices to the Interaction Server and received back as a mixed multi-party virtual presence stream (audio, video and text interaction). At the same time, the content to be watched collaboratively is received directly from the CP by the mobile device (see Figure 18B). Consequently, the end-point or the mobile device receives two streams which are mixed and rendered locally.

The advantage of this scheme is that it decouples the Interaction Server (IS) from the Content Provider (CP), which can provide higher degree of flexibility and choice for individual or group of users in a collaborative watching scenario. Furthermore, the localized mixing of the virtual presence stream and the content provider stream allows for individualized flexibility in arranging the rendering layout. In the end point mixing scheme, the mobile device is required to decode one additional stream compared to the centralized mixing approach. The primary challenge in this scheme is to maintain playback synchronization between the different mobile devices for content stream playback and the multi-party virtual presence stream received from the IS. The work in [19] presents schemes for inter-client synchronization.

Running two sets of decoders on the mobile device resulted in rapid draining of the battery. Implementation and performance details can be seen from section 4.6 and 4.7 in publication [26].

6.3 Proof-of-concept system

The proof of concept system is presented in section 3 of publication [P8]. This centralized mixing approach is described in this section. This is also the system used as a prototype system to study the user experience aspects in the subsequent section. The implementation approach is a fusion two signaling paradigms. First is the typical SIP [108] and SDP [55] based multiparty conferencing session negotiation and RTP [120] based media transport. Second is the HTTP [43] based session control module for implementing the shared playback control as well as content selection from an EPG (see Figure 20 for protocol stack).

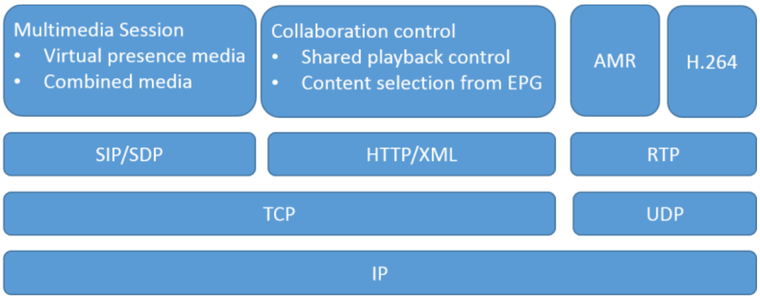


Figure 19. Protocol stack overview of POC system.

The common shared context, also referred to as *virtual shared space* (VSS) in publication [P8], is the facilitator for the *watching together* or the collaborative watching experience. Figure 4 in publication [P8] shows the sequence of initiating the collaborative watching session and subsequent interactions. In such a collaborative session, the participating users can talk, see and message the other participating users. The initiation of such a session involves inviting one or more people of interest using a SIP URI (which can be retrieved from the initiator device's phone book). The participants can join in by accepting the invitation. Joining in at a later point in time is also possible by starting the client entering any of the on-going collaborative watching sessions.

The session starts like a conventional multiparty video conference, where the users can talk and discuss before selecting the content to be watched. On selecting the content to be watched, all the users receive the content such that it is synchronized between the participants. The users can speak with the other participants or make gestures by popping-in with their video on the screen. There is a need to optimize the precious screen real estate and avoid obstructing the users' view of the watched content. Consequently, the participant video rendering is voice activated to grow in size. In absence of voice activity, the participant video thumbnail is kept small to provide a sense of presence without occupying excessive space on the screen (Figure 5 in publication [P8]). There is a shared control of content playback between the participants. This ensures that content selection as well as playback control interactions (SELECT, PLAY, PAUSE, STOP) are applied to the common shared context. This is an important aspect to maintain a cohesive experience for all the participants in the collaborative watching session.

The proof of concept system was tested on WLAN as well as cellular networks. For the WLAN bearer, the general feedback was positive and the response time experienced by the participants for response to interactions (like playback control, participant video activation, etc.) was observed to be about half second. For the 3.5G bearer, for a test setup with mobile clients in different locations (one mobile device was in Bristol, UK and the other was in Tampere, Finland), the response time for user interactions was less than one second.

6.4 User experience requirements

In this section we summarize the findings about the user experience impact of mobile based collaborative watching from publication [P9]. The feeling of social presence of collaborating participants was found to add value by all the study subjects. The key factors for influencing the level of virtual presence were relationship between the participants and the type of content being watched together.

The desired level of virtual presence affected the choice of the interaction modality that was employed by the participant. Users considered audio interaction based virtual presence engaging but distracting for some types of content. Interestingly, some users expressed preference for using asymmetric interaction modalities (for example, using audio as input interaction but receive the other participants' audio as text). The audio-visual interaction was considered to provide a higher degree of virtual presence compared to audio only and text. Consequently, its use was considered to be more sensitive and context dependent. The work in [112] validates the descending virtual presence for audio-visual, audio-only and text based interaction. Significantly, the [P9] study suggests an implicit feeling of etiquette which gets transferred from face to face collaboration to the virtual co-watching space.

The type of content that was preferred by the users was influenced by two factors, mobility and collaborative watching. Long format content was less preferred compared to short duration content. User generated content (home and family videos, short clips, funny clips, etc.), sports content, short TV episodes (TV shows, celebrities, etc.) and news content were considered most suitable type of content for watching collaboratively on a mobile.

Thus, in summary, although users desire rich interaction capabilities, they do not want all of it enabled all the time. Privacy needs, inter-personal relationship between the participants and the type of content has a direct impact on the acceptability of the system. The key requirement was found to be the easy personalization and customization of collaborative watching session based on user preferences.

6.5 Movable multimedia sessions

The collaborative watching has been discussed in the mobile device context. The earlier collaborative watching systems were primarily static scenarios, with TV being the primary video consumption device. Users could be interested in the possibility to shift from a mobile device to a TV or vice versa during a collaborative watching session. Considering the same possibility at a more general level, the ability to transfer any on-going multimedia session from one device to another without the need to restart the session provides many advantages. This aspect is analyzed in publication [P10] and forms the basis of the discussion in this section. In spite of the availability of multiple Internet enabled multimedia devices, the user often ends up either continuing the particular multimedia session from the original device or restarts the session from a suitable device.

6.5.1 Related work

A SIP based third party call control in [107] presents best practices for controlling media flow between two devices. The SIP based Session Mobility describes the signaling and media flow examples for transferring a communication session from one device to another [119][121]. A seamless application layer handoff for media delivery across different devices is presented in [30], with a middleware focused approach. An example of session state transfer can now be observed in consumer web services such as YouTube [52], although not in a real-time handover context. In this service, if a user is logged into the service, moving from consumption on an Internet TV to a tablet device, already indicates the video which was being viewed earlier (and also saves the playback position). This system is still not connected with a device discovery and handoff initiation mechanism. There have been recent developments in fusion of web browsers and SIP protocol support, which enables session mobility between browsers [2].

6.5.2 Session mobility

The traditional physical mobility and the service mobility is that while former keeps the service uninterrupted even as the consumption device moves. On the other hand, the service mobility continues the service experience, even if it is consumed from a different device. In the context of multimedia sessions, the mobility of multimedia sessions envisages service continuity despite of changing device through which the user consumes media. This requires seamless transfer of multimedia session from one device to another. Transfer of multimedia sessions (or session mobility in [P10]) can either be complete or partial. Either type of the session transfers can happen from one or more originating device to one or more target device.

