Semi-Automation in Video Editing

Than Htut Soe

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2022

UNIVERSITY OF BERGEN

Semi-Automation in Video Editing

Than Htut Soe

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 31.10.2022

Year: 2022

Title: Semi-Automation in Video Editing

Name: Than Htut Soe

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

This thesis was written while I was working as a PhD Research Fellow at the Department of Information Science and Media Studies at the University of Bergen. This thesis is part of the cooperative research project "Better Video Workflows via Real-Time Collaboration and AI-Techniques in TV and New Media" and addresses the *AI-Techniques* aspect within the project. The project is a collaboration between the Interaction Research Group at the Department of Information Science and Media Studies, and Vizrt[1], a global broadcasting and media technology company which headquarter is in Bergen. The project is funded by the Research Council of Norway under GrantNo.: 269790.

As this thesis is part of a joint project, the author worked under the supervision of the academic two supervisors from the department while working closely with Vizrt. The supervisor for the thesis is Prof. Marija Slavkovik and the co-supervisor is Prof. Frode Guribye. The former's work is mostly in AI and the latter's in HCI. So, this thesis is conducted at the interaction of HCI and AI. The results of this thesis were also presented to Vizrt employees throughout the PhD period. One example of the collaborative part of the project can be seen in the assisted subtitling work. Vizrt has developed Viz Story, the easy-to-use web-based video editing system. In the assisted subtitling work, Viz Story is modified to create a prototype for assisted subtitling.

[1] https://www.vizrtgroup.com/about/

Acknowledgements

This thesis would not be possible without the help and efforts from many people. First, I would like to express my gratitude for Ronan Huggard. He was involved in submission of the project proposal and establishing the project. In addition, his guidance and input as the project owner is critical for the progression and continuation of my thesis. His unique expertise in both academia and industry meant that he was able to provide important supervision for the thesis. And I would like to extend my thanks to everyone from Vizrt, especially, Even Normann, Senior Product Specialist, Roger Sætereng, R&D manager, Brage Breivik, Global Head of UX, and the entire UX team.

Special thanks to my supervisor, Marija Slavkovik, without her academic guidance, encouragement, and management, I would not be able to complete the thesis. She helped me above and beyond her responsibility so that I was able to continue progressing on the thesis. Thank you Marija for also giving me a chance to work on interesting AI research topics, one of which lead to my first paper published. Thank you very much for all that you have done to develop me so far as an academic and researcher, and showing me the wonders and how to deal with the frustrations of being in academia.

I am very lucky to have not just one, but two amazing supervisors. My co-supervisor, Frode Guribye, helped me tirelessly for this thesis. His input is crucial for the work that I have produced in my thesis, and without his help I would not be able to get up to speed on Human-Computer-Interaction aspect of the thesis. He also went out of his way to ensure that I can keep being productive despite the challenges. I would like to also thank Oskar Juhlin, for his input and advice for the project and the thesis.

I would like to express my gratitude to everyone at UIB who has helped me. Pavel Okopnyi, my colleague on the same project, for his help. Truls Pedersen and Bjarte Johansen for their friendship when I just arrived at UIB. I really appreciate all the students in my two classes and my teachers at various courses that I have taken at UIB, they taught me to be a better learner and a better teacher. Special thanks to the students from interaction design class who participated in the evaluation of my assisted subtitling prototype.

I would like to share my appreciation for all the people in the research communities I have participated in. Thank you to all the reviewers of my papers who provided me with the feedbacks necessary to improve the quality of my works. All the conferences and workshop organizers for their work to keep the research community flourishing even during the pandemic. Without all the research works that I have read and learned from, even those that were not cited in this paper, this research would not be possible.

Sincere thanks to my family and friends for the encouragements and support. Doing a PhD is a stressful endeavor, and their presence in my life is essential to help me deal with the challenges. My parents Hla Myint and Thuzar for their understanding and giving me the freedom to choose my own path since I was an adolescent. My aunt Phyu Phyu Aye for her selfless support for my education. Thanks to Bu Saw her encouragements and support. Special thanks to Ken and Karina, without their help, some of the experiments I have performed would not be possible. Thanks to Jing for her positivity, encouragements and giving me the belief that I can finish this thesis. Lastly, thank you, Norway and all the fantastic people that I have met in Bergen.

I would like to dedicate this thesis to my late brother Myint Thu Aung who was the kindest person I knew.

Preface

This is an article-based thesis. In accordance with the UIB guidelines for an article-based thesis, this thesis has three articles where I am the first author. In addition to the three papers, this thesis includes the framing of the research and how these papers form a cohesive attempt to address the challenges of using AI to support video editing. The three papers included in this thesis are described in the next three paragraphs.

Paper I is titled "AI video editing tools. What do editors want, and how far is AI from delivering them?". It is written by me as the first author, with Marija Slavkovik as the second author. An earlier version of this paper was presented at the IJCAI 2021 Workshop "AI and Product Design" [2] on 19th August 2021. The article is in Chapter 4. The paper will be submitted to Artificial Intelligence Review journal and the pre-print of it is available at https://arxiv.org/abs/2109.07809.

Paper II, titled "Evaluating AI Assisted Subtitling", is written by me as the first author with, Frode Guribye and Marija Slavkovik as the second and third authors. This paper has been published in the peer-reviewed conference proceeding, ACM International Conference on Interactive Media Experiences (IMX 21) in June 2021 and it is available in ACM Digital Library. The results of the paper were presented during the IMX 21 conference virtually and to Vizrt employees internally. The pre-print version of the paper is included in Chapter 5.

Paper III, "A content-aware tool for converting videos to narrower aspect ratios" is written by me and Marija Slavkovik as the first and second authors respectively. This paper has been published in the peer-reviewed conference proceeding, ACM International Conference on Interactive Media Experiences (IMX 2022) in June 2022. This paper was presented virtually during the IMX 2022 conference on 24th July and it is available in ACM Digital Library. The pre-print version is included in Chapter 6.

In addition to the three papers included in the thesis, two more papers were produced during the PhD period. However, they are not included in this thesis. These two pa-

[2] https://ijcai-21.org/workshops/

pers focused on a somewhat under-explored but still critical aspect of human–computer cooperation in semi-automated tasks — the problem of ethical human–AI user interface design. The designers of semi–automated processes have greater power in using psychology and interface design to nudge the human editors or users into making one choice over another. These nudges can be a powerful tool that improves the efficiency of the overall human–machine system, but they can also be abused to exploit human users. These two papers explore the problem of dark patterns in interfaces — the existence of dark patterns and the possibility of using machine learning to automatically detect the existence of dark patterns. The first paper [Soe et al., 2020] deals with the existence, identification, and categorization of dark patterns in cookie consent notices. The second paper [Soe et al., 2022] explores the feasibility and challenges of using automation to detect dark patterns in these cookie consent notices.

I started working on this PhD thesis with a proposal that focused on AI technology and tools. As I dived in and worked further on this thesis, I discovered something that I consider more important and ended up becoming the key elements of this thesis. It is about the user experience and Human–AI interaction challenges. In addition, having the chance to work with my supervisor, Marija Slavkovik, on the impact of AI on society, and other facets of AI which are fairness, accountability, transparency, and explainability influenced and shaped the work on this thesis.

The second half of the thesis was done during the pandemic, that caused unexpected disruptions in the work environment. However, the generous extensions offered from the UIB for the PhD students as well as the efforts from my supervisor and co-supervisor helped me stay productive and to be able to complete the thesis within a reasonable time.

Abstract

How can we use artificial intelligence (AI) and machine learning (ML) to make video editing as easy as "editing text"? In this thesis, this problem of using AI to support video editing is explored from the human–AI interaction perspective, with the emphasis on using AI to support users. Video is a dual-track medium with audio and visual tracks. Editing videos requires synchronization of these two tracks and precise operations at milliseconds. Making it as easy as editing text might not be currently possible. Then how should we support the users with AI, and what are the current challenges in doing so?

There are five key questions that drove the research in this thesis. What is the start of the art in using AI to support video editing? What are the needs and expectations of video professionals from AI? What are the impacts on efficiency and accuracy of subtitles when AI is used to support subtitling? What are the changes in user experience brought on by AI-assisted subtitling? How can multiple AI methods be used to support cropping and panning task?

In this thesis, we employed a user experience focused and task-based approach to address the semi-automation in video editing. The first paper of this thesis provided a synthesis and critical review of the existing work on AI-based tools for videos editing and provided some answers to how should and what more AI can be used in supporting users by a survey of 14 video professional. The second paper presented a prototype of AI-assisted subtitling built on a production grade video editing software. It is the first comparative evaluation of both performance and user experience of AI-assisted subtitling with 24 novice users. The third work described an idiom-based tool for converting wide screen videos made for television to narrower aspect ratios for mobile social media platforms. It explores a new method to perform cropping and panning using five AI models, and an evaluation with 5 users and a review with a professional video editor were presented.

Abstrakt

Semi-automasjon i video redigering[3]

Hvordan kan vi bruke kunstig intelligens (KI) og maskin læring til å gjøre videoredigering like enkelt som å redigere tekst? I denne avhandlingen vil jeg adressere problemet med å bruke KI i videoredigering fra et Menneskelig-KI interaksjons perspektiv, med fokus på å bruke KI til å støtte brukerne. Video er et audiovisuelt medium. Redigere videoer krever synkronisering av både det visuelle og det auditive med presise operasjoner helt ned på millisekund nivå. Å gjøre dette like enkelt som å redigere tekst er kanskje ikke mulig i dag. Men hvordan skal vi da støtte brukerne med KI og hva er utfordringene med å gjøre det?

Det er fem hovedspørsmål som har drevet forskningen i denne avhandlingen. Hva er dagens "state-of-the-art" i KI støttet videoredigering? Hva er behovene og forventningene av fagfolkene om KI? Hva er påvirkningen KI har på effektiviteten og nøyaktigheten når det blir brukt på teksting? Hva er endringene i brukeropplevelsen når det blir brukt KI støttet teksting? Hvordan kan flere KI metoder bli brukt for å støtte beskjærings- og panoreringsoppgaver?

Den første artikkelen av denne avhandlingen ga en syntese og kritisk gjennomgang av eksisterende arbeid med KI-baserte verktøy for videoredigering. Artikkelen ga også noen svar på hvordan og hva KI kan bli brukt til for å støtte brukere ved en undersøkelse utført av 14 fagfolk. Den andre studien presenterte en prototype av KI-støttet videoredigerings verktøy bygget på et eksisterende videoproduksjons program. I tillegg kom det en evaluasjon av både ytelse og brukeropplevelse på en KI-støttet teksting fra 24 nybegynnere. Den tredje studien beskrev et idiom-basert verktøy for å konvertere bredskjermsvideoer lagd for TV til smalere størrelsesforhold for mobil og sosiale medieplattformer. Den tredje studien utforsker også nye metoder for å utøve beskjæring og panorering ved å bruke fem forskjellige KI-modeller. Det ble også presentert en evaluering fra fem brukere. I denne avhandlingen brukte vi en brukeropplevelse og oppgave basert framgangsmåte,

[3]Title and abstract in Norwegian

for å adressere det semi-automatiske i videoredigering.

List of publications in Thesis

1. Than Htut Soe, Marija Slavkovik, *AI video editing tools. What do editors want, and how far is AI from delivering them?*, planned submission to Artificial Intelligence Review journal, Preprint available on: https://arxiv.org/abs/2109.07809

2. Than Htut Soe, Frode Guribye, Marija Slavkovik, *Evaluating AI assisted subtitling*, IMX '21: ACM International Conference on Interactive Media Experiences **21**, 6, 2021.

3. Than Htut Soe, Marija Slavkovik, *A content-aware tool for converting videos to narrower aspect ratios*, IMX '22: ACM International Conference on Interactive Media Experiences (IMX 2022) **22**, 6, 2022.

Additional publications during the PhD

1. Than Htut Soe, Oda Elise Nordberg, Frode Guribye, Marija Slavkovik, *Circumvention by design - dark patterns in cookie consent for online news outlets*, NordiCHI '20: Proceedings of the 11th Nordic Conference on Human–Computer Interaction: Shaping Experiences, Shaping Society **25**, 10, 2020.

2. Than Htut Soe, Cristiana Santos, Marija Slavkovik, *Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way*, submitted to ACM FAccT Conference 2022

Contents

Chapter 1

Introduction

Video is the most popular form of content on the Internet, in terms of the amount of data traffic. According to the Cisco visual networking index [Cisco, 2020], 75% of the Internet traffic in 2017 was video content, and it is projected to reach 82% by 2022. It is easier and quicker than ever to capture and publish videos. For example, mobile phones with capable cameras together with video sharing and social media platforms means that videos can be made and distributed on a whim. Editing videos, however, is still a labor-intensive task.

Video remains a challenging and time-consuming medium to edit for two reasons. First, video is a dual-track medium with both audio and video tracks, and both these tracks are required to be edited and synchronized. Second, editing video involves performing precise operations at individual frames. Can these barriers to video editing be lowered or removed using AI technology? *Can we make video editing as easy as editing text?*

The main theme of the thesis is *AI-assisted video editing*. In this thesis, the key question driving the research is "How can we use artificial intelligence (AI) to assist in video editing?" To introduce the thesis, first, the concepts of video editing and video workflows will be presented, followed by related topics in AI. After that, the arguments for combining both human intelligence and AI in video editing will be presented. The human–AI interaction issues will be presented from the human–computer interaction (HCI) perspective. The rest of the chapter will summarize the research questions, the challenges in combining human intelligence and AI, the overview of the works in this thesis, and the structure of the thesis.

1.1 Video editing workflows

Okun et al. [2015] define video editing as the act of cutting and joining pieces of one or more sources together to make one edited video. *Video editing tools* can be generally defined as (computer) programs that people can use to perform the task of video editing. All the popular video editing tools we have today are created for non-linear editing (NLE). NLE is defined as a form of video editing that does not require that sequences to be worked on sequentially [Okun et al., 2015] [1]. As of July 2021, the five most popular video editing tools according to Google Search trends[2] are Adobe Premiere Pro, DaVinci Resolve, iMovie, Lightworks, and Shortcut. These are all digital non-linear video editing tools. In this thesis, the term video editing refers exclusively to NLE.

Where does video editing fit in the entire video production process? A video production workflow consists of three stages, which are planning and preparation, production, and post-production [Owens and Millerson, 2011]. Production is "actually shooting the production" [Owens and Millerson, 2011]. In the entire video production workflow, video editing is part of the post-production where the raw video footage, captured during the production, is edited. This thesis, as the title implies, only addresses video editing.

What does a typical video editing workflow look like, and what tasks are included in video editing? There is no definitive answer to these two questions. In this thesis, we use a simplified video editing workflow, as shown in Figure 1.1. This workflow is created from reviewing the literature on intelligent video editing tools, and it was revised after consultations with a video editor and a video editing product manager [Soe and Slavkovik, 2021]. Mapping out tasks, as in Figure 1.1, helps in understanding what is usually involved in a video editing workflow, and it serves as a reference for us to estimate how far we are from automating the entire video workflow. This model for tasks in video editing is also used to create a comparison of different intelligent video editing tools in Chapter 2.

1.2 Supporting video editing tasks with AI

Artificial Intelligence (AI) can be loosely defined as intelligent behavior in artifacts [Nilsson, 1998]. This intelligent behavior can be perception, reasoning, learning, communication, or acting in complex environments [Nilsson, 1998]. Machine learning (ML) is a

[1] In contrast to linear editing, where editing has to be performed from the beginning of the video to the end in the exact sequential order.

[2] The search keyword is "most popular video editing software"

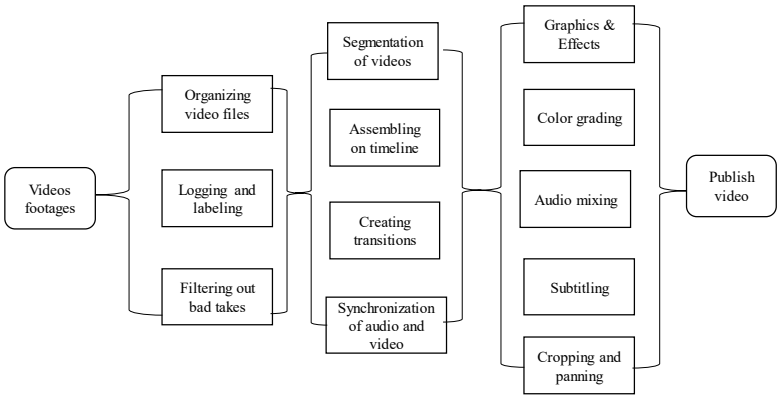

Figure 1.1: Tasks involved in video editing

subfield of AI and is concerned with how a machine can improve its future performance based on inputs or in response to external information [Nilsson, 1998]. Deep learning(DL) is a type of ML that uses neural networks with more than three layers of neurons [Goodfellow et al., 2016]. It is DL that has resulted in fairly recent breakthroughs in perception tasks such as speech and object recognition [Bengio et al., 2021].