- In case of a complete session transfer, the originating device will transfer all the individual media sessions to the target device.

- In case of a partial session transfer, the originating device will transfer only a part of the media session to the target device.

The main motivations for enabling movable multimedia sessions are physical mobility, optimal content consumption experience and lower costs. The first advantage is visible when transferring a multimedia session (e.g., a video call) from a desktop to a mobile device, when the user needs to leave the location. The second advantage allows a user to transfer the content consumption from her moble device to a high speed broadband connected Internet TV. In this case, the improvement in the viewing experience may be helped by using a better display as well as an improved bitrate for the content. The cost option is applicable while leveraging the optimal bearer (for e.g., using home WLAN instead of a cellular network connectoin).

6.5.3 Session mobility solution

Session mobility aims to achieve a seamless application layer handoff from the originating device to the target device. An application session can be abstracted into its *context* and *state information*. For a video receiver and playback application, the context and state consists of the video codec, the last rendered frame number, the receiver buffer state. Handoff of the multimedia session at application layer provides access to application context and state information [119]. The context and state information can be used by the target device to prepare it for receiving the media and consequently minimize the discontinuity. Discontinuity interval is a critical measure for the perceived effectiveness of the mechanism.

Session mobility mechanism is deeply influenced by the characteristics of the multimedia application. For a streaming application such as Video on Demand (VOD), the challenge is to minimize the initial buffering delay for the target device before rendering on the one hand and to synchronize the device switch (when transferring the media from one device to another). For a conversational application such as video telephony, on the other hand, has low latency requirements that require very small buffering at the receiver (often just to handle jitters caused by the underlying network or the nature of the media and audio-video synchronization). Furthermore, media specific requirements also influence session mobility mechanism. For example, a transfer of the H.265 video streams necessitates the H.265 sender to re-initiate the media stream from an IDR (Instantaneous Decoding Refresh) to facilitate decoding of the video stream by the target device.

In the following sub-section, we present the proposed architectures in publication [P10].

6.5.4 Session mobility architecture

We propose the architecture options for enabling session mobility and examine the benefits and drawbacks of the same. An important characteristic to evaluate the different options, is whether the entity involved in the session transfer is “session mobility aware”. A “session mobility aware” entity is expected to be able to distinguish between a new session being started and an on-going session being transferred from an originating device to the target device. The architectures could be device centric or network centric or hybrid.

A device centric approach requires minimal support from the network infrastructure, but depends on the incorporation of session mobility support in the devices involved in session transfer. The network centric approach, on the other hand, relies on the network based services for enabling the session transfer as well as choosing the optimal target device. In contrast to the device and network centric approaches, the hybrid approach attempts a compromise for situating the session mobility facilitation mechanisms. The right approach depends on the specific use case, the operating environment (whether SIP or HTTP or RTSP is used for session setup), the device capabilities and services available in the network infrastructure. The session mobility mechanism can be broadly divided into three steps.

Device and Service Discovery

This is a prelude to initiating the actual session transfer. For example, in (Universal Plug and Play) UPnP [97] based service advertisement and discovery mechanisms can be utilized to discover the target device and its capabilities. Another example of service discovery is (Service Location Protocol) SLP [54]. This step also forms an important part of the security mechanism during a session transfer. Security mechanisms are essential to identify if the participating user and device can be trusted. This is an important step before being authorized to proceed with the session transfer. Service advertisement and discovery mechanisms should include media capabilities advertisement and discovery as well. Device and media information are needed for capability negotiation when a session is transferred between devices.

Session state capture and representation

The session state capture of a multimedia session includes parameters like the media parameters like codec related information; the network parameters like IP address, bandwidth and transport protocol information; and application level parameters like buffer status and stream grouping for synchronization. This information can be represented using (Session Description Protocol) SDP [55] or a suitable (Extensible Markup Language) XML [133] format.

Session state transfer and capability exchange

After capturing the session state and representing it in a suitable format, the final step involves setting up the new session. This requires transferring the session state information to prepare the target device for continuing the session. The session transfer can be a hard hand-off or a soft hand-off, which is in principle, similar to the conventional handoff. In addition, the session transfer may involve session negotiation via capability exchange, if the goal is to optimize the session parameters.

6.6 Comparison with state of the art

There have been significant increases in the computational resource availability, display size as well as resolution and network bandwidth in the eight years since the proof of concept system was implemented. For example, if we compare Nokia N95 [90] and Samsung Galaxy S7 [113], the two devices which could be considered state of the art in their respective periods (see Table 6). The multimedia creation and consumption capability has increased significantly. Furthermore, it is accompanied by the upgrade in the network bandwidth availability (from the earlier HSPA to the current LTE). In spite of the increase in the hardware, software and network capability, the resource constraint together with user experience challenges continue to dominate collaborative watching experience. This is partly due to the increase in the users' expectations with respect to the media quality, which continues to consume significant network and computational resources. Easy adaptation of the rich interaction capabilities with the need to match the users' instantaneous contextual needs, continues to be a challenge.

TABLE 6. Specification comparison between two mobile devices.

Property	Nokia 95	Samsung Galaxy S7
Release year	2007	2016
Video playback resolution	640 X 480	3840 x 2160
Video recording resolution	640 X 480	3840 x 2160
RAM capacity	64MB	4000MB
Cellular connectivity	3G, HSDPA	4G, LTE

Collaborative watching in VR environment [91] has further expanded the envelope for providing a rich virtual presence to the collaboratively watching users. The VR platform from Oculus, leverages audio based interaction in combination with immersive omnidirectional content consumption to create rich virtual presence. Social interactions with VR is in its early days but it follows many of the key features present in the prototype system. For example, there is an initial staging area where the participants can interact with each other and discuss about the content to be watched. The integration with (Social Networking Services) SNSs like Facebook [42] indicate the possibility of leveraging different content servers.

There have been many recent advances which support various methods for leveraging of heterogeneous devices and networks. One such example of session mobility can now be observed in consumer web services such as YouTube [52]. In this service, if a user is logged into the service, moving from consumption on an Internet TV to a tablet device, already indicates the video which was in progress earlier (and also saves the playback position). This system is still not connected with a device discovery and handoff initiation mechanism. There have been recent developments in fusion of web browsers and SIP protocol support, which enables session mobility between browsers [2]. Google cast [49] provides the possibility to bridge the content consumption gap between a mobile device and a TV. This allows users to combine the benefits of consuming content with a large and high quality display afforded by a TV and other high quality audio speakers in the vicinity. In one mode of operation, the mobile device controls the Chromecast device to directly fetch content from Internet content services (thus relieving the mobile device from the media path). In another mode, the mobile device can directly transmit content to be consumed to the Chromecast or Google cast device. There are content streamers in the market from other companies, such as Roku [106], Amazon [4], and others. The streamers fulfill part of the session mobility use cases (of leveraging optimal hardware) in a localized scenario. However, consumer products for automatic seamless session transfer of in-progress video calls or video streaming sessions is not available. This suggests there are still challenges related to device discovery, security, NAT/Firewall issues and handoff orchestration, for ubiquitous session mobility.

7 Conclusions

Automatic co-creation of content from mobile videos and mobile based collaborative watching have many challenges such as meeting key stakeholder requirements, system design and implementation, algorithmic, among others. Some of these challenges have been analyzed and techniques presented to address them.

Firstly, thesis explores the novel aspect of end-to-end system design for automatic video remixing. A system for creating automatic remixes from crowdsourced sensor data enriched mobile video content is presented. Sensor enhanced source video content provides two advantages: sensor based analysis can achieve higher efficiency for semantic analysis; combining sensor and content analysis can deliver better semantic information. Consequently, a sensor enhanced automatic video remixing system can deliver higher quality remixes compared to a content only approach. The *sensor-enhanced video remixing prototype system* was designed without any specific operating parameter constraints, the goal was algorithmic verification and explore system feasibility to achieve a high overall user experience. However, the need for a proprietary client to record sensor data simultaneously with audiovisual content means that it is difficult to have a minimum critical mass of persons in an event who can contribute such source content. Also, there is absence of such sensor data aware social media services. This drives the need for adaptation of the system architecture such that it can improve the desired performance parameters while limiting the reduction in the overall user experience. The *sensor-less cloud based remixing system* removes the need to upload videos specifically for making remixes and solves the problem of *minimum critical user density*, since all users can contribute source content. On the other hand, the sensor-less approach compromises on computational efficiency as well as semantic information due to the absence of sensor augmented source content. The low footprint sensor-less AVRS system condenses the operating requirements to *“one user, one video and one device”*. The system architecture adaptation reduces the overall system complexity to an extent where any backend infrastructure is not required, enabling a single user to create a multi-camera remix experience from a single video. The presented system architecture adaptations exemplify the need for prioritizing performance parameters of interest in the system design. This is done in order to make the resulting system suitable for the chosen operating parameters with reduced compromise on other performance parameters.

The multiple studies of user experience impacts provided insights in both top down and bottom up manner. The user experience studies verify some of the top down design goals, highlight gaps and indicate which of the top down design choices have a negative impact on the user experience. Top down design choices such as use of automation to reduce complexity, crowdsourcing of source content, a continuous audio track were positively received by the end users. The emphasis on removing videos with poor illumination and switching camera angles in synch with the audio scene characteristics (e.g. music tempo, beat and downbeat) was highlighted in the first user study, which was subsequently incorporated and received positively. The need for advanced user control functionality which was not part of the initial system design, is an example of a bottom up user requirement. A linkage is observed between the user's preferences for the used switching regime and subjective visual quality assessment of the multi-camera remix from a single video in the low footprint remixing approach. This suggests a need for user control on modifying switching instance in addition to the view selection. A summary is presented in section 4.4 of the system design implications extracted from the user experience studies. The user studies were involving the sensor-enhanced video remixing methodology and the low footprint remixing approach.

The need for the system architecture adaptations described in the first chapter have been informed by the challenges and bottlenecks experienced by the users in the trials as well as the need to reduce the time to wait for the first video remix. For example, uploading large source video files involves waiting (due to the uplink speeds) which is further accentuated if this effort serves only one purpose (of creating a remix) and requires another upload to SMPs for social sharing. On the other hand, *instant gratification*, is appreciated by the users, as seen in interactive customization with low footprint remixing method. The possibility for instant preview after making the changes was positively received and considered to be very important by the users in the study.

After analyzing the system design aspects and the user experience impact, we next presented techniques for sport content summarization. The objective was to leverage the lessons learnt for video remixing and apply them for creating high quality sport summaries. The scenario pertaining to the unconstrained capture of basketball mobile videos, highlights the challenge with such type of content. Furthermore, the saliency detection method demonstrated the important role of sensors in reducing computational complexity and the value of multimodal analysis in improving the accuracy of saliency detection. The promising results from role based capture setup involving both mobile devices and professional equipment, indicated the importance of pragmatism for optimizing the desired performance parameters. The performance parameters to be optimized should be decided based on key stakeholder priorities. For example, in contrast to the unconstrained capture scenario which is suitable casual amateur recorders, the role based capture scenario was a better fit for professional and prosumer users.

In the previous discussion, the users' situational context (camera motion, location, etc.) is used in combination with her recorded content to create value added content such as video remixes and summaries. Subsequently, we analyzed the use of users' situation context via capture and sharing of rich virtual presence between the collaborating users. The architectural choices are directly impacted by the end-point device resource constraints and network latency, consequently a thin client approach is expected to scale more easily with increase in video resolution. The effect of interaction on media consumption was influenced by the type of content being consumed and the comfort level between the participants. Higher the closeness between the participants, greater openness for richer virtual presence was observed. The key challenge in future would be to develop a content and context adaptive system, which leverages SNSs to determine closeness between users to adjust the default presence sharing levels.

It can be seen from the user experience studies as well as the direction of the upcoming VR platforms, that collaborative consumption is still in its early stages and there is significant scope to develop. On comparing the proof-of-concept system presented in the thesis and the upcoming VR based collaborative consumption systems certain commonalities can be seen. Features such as a lounge or meet-up area, commonly consumed content and rich interaction between the users to infuse a *common shared context* can be seen. In addition, with the presence of multiple Internet enabled multimedia devices (mobile devices, tablets, laptops, desktops, TVs) the ground for multi-device content consumption with movable multimedia is stronger. Although implicit or automatic transfer of multimedia sessions is not yet common in consumer space, the analog of third party call control and screen sharing have made viewing content from optimal device, commonplace. The essential aspects of *session state capture and sharing via a device centric approach* has become more successful in a localized scenario by avoiding inter-network security, privacy and NAT/FW related complexities. The advances in IOT indicates a strong potential for further development of service mobility across multiple devices. The increase in multi-device ecosystem (mobile device, accessory cameras, VR headsets, etc.), the lines between collaborative creation and consumption systems are blurring. (Figure 1 in section 1.1).

7.1 Future developments

The future trends of increase in network speeds (e.g. 5G), mobile device multimedia capabilities (4K recording, omnidirectional content consumption) and IOT are key trends that will affect the trajectory of video creation and consumption ecosystem. Live video content is proving to be an important tool for bringing families together as well as for social media activism. In future boundaries between content creators and consumers will become fuzzier. The ability to contribute high quality content in real-time enables the use of such content for creating automatic remixes in real-time. Research to optimize the algorithmic latency, system latency, scalability are some of the research aspects which need further study. In addition, further research is needed to identify techniques for identifying user requirements for video remixes and evaluating them. Furthermore, mainstreaming of (Omni-directional content capture) OCC devices will have a significant impact on the video remixing and summarization techniques. For example, OCC devices coupled with appropriate person or object tracking methods has the potential to completely remove humans from the capture phase. Systems for low latency and jitter free transport of content from the multiple constituent cameras of OCC devices requires further research. Equally important is to understand the user experience impact of using such a method in different event scenarios by professional users, prosumers and consumers.

References

[1] J. Abreu, P. Almeid, V. Branco, "2BeOn – Interactive television supporting interpersonal communication", *Proceedings of the 6th Eurographics workshop on Multimedia*, 8-9 Sep. 2001, Machester, U.K., pp. 199-208.