Because DL has many potential applications for videos, it will be discussed further in this paragraph. A more comprehensive review of the history and the state of the art of AI is presented in Chapter 2. In 2009, a DL method proposed by Mohamed et al. [2011] outperformed other state-of-the-art speech recognition models on the task of recognizing spoken text in audio recordings. The breakthrough for DL in object recognition is the work by Krizhevsky et al. [2012], which won the 2012 ImageNet competition. The competition evaluates machines' ability to recognize objects in and classify images. Although there is a clear distinction between the terms AI, ML, and DL, they all tend to be referred to as "AI" in the industry and in the media.

State-of-the-art methods in AI are *narrow*. To put it in different words, AI techniques are developed for a performing a specific task. For example, Bengio et al. [2021] give examples of breakthroughs in AI in speech recognition and object recognition, both of which are narrow tasks. These tasks are considered narrow because an AI speech recognition model can transcribe the spoken words into text, but it cannot perform anything other than that particular task. This is where the breakdown of tasks in video editing workflows is helpful in the quest to (semi-)automate video workflows. It is more practical to apply existing narrow AI techniques to each smaller task in a video editing workflow than to come up with AI that handles the entire video editing, end to end. For example, an AI-based speech recognition model can be used directly to support a subtitling task without any modification to the existing speech recognition techniques.

Depending on the availability of suitable AI techniques for a video editing task, some tasks of the video editing workflow can be easier to automate than others. For example, verbatim subtitling is one of the easiest video workflow tasks to automate using machine speech recognition. This is because speech recognition is a mature application of AI, with many tools and models available to perform the recognition task. However, some video editing tasks, such as composing video and audio segments, do not have any suitable existing AI techniques to directly automate them. Developing AI techniques for composing video segments would require the machine to *understand* the content of both audio and video, as well as to be able to understand and reason with the story or meaning of the video.

DL-powered breakthrough advances in areas such as image processing, computer vision, and natural language processing [Bengio et al., 2021], have made the automation and augmentation of video editing tools possible. However, the dream of removing the drudgery of video editing is far from being accomplished. There are many obstacles to having quality automated video editing or very efficient semi-automated video editing. The main challenge comes from the need to pull together research and insights from three different fields of studies, namely, the study of workflows and practices in video editing, AI, and human–computer interaction. These fields need to be explored further before we can even discuss the full automation of video editing. However, as suggested in the previous section, AI (especially narrow AI) can be used in supporting or augmenting the video editing workflow.

The lack of closely aligned AI techniques does not mean that some of the video editing tasks, such as composing videos segments, cannot be supported with AI. For example, the work by Leake et al. [2017] uses dialogue scripts, speech recognition, and face detection to support composing video segments for dialogue-driven videos. Therefore, with an abundance of AI techniques and tools, the main challenge is perhaps not automation itself but *coordination between the individually automated tasks (that can be easily automated) and those done by a human editor*. These design challenges and human–AI interaction issues in the context of video editing tasks became important issues for supporting video editing with automation when AI techniques are embedded into existing video editing workflows. Semi-automating video editing is a very narrow field; however, there are plenty of relevant research in automating human tasks and human–AI interaction. These two topics are discussed in the following sections.

1.3 Automating human tasks: what can be done, what can be done well

Let us briefly consider the state of the art in automating human tasks, and particularly the limitations we face in automating some of these of tasks. A task in this section is loosely defined as a piece of work that was usually done by a human. Such tasks can range from perception tasks (speech to text, face identification, medical image diagnosis) to making decisions (automated essay grading, spam detection), making a predictions (predicting recidivism, predicting stock prices), to navigate in the physical world (driving a car, walking). When it comes to automation, tasks can be fully or partially automated. Full automation means that the AI can execute the human tasks with no supervision. However, full automation in some types of tasks is often not possible due to limitations of current AI technology. In these cases, semi-automation is more commonly used. Semi-automation (partial automation) combines machine automation with human input or oversight.

Driving is an example of a task that has not been fully automated yet. For regularity reasons, the United States' National Highway Traffic Safety Administration (NHTSA) defines six formal levels [NHTSA, 2022] for degrees of automation in driving. There are six different levels of automation, from a manual driving car (Level 0) to a fully automated car (Level 5). Level 1 is driver assistance, which automates a single control, such as brake assist or steering. In Level 1, only one control is automated at a time. Level 2 is partial automation, where combined automated controls are used together for more functions, such as steering and accelerating on a highway. Level 3 is conditional automation, which does not require constant attention but requires the driver to be ready to take over on short notice. Level 4 is high automation, where the automation can take full control under certain conditions with the option to take over altogether. Level 5 is full automation under all conditions. These semi-automation levels can be used to measure progression towards the ultimate goal, which is fully automated driving [Casner et al., 2016].

However, some tasks, such as playing chess have been completely automated, and machines have surpassed the human level in these tasks. It means that a chess program has exceeded the performance of a human player and can easily and consistently beat even the best human players in a game of chess [Silver et al., 2017]. AI excels in playing/automating this type of games/tasks because the rules of these games are strictly defined, input is always without noises, and it is easy to accurately simulate these games entirely inside a computer. In fact, the same algorithm has been used to play both chess and shogi (a Japanese board game) [Silver et al., 2017], and human performance was

exceeded in both cases.

Relatively simple tasks for humans, such as recognizing speech, are greater challenges for AI than playing chess. For example, in converting speech to text the best–performing system, Google Speech API, has 9% error rate [Këpuska, 2017]. Converting speech to text is challenging to automate, as speech can have numerous variations in terms of voices, accents, volume, tone, and background noises. In addition, language can contain proper nouns, idioms, and variations that AI has not encountered (in the training data) before. In the cases where complete automation has fallen short of our expectations, as in driving a car or using speech recognition to create accurate subtitles, we can still combine automation and human input (semi-automation) to help human users perform better.

The main challenge of semi-automation is figuring out how to share and integrate human activities and machine activities seamlessly to perform a particular task. Semi-automation combines human input or human supervision with AI. In order to do so effectively, a human has to interact with, understand, and put enough trust in the AI technology. Consequently, semi-automation has human–AI interaction issues that need to be addressed. For semi-automated driving, some examples of the human–AI interaction problems can be the driver putting too much trust in automation, inadequate attention by the driver when the automation is engaged, and atrophied driving skills over time [Casner et al., 2016]. Semi-automation in video editing also has its own sets of human–AI interaction challenges, but they have not received much attention and thus largely unexplored. Retaining the creative control of the user is the only challenge discussed in the semi-automation of video editing [Girgensohn et al., 2000; Leake et al., 2017; Wang et al., 2019]. This topic of human–AI interaction issues in general is discussed in the next section.

1.4 Human–AI interaction

Human–AI interaction is becoming an important topic since more and more products and services in our society have AI embedded in them. In this section, key aspects of human–AI interaction will be introduced from the perspective of user experience(UX) and human–computer interaction(HCI) disciplines. Following, human-centered AI and design methods to address human–AI problems will be discussed. The section will conclude with a definition of human–AI interaction for this thesis and what makes a good human–AI interaction.

Horvitz [1999] introduced "12 principles for enhancing human–computer interaction through an elegant coupling of automated services with direct manipulation". This work is one of the earliest works to explore the idea of enhancing human–computer interactions with automation. The first principle from the paper is "developing significant value-added automation", which emphasizes that automated services should provide "genuine" value to the users. Heer [2019] argues for the design space of augmenting human intelligence with AI over purely automated methods, and proposes a method to perform that augmentation. The augmentation in Heer's uses the 12 principles from Horvitz [1999] to integrate human agency and automation in an intelligence system via shared representation.

Since AI has been integrated into increasingly more software applications, it has received more attention from the interaction design and User Experience (UX) [3] research communities. Their main question is how to integrate this AI into existing UX and design practices, and to identify the challenges in doing so. Thus, AI becoming a new design material to create products and to create applications is discussed in [Dove et al., 2017] and [Holmquist, 2017]. Dove et al. [2017] surveyed UX practitioners and pointed out AI-related challenges for the UX community, and discussed the limitations of the UX community regarding AI. Some challenges identified by Dove et al. [2017] are "difficulties in understanding AI and its capabilities", "AI as a design material", and "the purposeful use of AI". The limitations in the UX community include "lack of knowledge of AI" and "difficulties in creating prototypes with AI" [Dove et al., 2017]. Yang et al. [2020] built upon these two works and discussed reasons why and how human–AI interaction is difficult to design. They also proposed a framework to help in navigating the human–AI interaction issues. The proposed framework [Yang et al., 2020] emphasized AI technological advances, designing user–system co-evolvement, and designing adaptive interactions at scale, together with HCI design as usual.

To make the work of designing products and services with AI easier, design processes and guidelines have been proposed for designing human–AI interactions. Jin et al. [2021] extracted 40 design heuristics from AI patents to support UX designers and applied these heuristics into two case studies. Some examples of heuristics that are classified under the category personalization are creating interactions that are "adaptive, learning, natural and recommending" [Jin et al., 2021]. For each of the heuristics, one example of the work is listed, but the summary about the example of the work is not available in the paper. Since, as the summaries are not available, it is a challenge to navigate the examples in [Jin et al., 2021].

[3] UX is usability as the main attribute but includes, for example, accessibility, fun, and aesthetics, depending on the context [Turner, 2017]

A similar study by Amershi et al. [2019] created 19 general guidelines for human–AI interaction and verified them with 49 design practitioners. The guidelines include, for example, "make clear what the system can do" and "learn from user behavior". In contrast to Jin et al. [2021], Amershi et al. [2019] provide descriptions and examples of the guidelines, making them easier to apply. However, it can be argued that some guidelines are not specific to AI, such as "Make clear what the system can do" can be applied to any software. As mentioned before, [Yang et al., 2020] also proposed a conceptual framework for mapping human–AI interaction issues. Although a few guidelines and frameworks has been proposed, human–AI interaction designs and studies do not have any widely adopted guidelines and frameworks. In addition, the nature of human–AI interaction is dependent on the context and the environment, meaning that guidelines distilled from one category of products might not be relevant to another.

From the AI research perspective, emphasizing the importance of human–AI interaction is done under the term "human-centered-AI (HAI)". The three goals of HAI according to the Stanford HAI initiative [HAI, 2022] are: AI to incorporate versatility, nuance, and depth of human intellect; development of AI should be guided by its impact on human society; and the purpose of AI should be to enhance our humanity. Xu [2019] discusses HAI from the perspective of HCI and uses the term "third-wave AI" in somewhat of a parallel to third-wave HCI [Bødker, 2006]. The third wave of AI, according to Xu [2019], should provide useful and real problem-solving AI solutions; should focus on ethical design, technological enhancement, and human factors design; and should take a human-centered approach.

Because human–AI interaction can happen in a variety of domains and for different purposes, it is necessary to focus on a specific definition of human–AI interaction. van Berkel et al. [2021] defined human–AI interaction as "the completion of a user's tasks with the help of AI support". This is the definition we will use in this thesis. At the beginning of this chapter, we have introduced the tasks in video editing. These tasks are normally done by the video editor and AI can be used to support the users in completing them. For example, AI can be used to support the task of creating subtitles. Supporting video editing tasks with AI requires determining what tasks are important, how automation can be used to support users in these tasks, and what kind of automation is useful.

The definition of human–AI interaction for this thesis was presented in the previous paragraph. The next question is to explore what makes a quality or successful human–AI interaction. One aspect of a quality human–AI interaction is usefulness. Xu [2019] defined useful AI as an "AI solution that can provide the functions required to satisfy target users' needs in the valid usage scenarios of their work and life". Another aspect of a quality human–AI interaction is efficiency. AI can be used to support a user's task in

a way that leads to more efficient completion of the task. Efficiency can be measures by comparing the time it takes to complete a task without AI assistance and the completion time with AI assistance.

1.5 Summary of the research design

The goal of this thesis is to explore how AI can be used to support users in video editing tasks. Based on the goal of the thesis, the following more specific research questions are posed:

- *RQ1*: What is the state of the art in AI-assisted video editing tools?
- *RQ2*: What are the opinions and expectations of video professionals regarding AI in video editing tools?
- *RQ3*: What is the impact of introducing AI assistance on the efficiency and quality of subtitles?
- *RQ4*: What are the changes in the user experience when AI assistance is added to the subtitling task?
- *RQ5*: How can we use AI to create a new way of performing cropping and panning in video editing?

The purpose of *RQ1* and *RQ2* is to lay the groundwork for the thesis. The goal of *RQ1* is to define the field of study based on the literature, and the goal of *RQ2* is to frame the thesis based on video professionals' opinions about AI. Afterwards, two tasks were chosen as tasks in which to explore how AI can be used to support the users performing these tasks. The two tasks are subtitling (*RQ3, RQ4*) and cropping and panning (*RQ5*). For each of the two tasks, separate AI-assisted prototypes were implemented, and these prototypes were used in performing user studies. The rationale for the selection of the tasks and what is involved in answering each research question is addressed in this section.

RQ1 and *RQ2* are addressed in the first paper [Soe and Slavkovik, 2021]. *RQ1* is answered using a systematic survey of the literature on AI-assisted video editing tools. A synthesis, comparison and critique of existing work were used in our attempt to answer *RQ1*. *RQ2* is addressed by surveying people in the industry to obtain opinions on AI and to identify what they would like AI to do for them in video editing, as well as what AI should not be doing.

The second paper [Soe et al., 2021] answers *RQ3* and *RQ4*. These two questions are about exploring how a single AI model can directly support subtitling task. The first reason for selecting the subtitling task is that it can be supported using just one AI model. That means that the impact of AI performance (mistakes in particular) on users can be studied. The second reason is that the quality of the work in subtitling can be measured objectively and accurately. A prototype that evaluates the AI-assisted subtitling against the subtitling without any AI assistance was built, and a user study was performed. *RQ3* is answered using the differences in the quality and efficiency of creating subtitles based on the data from the user study. The user feedback and observations from the study are used to answer *RQ4*.

Some tasks in video editing cannot be supported using a just single AI model yet. The task of cropping and panning is one of these tasks. However, it is still possible to create easier workflows for such tasks using multiple AI models. The cropping and panning task was selected because it can be supported using AI techniques for video perception. Video perception is a well-developed application area of AI. Most importantly, the cropping and panning task isolated can be framed as video retargeting, editing a video to a different aspect ratio. A prototype was built using five AI models, and a user study is performed to answer *RQ5* in the last paper [Soe and Slavkovik, 2022]. We created a prototype for video retargeting and asked the users to use the prototype to create videos that are in different aspect ratios from the original video. The user study is used to explore how cropping and panning should be supported with AI and to determine the usefulness of the prototype we have built.

1.6 Challenges of semi-automation

The main challenges of semi-automation stem from the fact that it requires a human to interact with automation. The five research questions introduced in the previous section are specific to this thesis and are aimed towards semi-automation in video editing. However, there are some overarching concerns regarding semi-automation in general. These concerns will be presented in this section, followed by the challenges of semi-automation in video editing.

The efficiency and usefulness of using automation or semi-automation in video editing workflows depend on crafting effective human–AI interactions and well-designed user interfaces that facilitate those interactions. The challenges involved in crafting human–AI interactions and interfaces are discussed in works such as [Amershi et al., 2019; Horvitz, 1999]. The guidelines in Amershi et al. [2019] cover topics such as building

user experience, clarifying user expectations, matching social norms, and learning from users' behavior. Guidelines like these can assist in identifying possible issues. But before applying them to any specific scenario, it is necessary to reevaluate them in that specific environment and context. Context is important in HCI and Human–AI interaction is no different in this regard.

Improvements in efficiency can be a good measure to evaluate semi-automation. If we define *efficiency* as the time it takes to complete a task, measuring efficiency becomes an easy task. However, when comparing semi-automated tasks to manual tasks, measuring efficiency alone might not be sufficient. The follow-up question to ask could be whether improvements in efficiency comes at the cost of the quality of the work being done. Unlike efficiency, measuring the *quality* of a task (semi-automated or not) is difficult in video editing.