[2] M. Adeyeye, P. Bellavista, "Emerging research areas in SIP-based converged services for extended Web clients", *World Wide Web*, Vol, 17, No. 6, November 2014, pp. 1295-1319. [Online]. Available: http://dx.doi.org/10.1007/s11280-013-0238-0

[3] A. Alahi, Y. Boursier, L. Jacques, P. Vandergheynst, "Sport players detection and tracking with a mixed network of planar and omnidirectional cameras", *Proceedings of the 3rd ACM/IEEE International Conference on Distributed Smart Cameras*, ICDSC 2009, 30 Aug – 2 Sep. 2009, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/ICDSC.2009.5289406

[4] Amazon, "Fire TV", https://www.amazon.com/Amazon-Fire-TV-Family/b?ie=UTF8&node=8521791011 (Accessed on July 5 2016).

[5] I. Arev, H. Park, Y. Sheikh, J. Hodgins, A. Shamir, "Automatic Editing of Footage from Multiple Social Cameras", *ACM Transactions on Graphics*, Vol. 33, No. 4, Article 81, July 2014. [Online]. Available: http://dx.doi.org/10.1145/2601097.2601198

[6] Apple, "iTunes App Store", https://itunes.apple.com/en/genre/ios/id36?mt=8 (Accessed on July 13 2016).

[7] X. Bao, R. Choudhury, "Movi: mobile phone based video highlights via collaborative sensing", *Proceedings of the 8th ACM International Conference on Mobile systems, applications, and services*, MobiSys 2010, 15-18 June 2010, San Francisco, CA, U.S.A., pp. 357–370. [Online]. Available: http://dx.doi.org/10.1145/1814433.1814468

[8] E. Boertjes, "ConnecTV: Share the experience", *Proceedings of the 5th European Conference, Interactive TV: A Shared Experience*, EuroITV 2007, Amsterdam, the Netherlands, 24-25 May. 2007, pp. 139-140. [Online]. Available: https://dx.doi.org/10.1007/978-3-540-72559-6

[9] T. Bouwmans, F. El Baf, B. Vachon, "Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey", *Recent Patents on Computer Science*, Bentham Science Publishers, 2008, Vol.1, No.3, pp. 219-237. Available: http://dx.doi.org/10.2174/2213275910801030219

[10] D. Brezeale, D. Cook, “Automatic video classification: A survey of the literature”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, May 2008, Vol. 38, No.3, pp. 416–430. [Online]. Available: http://dx.doi.org/10.1109/TSMCC.2008.919173

[11] A. Carlier, R. Guntur, V. Charvillat, T.W. Ooi, “Combining content-based analysis and crowdsourcing to improve user interaction with zoomable video”, *Proceedings of the 19th ACM Internation[al Conference on Multimedia*, MM 2011, Scottsdale, AZ, U.S.A, 28 Nov.-1 Dec. 2011, pp. 43–52. [Online]. Available: http://dx.doi.org/10.1145/2072298.2072306

[12] A. Carlier, V. Charvillat, W.T. Ooi, R. Grigoras, G. Morin, “Crowdsourced automatic zoom and scroll for video retargeting”, *Proceedings of the 18th ACM international conference on Multimedia*, MM 2010, Firenze, Italy, 25-29 Oct. 2010, pp. 201-210. [Online]. Available: http://dx.doi.org/10.1145/1873951.1873962

[13] A. Carlier, R. Guntur, W.T. Ooi, “Towards characterizing users’ interaction with zoomable video”, *Proceeings of the ACM workshop on Social, adaptive and personalized multimedia interaction and access*, SAPMIA 2010, Florence, Italy, 29 Oct. 2010, pp. 21–24. [Online]. Available: http://dx.doi.org/10.1145/1878061.1878069

[14] R. Chalfen, *Snapshot versions of life*. Bowling Green State University, Bowling Green Ohio, 1987.

[15] F. Chen, C. De Vleeschouwer, “Personalized production of basketball videos from multi-sensored data under limited display resolution”, *Computer Vision and Image Understanding*, Vol. 114, No. 6, Jun. 2010, pp. 667–680. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2010.01.005

[16] X. Chen, A. O. Hero, S. Savarese, “Multimodal video indexing and retrieval using directed information”, *IEEE Transactions on Multimedia*, Vol. 14, No.1, Feb. 2012, pp. 3–16. [Online]. Available: http://dx.doi.org/10.1109/TMM.2011.2167223

[17] K.-Y. Cheng, S.-J. Luo, B.-Y. Chen, H.-H. Chu, “Smartplayer: User-centric video fast-forwarding”, *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI 2009, Boston, USA, 4-9 Apr. 2009, pp. 789-798. [Online]. Available: http://dx.doi.org/10.1145/1518701.1518823

[18] T. Coppens, K. Handekyn, F. Vanparijs, “AmigoTV: A Social TV Experience through Triple-Play Convergence”, Alcatel Technology white paper, 2005.

[19] F. Cricri, Media Mixing and Inter-Client Synchronization for Mobile Virtual TV-Room. M.Sc. Thesis. Tampere University of Technology, 2008.

[20] F. Cricri, Multimodal analysis of mobile videos, Ph.D. Thesis, Tampere University of Technology, Tampere, Finland, Publication 1207, May 2014.

[21] F. Cricri, K. Dabov, I.D.D. Curcio, S. Mate, M. Gabbouj, "Multimodal event detection in user generated videos", *Proceedings of the IEEE International Symposium on Multimedia*, ISM 2011, Dana Point, CA, USA, 5-7 Dec.2011, pp. 263-270. [Online]. Available: http://dx.doi.org/10.1109/ISM.2011.49

[22] F. Cricri, M. Roininen, J. Leppänen, S. Mate, I.D.D. Curcio, S. Uhlmann, M. Gabbouj, "Sport type classification of mobile videos", *IEEE Transactions on Multimedia*, Vol. 16, No. 4, February 2014, pp. 917-932. Available: http://dx.doi.org/10.1109/TMM.2014.2307552

[23] F. Cricri, I.D.D. Curcio, S. Mate, K. Dabov, M. Gabbouj. "Sensor-based analysis of user generated video for multi-camera video remixing", Proceedings of the 18th International Conference on Advances in Multimedia Modeling, MMM 2012, Klagenfurt, Australia, 4-6 Jan. 2012, pp. 255-265. [Online]. Available: https://dx.doi.org/10.1007/978-3-642-27355-1_25

[24] F. Cricri, K. Dabov, I.D.D. Curcio, S. Mate, M. Gabbouj, "Multimodal extraction of events and of information about the recording activity in user generated videos", *Multimedia Tools and Applications*, Vol. 70, No. 1, May 2014, pp. 119–158.[Online]. Available: https://dx.doi.org/10.1007/s11042-012-1085-1

[25] F. Cricri, K. Dabov, M. J. Roininen, S. Mate, I. D. D. Curcio, M. Gabbouj. Multimodal semantics extraction from user-generated videos. Adv. MultiMedia, Vo1, No.1, Jan. 2012. [Online]. Avaiable: http://dx.doi.org/10.1155/2012/292064

[26] F. Cricrí, S. Mate, I.D.D. Curcio, M. Gabbouj, "Mobile and Interactive Social Television – A Virtual TV Room", Proc. 10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '09), 15-19 June 2009, Kos, Greece. [Online]. Available: http://dx.doi.org/10.1109/WOWMOM.2009.5282411

[27] I.D.D. Curcio, S. Mate, "Method and apparatus for smart video rendering", Patent application, United States, US , US20150078723A1, 19 Mar. 2015, priority date 21 Jun. 2013.

[28] I.D.D. Curcio, S. Mate, F. Cricri, “Method and apparatus for providing media mixing with reduced uploading”, Patent application, United States, US , US8805954 B2, 12 Aug. 2014, priority date 7 Feb. 2011.

[29] I.D.D. Curcio, S. Mate, K. Dabov, F. Cricri, “Method and apparatus for selecting content segments”, Patent, United States, US , US8565581B2, 22 Oct. 2013, priority date 12 Nov. 2010.

[30] Y. Cui, K. Nahrsetedt, D. Yu, “Seamless User-leel Handoff in Ubiquitous Multimedia Service Delivery”, *Multimedia Tools and Applications*, Vol. 22, No. 2, Feb. 2004, pp. 137-170. [Online]. Available: https://dx.doi.org/10.1023/B:MTAP.0000011932.28891.a0

[31] J. Daemen, J. Herder, C. Koch, P. Ladwig, R. Wiche, K. Wilgen, "Semi-Automatic Camera and Switcher Control for Live Broadcast", *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX 2016, Chicago, IL, USA, 22 – 24 June 2016, pp. 129-134. [Online]. Available: http://dx.doi.org/10.1145/2932206.2933559

[32] M. De Sa, D. Shamma, E.F. Churchill, “Live mobile collaboration for video production: design, guidelines, and requirements”, *Personal and Ubiquitous Computing*, Vol.18, No. 3, Mar. 2014, pp. 693-707. [Online]. Available: https://dx.doi.org/10.1007/s00779-013-0700-0

[33] N. Diakopoulos, K. Luther, Y. Medynskiy, I. Essa, “The evolution of authorship in a remix society”, *Proceedings of the 18th conference on Hyptertext and hypermedia*, HT 2007, Manchester, UK, 10-12 Sep. 2007, pp. 133-136. [Online]. Available: http://dx.doi.org/10.1145/1286240.1286272

[34] D. Douglas, T. Peucker, "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature", *The Canadian Cartographer*, Vol. 10, No. 2, Dec. 1973, pp. 112–122. [Online]. Available: http://dx.doi.org/10.3138/FM57-6770-U75U-7727

[35] Dropbox, “Dropbox”, https://www.dropbox.com/ (Accessed on July 14 2016).

[36] H. El-Alfy, D. Jacobs, L. Davis, “Multi-scale video cropping”, *Proceedings of the 15th ACM international conference on Multimedia*, MM 2007, Augsburg, Germany, 23-28 Sep. 2007, pp. 97–106. [Online]. Available: http://dx.doi.org/10.1145/1291233.1291255

[37] D. Ellis, “Beat Tracking by Dynamic Programming”, *New Music Research*, Vol. 36 No. 1, Mar. 2007, pp. 51-60. [Online]. Available: https://dx.doi.org/10.1080/09298210701653344

[38] A. Engström, M. Esbjörnsson, P. Juhlin, “Mobile collaborative live video mixing”, *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, MobileHCI 2008, Amsterdam, the Netherlands, 2-5 Sep. 2008, pp. 157–166. [Online]. Available: http://dx.doi.org/10.1145/1409240.1409258

[39] A. Engström, M. Esbjörnsson, O. Juhlin, “Nighttime visual media production in club environments”, *Night and darkness: Interaction after dark-Workshop*, CHI 2008. [Online]. Available: http://research.microsoft.com/en-us/um/people/ast/chi/darkness/papers/engstrom_et_al.pdf

[40] A. Engström, M. Perry, and O. Juhlin, “Amateur vision and recreational orientation: creating live video together”, *Proceedings of the ACM 2012 conference on Computer Supporter Cooperative Work*, CSCW 2012, Seattle, WA, USA, 11-15 Feb. 2012, pp. 651-660. [Online]. Available: http://dx.doi.org/10.1145/2145204.2145304

[41] A. Eronen, A. Klapuri, “Music Tempo Estimation with k-NN regression”, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 1, Jan 2010, pp. 50-57. [Online]. Available: http://dx.doi.org/10.1109/TASL.2009.2023165

[42] Facebook, “Facebook Home page”, https://www.facebook.com/ (Accessed on July 11 2016).

[43] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee, “Hypertext Transfer Protocol – HTTP/1.1”, IETF Request for Comments 2616, Jun. 1999.

[44] J. Foote, M. Cooper, A. Girgensohn, “Creating music videos using automatic media analysis”, *Proceedings of the 10th ACM international conference on Multimedia*, MULTIMEDIA 2002, Juan Les Pins, France, 1-6 Dec. 2002, pp. 553–560. [Online]. Available: http://dx.doi.org/10.1145/641007.641119

[45] B. Ghanem, M. Kreidieh, M. Farra, T. Zhang, “Context-Aware Learning for Automatic Sports Highlight Recognition”, *Proceedings of the 21st International Conference on Pattern Recognition*, ICPR 2012, Tsukuba, Japan, 11-15 Nov. 2012, pp. 1977-1980.

[46] A. Ghittino, A. Iatrino, S. Modeo, F. Ricchiuti, “Living@room: a Support for Direct Sociability through Interactive TV”, *Adjunct Proceedings of the 5th EuroITV Conference*, Amsterdam, May 2007, pp.131-132. [Online]. Available: https://soc.kuleuven.be/com/mediac/socialitv/Living@room%20-%20a%20Support%20for%20Direct%20Sociability%20through%20Interactive%20TV.pdf

[47] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, L. Wilcox, “Semi-automatic approach to home video editing”, *Proceedings of the 13th annual ACM symposium on User interface software and technology*, UIST 2000, San Diego, CA, USA, 6-8 Nov. 2000, pp 81–89. [Online]. Available: http://dx.doi.org/10.1145/354401.354415

[48] A. Girgensohn, S. Bly, F. Shipman, J. Boreczky, L. Wilcox, “Home video editing made easy—balancing automation and user control”, *Proceedings of Human Computer Interaction*, INTERACT 2001, Tokyo, Japan, 9-13 Jul. 2001, pp. 464–471. [Online]. Available: http://www.fxpal.com/publications/home-video-editing-made-easy-balancing-automation-and-user-control.pdf

[49] Google, “Google Cast”, http://www.google.com/cast/ (Accessed on July 5 2016).

[50] Google, “Google Drive”, https://www.google.com/intl/en/drive/ (Accessed on July 14 2016).

[51] Google, “Google Play”, https://play.google.com/store/apps?hl=en (Accessed on July 13 2016).

[52] Google, “YouTube”, http://www.youtube.com/ (Accessed on July 5 2016).

[53] J. Guo, D. Scott, F. Hopfgartner, C. Gurrin, “Detecting complex events in user-generated video using concept classifiers”, *Proceedings of 10th International Workshop on Content-Based Multimedia Indexing*, CBMI 2012, Annecy, France, 27-29 Jun. 2012, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/CBMI.2012.6269799

[54] E. Guttman, “Vendor Extensions for Service Location Protocol, Version 2”, IETF Request for Comments 3224, Jan 2002.

[55] M. Handley, V. Jacobson, C. Perkins, “SDP: Session Description Protocol”, IETF Request for Comments 4566, Jul. 2006.