The difficulty of measuring video editing tasks objectively is discussed in Niu and Liu [2012] and Radut et al. [2020]. For example, let us look at the task of cutting video clips. It is challenging to judge how well a video segment is cut from a video file, or whether one cut is better than another. The subtitling is one of the few tasks where both efficiency and quality of the task can be measured with ease. In our work [Soe et al., 2021], we use word error rate (WER) to measure the quality of the semi-automated subtitling task. WER can be measured by dividing the number of corrections required (substitution, insertion, deletion) by the total number of words. However, the WER measure lacks details on the nature of errors. Therefore, we also discussed flaws of the WER, and pro and cons of other measures in our paper [Soe et al., 2021].

Automation itself is a double-edged sword when it comes to improving performance and reducing workload. In Soe et al. [2021], some participants pointed out that automation can make people "lazy"; that is, they did not check the automated subtitles properly. While not particular to video editing, in the automation literature, the lumberjack effect was confirmed by meta-analyses [Onnasch et al., 2014]. What is meant by the lumberjack effect is that a higher degree of automation improves performance when automation functions as intended. When automation fails, however, the performance degrades more as a result of a loss of situation awareness in humans due to the use of automation. While this lumberjack effect might serve as a warning about introducing automation to video editing tools, automation's impact in this very complex domain of video editing is largely unexplored. Unlike other rigid workflows, video editing workflows are often flexible processes, and they do vary depending on who is editing the video, where it is being worked on, and what type of video it is. AI and Machine Learning can be seen as both a challenge and opportunity from HCI and UX practices [Dove et al., 2017]. In a very flexible and creative workflow such as video editing, the balance between "control"

and "automation" is harder to find.

1.6.1 Semi-automation challenges in video editing

Based on the five research questions of this thesis and the challenges of semi-automation detailed in the literature discussed so far, specific challenges to this thesis can be discussed. These challenges serve as an extrapolation of the research questions into more general questions about semi-automation in video editing tools. The challenges involved in using semi-automation to create intelligent video editing tools can be summarized as follows:

- Understanding existing video workflows
- Mapping and comparing existing work on intelligent video editing tools
- Understanding video professionals' perception and needs for automation
- Exploring the different applications of AI in creating tools for intelligent video editing
- Designing and creating prototypes of intelligent video editing tools
- Understanding the impact of automation by evaluating the prototypes
- Summarizing user experiences and identifying implications for the further work in intelligent video editing tools.

1.7 Overview of the thesis work

This thesis has three research papers, each addressing different research questions. This section contains brief overviews of these papers. In addition to these three papers, two more research papers were produced during the thesis. Those two papers are not included in the thesis, but were briefly discussed in Preface.

AI video editing tools. What do editors want, and how far is AI from delivering them? [Soe and Slavkovik, 2021] There is a need to synthesize the existing literature on using AI to assist with video editing and to understand the needs of video professionals to inform further work on intelligent video editing tools. Before engaging in finding solutions regarding the automation of different tasks in the video workflow, it is important to understand not only what we can automate, but also what users would

like to have automated. The users are, of course, the human video editors who will use the semi-automated video editing tools. In the first included paper [Soe and Slavkovik, 2021], we did exactly this; we identified and summarized existing literature on intelligent video editing solutions, we surveyed 13 video professionals about what their expectations and requirements for automation, and we specified what automation requirements have been met and how to work towards meeting the remaining unaddressed needs.

This work [Soe and Slavkovik, 2021] makes three main contributions to the thesis. First, it lays the foundation for this thesis by summarizing the current state of the art in intelligent video editing tools and identifying themes in this field of research. Second, to address the question from the paper's title, a survey of 14 video editors from the industry was conducted, and the responses were summarized. The survey results clarified the participants' expectations of AI, the need for automation in their workflows, and their opinions on automation in video editing. We used the summary of the current work on AI in video editing and the survey results to inform further work on intelligent video editing tools and suggested what AI technology might be suitable for some video workflow tasks.

The remaining two papers address the semi-automation of two different video workflow tasks, subtitling and cropping and panning. The first work contains both quantitative and qualitative evaluations, and while in the second work employs only qualitative measures.

Evaluating AI-assisted Subtitling [Soe et al., 2021] address semi-automated subtitling. This empirical work is motivated by the lack of focus on the human factor and human input in automated subtitling tools. With the fact that speech-to-text technology is not completely error-free, a semi-automated subtitling tool where human users correct the machine errors is proposed. The work evaluates two key hypotheses. i) Assisted subtitling helps novice users create more accurate subtitles, and ii) assisted subtitling helps novices users make subtitles faster. This work also addresses the efficiency and effectiveness of semi-automated subtitling compared to baseline subtitling, how the UX changes when automation is introduced, and what are and how to solve the usability issues caused by automation in facilitating subtitling.

This paper [Soe et al., 2021] makes two key contributions. The first one is a quantitative exploration of the impact of introducing automation into a subtitling workflow. This was achieved by building an AI-assisted subtitling prototype on top of a production-grade video editing tool and running an experiment with 24 participants. The measured data from this experiment proved that the AI-assisted subtitling prototype led to more accurate and faster subtitling. The second is identifying the changes in UX and discovering new types of interactions when automation is introduced. This was achieved by collect-

ing data regarding the user experiences from our experiment and using thematic analysis to identify UX issues. Potential solutions to the discovered issues are also suggested. Another contribution of this work is that it highlights the importance of user experience and using human intelligence and reasoning to fix the errors of AI. We also used and commented on the suitability of using word error rates (WER) as a qualitative measure to access verbatim subtitles.

Another important contribution of this work [Soe et al., 2021] is the investigation of how semi-automation or having to work with automation in subtitling, changes the performance, behavior, and experiences of novice users. The results of the experiment confirm that the assisted subtitling prototype with speech-to-text enables the novice users to create slightly more accurate subtitles much more quickly. However, the users rate the experience with assisted subtitling as being more difficult than starting from scratch. This paper also addresses the UX challenges involved in introducing automation into the subtitling workflow, and the usability problems that need to be addressed for efficient human–machine collaboration in subtitling. In addition, the possibility of retraining the state-of-the-art machine learning based speech-to-text systems with user corrections in subtitling is discussed.

A content-aware tool for converting videos to narrower aspect ratios [Soe and Slavkovik, 2022] explores using five AI models to support the cropping and panning task. ML-based visual perception uses ML to imitate human visual perception. However, ml-based visual perception models are specific to a particular task such as detecting faces, detecting texts, or detecting salient regions (areas of interests in a video frame). In our work, we use these different ML models together to create idioms-based video retargeting workflow.

The semi-automated cropping and panning task in our paper is presented as video retargeting. Video retargeting in our context is defined as creating a different aspect ratio of the same video. For instance, converting a video created for a 16:9 wide screen format to a 1:1 square format to upload to Instagram is considered a video retargeting problem. The main intention of this work is to explore a new interactive cropping and panning workflow using ml-based video processing.

This work [Soe and Slavkovik, 2022] explores how we can design semi-automated cropping and panning using ML-based video processing, what are the design challenges, and the user experience issues. We employed six cinematic idioms for the users to control cropping and panning in this work. We performed a user study with 5 users to determine the feasibility and to explore user experience issues with the tool. We analyzed responses from the user evaluation to explore the challenges involved in using ml-based

perception to create a semi-automated cropping and panning tool. In addition, we performed a review of the output of the tool with a professional video editor and made a short comparison with results of manual retargeting by the video editor.

The contributions from [Soe and Slavkovik, 2022] are as follows. First, we presented a design and implementation of a prototype that uses semi-automation to perform cropping and panning. Second, we confirmed the feasibility of the tool for exploring different possibilities of video retargeting by performing a usability evaluation. Lastly, we identified which areas of the proposed approach should be improved based on the results of the evaluation. In this study, we reported issues concerning the ML models used, the user interactions created, the interface design of the tool, and what we have learned from the user study to inform further work on semi-automated cropping and panning.

The three papers included in this thesis focus on the interaction and interplay between human users and semi-automated tools in the context of editing videos.Throughout the thesis, we emphasized the importance of user experience and interactions, as the ultimate intention is to explore the use of semi-automation to create useful and better tools for video editing workflows.

1.8 Structure of the thesis

The remainder of this thesis is structured as follows. In Chapter 2, background topics on HCI, AI, user experience with AI, video workflows, and intelligent video editing tools are discussed in details. Chapter 2 also contains a comprehensive review of the literature to situate the contributions of this thesis in the literature and to help understand the three papers in this thesis.

Chapter 3 discusses the research methods used in this thesis and how various research methods fit together in our works. For each research method utilized throughout the thesis, the following details are provided: the definition and summary of the method, the rationale for why a particular method is used, how they are used, and instances where a method is employed.

The next three chapters comprise the three papers included in this thesis. They are: Paper I: “AI video editing tools. What do editors want and how far is AI from delivering” in Chapter 4; Paper II: “Evaluating Assisted Subtitling” in Chapter 5; Paper III: “A content-aware tool for converting videos to narrower aspect ratios” in Chapter 6.

Chapter 7 is the final discussion and the conclusion of this thesis. In this chapter, the

answers to the research questions are summarized and the contributions this thesis has made are presented. In addition, the lessons learned from the attempts to answer the research questions are discussed. This chapter and the thesis are concluded with sections on the future work made possible by this thesis and limitations of the thesis.

Chapter 2

Background

This chapter places the works in this thesis in the scientific context and reviews relevant research. In addition, it gives an overview of the research areas, concepts, definitions, methods, and tools that are relevant for understanding the remaining chapters of the thesis. It starts with a comprehensive summary of intelligent video editing tools. Afterwards, topics that are foundations of intelligent video editing tools are discussed. The first foundation is AI, the emphasis is on AI relevant to videos and video editing. The second foundation is HCI. It includes an overview of human–computer interaction (HCI), user experience (UX), and HCI's & UX's perspective on designing tools that includes human–AI interactions. In addition, topics more specific to the thesis such as subtitling, semi-automated subtitling, video aspect ratios, and video retargeting are discussed. The last topic is Viz Story, the web-based video editing tool from Vizrt, that was used to built the prototype in Chapter 5.

2.1 Intelligent video editing tools

All the popular video editing tools today are created for *non-linear editing* (NLE). Non-linear editing is defined as a form of "video editing that does not require the sequence to be worked on sequentially" [Okun et al., 2015]. Examples of some popular NLEs are, Adobe Premier Pro[1], Avid Media Composer[2], DaVinci Resolve[3] and OpenShot[4]. In addition to supporting non-linear editing, video editing software usually contains features such as subtitling and captioning, color correction and grading, graphics and

[1]https://www.adobe.com/products/premiere.html
[2]https://www.avid.com/media-composer
[3]https://www.blackmagicdesign.com/products/davinciresolve/
[4]https://www.openshot.org/

animations, video effects and transition, and audio adjustments. NLEs are powerful and mature software for editing video, but they do not solve the high skill requirements or labor-intensiveness of video editing.

As stated in the previous chapter, intelligent video editing tools are video editing tools that employ semi-automation to support users in video editing workflows. Intelligent video editing tools exists between two approaches to video editing. On one hand, there is non-linear editing (NLE) software where videos must be edited by the users at frames. On the other hand, there is completely automated video editing, in which the entire editing process is done by automation. In contrast to completely automated video editing, intelligent video editing tools shares the editing tasks between users and automation. As intelligent video editing tools uses semi-automation, some tasks are automated while the rest of the tasks are placed under the control of the users. In the next section, we provide a comprehensive summary of intelligent video editing tools and discusses important topics in these tools, such as shared control of tasks between user control and automation.

2.1.1 Fully automated video editing

It is a challenging task, using current AI technology, to completely automate video editing. Currently, fully automated video editing is limited to simple and narrow editing, such as creating highlights or video mashups. Here, computation is used to automatically process one or more video clips into a modified video without any user input. The rest of this paragraph will cover automated highlight generation. One popular example of AI being used to create highlights is "Made by Machine When AI met the Archive"[5]. In this work, the BBC created short complications from the BBC archive of 270,000 programs. This work employed AI technologies to create highlights from the BBC archive, namely, object and scene recognition, subtitle analysis, and visual energy. For another example, this work [Wu et al., 2020] presented a system for automated editing of corporate meeting videos. This work employed heuristics from how a human editor edited these type of videos and uses two attention models (for audio and video) to automatically edit meeting videos.

Generation of mashups is another area where completely automated video editing has been applied. A mashup is a video compiled from a combination of video clips from different sources about a single event or topic. For example, Virtual Director [Shrestha et al., 2010] is an automation method created for mashups of concert recordings. The mashup generation rules in Virtual Director was created by interviewing video editors

[5]https://www.bbc.co.uk/rd/blog/2018-09-artificial-intelligence-archive-made-machine

and film literature. These rules are then translated to computable rules, and a rule-based optimization method is used to select recordings and create a synchronized compilation of a concert. There are other mashup generation methods such as Hua et al. [2004] for creating highlight of home videos that matches a music, MoViMash [Saini et al., 2012] for concert recordings, Jiku Director 2.0 [Nguyen et al., 2014] for mobile videos, and WeMash [Wang et al., 2014] for online web videos about events.

Another use of completely automated video editing is that using AI to automatically edit live events' coverage. The methods used in automatically editing live events and computationally measuring the quality of automatically edited videos from these systems is discussed in Radut et al. [2020]. In the same work [Radut et al., 2020], an interesting concept of "good enough" is discussed. Though the automated live editing of events might be inferior to a production being done by a skilled crew, some events might not have the means to hire a crew for live broadcast. In such cases, automation can be deployed for live editing, but it has to be of sufficient quality to use it over just a simple single camera setup. Discovering the quality requirements necessary for automation to considered *good enough* to use it over simple solutions (i.e. without automation) can be the key to refining live automated editing [Radut et al., 2020]. This concept of being good enough can be the key to understanding requirements for other types of automated video editing, such as highlights and mashups.

In contrast to intelligent video editing tools, fully automated video editing does not require any user input, human oversight, or feedback during the video editing process. The lack of user input means that user experience issues in fully automated video editing is limited to perceived quality of the edited video.

2.2 Semi-automating video workflows

The key problem studied in this thesis is how can we use artificial intelligence (AI) to support video editing? Full automation of video editing is limited to the very narrow applications as described in the previous section. Semi-automation in video workflow can be used to make easier, faster and more accessible video editing tools for the human editors, but this too is not without challenges. The key challenge of semi-automation in video editing workflow is human–AI interaction issues in the context of video editing. This is because semi-automation meant that AI is embedded into the video editing tools. Human editors who use these editing tools are required to interact with AI assistance or output of an AI model in their video editing tools. In this section, the topic of semi-automated video editing tools is introduced, followed by the challenges in this field.

Creating better video editing workflows through some form of semi-automation is a decades old problem. In this thesis, semi-automated video editing tools will be referred to as *intelligent video editing tools*. Intelligent video editing tools are video editing tools that employ semi-automation to create easier and faster video editing workflows. The current body of literature on intelligent video editing tools consists mostly of works that are created for editing only a specific type of video. For example, a tool for placing cuts and transitions in interview videos, proposed by Berthouzoz et al. [2012], is created for editing of monologue interview view videos. A tool for editing instructional videos was presented by Truong et al. [2016]. Another example is a tool for editing dialogue-driven videos [Leake et al., 2017], where the tool requires the video script of dialogues and videos has to be different takes of the script. A possible reason for having video specific tools is video editing requirements and workflows vary depending on the type, the purpose, and the context of the video being edited. It is also a lot easier to create workflows for a specific type of video.

One of the most common similarities among different intelligent video editing tools is to remove the need to do frame by frame adjustment by offering manipulations from a higher level of abstractions (shots, scene, dialogs, words, etc.). For example, Silver [Casares et al., 2002], provides smart manipulation of video where the users can edit with clips and shots. In addition, this tool enabled abstract views of video editing. It provided these functions by using the metadata created for the videos. Another example is Roughcut [Leake et al., 2017] where the users can create edits by just selecting a set of available video editing idioms.

Editing text is much easier than editing videos, and for this reason, text has been used as a proxy to edit videos. How to edit the video automatically as using the changes in the corresponding text is explored in many research works [Berthouzoz et al., 2012; Wang et al., 2019] and in the industry tools - Descript[6] and AutoEdit[7]. A summary of intelligent video editing is presented in Table 2.1.

Table 2.1 lists previous works on intelligent video editing tools and classify them in terms of video type, abstractions used, and tasks that are semi-automated. Video type is the type of the video that the tools are created for editing. For instance, the tool proposed by Pavel et al. [2014] is for editing video lectures. Abstraction used describes the unit of video that the users can edit the video with. In traditional NLEs, the basic unit of editing is frames, and video editing is done by manipulating the (a group of) frames in a video. However, by mapping a group of frames to spoken words, editing can be done by manipulating words [Berthouzoz et al., 2012]. Tasks semi-automated are the tasks

[6]https://www.descript.com
[7]https://pietropassarelli.com/autoEdit.html

that are supported by AI in each of the tools. The most common task that is supported across the tools is cutting and composing video segments.

Work	Video type	Abstraction used	Tasks semi-automated
Casares et al. [2002]	Videos with annotated metadata	Clips, shots, frames	Cutting segments, joining segments
Leake et al. [2017]	Videos taken for a dialog script	Shots, dialog lines, editing idioms	Cutting segments, composing segments, selection of alternative video takes
Chi et al. [2013]	Video tutorials with voice annotations	Annotated steps in a tutorial	Cutting segments, composing segments
Berthouzoz et al. [2012]	Interview videos	Spoken words	Removing undesired video segments, Cutting and composing segments
Truong et al. [2016]	Narrated videos with audio annotations	Lines of sentences in narration, static and kinetic segments	Cutting segments, composing segments
Pavel et al. [2014]	Video lectures	Chapters, sections	Creation of chapter and sections summaries
Wang et al. [2019]	Video montage	Keywords, editing idioms	Cutting segments, composing segments
Descript.com	Videos with narration	Words, Sentences	Cutting segments, composing segments
Passarelli [2019]	Video interviews	Words	Cutting segments, composing segments

Table 2.1: Summary of intelligent video editing tools

Which tasks are automated and what the users are required to do varies across the intelligent video editing tools reviewed in this thesis. Answers are often framed by the purpose of the tool, the type of users video the tool was created for, and limitations of the AI that was being employed. For example, in Democut [Chi et al., 2013] users are only allowed to: create and edit annotations, apply different effects to segments, modify any visual effects, edit subtitles, resize the cropped region, or add/delete highlights. Democut is created for editing only single take demonstration videos. To elaborate, the input is a single video file and output is a shortened, edited version of that single video file. The task of applying effects, applying transitions, removing silent-segments, adjusting segment boundaries and compiling the video is done by automation.

Here is another example to demonstrate the differences in allocation of tasks between users and automation. In Roughcut [Leake et al., 2017], the only method of manipulating the video for the users is by applying a set of video editing idioms from 13 cinematic

idioms made available to the users. Segmenting the videos, labelling them and selecting the best segments for the user provided list of idioms is done automatically. In Chi et al. [2013] and Leake et al. [2017], frame-level adjustments are not possible at all. However, they do allow exporting the results in edit decision list (EDL) files, which can be used to further fine tune the edit on other NLEs. EDL is a popular file format for NLE that can encodes the results of segmenting video sources files (with timestamps) and compositions of video segments on a timeline.

In some other tools, there is more user control, or to put it in different words, the level of automation is lower. For example, in Quickcut [Truong et al., 2016], the user can do frame level selection of the footage for segmentation of clips as well as select the order segments appear. What is being automated in Quickcut is suggesting relevant footage for story segments and selecting suitable cut point for transitions and applying aesthetic pleasing cinematic effects in transitions. In Silver [Myers et al., 2001], the automation is limited to providing abstract views of the video (i.e. text transcript, storyboard, subject, outline and source views) and smart selection of clips using shots and scene boundaries.

Intelligent video editing tools from industry. In contrast to tools proposed in research, limited information is available on the tools from the industry. The most common type of intelligent video editing offered in industry is "editing videos by editing text". It is an application where users can edit the corresponding text aligned with the video and the video will be automatically edited to reflect the changes in the text. For example, both autoEdit[8], Descript[9] supports editing video by editing text. Rev.com[10], a popular online service for subtitling and transcripts, also allows selection of video by selecting text. How does editing videos by editing text works? First, the text transcript (which can be automatically generated with speech-to-text or human labor) is aligned to the video automatically. The alignment of the transcript to video is done using speech to text. A method for aligning text to video using speech-to-text is described in Huang [2003]. The users are presented with both the video and the transcript of the video. By cutting and moving the text in the text transcript, the video will be edited accordingly.

There is a commercial product from Adobe Experience Cloud[11] that offers users ML-based area of interest detection for cropping and panning of images and videos. However, how the area of interest is computed is not available to the public.

One of the biggest challenges in semi-automation of video workflow is the way users perceive and use the tool can change when automation is embedded into the existing

[8]https://pietropassarelli.com/autoEdit.html
[9]https://www.descript.com/video-editing
[10]https://www.rev.com/blog/edit-videos-faster-pull-selects-from-your-transcripts-with-an-edl
[11]https://business.adobe.com/products/experience-manager/assets/smart-crop.html

workflow. We found out that users interact with the same subtitling interface very differently with and without automation in our semi-automated subtitling work [Soe et al., 2021]. Therefore, examining user experience for semi-automated video editing tools is necessary. Understanding user experiences in intelligent video editing tools and thus creating better tools is limited by the lack of literature on this topic and the lack of attention on user experience issues in works that discusses intelligent video workflows.

Simply plugging semi-automation into a video editing tool does not help the human video editor to understand, trust and utilize what AI could do and correct when it fails to deliver. So, what are the challenges for semi-automation in video workflow? Several HCI studies Dove et al. [2017]; Yang et al. [2020] had attempted to map out key human–AI interactions design challenges in creating new applications with AI.

Easy to use and efficient user interfaces and interactions are necessary to successfully create semi-automated video workflows. According to Dove et al. [2017], the challenges of designing UX with automation are the lack of understanding of ML in UX community, the data dependent nature of ML black-boxes and the difficulty of making interactive prototypes with ML. Additionally, there is the automation vs control trade-off, in which giving users more creative control over video editing workflow might come at the cost of a loss in efficiency.

Since video workflows can be very different depending on the type of video being edited, it is difficult to build upon the literature on intelligent video editing tools as they are created for different types of videos and thus have different workflows. Instead of creating semi-automated video workflows for every different types of videos, semi-automation can be introduced for individual video tasks. Semi-automation of individual tasks will allow each tasks to be executed in the order desired by the editor, leading to more flexible tools. On the other hand, automating the entire video workflow means that the automation will dictate the order the tasks are executed, which might not be desirable by the users[12] in most cases.

Task level semi-automation also enables measurements and evaluations of the task and makes possible the study of the changes that might occur in a task when automation is introduced. In Soe et al. [2021], we studied semi-automation of subtitling task, and studied the impact of introducing automation into the workflow using an AI-embedded prototype. Moreover, current ML techniques are only applicable to a single task (narrow AI). Therefore, automating the whole video editing process is a task that is not suitable for current approaches in machine learning-based automation, as it is difficult to create a dataset and ML-based automation for the correct way of editing a whole video.

[12]the users who are already using some type of non-linear video editing tools

2.2.1 Evaluation methods used in intelligent video editing tools

How were the existing work on intelligent video editing tools evaluated, and what were the results? The evaluation method used and results of the study are summarized in Table 2.2. The most popular evaluation method in use is *output evaluation*. In output evaluation, intelligent video editing tools were used to edit a few videos and the output videos from the tool is evaluated. A better evaluation method used in the studies is *user study* found in two of the works. This method is better for studying UX, as a user study can reveal the user perception of the tools and possible user interface issues.

Work	Evaluation	Results
Casares et al. [2002]	Pilot user study	Number of participants: seven Study design: within-subjects Treatment: one with smart editing and one without Results: no significant differences in editing time
Leake et al. [2017]	Output evaluation	Using a professional video editor to edit eight scenes. Using the system to produce edits for the same set of scenes, and compare the results.
Chi et al. [2013]	User study	Number of participants: eight Treatment: none Task: create a how-to video Results: measure for time taken and perceived qualities of videos
Berthouzoz et al. [2012]	Output evaluation	Using the tool to create transitions for five videos Results: computational processing time and informal input from journalists
Truong et al. [2016]	Output evaluation	Using the tool to compose five videos from five sets of video takes Measures: time taken to create videos using the tool Results: informal feedback from ten novices and one video editor
Pavel et al. [2014]	Output evaluation	Using the tool to create four video summaries and manually edit the summaries Measures: evaluate time taken to refine the summaries against time taken for manual edit
Wang et al. [2019]	Output evaluation	Using the tool to create 20 video montages of five types. Evaluate the visual and text matching quality and quality of shot assembly. Comparison of one video with professional editing.

Table 2.2: Summary of Evaluation methods used in intelligent video editing tools

Most of the tools use output evaluation as the sole evaluation method. And the key problem with reporting just output evaluation is the evaluation of the process and user

experience is ignored. For instance, Leake et al. [2017] presents the output of 8 dialogue-driven scenes and compared the output videos from their tool with editing done by a professional video editor. Performing only *output evaluation* meant systematic reporting of usability issues and user feedback is not published. If the evaluation results are just "this tool was used to create good videos", the contribution of the study towards strengthening the knowledge on human–AI interaction in video editing tools is limited. Only one study [Casares et al., 2002] uses experimental treatment with within-subjects design. Usage of experimental treatment can expose the influence of the automation into the tool by comparing the results with automation and without automation among the users.

2.3 Artificial Intelligence for Video and Video Editing

Intelligent video editing tools are tools that utilize both AI and human input to create semi-automated video workflows. In this section, we will present an overview of AI technology relevant to video editing.

Artificial Intelligence (AI) is loosely defined as intelligent behavior in computers, and perception is one of such intelligent behaviors [Nilsson, 1998]. Video is a time-based medium containing both audio and visual information. Therefore, perception of visual and auditory signals is what is of interest when it comes to video editing. Besides perception, natural language processing, making computers *understand* language, is also an important AI technology for semi-automating video workflows as most of the videos contain human speech.

The state-of-the-art breakthroughs in AI for computer vision and speech has brought on by machine learning and deep learning in particular [LeCun et al., 2015]. Machine Learning (ML) is how a machine can improve its future performance based on inputs or in response to external information [Nilsson, 1998]. Unlike humans who can learn from a few examples, machines require many examples (from hundreds to millions of examples) to learn from. Such examples, known as datasets, are essential for research in machine learning to both develop and evaluate new methods. One of the important dataset for computer vision is ImageNet [Deng et al., 2009], which has 3.2 million images of different classes. This dataset enabled many break through works in image classification such as Krizhevsky et al. [2012] which trained a deep convectional neural networks for image classification. As datasets play such an important role in development of AI, relevant datasets will be introduced when discussing AI techniques.

The breakthroughs in AI for visual and audio processing meant that we can automatically process the content of the videos. Processing the content of the video provides a lot of utility for creating intelligent video editing workflows. For example, face recognition is used in these works [Leake et al., 2017; Soe and Slavkovik, 2022; Wang et al., 2019] to detect the speakers in a video and their facial actions. Based on the faces and facial actions detected, abstract editing decisions such as making "speaker visible" can be automated. Furthermore, using AI to suggest music suitable for a video [Lin et al., 2021] or judge the quality of a video clip [Niu and Liu, 2012] can be used to suggest sound effects in video editing or rank video takes according to the computed quality of the shots. The discussions on AI techniques of interests in visual perception and generating videos with AI is expanded in the rest of this section.

2.3.1 AI for computer vision

AI has been used to emulate our vision to "see" things in the world via images and moving images captured by cameras. However, unlike vision in nature, AI methods for computer vision are created for each narrow specific tasks. For example, a facial recognition model will only recognize faces and will not work for seeing anything else. AI for visual perception, widely known as computer vision, has received a lot of attention recently. In this subsection, we will provide an overview of computer vision techniques relevant for this thesis, namely, face detection, object detection, object tracking, scene detection, sentiment analysis, video reasoning, and video captioning. An introduction to each of the topics and overview of the work in these topics are discussed.

Face detection is using machine learning to detect faces in images or videos. An open-source toolkit named OpenFace 2.0 [Baltrusaitis et al., 2018], is created for detecting faces, facial landmarks (eyes, nose mouth and face shape), gaze and head orientation. In addition, OpenFace 2.0 model can also recognize 20 facial actions units (e.g. blinking, jaw dropping, lip sucking). The techniques and datasets used for detecting face and facial actions are listed in the original paper [Baltrusaitis et al., 2018]. One of the most popular datasets for face detection is Labeled Faces in the Wild [Huang et al., 2008] (which is also one dataset used in creating Openface 2.0) consists of 13,233 images of 5749 people taken from the web with labels that can be used to train face detection algorithms. Another popular dataset for face detection and recognition is VGGFace2 [Cao et al., 2018]. It consists of 3 million images of 9131 subjects.

Object detection is detecting the presence of objects and classifying the types or classes of these objects. Object tracking takes the object recognition further and tracks the movement of objects across frames in a video. Two of the most popular methods of

object detection and tracking are You Only Look Once (YOLO) [Redmon et al., 2016] and Single Shot multibox Detector (SSD) [Liu et al., 2016]. YOLO is a real-time capable object detection method that uses a single neural network for both predicting bounding boxes and class probabilities, instead of using one neural network for each of the tasks. SSD is also a single shot detector (that uses just a single neural network like YOLO) and it works slightly faster than YOLO. Prominent datasets for object recognition are ImageNet [Deng et al., 2009], The Pascal Visual Object Classes (VOC) [Everingham et al., 2010] and Microsoft COCO: Common Objects in context [Lin et al., 2014]. Microsoft COCO [Lin et al., 2014] consists of 2.5 million labelled objects in 238,000 images.

Scene detection is finding semantically or visually related segments (scenes) in a video. A shot is an uninterrupted sequence of images from a single camera take [Okun et al., 2015]. A scene is thus defined as a collection of shots that share a common setting or theme [Okun et al., 2015]. With this definition of shots and scenes, scene detection is usually done by grouping the shots. Baraldi et al. [2015] combined scene detection method that shot detection using Shot Transition Graph followed by scene detection using hierarchical clustering for re-using old broadcast videos. MovieScenes dataset [Rao et al., 2020] is a large-scale scene segmentation dataset which contains 21K scenes derived by grouping over270K shots from 150 movies. In the same work [Rao et al., 2020], a local-to-global scene segmentation framework is presented, and the framework achieves better results than other segmentation models on the MovieScenes dataset.

Sentiment analysis in videos is the task of identifying the sentiment conveyed in a given content. Since video is a dual-tracked medium, multi-modal analysis is the most common method for identifying sentiments. Morency et al. [2011] present a proof-of-concept tri-modal (audio, visual and language) approach to sentiment analysis of videos and a small dataset of clips and sentiments from 45 videos from YouTube. Poria et al. [2015] present a multi-modal sentiment analysis on short video clips using features from text, visual and audio to train a classifier for sentiment analysis.

Visual reasoning is using machine intelligence to reason about temporal and casual events from videos. Yi* et al. [2020] created a dataset called CLEVER (Collision Events for Video Representation and Reasoning) which contains videos and four types of questions: descriptive, explanatory, predictive and counterfactual. They also create a benchmark based on their dataset and evaluated various visual reasoning models on their benchmark. The dataset contains videos of different objects of shapes and colors moving and colliding.

Video captioning is generating descriptive, natural sentences to capture the dynamics in videos. Wu et al. [2016] create a common architecture for video captioning using sequence learning. This work also listed the datasets of video captioning and the biggest

dataset is MSR-VTT-10K [Xu et al., 2016] which contains 10k web video clips with 200k sentences describing these clips (each clip is annotated with 20 natural sentences). The Kinetics human action dataset [Kay et al., 2017] consists of 700 classes of human actions and at least 700 videos for each action. It can be used for developing methods to detect human actions in videos.

Video quality. In the first paper of this thesis [Soe and Slavkovik, 2021], we identify automation needs from the industry for video editing supporting tasks such as filtering out bad takes or bad quality videos. There are some datasets that can be used to support those tasks. These are briefly listed here.