[56] C. Holz, M. Bentley, K. Church, M. Patel, ""I'm just on my phone and they're watching TV": Quantifying mobile device use while watching television", *Proceedings*

of the ACM International Conference on Interactive Experiences for TV and Online Video, TVX 2015, Brussels, Belgium, 3-5 June 2015, pp. 93-102. [Online]. Available: http://dx.doi.org/10.1145/2745197.2745210

[57] W. Hu, N. Xie, L. Li, X. Zeng, S. Maybank, "A survey on visual content-based video indexing and retrieval", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 41, No. 6, Nov. 2011, pp. 797–819. Available: http://dx.doi.org/10.1109/TSMCC.2011.2109710

[58] U. Iqbal, I. D. D. Curcio, M. Gabbouj, "Who is the hero? semi-supervised person re-identification in videos," *Proceedings of International Conference on Computer Vision Theory and Applications*, VISAPP 2014, Lisbon, Portugal, 5-8 Jan. 2014, pp. 162-173.

[59] G. Jacucci, A. Oulasvirta, A. Salovaara, "Active construction of experience through mobile media: A field study with implications for recording and sharing", *Personal and Ubiquitous Computing – Memory and sharing of experiences*, Vol. 11, No. 4, April 2007, pp. 215-234. [Online]. Available: http://dx.doi.org/10.1007/s00779-006-0084-5

[60] G. Jacucci, A. Oulasvirta, A. Salovaara, R. Sarvas, "Supporting the Shared Experience of Spectators through Mobile Group Media", *Proceedings of the 2005 International ACM SIGGROUP conference on Supporting group work*, GROUP 2005, Sanible Island, FL, USA, 6-9 Nov. 2005, pp. 207-216. [Online]. Available: http://dx.doi.org/10.1145/1099203.1099241

[61] JSON, "Introducing JSON", http://www.json.org/ (Accessed on July 13 2016).

[62] O. Juhlin, A. Engström, E. Önnevall,"Long tail revisited: from ordinary camera phone use to pro-am video production", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2014, Toronto, Canada, 26 Apr. – 1 May 2014, pp. 1325-1334. [Online]. Available: http://dx.doi.org/10.1145/2556288.2557315

[63] O. Juhlin, G. Zoric, A. Engström, E. Reponen, "Video interaction: a research agenda", *Personal and Ubiquitous Computing*, Vol. 18, No. 3, Mar. 2014, pp. 685-692. [Online]. Available: http://dx.doi.org/10.1007/s00779-013-0705-8

[64] L. Kennedy, M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos". *Proceedings of the 18th International conference on World Wide Web*, WWW 2009, Madrid, Spain, 20-24 Apr. 2009, pp. 311-320. [Online]. Available: http://dx.doi.org/10.1145/1526709.1526752

[65] D. Kirk, A. Sellen, R. Harper, K. Wood, “Understanding videowork”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2007, San Jose, CA, USA, 28 Apr. – 3 May 2007, pp. 61–70. Available: http://dx.doi.org/10.1145/1240624.1240634

[66] T. Koponen, H. Väätäjä, "Early adopters' experiences of using mobile multimedia phones in news journalism", *Proceedings of the European Conference on Cognitive Ergonomics: Designing beyond the Product --- Understanding Activity and User Experience in Ubiquitous Environments*, ECCE 2009, Helsinki, Finland, 30 September – 2 October 2009, Article No. 2.

[67] B. Lee, R. S. Lee, “How and Why People Watch TV: Implications for the Future of Interactive Television”, *Journal of Advertising Research*, Vol. 35, No. 6, 1995, pp. 9-18.

[68] A. Lehmuskallio, R. Sarvas, “Snapshot Video: Everyday Photographers Taking Short Video-Clips”, *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, NordiCHI 2008, Lund, Sweden, 20-22 Oct. 2008, pp. 257-265. [Online]. Available: http://dx.doi.org/10.1145/1463160.1463188

[69] M. Lew, N. Sebe, C. Djeraba, R. Jain, “Content-based multimedia information retrieval: State of the art and challenges”, *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 2, No. 1, Feb. 2006, pp. 1–19. [Online]. Available: http://dx.doi.org/10.1145/1126004.1126005

[70] F. Liu, M. Gleicher, “Video retargeting: automating pan and scan”, *Proceedings of the 14th annual ACM international conference on Multimedia*, MM 2006, Santa Barbara, CA, USA, 23-27 Oct. 2006, pp. 241–250. [Online]. Available: http://dx.doi.org/10.1145/1126004.1126005

[71] N. Luhmann, *Familiarity, confidence, trust: Problems and alternatives*. Basil Blackwell, Oxford, 1990.

[72] J. Makonen, R. Kerminen, I.D.D. Curcio, S. Mate, A. Visa, “Detecting events by clustering videos from large media databases“, *Proceedings of the 2nd ACM international workshop on Events in multimedia*, EiMM 2010, Firenze, Italy, 25-29 Oct. 2010, pp. 9–14. [Online]. Available: http://dx.doi.org/10.1145/1877937.1877942

[73] S. Malinen, J. Ojala, “Maintaining the instant connection - Social media practices of smartphone users”, *Proceedings of the 10th International Conference on the Design of Cooperative Systems From research to practice: Results and open*

challenges, COOP 2012, Marseille, France, 30 May – 1 Jun. 2012, pp. 197-211. [Online]. Available: http://coop-2012.grenoble-inp.fr/pdf_papers/Malinen_34.pdf

[74] C. Marshall, F. Shipman, "The ownership and reuse of visual media", Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, JCDL 2011, Ottawa, Canada, 13-17 Jun. 2011, pp.157-166. [Online]. Available: http://dx.doi.org/10.1145/1998076.1998108

[75] J. Martin, *Systems Engineering Guidebook*. CRC Press, Boca Raton, Florida, 1996.

[76] S. Mate, I.D.D. Curcio, "Method and apparatus for mobile assisted event detection and area of interest determination", Patent, United States, US ,8335522, 18 Dec. 2012, priority date 15 Nov. 2009.

[77] S. Mate, I.D.D. Curcio, "Video remixing System", Patent, United States, US9380328 B2, Priority date 28 Jun. 2011.

[78] S. Mate, I.D.D. Curcio, A. Lehtiniemi, "Video editing", Patent application, European Union, EU, EP2816563A1, 24 Dec. 2014, priority date 18 Jun. 2013.

[79] S. Mate, I.D.D. Curcio, F. Cricri, K. Dabov, "Method and apparatus for video synthesis", Patent, United States, US , US8874538B2, 28 Oct. 2014, priority date 8 Sep. 2010.

[80] S. Mate, I.D.D. Curcio, V. Malamal Vadakital, "Method and apparatus for generating a media compilation based on criteria based sampling", Patent, United States, US , US8880527 B2, 4 Nov. 2014, priority date 31 Oct. 2012.

[81] Merriam-Webster, "Merriam-Webster Online Dictionary", http://www.merriam-webster.com/ (Accessed on July 16 2016).

[82] Microsoft, "OneDrive", https://onedrive.live.com/ (Accessed on July 11 2016).

[83] Microsoft, "Windows Phone Store: Smart Resize", https://www.microsoft.com/en-us/store/apps/smart-resize/9wzdncrfhxfh (Accessed on July 14 2016).