Video blooper dataset [13] consists of 600 monologue videos of individuals talking to a fixed camera and labels for blooper and non-blooper videos. The goal of this dataset is to allow detection of blooper (bad takes) so that they can be filtered out without having to perform editing. AutomEditor [14] is created based on the video blooper dataset and multi-modal utterance level analysis technique from this paper [Deng et al., 2018]

Video quality assessment is another area AI has been used to judge the quality of videos. Ying et al. [2020] created a dataset of user-generated video containing 39,000 real world distorted videos and 117,000 space-time localized video patches ('v-patches'), and 5.5M human perceptual quality annotations. In the same paper, they also presented two new approaches to train video quality assessment models based on the data-sets.

Video summarization is creating a concise version of an original video containing the most interesting or important parts from the original. SumMe dataset [Gygli et al., 2014] consists of 25 videos covering holidays, events and sports and allows benchmarking of video summarization methods. In the same work, superframes-based [15] summarization for user videos is also presented. There is another benchmark for video summarization called, TVSum [Yale Song et al., 2015] consists of 50 videos and 1000 annotation of shot level importance scores.

Text detection or optical character recognition (ocr) is converting images of text into encoded text. OCR can be useful in video editing as readable text in the video can usually be an object of interest, or it can be used to detect text graphics in a video. OCR techniques like many other computer vision approaches is solved using hand-crafted methods in the early days [Mori et al., 1992]. However, like many computer vision challenges, OCR also became of the problem that is best solved with deep learning and neural networks [LeCun et al., 2015]. The state-of-the art OCR in this work [Lee and

[13] https://www.kaggle.com/toxtli/video-blooper-dataset-for-automatic-video-editing
[14] https://github.com/toxtli/AutomEditor
[15] A superframe is a frame where the video can be cut into segments

Osindero, 2016] used convolutional neural networks for image encoding and recurrent neural networks for language modelling. A well-known dataset for OCR is International Conference on Document Analysis and Recognition (ICDAR) Robust Reading Challenge. The ICDAR Challenge 4 dataset [Karatzas et al., 2015] consists of 1,670 images.

Saliency is the quality of being noticeable or prominent. Saliency prediction in images is the problem of detecting the areas that are most likely to attract the attention or interests of a viewer. As explained by Kummerer et al. [2017], saliency prediction is the task of predicting fixation locations given the image the observer is viewing. Saliency prediction has been used to crop images for creating gallery views on social media that is to create representative thumbnails of full-sized images. The Deepgaze II [Kummerer et al., 2017] model used in our work on semi-automated panning [Soe and Slavkovik, 2022] is trained using the Saliency in Context (SALICON) dataset, [Jiang et al., 2015] which contains human "free-viewing" data on 10,000 images from the Microsoft COCO dataset.

2.3.2 AI for video editing and generating videos

There are AI techniques that are developed for extracting video editing rules. Matsuo et al. [2002] present data mining techniques to classify editing patterns in terms of three types of shots (loose, medium, tight shots) from videos with the goal of creating reproducible editing patterns. Earlier work by Butler and Parkes [1997] presented a rule- and query-based approach to automate video editing, whereby rules were developed from cinematic theory. Automated video editing by modelling the editing process and using semantics is presented in Nack and Parkes [1997]. In more recent work, Wu et al. [2020] present a system for automated editing of meeting videos using faces, poses and gazes data of people in the video.

AI techniques have been developed to manipulate or synthesize video as well. Some of the earliest work on this topic, Video Rewrite [Bregler et al., 1997], uses existing footage of a person to automatically create a video of said person speaking to a different audio track. This work was done intending to facilitate movie dubbing. AI synthesizing videos from existing footage became popular once again after discovering techniques to train deep neural networks to synthesize fake videos — the videos created this way are known as *deep fakes*. According to [Mirsky and Lee, 2021], a deep fake is a content generated by AI that is authentic according to a human observer. Deep fakes mostly received negative concerns in the media, however, they could also have potential applications in generating or adapting video content.

Borgo et al. [2012] summarized the methods available for generating video-based computer graphics, which has a lot of potential application in adaptive graphics and applying editing effects in video editing. Another form of automated video editing is improving aesthetic qualities of videos such as motion stabilization, shots removal and color adjustment [Choi and Lee, 2015]. Bai et al. [2009] propose a background cut out method for videos using multiple local classifiers.

What do people want from AI in video editing? The opinions on automation and requirements from automation in video editing have never been reported in research studies. One study [Girgensohn et al., 2001] explores the balance of automation and user control and reports a finding similar to lumberjack effect (i.e. when automation doesn't work as expected, it degrades performance more). They reported negative experiences when people can't understand the workings of automation and frustrations from inability to overwrite the automation.

2.4 HCI and User Experience with AI

Intelligent video editing tools should be approached with the Human–Computer Interaction (HCI) perspective. HCI, in particular the user interfaces designs and user experiences in video editing tools with AI/automation, is an essential to ensure that automation helps the users more than it hinders them in intelligent video editing tools. In this section, overview of HCI and user experience (UX) topics in human–AI interactions presented.

2.4.1 Human Computer Interaction

Human–computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them [Hewett et al., 1992]. HCI is an interdisciplinary area and for the context of this thesis we use the perspective of HCI from computer science discipline. Human (the user) is the first and most important part of HCI. The second factor is the computer technology which ranges from screen, keyboard, mouse interfaces to touchscreen, virtual reality, automation and beyond. HCI focuses on the interaction between one or more humans and one or more computational machines. Bødker [2006] presented the evolution of HCI from second wave, with focus on work settings and established practices to third wave, which is broader in context, application types, and it includes cultural, emotional aspects. The third wave of HCI is somewhat

similar to the expansion of AI research in that AI research has also broadened up towards non-technical aspects with the wide-spread adoption of AI.

Usability is defined in ISO 9241-11 as "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". By that definition, when evaluating usability, the goal should be to emphasize effectiveness, efficiency, and satisfaction. On the other hand, *user experience* (UX) is a wider term. User experience is defined in the same document (ISO 9241-210) as "a person's perceptions and responses that result from the use and/or anticipated use of a system, product or service". Turner [2017, p. 15] proposed UX as an ad hoc category with three core attributes namely, involvement, affect, and aesthetics. Involvement can be, for example, the daily use of digital products. Affect includes emotions, feeling and moods. Aesthetics is the appearance of a digital product.

2.4.2 User Experience in semi-automated and AI-embedded tools

Simply plugging AI into an existing video editing tool does not help the user to understand and utilize what AI could do and what it could not do. Study of human factors for automation is necessary to explore how to help users understand and use automation effectively. These human factors are usability, user interaction designs, and user experience. In addition, affordances provided by automation allows designers to create new type of interactions. But what are those new interactions that work, and which of those are desirable for the users? Dove et al. [2017] say in their study that machine learning (ML) is both under-explored opportunity for HCI researchers and has an unknown potential as a design material. The authors also pointed out the challenges in creating a better user experience with AI. The challenges are: the lack of understanding of ML in UX community, the data dependent nature of ML black boxes (i.e. unlike written code, inspecting ML models do not reveal much information), and the difficulty of making interactive prototypes with ML.

There are some previous works that exist which explores the UX and HCI issues of tools with automation. One of the classic example of integrating automation to an existing tool is principles of mixed-initiative user interfaces [Horvitz, 1999]. Mixed-initiative user interfaces are defined as "interfaces that enable users and intelligent agents (AI) to collaborate efficiently", and this concept is demonstrated using automated calendar scheduling from emails using Support Vector Machine (SVM) text classification [Horvitz, 1999]. In more recent work by Amershi et al. [2019], some general guidelines and heuris-

tics for designing AI infused user experiences are presented. The guidelines were created from studies with 49 design practitioners on 20 AI embedded products. The guidelines however understandably focuses on mature ML tasks such as recommendation, search and filtering. The challenges of Human-AI interaction are explored in Yang et al. [2020], centering around two key issues: uncertainties surrounding AI capabilities and AI output complexity.

2.4.3 AI techniques with human input

There are works from the AI community that focuses on how to integrate human input into the automation process. One example of such approach is the human-in-the-loop (HITL) machine learning. HITL machine learning attempts to combines or uses human inputs in improving ML models or fixing mistakes of the ML outputs. Another term coined recently in 2018 is machine-in-the-loop by Clark et al. [2018] where the goal of the system is to improve the ability or performance of human users with machine playing just a supporting role. How do we build efficient human–AI systems? This question requires approaches from both AI and UX Design disciplines. AI techniques that integrate user interactions are discussed in the paragraphs below.

Active learning's key hypothesis is that if a learning algorithm is allowed to choose which data it learns from (curiosity) then it will perform better with less training data. Active learning (known as optimal experiment design in statistics) is well motivated in ML problems where labels are scarce and expensive, but data is in abundance [Settles, 2009]. In active learning, the (machine) learner actively selects or generates a sample instance and request a label from the human to learn from it with the goal of efficient learning with less labelled data. As such, active learning is a human-in-the-loop approach which tries to minimize the number of samples required by asking for the most informative input to be labelled.

There are different types of generating samples in active learning which are: membership query synthesis, stream-based selective sampling, and pool-based sampling. In membership query synthesis settings, the learner generate new samples/queries from the input space. Stream-based selective sampling approach assumes that getting the data instance is free and the learner can sample data and then decide whether to query for label based on informativeness. Pool-based active learning assumes that the data can be collected at once and then from the pool, which has a small set of labelled data and large set of unlabeled data. The main differences between stream-based and pool-based approaches are that the former scans the data sequentially, evaluates each sample individually, while the latter processes and ranks the entire pool.

An active learning approach that might be useful in creating tools (as in software applications) with ML is *cost-sensitive* active learning, which considers that each instance/actions has varying cost of labelling. Value of information approach to this considers estimates of both labelling cost and cost of misclassification. Active learning is difficult with deep learning methods because active learning relies on the ability to learn and update models from small amounts of data. Active learning techniques with deep Bayesian convectional neural networks for image data on MNIST dataset and cancer diagnosis images is presented by Gal et al. [2017] .

Interactive machine learning (iML) also known as mixed-initiative or human-in-the-loop systems — incorporates human input to produce an output or a decision. By putting the human in-the-loop, a human kernel, as defined in Wilson et al. [2015], iML looks for "algorithms which interact with agents and can optimize their learning behavior through this interaction – where the agents can be humans". Human-in-the-loop means a model requiring human interaction which leads to changes in the outcome of an event or process. Human-in-the-loop leverages both human intelligence and machine intelligence. However, placing a human in the loop is not always desirable. In secure system designs, the designers often try to keep human out of the loop as a way to eliminate human errors leading to security vulnerabilities. The framework proposed by Cranor [2008] provides a reasoning framework for human-in-the-loop secure systems design.

HITL ML uses humans to correct inaccuracies of ML algorithms for improving accuracy (correction of the output) and/or and providing new training samples. Both supervised learning and active learning methods can be used in human-in-the-loop machine learning settings. Crayon [Fails and Olsen, 2003], interactive machine learning for training image classifiers using decision trees, is one of the first human-in-the-loop ML and emphasized rapid correction of prediction mistakes by humans. Smith et al. [2018] explored human-in-the-loop topic modelling and also examines user experiences - how users are affected by issues such as unpredictability, latency, trust and lack of control and how to address them. Bias from humans that are in the algorithmic loop is different from that of bias in training data. Holzinger et al. [2018] uses iML with ant colony optimization problem to enhance results provided by travelling salesman problem solver using a gamified interface.

Terms such as human is the loop and machine in the loop are used to indicate the *emphasis on the human side of the loop.* The term "human is the loop" is used to propose a shift in the focus [Endert et al., 2014] in visual analytics field. In human is the loop, the focus is on fitting algorithms into human analysts work processes, sort of human-centered approach to HITL. Visual analytics is the science of marrying interactive visualizations and analytic algorithms to support exploratory knowledge discovery in large datasets. Clark et al. [2018] presented machine in the loop creative writing system and the authors

used the term to clarify that the goal of the system is to improve the ability of humans with machine playing a supporting role. In contrast, human-in-the-loop machine learning includes humans in the process for the sole purpose of training machine learning models by asking humans to provide feedback or improve accuracy.

In certain use cases, it makes more sense to crowdsource the human input. Russakovsky et al. [2015] framework for HITL large-scale object annotation that uses a Markov Decision Process to integrate multiple computer vision models and crowdsourced human input. This framework also includes a trade-off model of desired precision, utility and cost (human). The combination of HITL crowdsourcing with active learning strategy can be seen in Active Crowd Translation system [Ambati, 2012], which aims to reduce the labelling cost of language annotations.Active learning aims at reducing cost of label acquisition by prioritizing the most informative data for annotation (sentence selection), while crowdsourcing reduces cost by using the power of the crowds.

There are also attempts to just make AI more understandable. Opaqueness of the AI is one of the reasons why crafting UX with AI is difficulty [Dove et al., 2017]. Explainable Artificial Intelligence (XAI) is an emerging field in AI to come up with techniques that makes AI are more explainable to human users [Gunning et al., 2019].

2.5 Subtitling and semi-automation in subtitling

2.5.1 Subtitling

A subtitle is text that describes human speech (monologue or dialogue) in a video. A closed captioning, similar to a subtitle, includes both textual description of speech and other audio components of the video. In this thesis, only subtitling is explored as it can be automated using state-of-the-art speech to text models. Subtitles are used to make videos available to viewers in foreign languages, those with hearing impairments and viewers in environments where audio is not accessible (e.g. noisy environments and environments where silence has to be maintained).

Subtitles can be clarified, shortened or rephrased from the original speech. However, in this thesis only verbatim (exactly as spoken) subtitling is used as it is easier to measure the quality of verbatim subtitles using word error rates (WER). Unlike transcripts, subtitles needs to be synchronized with the video and segmented into readable chunks. How the subtitles should be created varies among different languages and regions. There are many guidelines available for those new to subtitling and to make subtitling more con-

```
WEBVTT

00:00:00.000 --> 00:00:04.087
Hi, my name is Mark and with Vizrt and today I am with Patrick from CNN

00:00:04.474 --> 00:00:06.200
Patrick.  Welcome to Bergen.

00:00:06.200 --> 00:00:07.059
 Thank you.
```

Figure 2.1: A simple WebVTT file format

sistent. One example is of such guidelines is the BBC guideline[16] for English language online content. This guideline includes recommendations such as : verbatim subtitles instead of edited ones, and length of subtitles should be no longer than two lines.

There are many tools created for only subtitling, and many NLEs support subtitling. Most subtitling tools support subtitle encoding standards, such as TTML[17], WebVTT (Web Video Text Tracks)[18], and SRT[19]. These subtitle-encoding standards specify how the subtitles should be stored in files and how should they be delivered over the internet. These standards ensure that subtitles can be created with a variety of subtitling tools, and they area all compatible with multiple video players. What does a subtitle file look like? A simple WebVTT file can be seen in the Figure 2.1. In the WebVTT standard, each subtitle is separated by a new empty line. A basic subtitle unit consists of starting and end time of subtitles separated by "– –>" and on the next line, the text content of the subtitle. Additional features of WebVTT includes styling of captions, comments, chapters, and metadata.

2.5.2 Speech-to-text

Automated speech recognition, also known as speech-to-text, is a machine conversion of spoken language (audio waveforms) into text. Earlier speech-to-text models use Hidden Markov Model (HMM). One of the first speaker-independent continuous speech-recognition systems was the SPHINX Huang et al. [1989].It uses a HMM approach and has achieved 93% accuracy on a 997-words recognition task. Earlier methods were limited in the vocabulary that they can recognize (just 997 words for SPHINX). One of the most popular open-source dictionary for training speech-to-text system is the CMU library, [CMU, 2014] which contains 134,000 words in North American English. Though current state-of-the-art speech-to-text methods has much larger dictionaries, recognizing

[16] https://bbc.github.io/subtitle-guidelines/
[17] https://www.w3.org/TR/2018/REC-ttml1-20181108/
[18] https://w3c.github.io/webvtt/
[19] https://www.matroska.org/technical/subtitles.html#srt-subtitles

words outside their dictionaries remains impossible.