[84] Mobile Business Insights, "Twenty surprising mobile stats for 2016: The smartphone takeover", http://mobilebusinessinsights.com/2016/06/twenty-surprising-mobile-stats-for-2016-the-smartphone-takeover/ (Accessed on July 12 2016).

[85] A. G. Money, H. Agius, “Video summarisation: A conceptual framework and survey of the state of the art”, *Journal of Visual Comunication and Image Representation*, Feb. 2008, Vol. 19, No. 2, pp. 121–143. [Online]. Available: http://dx.doi.org/10.1016/j.jvcir.2007.04.002

[86] A. Monroy-Hernández, B.M. Hill, J. Gonzalez-Rivero, D. Boyd, “Computers can’t give credit: how automatic attribution falls short in an online remixing community”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2011, Vancouver, BC, Canada, 7-12 May 2011, pp. 3421-3430. [Online]. Available: http://dx.doi.org/10.1145/1978942.1979452

[87] G. Muller, *CAFCR: A multi-view method for embedded systems architecting; balancing genericity and specificity*, Doctoral Thesis, Delft university of technology, Delft, Netherlands, 2004.

[88] Nest, “Nest Cam Indoor”, https://nest.com/camera/meet-nest-cam/ (Accessed on July 8 2016).

[89] K. Ngo, R. Guntur, A. Carlier, T. Ooi, “Supporting zoomable video streams via dynamic region-of-interest cropping”, *Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, MMSys 2010, Scottsdale, AZ, U.S.A., 22-23 Feb. 2010, pp. 259–270. Available: http://dx.doi.org/10.1145/1730836.1730868

[90] Nokia, “N95 specifications”, https://en.wikipedia.org/wiki/Nokia_N95 (Accessed on July 5 2016).

[91] Oculus,”Join Friends in VR with New Oculus Social Features”, https://www.oculus.com/en-us/blog/join-friends-in-vr-with-new-oculus-social-features (Accessed on 5 July 2016)

[92] J. Ojala, K. Väänänen-Vainio-Mattila, A. Lehtiniemi, ”Six Enablers of Instant Photo Sharing Experiences in Small Groups Based on the Field Trial of Social Camera”, *Proceedings of the 10th International Conference on Advances in Computer Entertainment Technology*, ACE 2013, Boekelo, The Netherlands, 12-15 Nov. 2013, pp. 344-355. [Online]. Available: https://dx.doi.org/10.1007/978-3-319-03161-3_25

[93] J. Ojala, “Personal Content in Online Sports Communities: Motivations to Capture and Share Personal Exercise Data”, *International Journal of Social and Humanistic Computing*, 2013, Vol. 2, No. 1-2, pp. 68–85. [Online]. Available: https://dx.doi.org/10.1504/IJSHC.2013.053267

[94] J. Ojanperä, “Audio scene selection apparatus”, Patent, United States, US9195740B2, 24 Nov. 2015, priority date 18 Jan. 2011.

[95] J. Ojanperä, “Methods and apparatuses for time-stamping media for multi-user content rendering”, Patent, United States, US8909661B2, 9 Dec. 2014, priority date 18 Sep. 2012.

[96] T. Olsson, “Understanding Collective Content: Purposes, Characteristics and Collaborative Practices”, *Proceedings of the fourth international conference on Communities and technologies*, C&T 2009, University Park, PA, USA, 25-27 Jun. 2009, pp. 21-30. [Online]. Available: http://dx.doi.org/10.1145/1556460.1556464

[97] Open Connectivity Foundation, “UPnP”, https://openconnectivity.org/upnp (Accessed on July 18 2016).

[98] C. Poppe, S. De Bruyne, R. Van de Walle, “Generic Architecture for Event Detection in Broadcast Sports Video,” *Proceedings of the 3rd ACM International Workshop on Automated Information Extraction in Media Production*, AIEMPro 2010, Firenze, Italy, 25-29 Oct. 2010, pp. 51–56. [Online]. Available: http://dx.doi.org/10.1145/1877850.1877865

[99] H. Papadopoulos, G. Peeters, “Joint Estimation of Chords and Downbeats From an Audio Signal”, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, Issue 1, Jan. 2011, pp. 138-152. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2045236

[100] M. Pietikäinen, Local Binary Patterns. Scholarpedia, Vol. 5, pp. 9775, 2010. [Online]. Available: http://dx.doi.org/10.4249/scholarpedia.9775

[101] A. Quitmeyer, M. Nitsche, "Documatic: participatory, mobile shooting assistant, pre-editor, and groundwork for semi-automatic filmmaking, *Proceedings of the 10th European conference on Interactive tv and video*, EuroITV 2012, Berlin, Germany, 4-6 July 2012, pp. 135-138. [Online]. Available: http://dx.doi.org/10.1145/2325616.2325643

[102] E. Rechtin, M. Maier, *The Art of Systems Architecting*. CRC Press, Boca Raton, Florida, 1997.

[103] K. Reddy, M. Shah, “Recognizing 50 human action categories of web videos”, *Machine Vision and Applications*, Vol. 24, No. 5, Jul. 2013, pp. 971–981, July 2013. [Online]. Available: http://dx.doi.org/10.1007/s00138-012-0450-4

[104] R. Ren, J. Jose, “Temporal Salient Graph for Sports Event Detection”, *Proceedings of the 16th IEEE International Conference on Image Processing*, ICIP 2009, Cairo, Egypt, 7-10 Nov. 2009, pp. 4313–4316. [Online]. Available: https://dx.doi.org/10.1109/ICIP.2009.5419129

[105] M. Roininen, J. Leppänen, A. Eronen, I. D. D. Curcio, M. Gabbouj, "Modeling the timing of cuts in automatic editing of concert videos", *Multimedia Tools and Applications*, Feb. 2016, pp. 1-25. Available: https://dx.doi.org/10.1007/s11042-016-3304-7

[106] Roku, “Roku TV”, https://www.roku.com/roku-tv (Accessed on July 5 2016).

[107] J. Rosenberg, J. Peterson, H. Schulzrinne, G. Camarillo, “Best Current Practices for Third Party Call Control (3pcc) in the Session Initiation Protocol (SIP)”, IETF Request for Comments 3725, April 2004.

[108] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, “SIP: Session Initiation Protocol”, IETF Request for Comments 3261, Jun. 2002.

[109] M. Rubinstein, A. Shamir, S. Avidan, "Multi-operator media retargeting", *ACM Transaction on Graphics*, Vol. 28 No. 3, Article 23, Aug. 2009. [Online]. Available: http://dx.doi.org/10.1145/1531326.1531329

[110] M. Sa, D. Shamma, E. Churchill, "Live mobile collaboration for video production: design, guidelines, and requirements", *Personal and Ubiquitous Computing*, Vol. 18, No. 3, March 2014, pp. 693-707. [Online]. http://dx.doi.org/10.1007/s00779-013-0700-0

[111] M. Saini, R. Gadde, S. Yan, W. Ooi, “Movimash: online mobile video mashup”, *Proceedings of the 20th ACM International Conference on Multimedia*, MM 2012, Nara, Japan, 29 Oct. – 2 Nov. 2012, pp. 139–148. [Online]. Available: http://dx.doi.org/10.1145/2393347.2393373

[112] E. Sallnas, *The Effect of Modality on Social Presence, Presence and Performance in Collaborative Virtual Environments*, Doctoral Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2004.