Earlier speech recognition techniques requires separate training of three models for acoustic, pronunciation and language. But recently, end-to-end (sequence-to-sequence) training models became the norm. In end-to-end training, all three components are trained together at the same time. One of the first in end-to-end speech-to-text used two recurrent neural networks [Graves, 2012]. Encoder-decoder architecture is one of the most popular end-to-end methods. Using single neutral network with attention-based encoder-decoder architecture, Chiu et al. [2017] achieved 5.6% WER on a dictation task consisting of 15.7K utterances that have added synthetic noise. There are open source end-to-end speech recognition toolkits which enable developers to use and customize speech-to-text systems. One of the open-source toolkit is the ESPnet [Watanabe et al., 2018] allows development of sequence-to-sequence speech recognition systems with optional language model integration. Evaluation of the ESPnet achieves 7.3% WER on WSJ task (artificially mixed speech taken from the Wall Street Journal database) on the best configuration.

2.5.3 Measuring performance of subtitling quantitatively

The common performance measure for speech-to-text and transcription systems is *word error rate* (WER). The WER measure is the word-level edit-distance between two text sequences (can be of different lengths), or the number of changes required to correct all the mistakes. The formula for calculating the WER is given below:.

$$\text{WER} = \frac{I + S + D}{N}$$

I is the number of insertions, S is the number of substitutions, D is the number of deletions, and N is the total number of words in the reference text. Though WER is a common measure, it has its own blindsides, stemming from the fact that it only counts syntax level errors. However, in practice, some types of errors alters the meaning of a sentence more than others. Let's take a reference sentence, "I love eating cookies.". Two transcriptions "I like eating cookies" and "I loath eating cookies" will have the same WER 75%. However, the former sentence is semantically much closer to the original than the latter sentence.

Although alternative measures that address some weaknesses of the WER has been proposed [Morris et al., 2004], the WER still remains a standard measure for speech-to-text and automated subtitling. Therefore, in order for the results in our semi-automated

subtitling [Soe et al., 2021] to be comparable with previous works, we used the WER measure.

2.5.4 Automated subtitling and correction methods for speech-to-text

Automated subtitling has been discussed in previous research works [Brousseau et al., 2003; del Pozo et al., 2014; Obach et al., 2007] and in commercial software and services. Automated subtitling uses speech-to-text to create transcripts and then create subtitles from those transcripts. However, automated subtitling is far from generating subtitles that are good enough. For example, YouTube recommends that their automated captions to be checked by a human. To be exact, it is stated here that [20] "You should always review automatic captions and edit any parts that haven't been properly transcribed." Therefore, to create subtitles, the remaining problem with automated subtitling is how to help people correct errors in automation in subtitling. We argue that soon the solution will involve, semi-automated subtitling, in which a human is correcting machine mistakes.

Error correction methods and interfaces in speech-to-text has been explored in terms of multi-modal interfaces [Suhm et al., 2001]. Highlighting low confidence words is one of the methods that is explored to assist correcting errors. In one study [Vertanen and Kristensson, 2008] , participants take advantage of the highlighted words, but the effectiveness depends on highly accurate confidence scores. Another study [Suhm et al., 2001] concluded that the effectiveness of highlighting is inconclusive. In our own work [Soe et al., 2021], we did not evaluate confidence highlighting as an experiment condition. We made the observation that most of the highlighted errors were fixed. We also asked the participants if they think confidence highlighting helped them correct the error, and this question received mixed responses.

2.6 Cropping and panning for video retargeting

2.6.1 Video platforms and aspect ratios

Aspect ratio of a video is the ratio of the width of a video to its height. International standard format for wide screen television is 16:9 (read sixteen by nine). A common format for 16:9 widescreen format is 1080p, which has 1920 pixels on its width and 1080

[20] https://support.google.com/youtube/answer/6373554?hl=en

pixels on its height. The biggest video platform, YouTube, has 16:9 ratio as its standard ratio for computer [21]. However, the video landscape on the internet is characterized by a variety of devices and platforms demanding different aspect ratios. For example, videos on mobile devices are viewed in both landscape 16:9 and portrait 9:16 ratio. Some video platforms may have their own requirements, such as 1:1 square video format for Instagram. In our work on video panning [Soe and Slavkovik, 2022] we explore 16:9, 4:3, 1:1 and 9:16 aspect ratios which covers most of the aspect ratios in use today.

2.6.2 Video retargeting and cropping and panning

Video retargeting involves modifying videos for better viewing experiences for difference screen sizes (in aspect ratios and resolutions) on different devices. When mobiles phones had limited screen sizes with significantly lower resolutions, video retargeting approaches are concerned with down-scaling, cropping and transforming the video in a way that distorts the images [Kopf et al., 2011; Yo et al., 2013]. However, today the problem of video retargeting has become only about aspect ratios since mobile phones has screen resolutions that matches that of larger screen sizes.

Cropping is defined as removal of parts of an image that is outside a specific boundary [Okun et al., 2015]. Cropping images is relatively easy. However, cropping out videos is difficult, as video contains movements across the frames that the crop has to keep up with. To crop in a video not only the movement but the transition between scenes has to be considered. To retarget a video to narrower aspect ratios (e.g. convert a video in 16:9 to 1:1) some areas have to be cropped out. Outside retargeting, cropping and panning in a video is a common video editing task. How can we use semi-automation to make cropping of videos easier is explored in our work [Soe and Slavkovik, 2022].

2.7 Viz Story

Viz Story [22] is a web-based NLE developed by Vizrt. Viz Story is a browser-based package of tools for video editing and publishing. Viz Story package contains full set of features to support the process of creating video stories and publish multiple versions of them to different platforms. The current release version (as of 2021 September) of Viz Story has features such as keyframe-based panning, multi-video editing, text on video and support for subtitling. In fact, the Viz Story tool was used in the semi-automated subtitling

[21] `https://support.google.com/youtube/answer/6375112?hl=en&ref_topic=9257782`
[22] `https://www.vizrt.com/products/viz-story`

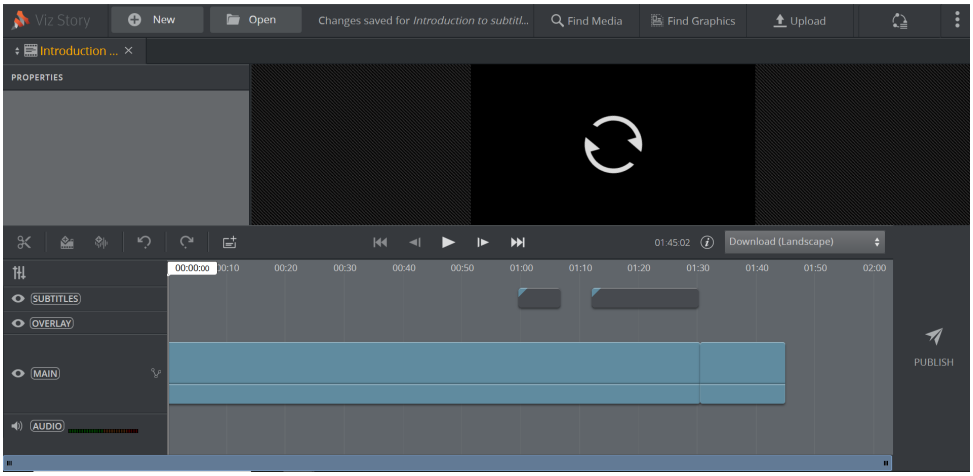

Figure 2.2: The main user interface of Viz Story

paper [Soe et al., 2021] to build a subtitling workflow with automation embedded. The primary user interface of Viz Story (version 1.7 which is used in this thesis) can be seen in the Figure 2.2.

2.8 Summary

Throughout this chapter, a comprehensive overview of the relevant research areas which serve as the literature context of the thesis are presented. First, the research area of intelligent video editing tools was presented and contrasted with fully automated video editing. Afterwards, semi-automation in video workflows and the pillars for this semi-automation, namely, AI and HCI topics are summarized. It is followed by the research works that tried to bridge the AI–HCI research areas. At the end of this section, two topics, subtitling and video retargeting, that are the video tasks explored in the papers in this thesis are discussed. A brief paragraph on, Viz Story, a web-based video editing tool, used in the subtitling paper is also included to provide a complete background.

The first chapter introduced the thesis and provided the goals of this thesis. In this chapter, the literature necessary to both situate and understand the thesis has been presented. The following chapter will provide the research methods and answer the question of how all the research in this thesis was conducted.

Chapter 3

Research Methods

In this chapter, all the research methods employed in the thesis are described in details. The description of each research method includes — the need for a research methodology, why a particular method was selected, how a method was used, and what outcomes that an application of a method produced. A summary of all the research methods used in each paper and the purposes are given in the Table 3.1. This thesis has two works [Soe and Slavkovik, 2022; Soe et al., 2021] in which two functioning prototypes were designed and built in order to run a user study and a usability evaluation. A summary of work involved in running these studies and constructing the prototypes are also presented in this chapter.

3.1 Systematic literature review

The goal of this thesis is to explore how to use AI to support video editing workflows. Building a research foundation for this thesis required a comprehensive overview of the state of the art and history of using AI to create an easier to use video editing workflows. So, the first step is to identify and synthesize the existing literature concerning the question, "How has AI been used to support video editing workflows?". The systematic literature review method was used to answer this question. This method ensures that the search, survey and synthesis of the literature is done in a reproducible and systematic way.

The specific systematic literature review guidelines used were from Kitchenham and Charters [2007], which describes the process of the systematic literature review for software engineering [1]. According to the guidelines, the three steps of the systematic liter-

[1]Software engineering is one of the few sub-disciplines in computer science that has published guide-

Paper	Research Methods	Purposes
AI video editing tools. What do editors want, and how far is AI from delivering them?[Soe and Slavkovik, 2021]	Systematic literature review	To explore and synthesize the literature in AI-supported video editing tools.
	Surveys	To gather video professionals' opinions on using AI in video editing.
	Thematic analysis	To discover themes in the textual data from surveys' open-ended questions.
Evaluating AI Assisted Subtitling[Soe et al., 2021]	User study	To perform a user experiment comparing AI-assisted against baseline subtitling.
	Quantitative analysis	To measure the quality of subtitles with Word Error Rates and efficiency with Words Per Minute (WPM).
	Surveys	To gather feedback during the user experiment.
	Thematic analysis	To discover themes regarding usability issues from open-ended questions from the survey for the AI-assisted subtitling prototype.
A content-aware tool for converting videos to narrower aspect ratios [Soe and Slavkovik, 2022]	Usability evaluation	To study the feasibility of using semi-automated cropping and panning with cinematic idioms. To understand the semi-automated cropping and panning better.
	Surveys	To gather feedback on the novel interface for cropping and panning.
	Thematic analysis	To discover the themes in the usability issue from open-ended questions and think-aloud feedback.

Table 3.1: Summary of Research Methods

ature review are planning the review, conducting the review, and reporting the result.

In planning of the review, important tasks are defining the search terms, literature databases to search, and setting the inclusion criteria. In our work [Soe and Slavkovik, 2021], the search terms used were *(intelligent OR smart OR automated OR AI) AND (video editor OR video editing)* and we searched on the computer science literature

lines for systematic literature review.

databases which are DBLP, ACM digital library, and Google Scholar[2]. The inclusion criteria we used in our work was the paper must describe a usage of AI/Automation to make video editing tools for the user, and the paper must also include description of the user interface. The search terms, the database, and the inclusion criteria were defined before conducting the search and review.

We then performed the search and collected the titles, abstracts, and the full-texts of papers. Afterwards, the inclusion criteria was used to filter out the collected papers. The remaining papers that met the inclusion criteria were read in details and categorized according to video editing tasks, mode of interaction with automation, and AI technology used. The result of this systematic literature review on AI-supported video editing tools is reported in our first paper [Soe and Slavkovik, 2021]. The plan of the systematic literature review and the data generated during the review is archived and will be published upon publishing of the paper.

3.2 Surveys and their usage

"Surveys are well-defined and well-written sets of questions to which an individual is asked to respond to" [Lazar et al., 2017]. Surveys are frequently used to describe populations, explain behavior and explore uncharted waters [Babbie, 1990]. In this thesis, we are particularly concerned with valid usages of surveys in HCI. On the appropriate usage of the surveys in HCI, Müller et al. [2014] stated that they can be useful for collecting "attitudes and perception towards an application in the context of usage" as well as for gathering "user experience feedback". In this thesis, we used surveys precisely for these two of the purposes.

In the first paper [Soe and Slavkovik, 2021], we used a survey to learn opinions and intents regarding using AI to support video editing from the subset of people who work in the video and broadcasting industry. The questions in this survey were created under three categories, namely, the participants' previous experiences on video editing and AI, what an ideal AI editor is in their opinion, and what do they want to automate in their video editing work. The survey was distributed anonymously via email using the survey tool called SurveyXact[3]. SurveyXact is a survey tool recommended by the University of Bergen, and their data management practices has been vetted by the university. Therefore, this tool was used in all the surveys to safeguard the participants' data. The usage of surveys in conjunction with our user study and our usability evaluation are

[2]Google Scholar was used as a backup to ensure that no work was missed
[3]`https://www.surveyxact.no/`

discussed in the following subsection.

3.2.1 Usage of surveys in the user experiment and the usability evaluation

Surveys should not be used as the sole method for observation and measure in a user study or usability evaluation. In this thesis, they were used together with other methods such as screen captures, accuracy measures, efficiency measures, and think-aloud method.

Two surveys were used to collect background information and user experience feedback in the user study for the AI-assisted subtitling prototype [Soe et al., 2021]. The background information survey was used to collect only the participants' non-identifiable information relevant to the study and their previous experience with subtitling and video editing. It also includes the typing speed in English measured right before the user experiment. The survey contained the form for informed consent regarding the participation in the user experiment.

The informed consent forms in the surveys were created using the recommendation and consultation from the Norwegian Center for Research Data (NSD[4]) which UIB is a member of. The informed consent forms in each of our surveys included a clear explanation on what information will be recorded, what will be published, and an instruction on how to retract participation from the experiment and request the data to be deleted at any time. We performed the surveys and data collection within the strict guidelines on piracy and data management from the NSD. For example, the data management plan for the work on AI Assisted subtitling [Soe et al., 2021] was submitted to the NSD, went through the approval process for privacy and data management practices by the NSD.

We used two types of survey questions in both of our user evaluations. First, we used 5-points Likert Scale questions to get ratings of the different aspects of our tools. Second, opened-ended questions were used so that the users can elaborate further on the user experiences. In the user study of AI-assisted subtitling [Soe et al., 2021], we used Likert Scale questions about the perceived difficulty of baseline subtitling, perceived difficult of assisted-subtitling and the quality of subtitle segmentation. The same survey also included open-ended questions covering what the participants like and dislike about the assisted-subtitling.

The usage of surveys in our usability evaluation [Soe and Slavkovik, 2022] of semi-automated cropping and panning is very similar to what we used in the user study of

[4] `https://www.nsd.no/en`

AI-assisted subtitling. Surveys were used in the usability evaluation to gather informed consent, background information, and user experience feedback. We used a survey form to collect informed consent and to ask the participants about what video aspect ratios that they had worked with. A different survey form was used post-experiment to collect feedback on user experience. The post-experiment survey included Likert scale questions about the usability of the tool. They are — results of the tool for 4:3, 1:1 and 9:16 ratios, usefulness of the tool for exploring different cropping and panning possibilities, how easy it is to understand the cinematic idioms, and if they would publish the results of the tool. In the same survey, open-ended questions were used to ask feedback on what did they like and dislike about the tool and suggestions on new cinematic idioms they would like to have in our tool.

3.3 User study

The most impactful research method used in Soe et al. [2021] is experimental research. It is a method in which controlled experiments are conducted to acquire knowledge. In HCI, the word *user study* [MacKenzie, 2013] is used to imply these controlled experiments where different configurations of a user interface are tested and compared. A user study is usually planned and conducted with the purpose to answer a set of hypotheses. It is then performed with participants to answer the hypotheses. Measurements and observations recorded from the user study are analyzed to either confirm or reject the hypotheses.

In our work on AI-assisted subtitling [Soe et al., 2021], for example, we started with the hypothesis that "semi-automated subtitling can help novice users perform subtitling faster". To evaluate this hypothesis, we planned and performed the user study, where the speed of the subtitling of each participant's is measured in words per minute (WPM).We measured the differences in the WPM between two configurations of the subtitling interface, namely, baseline and AI-assisted subtitling. The presence of AI-assisted subtitling is the independent variable for this user experiment, and the dependent variable is the speed measured in WPM. The participants performed the subtitling tasks using both configurations, and the order of the two configurations was alternated among the participants to mitigate learning effects. In our user study with 24 participants, the measured WPM is much higher for AI-assisted subtitling, and ANOVA analysis of the WPM confirmed the hypothesis.

In this user study of AI-Assisted subtitling, the post-experiment survey was used together with the measurements and screen recordings in the user experiment. The measurements in speed and accuracy of subtitles provided a quantitative analysis of the differences

between two user interface configurations. The survey with Likert Scale and open-ended questions added the knowledge about user feedback and their perceived experience with both configurations of the tool. These qualitative feedbacks from the survey helped in discovering the user experience issues and ways to improve the user experience in AI-Assisted subtitling.

3.4 Usability evaluation

In our work on semi-automated cropping and panning [Soe and Slavkovik, 2022], the main research method used in conducting the evaluation was *usability evaluation*. Usability evaluation involves "accessing a single user interface for feasibility, strengths and weaknesses" [MacKenzie, 2013]. In a usability evaluation, only a single user interface is studied, and the participants provide feedback on their experience of using the user interface. In contrast to a user study, a usability evaluation does not evaluate any hypotheses and only a single user interface is involved. In our paper [Soe and Slavkovik, 2022], the usability evaluation is used to explore feasibility of the new semi-automated method to crop and pan with cinematic idioms. In addition, usability evaluation provided valuable observations on what parts of the novel approach worked, what aspects should be improved, and what should be explored further. In this usability evaluation, surveys were used together with screen recordings and think-aloud observation method to explore the participants' experiences and collect their feedback.

3.5 Thematic analysis

Thematic analysis was used for qualitative analysis of textual and verbal responses in all our work. According to Braun and Clarke [2012], a thematic analysis is "a method for identifying, analyzing, and interpreting patterns of meanings or themes in qualitative data". The usage of thematic analysis for qualitative research relating to HCI practices is discussed in detail in McDonald et al. [2019]. For example, we used thematic analysis in our paper [Soe and Slavkovik, 2021] to identify different categories of automation needs from the open-ended questions. One of these questions asked the participants on "what would they want an AI editor to help them in video editing?". In another paper [Soe et al., 2021], there were open-ended questions on what the participants like and dislike about semi-automated subtitling and thematic analysis was used to identify different issues with semi-automated subtitling from the responses. It is used in the similar manner in the work for cropping and panning [Soe and Slavkovik, 2022], in

which responses to open-ended questions and data from think-aloud transcriptions were analyzed to identify patterns in the issues with the new interface proposed in the paper.

3.6 On designing, building and evaluating AI-assisted prototypes

When planning to perform a user study or a usability evaluation, a functional or interactive prototype has to be designed and built. Designing and building prototypes can be more challenging when AI is involved in building these prototypes. In fact, the difficulties in designing prototypes with AI for the UX designers is mentioned as a challenge for exploring AI-powered tools in Dove et al. [2017]. Some reasons for prototypes with AI being challenging to design are the dynamic, data-dependent, and unpredictable nature of AI as well as the UX designers lacking knowledge about AI abilities and limitations[Dove et al., 2017]. However, the author of the thesis has experience and education on AI and thus able to perform the task of making these prototypes without major issues.

Two fully functional AI-assisted prototypes were designed and built in this thesis for AI-Assisted subtitling and semi-automated cropping and panning [Soe and Slavkovik, 2022; Soe et al., 2021]. There were interesting research and design issues that were discovered while building these two prototypes. The goal of the prototypes, the challenges involved in building the prototypes, design choices, implementation, and lessons learned from the experience of implementing them will be discussed in this section.

AI-assisted subtitling tool [Soe et al., 2021] — this tool was built as a part of the user study to investigate the impact of introducing ML-based speech-to-text into a subtitling workflow. As mentioned before, two different versions of the user interface are evaluated, specifically, the baseline subtitling and AI-assisted subtitling. The existing subtitling UI on Viz Story was used as the baseline interface and this subtitling UI has elements found in many other subtitling tools, which are, subtitle controls, subtitling text entry box, the timeline, and video preview [Soe et al., 2021]. When designing this prototype, we considered some characteristics of ML-based speech-to-text such as the confidence values of detected words, timing of detected words and performance of the ML-based speech-to-text in words error rates (WERs). Using the previous literature on assisting correction of speech-to-text systems [Suhm et al., 2001; Vertanen and Kristensson, 2008], we added highlighting of low confidence words. In addition, sentence detection and heuristics derived from subtitling guidelines were used to segment the text transcript into subtitles.

As a baseline subtitling interface was required, the subtitling prototype was built by modifying the subtitling UI of Viz Story. Because it was a modification of an existing tool, the system server programming language and UI language choices were already made. The Viz Story is built using a client-server architecture, with the client running on a web browser. The server-side of the tool was built using C# language compiled into Windows executable servers and the client slide was built using Java-based web framework called Google Web Toolkit (GWT). The GWT automatically translates the Java code into web languages, specifically, HTML, JavaScript, and CSS. Regarding the selection of the speech-to-text model, a quick comparison was done and Google Cloud Speech-to-text [Cloud, 2022] was used. Google Cloud Speech-to-text is a web service that converts uploaded audio files into transcribed text using AI. Conversion of generated text transcript to subtitles is done using a sentence detection model and by implementing heuristics derived from the subtitling guidelines. Both the server and the client of the Viz Story tool were modified to be able to accommodate two configurations of the user interface for baseline and AI-assisted subtitling.

Two measures used in this user study of this prototype are words per minute (WPM) — how fast the participants can make subtitles measured in average number of words per minute, and words error rate (WER) — how accurate are the subtitles produced. WPM is measured using the output subtitles, application logs and screen recordings of the sessions. WER is measured using a Python package called jiwer[5]. The measured values are analyzed with one-way ANOVA to confirm the impact of AI-Assisted subtitling on the speed and accuracy of the subtitles produced by the participants. In addition, the data is visualized into graphs using matplotlib[6] to make additional observations on the possibility of higher machine transcription error in one video leading to a higher error rates. The data analysis for this evaluation is performed using pandas[7], a python data analysis toolkit. The subtitles created with and without AI assistance is collected from the study and has been publicly released as a part of the publication.

AI-assisted video retargeting tool [Soe and Slavkovik, 2022] — this tool explored a new interface for using AI to support easier panning and cropping tasks in videos. There were two main ideas from existing research integrated in this new interface. The first one was using cinematic idioms to control the cropping and panning. The usage of cinematic idioms in video editing is built upon cinematic idioms for rough edits [Leake et al., 2017] and for controlling video transitions [Wang and Moulden, 2021]. The second idea is computational video aesthetics [Niu and Liu, 2012], which concerns with computationally measuring and evaluating the quality of videos edits. In this prototype, the data

[5] `https://pypi.org/project/jiwer/`
[6] `https://matplotlib.org`
[7] `https://pandas.pydata.org/`

created by computer vision models are translated into what would make a good edit using heuristics from computational video aesthetics. Examples of heuristics are, what is a good shot length, and what is a good camera movement.

The prototype uses a cinematic idiom-based[8] interface for cropping and panning in videos. Each idiom is implemented by performing calculations from the AI-generated data, and it is descried using equations in the paper [Soe and Slavkovik, 2022]. The prototype was built from scratch using Python for the server and web languages for the client. The user interface of the prototype runs on a web browser. Python was selected because of the availability of both AI libraries and web services libraries in this language. The Python packages used are Flask[9] for the web server, NumPy[10] and SciPy[11] for the numbers and data processing, and TensorFlow[12] for machine learning. This prototype also includes a video labelling pipeline using AI for both structural information and the content analysis of videos. Labelling videos with structural information was accomplished using AI to detect shots and scenes and the type of shots. The content analysis was performed using various AI models for face detection, detecting interesting areas, and text detection. The source code of both the prototype and the documentation has been released with the publication.

When using AI to create prototypes, AI can be used from different forms, such as trained AI models, AI as a service, and training a new AI model using a dataset. Trained models are ML-based AI models that have already been trained with datasets and are ready to be used via a programming language interface. For example, a trained model of face detection, OpenFace[Baltrusaitis et al., 2018], can take images and generate the location of the faces in those images. The advantage of using a trained AI model is they are available to use without requiring the time, resources and expertise to train a model. However, a trained model still must be deployed with the correct computation environment for it to work. AI as a service is a web service such as Google Cloud speech-to-text[13] where the input voice is uploaded to a web service and a transcription of the voice is returned from the service. The benefit of using AI as a service is that it is the easiest and fastest way to use AI, but it comes at the cost of paying for such services and not having any control over the way these AI services work. Both the aforementioned methods are time- and cost-effective way to use AI. Training AI using a dataset is another way of using AI. However, it does require expertise, time, and significant computation

[8] Cinematic idioms in this work are video editing jargon that describe how the cropping and panning should be performed.

[9] `https://flask.palletsprojects.com/`

[10] `https://numpy.org/`

[11] `https://scipy.org/`

[12] `https://www.tensorflow.org/`

[13] `https://cloud.google.com/speech-to-text/`

resources to train an AI model. The dataset for training has to be developed if there is no existing dataset that satisfy the need. However, a very significant cost and effort are required to create a new dataset, and developing a new dataset alone usually constitute enough effort to be published as an academic contribution.

3.7 Summary

This thesis was not driven by a design-led approach. However, designing interactions for AI-embedded video editing tools is one of the problems that we have addressed in this thesis [Soe and Slavkovik, 2022; Soe et al., 2021]. Zimmerman et al. [2007] introduced research through design as a method for design researchers to make design research contributions towards HCI research and practices. In this thesis, we explored human–AI interaction through the lens of usability and user experience. Although the work in this thesis involved designing human–AI interfaces, the AI-embedded prototypes we have constructed in this thesis are more of system prototypes than design artifacts. Addressing the human–AI interaction challenges with design-led inquiry as in research through design [Zimmerman et al., 2007] could also be another promising approach. For instance, Lindley et al. [2020] uses research through design process to explore the problem of explaining AI roles and capabilities in a system to the users.

The purpose of this chapter is to present the research methods used in this thesis and to provide discussions on how the research work was executed in this thesis. First, the research methods employed in three papers of this thesis were listed together with purposes for their usage. Afterwards, the definitions and typical applications for each research method were summarized from the literature. In each of the papers, a set of research methods was used together in combination to create new knowledge. Therefore, how some research methods complimented each other was also discussed. Two of the papers involved building and designing fully functional prototypes. These prototypes were built for two different evaluation methods, which also influenced how they were designed. The differences in these two prototypes, the critical design choices, and how each prototype was engineered were also presented. To conclude this chapter, the key takeaways from building prototypes with AI/ML and how to use automation capabilities offered by AI/ML were presented.

The first three chapters, including this chapter, has provided all necessary introduction, literature context and research process for this thesis. These three chapters also provided the overarching theme of this thesis and how the papers of this thesis contributes to the intention of exploring using AI to assist video editing. Since, the framing for the thesis

has been provided, the next three chapters will be the papers of this thesis followed by the final chapter.

Chapter 4

Paper I: AI video editing tools

Chapter 5

Paper II: Evaluating AI Assisted Subtitling

Chapter 6

Paper III: A content-aware tool for converting videos to narrower aspect ratios

Chapter 7

Discussion and Conclusion

The overarching question of this thesis is how can AI be used to support human editors in video editing. In this thesis, semi-automated tools for video editing known as intelligent video editing tools are explored. The research area of intelligent video editing tools is identified and summarized in this thesis. In addition, we have surveyed the need for AI support in video editing, and proposed a task-based approach to studying semi-automation in video editing. We have also published evaluations of AI-assisted subtitling and video retargeting in this thesis. As this chapter concludes the thesis, the work done in the entire thesis will be summarized, and the key takeaways will be presented.

A summary of the thesis is introduced by listing the contributions of this thesis and connecting contributions with the five research questions. In addition, the answers from this thesis towards the challenges of semi-automated video editing in the Chapter 1 is revisited. Afterwards, the impact of this thesis on semi-automation in video editing and human–AI interactions in general is presented. Lessons learned, and the procedures used in this thesis, are also discussed with suggestion on how they could have been done better. This chapter ends with discussions on the limitations, the future work, and what this thesis has done to shape the future work.

The work in the thesis is done with the intention to answer the five research questions introduced at the beginning of the thesis.

- *RQ1*: What is the state of the art in AI-assisted video editing tools?
- *RQ2*: What are the opinions and expectations of video professionals regarding AI in video editing tools?
- *RQ3*: What is the impact of introducing AI assistance on the efficiency and quality of subtitles?

- *RQ4*: What are the changes in the user experience when AI assistance is added to the subtitling task?
- *RQ5*: How can we use AI to create a new way of performing cropping and panning in video editing?

Three papers were written to answer these research questions. The contributions made from the attempt at answering the research questions are discussed in the next section.

7.1 Contributions

The contributions of the thesis and related research questions for each of the contributions are discussed in this section. Some contributions are directly answering the research questions, while others are the lessons learned from our efforts to answer the research questions.

We have identified the research area of intelligent video editing tools and synthesized of existing work in the field using a systematic literature review [Soe and Slavkovik, 2021] (*RQ1*). This paper enabled comparisons of existing work and classified them based on video editing tasks that were automated in the tools, and AI techniques used in the tools to automate these tasks. In addition, the evaluation methods used in existing works and results of these evaluations are summarized in Table 2.2.

In the same paper [Soe and Slavkovik, 2021], we presented the result of our survey of opinions, preferences and expectations on AI from the video professionals (*RQ2*). The analysis of the survey data and the review of the literature were used to identify unexplored AI applications areas in video editing. The common applications across both research and survey data included tasks such as composing video segments, and synchronization of audio and video. Unexplored applications we have identified are aesthetic improvements, video pre-editing tasks (such as filtering out bad takes and organizing video files), and providing recommendations for video editing with AI. The state-of-the-art AI that can be used to support the identified area were also suggested.

We have published the first empirical experiment [Soe et al., 2021] on AI assisted subtitling, which involved both quantitative and qualitative measures (*RQ3, RQ4*). The results indicated that AI assistance in subtitling helped novice users create subtitles significantly quicker and a little more accurate than both baseline and AI-generated subtitles (*RQ3*). It meant that AI-generated subtitles were improved by the human users using our user study. In addition, several usability issues and areas to improve in

designing AI-assisted subtitling user interfaces were identified as a result of our study (*RQ4*). The user study also produced the dataset on AI-assisted subtitling evaluation, containing the subtitles that the participants generated from the scratch and with AI assistance. This dataset can be used, for example, for exploring differences in the nature of errors made with and without AI assistance.

In the same subtitling paper [Soe et al., 2021], the design and implementation artifacts for semi-automation of AI assisted subtitling is described in details. Some design and implementation artifacts are — the prototype built on top of a production grade video editing tool, design of AI-assisted subtitling interface, selection of speech-to-text models, implementation of subtitling guidelines, and highlighting of low confidence words. Some important insights from the study are, the need to control the delivery of machine-generated subtitles not to make the users feel overwhelmed and most of the highlighted errors ended up being corrected. In addition, the details on how the prototype was implemented such as programming languages used, and the needs for speech-to-text models to learn from user corrections is described.

We have also performed and published another empirical experiment which is about using multiple AI models to support cropping and panning task [Soe and Slavkovik, 2022]. The experiment used qualitative measures to explore idiom-based interface for cropping and panning (*RQ5*). The results from this paper suggested that it is feasible to use an idiom-based interface in cropping and panning. In the same work, design and implementation of the idiom-based interface is also laid out. The design choices made for the idiom-based interface and implementation of it using AI methods were fully described in the paper[Soe and Slavkovik, 2022]. In this work, we proposed six cinematic idioms to control what areas to focus on in a video and how should the camera move. Three machine learning models for computer vision and a shot and scene detection method were used to enable all these interactions. In addition, we explored how ordering of idioms that can be used to represent priorities in which part of the image should be focused[1]. The source code, the output of the tool and cropping and panning done manually by a professional video editor has been published together with the paper.

We have also performed two thematic analyses using the data from each of the empirical experiments. These two thematic analyses highlighted the important user experience issues in semi-automated subtitling and video retargeting workflows, respectively (*RQ4,RQ5*). Here are some of the issues we have identified using thematic analysis on assisted subtitling[Soe and Slavkovik, 2021] — the participants like that it is easier, it is quicker, and generated subtitles are relatively good. However, we also found that au-

[1]For example, the users can put speaker-visible idiom as the first in the list followed by make-text-visible to ensure faces are shown over the text if both are present in a shot

tomation can make people unfocused, rearranging subtitles is harder, mistakes could go unnoticed, and it is more costly to notice and fix error.