[113] Samsung, “Galaxy S7 specifications”, https://en.wikipedia.org/wiki/Samsung_Galaxy_S7 (Accessed on July 5 2016).

[114] Samsung, "Gear VR", http://www.samsung.com/global/galaxy/wearables/gear-vr/ (Accessed on July 12 2016).

[115] R. Sarvas, M. Viikari, J. Pesonen, H. Nevanlinna, "MobShare: Controlled and Immediate Sharing of Mobile Images", *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA 2004, New York, USA, 10-16 Oct. 2004, pp. 724-731. [Online]. Available: http://dx.doi.org/10.1145/1027527.1027690

[116] R. Schatz, S. Egger, "Social Interaction Features for Mobile TV Services", *Proceedings of the 2008 IEEE International Symposium on Broadband Multimedia Systems and Broadcast*, Las Vegas, NV, U.S.A, 31 Mar. - 2 Apr. 2008, pp. 1-6. [Online]. Available: http://dx.doi.org/10.1109/ISBMSB.2008.4536629

[117] R. Schatz, S. Wagner, S. Egger, N. Jordan, "Mobile TV becomes Social – Integrating Content with Communications", *Proceedings of the IEEE 29th International Conference on Information Technology Interfaces*, ITI 2007, Cavtat, Croatia, 25-28 25-28 Jun. 2007, pp. 263-270. [Online]. Available: http://dx.doi.org/10.1109/ITI.2007.4283781

[118] R. Schatz, S. Wagner, N. Jordan, "Mobile Social TV: Extending DVB-H Services with P2P-Interaction", *Proceedings of the 2007 Second International Conference on Digital Telecommunications*, ICDT 2007, San Jose, CA, USA, 1-5 Jul. 2007. [Online]. Available: http://dx.doi.org/10.1109/ICDT.2007.61

[119] H. Schulzrinne, E. Wedlund, "Application-Layer Mobility Using SIP", *ACM SIGMOBILE Mobile Computing and Communications Review*, Vol. 4, No. 3, Jul. 2000, pp.47-57. [Online]. Available: http://dx.doi.org/10.1145/372346.372369

[120] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", IETF Request for Comments 3550, Jul. 2003.

[121] R. Shacham, H. Schulzrinne, S. Thakolsri, W. Kellerer, "Session Initiation Protocol (SIP) Session Mobility", IETF Request for Comments 5631, Oct. 2009.

[122] D. Shamma, A., Shawn, R., P.L. Shafton, Y. Liu, "Watch what I watch: using community activity to understand content", Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2007, Augsburg, Germany, 28-29 Sep. 2007, pp. 275-284. [Online]. Available: http://dx.doi.org/10.1145/1290082.1290120

[123] T. Sheridan, *Humans and automation: System design and research issues*. Wiley Inter-Science, 2002

[124] P. Shrestha, M. Barbieri, H.Weda, “Synchronization of multi-camera video recordings based on audio”, *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA 2007, Augsburg, Germany, 24-29 Sep. 2007, pp. 545–548. ACM. [Online]. Available: http://dx.doi.org/10.1145/1291233.1291367

[125] P. Shrestha, H. Weda, M. Barbieri, D. Sekulovski, "Synchronization of multiple video recordings based on still camera flashes", *Proceedings of the 14th ACM international conference on Multimedia*, MM 2006, pp. 137-140. [Online]. Available: https://dx.doi.org/10.1145/1180639.1180679

[126] P. Shrestha, P. H. de With, H. Weda, M. Barbieri, E. H.L. Aarts, “Automatic mashup generation from multiple-camera concert recordings”, *Proceedings of the 18th ACM International Conference on Multimedia*, MM 2010, Firenze, Italy, 25-29 Oct. 2010, pp. 541–550. [Online]. Available: http://dx.doi.org/10.1145/1873951.1874023

[127] C. Stauffer, W. Grimson, “Adaptive Background Mixture Models for Real-Time Tracking”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 1999, Fort Collins, CO, USA, 23-25 Jun. 1999, pp. 246-252. [Online]. Available: https://dx.doi.org/10.1109/CVPR.1999.784637

[128] W.-T. Tan, G. Cheung, A. Ortega, B. Shen, “Community Streaming With Interactive Visual Overlays: System and Optimization”, *IEEE Transactions on Multimedia*, Vol. 11, No. 5, Aug. 2009, pp. 987-997. [Online]. Available: http://dx.doi.org/10.1109/TMM.2009.2021797

[129] Twitter, “Twitter Homepage”, https://twitter.com (Accessed on July 11 2016).

[130] N. Ukita, T. Ono, M. Kidode,“Region extraction of a gaze object using the gaze point and view image sequences”, *Proceedings of the 7th international conference on Multimodal interfaces*, ICMI 2005, Trento, Italy, 4-6 Oct. 2005, pp. 129–136. [Online]. Available: http://dx.doi.org/10.1145/1088463.1088487

[131] K. Väänänen-Vainio-Mattila, M. Wäljas, J. Ojala, K. Segerståhl, “Identifying Drivers and Hindrances of Social User Experience in Web Services”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2010, Atlanta, GA, USA, 10-15 Apr. 2010, pp. 2499–2502. [Online]. Available: http://dx.doi.org/10.1145/1753326.1753704

[132] N. Van House, “Collocated photo sharing, story-telling and the performance of self”, *International Journal of Human-Computer Studies*, Vol. 67, No. 12, Dec. 2009, pp. 1073-1086. [Online]. Available: http://dx.doi.org/10.1016/j.ijhcs.2009.09.003

[133] W3C, Extensible Markup Language (XML) 1.0 (Fifth Edition), https://www.w3.org/TR/xml/ (Accessed on July 18 2016).

[134] J. Wang, C. Xu, E. Chng, H. Lu, Q. Tian, “Automatic Composition of Broadcast Sports Video”, *Multimedia Systems*, Vol. 14, No. 4, Sep. 2008, pp. 179–193. [Online]. Available: https://dx.doi.org/10.1007/s00530-008-0112-6

[135] E. Weilenmann, R. Säljö, A. Engström, “Mobile video literacy: negotiating the use of a new visual technology”, *Personal and Ubiquitous Computing*, Vol. 18, No. 3, Mar. 2014, pp. 737-752. [Online]. Available: https://dx.doi.org/10.1007/s00779-013-0703-x

[136] X. Xie, H. Liu, S. Goumaz, S., W.-Y. Ma, “Learning user interest for image browsing on small-form-factor devices”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2005,Portland, OG, 2-7 Apr. 2005, pp. 671–680. [Online]. Available: http://dx.doi.org/10.1145/1054972.1055065

[137] V. Zsombori, M. Frantzis, R. L. Guimaraes, M. F. Ursu, P. Cesar, I. Kegel, R. Craigie, D. Bulter-man, “Automatic generation of video narratives from shared UGC”, *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, HT 2011, Eindhoven, The Netherlands, 6-9 Jun. 2011, pp. 325-334. [Online]. Available: http://dx.doi.org/10.1145/1995966.1996009

Publications

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3901-5 (printed)
ISBN 978-952-15-3902-2 (PDF)
ISSN 1459-2045