As for our idiom-based cropping and panning work [Soe and Slavkovik, 2022], we found out that the participants enjoy using the tool for — the simple interface, quick response time, providing an overview of retargeting, and being easier than doing so with video editing tools. On the other hand, understanding the idioms used, understanding ordering and combinations of idioms can be a challenge for the users. Additional problems are not being able to overwrite automation, visualization of crop area, application of idioms to scenes and crop quality issues. We also suggested solutions for some issues, such as using icons and color codes to explain the meanings of idioms in the paper.

Across the two thematic analyses, there were some common user experience issues with semi-automation. The users will be satisfied with the results of the automation assistance when it worked as they expected. However, when it does not work as expected it is considered a mistake and that mistake has to be noticed and then corrected by the users. But correction can be costly in semi-automation. For instance, if subtitles were to be segmented differently than what the machine has generated, it is harder than segmenting them from scratch[2]. In our video retargeting prototype, all the users pointed out that they would like to overwrite some part of the automation.

Across all three papers, we proposed a task-based to approach semi-automation of video editing. Our review of intelligent video editing tools [Soe and Slavkovik, 2021] indicated that most tools are created for a single type of video [3]. What is meant by task-based approach is that automation is applied to individual tasks in video editing and thus the results are not for specific video types but can be applied to a wide range of videos. We have also identified the tasks involved in previous works of intelligent video editing tools in Table 2.1.

7.2 On dealing with AI errors

When using semi-automation in a flexible, creative and end-user-led workflows, such as video editing, the user experience should be a priority. Among the possible user experience challenges, there is one type of user experience challenge that is inevitable when machine learning(ML)-powered AI is used. As the result of the probabilistic nature and uncertainties with machine learning [Dove et al., 2017; Yang et al., 2020], users will

[2]If a subtitle in the middle of two other subtitles has to be adjusted usually either one of the adjacent subtitles has to be adjusted as well.

[3]For instance, the video editing tool by Berthouzoz et al. [2012] is created just for interview videos

have to deal with AI errors[4]. What is meant by the probabilistic nature is that, there will be errors in ML outputs. For example, a highly accurate ML model in speech-to-text with 9% WER (word error rate) will contain 9 errors for every 100 words on average.

As a starter, the users should be provided with an interface to correct the errors. These interfaces require *clever user interface designs* that reduce the cost of correcting AI mistakes. When designing error correcting interfaces, one should consider the user experience, the task, and properties of the ML employed.

Another thing to consider when dealing with AI mistakes is users might put too much reliance on automation provided by AI and might fail to do proper reviews. Some participants in our assisted subtitling study [Soe et al., 2021] suggested that AI can make them non-attentive, and thus increasing the chances of failures to notice errors. The challenge is then in looking for clever user interface designs that can reduce both the cost of finding and also correcting errors. The goal is to create an overall good user experience despite the imperfections of AI. For instance, to aid in error detections, highlighting low confidence words[5] was used in our subtitling prototype [Soe et al., 2021]. I have also considered using alternative predictions of words[6] as a way to assist in correcting errors, but did not due to prioritization of the work in the paper.

So far, dealing with errors in AI predictions has not been considered in the publications on intelligent video editing tools. We can see that from the summary of evaluation results of intelligent video editing tools in Table 2.2. In that table, most of the results are validations of the tool and did not consider how the AI errors are handled. It is well established that designing human–AI interaction is difficult [Dove et al., 2017; Yang et al., 2020]. Reporting negative experiences, what designs worked, and what did not, is essential for progress in designing for human–AI interactions. For instance, a common negative experience reported in our assisted subtitling work[Soe et al., 2021] is that users felt *overwhelmed* by the amount and pace of the automated subtitles available to them at the beginning of the task. This meant that we should look at different design solutions, such as delivering automated subtitles one line by one line, instead of providing the entire subtitle at the beginning. Another negative experience we have encountered in semi-automated retargeting work[Soe and Slavkovik, 2022] is that the users would like to "overwrite" automation. Overwriting automation in video retargeting is not explored in our work, but that did not stop the users from providing comments that they would like to overwrite it. It could be stated that the ability to overwrite AI is essential for AI-assisted video editing tools, especially for the situations when AI failed to meet the

[4]The term error is loosely defined as the outputs that the user did not expect

[5]Words that the speech-to-text model is uncertain about its predictions

[6]ML speech-to-text usually have alternative predictions that are of lower probability than the final word predicted by the model.

users' expectation.

7.3 Human–AI interaction challenges in video editing

What are the human–AI interaction challenges in video editing that we have discovered during this thesis? Some challenges for human–AI interaction from the literature can also be applied to the domain of video editing. For example, Yang et al. [2020] reviewed human–AI interaction challenges discussed in the literature and provided a framework summarizing these challenges. The challenges from Yang et al. [2020] that are applicable to video editing and related reflections from what we have learned thought the thesis are discussed in this section.

The first challenge is "technical feasibility of a design idea is highly dependent on the data" [Yang et al., 2020, p. 2]. This applies to semi-automated video editing, in particular, very few datasets for video editing are available. In more popular research areas, such as in computer vision, a much larger number of higher quality datasets are available. For example, if there is a need to train a machine learning model for segmenting videos, creating video transitions, or composing videos segments, the only option is to create your own dataset. However, the cost of developing large datasets is both time and cost prohibitive for most projects. The limited availability of data in video editing constraints what can be semi-automated with machine learning.

Another challenge is - it can be "difficult to see the potential effects of AI" [Yang et al., 2020, p. 2] when AI is integrated into a workflow or tool. It can be a problem for video editing, which is a creative and open-ended task with unlimited scope. It is difficult to see all potential effects of using AI in video editing, and consider all cases in which AI might end up hindering the user. Moreover, there is an issue of monitoring the effects of using AI in video workflows towards the way in which videos are produced, distributed and viewed. A mitigation strategy against this challenge suggested during our survey[Soe and Slavkovik, 2021] is *turning off automation* feature, where the users can easily switch off automation or undo the work of automation. To enable turning off the automation feature, the tool has to be designed to work with and without AI.

A related challenge is that it can be "difficult to explain AI behaviors to users" [Yang et al., 2020, p. 2]. Some AI behaviors such as, correct tracking of objects in videos can be explained with relative ease. However, some behaviors such as AI segmenting videos

or selecting appropriate transition points can be harder to explain[7]. This challenge also goes hands in hand with the explainable AI research area [Gunning et al., 2019].

It is "difficult to design shared control between AI and the users" [Yang et al., 2020, p. 2]. This challenge of designing shared controls was discussed in details in Chapter 2. Designing which tasks should be automated and what user inputs should be necessary is a difficult and very critical challenge. In this thesis, we have explored shared control challenge for subtitling and cropping and panning tasks. The last challenge is: It is "difficult to anticipate/mitigate unpredictable AI behavior" [Yang et al., 2020, p. 2] which has been discussed in details in the previous section under the assumption that users perceive unpredictable AI behaviors as errors.

7.4 Evaluating semi-automated video editing tools

To be able to synthesize the results of the previous studies is essential for advancing a research area, and that of intelligent video editing tools is no exception. In this thesis, we have suggested a task-based approach [Soe and Slavkovik, 2021] to compare the work on intelligent video editing tools and summarized evaluation methods used in Table 2.2. When evaluating the intelligent video editing tools, there are two aspects to consider, evaluation of the process of using the tool and that of the outcome of the process. The evaluation of the process usually involves exploring the user experience, measuring the time it takes to complete the task, and observing how does the users use the tool to perform the task. The evaluation of the outcome accesses how good the output of the tool is. For example, the output of an editing tool can be a fully edited video, a raw cut of videos, or just subtitles.

What lessons have we learned in terms of evaluating intelligent video editing tools in this thesis? In our semi-automated subtitling work [Soe et al., 2021], we evaluated both the process of assisted subtitling and the output subtitles. In addition, we evaluated assisted subtitling as experimental treatment and used within-subjects study design. As a result, each user uses both baseline and semi-automated subtitling. Doing so allowed the users to experience the impact of added automation against the baseline, and thus able to provide their experiences associated with both the positives and negatives of automation.

Using automation as an experimental condition was possible because we built the semi-automated feature on top of the existing subtitling user interface in Viz Story. With the

[7] e.g. Why a segmentation is done at a particular frame not the next one?

baseline as a reference, we were able to conclude that semi-automated subtitling lead to significantly faster subtitling performance, and slightly more accurate subtitles measured in Word Error Rates(WER). In addition, measurement of quality of subtitles with WER meant that the comparison of our results with previous studies on both semi-automated and automated subtitling is possible.

Comparing automation assistance with a baseline in an evaluation, however, is not always feasible. One main barrier for planning evaluations with automation as experimental treatment is that the prototype tool must have user interfaces and features for both automation and baseline usage for the users. That will most likely double the amount of work required to construct the prototype. This is a very significant engineering effort, and it is usually not feasible for most research prototypes.

Implementing semi-automated designs on top of existing off-the-shelf tools is a possibility, like in our subtitling prototype. However, doing so requires access to source code of these tools and modifying them. Modifying existing code involves having a certain level of software engineering experience and requires building an understanding of how the existing tool was built. From my experience in this thesis, given the same set of features, the amount of effort involved in modifying an existing tool was much more than building a prototype from scratch. The amount of work we have committed to understanding and modifying Viz Story[Soe et al., 2021] was way more than building the second prototype[Soe and Slavkovik, 2022] from scratch.

We used the user study method to evaluate our semi-automated cropping and panning work [Soe and Slavkovik, 2022] without comparing with a baseline panning and cropping. One reason for this is the prohibitive cost of implementing both baseline and semi-automated interfaces, which has been discussed in the previous paragraph. Another reason is that comparing automation assistance with baseline does not make sense when the new workflow proposed is very different from baseline. For example, we used a set of idioms to control cropping and panning in our work, in contrast to baseline cropping and panning which uses marked crop areas and keyframes. In addition, the baseline task can be too difficult for the novice users and the purpose of the semi-automation is to make the workflow much easier. Since, the baseline cropping and panning workflow is very difficult for untrained users, comparing with a baseline was not possible within a reasonable time. The intention of this user study is to explore a novel design of doing a task with AI assistance, and the evaluation is used to answer feasibility and user experience issues. Comparison with a baseline is not a necessary condition for an evaluation to provide knowledge towards enabling semi-automation.

7.5 How have we used AI to support video editing

I have two main intentions in using AI to support video editing that is shared across three works in this thesis. The first is using AI to make tasks *easier and more accessible*. The second is to *automate away the boring and repetitive tasks*. These two purposes will be explored further in the context of both the thesis and in the literature.

Exploring how to use AI to make tasks easier and more accessible in video editing is one of the main intention of this thesis. The ultimate goal is to make video editing as easy and accessible as *editing text* for the novice users. However, this goal is far from being feasible in the near future. Even semi-automating a simpler and smaller task of subtitling had challenges that were not known before this thesis. In our evaluation [Soe et al., 2021], the participants responded that the semi-automated subtitling is a bit harder than baseline subtitling. Even though the assisted subtitling enabled them to complete the tasks much faster, we failed to make semi-automated subtitling feel easier for novice users. Using the survey responses and open-ended questions in our evaluation, we have suggested design changes to make semi-automated subtitling easier. These changes include: delivering the subtitles line by line on demand, automatically synchronize the edited or entered subtitles, and highlighting the subtitles words being spoken.

The semi-automated cropping and panning tool proposed an idiom-based user interface [Soe and Slavkovik, 2022] that is simpler and easier. The participants' feedback from the evaluation was that the tool indeed was quite fast and easy to use, and thus we achieved the goal of making the task easier in our second attempt. Many other tools in the literature were also made with the same purpose of making video editing easier for the novice users. For example, Casares et al. [2002] proposed smart interactions and lenses to make video editing more accessible to novices. Other works that also proposed easier video editing methods are Chi et al. [2013]; Leake et al. [2017]; Truong et al. [2016]. In all of these tools, the need to perform frame by frame editing is replaced with simpler interactions.

The second purpose is to *automate away the boring and repetitive tasks*. The survey in our first paper Soe and Slavkovik [2021] identified the tasks that video professionals would like to be automated. Some of those tasks are subtitling, logging of videos[8], organization of video editing projects, video aesthetic improvements, color grading, and content suggestions. This provided us with an overview of the tasks that video professionals might like to automate away. However, with current AI technology, completely

[8]logging using the metadata and AI content analysis of the video

automating the task is not always feasible.

Our semi-automated subtitling tool enabled much faster subtitling for the novice users [Soe et al., 2021]. Therefore, though it doesn't completely automate the task, it makes the task more time efficient for the users. Semi-automation can help professional video editors to be more productive as well. The professional video editor in semi-automated cropping and panning [Soe and Slavkovik, 2022] commented that he would rather let automation do most of the work and just fix where it *misses*.

What tasks are considered repetitive, and what are the creative tasks that the users should be in full control is yet to be answered. A related design choice is what form of control is given to the user in a task. In our cropping and panning work [Soe and Slavkovik, 2022], and in two other works [Leake et al., 2017; Wang et al., 2019], the form of control is via selecting a set of cinematic idioms. After the idioms are selected, the rest of the editing tasks are automated. Such coarse-grained control of video tasks offers the editor a quick way to explore different possibilities and lowers the barrier of using these tools for novice users. Though, there are differences among the semi-automated tools in video editing, the common goal is to use automation to make video editing easier and less tedious.

7.6 Conclusion

This section starts with a summary of how we have addressed the challenges of automating video workflow introduced in Chapter 1. The first challenge is *understanding existing video workflows*. We have consulted the literature to understand video production workflows. And we have proposed a limited but practical task breakdown of video editing. The breakdown in Figure 1.1 serves as a practical reference to address semi-automation in video workflows. However, actual video editing and production depends on the type of the videos, the production environment and who the editor is.

Towards understanding the *existing work on intelligent video editing tools*, we have used the systematic literature review to map and synthesize work on intelligent video editing [Soe and Slavkovik, 2021]. This work enabled a quick overview and critical review of this area of research. In addition, we have collected the *opinions on and requirements from AI* in video workflows from video professionals [Soe and Slavkovik, 2021] to help guide further discussions and research works.

In two very different video tasks, subtitling [Soe et al., 2021] and cropping and panning [Soe and Slavkovik, 2022], we have explored *applications of AI to support users* in these

tasks. Two prototypes for these tasks were designed, implemented and evaluated. We have used these prototypes to run experiments to better understand the impact and consequences of using automation, and used what we learned from each experiment to suggest how to create better human–AI interactions in video editing.

In this thesis, we have made contributions that proposed new approaches to use AI to assist users in video editing. In the first paper [Soe and Slavkovik, 2021], we proposed addressing the semi-automation in video editing from task-based approach instead of creating specialized tools that are designed for a specific type of video. In addition, we suggested unexplored applications of AI in video editing based on the survey.

Following up on the task-based approach to address AI-assisted video editing, two works were produced. The first work [Soe et al., 2021] explored AI-assisted subtitling, and the second work [Soe and Slavkovik, 2022] proposed a new way of doing semi-automated cropping and panning. In these works, the benefits of using the task-based approach were demonstrated. For example, in the subtitling work [Soe et al., 2021], the impact of AI in the subtitling workflow was examined as an experimental condition, and it allowed the results of the paper to be comparable with existing work on subtitling. Unlike video specific tools, contributions from two of our prototypes [Soe and Slavkovik, 2022; Soe et al., 2021] can be applied to any type of video[9].

Throughout the thesis, the user experience was also emphasized, and we ensured our user studies explored user experience issues. This is because our goal of crafting human–AI interactions in video editing is to ensure the *completion of a user's task with the help of AI* [van Berkel et al., 2021], with AI playing the supporting role. The goal of this thesis is not to create better AI for video editing, but how to make AI to be more helpful for the users in video editing. This changes the focus of AI in video editing from AI technology first approaches to user experience focused direction. This is because I believe that videos are made to tell stories, and the story telling process should still be human-led. There are similar approaches from AI research in human-centered AI, that emphasized the need to focus the development of AI for the interests of the users. I hope the work in this thesis will contribute towards the shift to a more user-centered and user experience-focused approach to semi-automation in video editing.

AI technology is still far from performing a complex task such as video editing without any human intelligence involved. The current usage of automated video editing is limited to creating video mashups[Saini and Ooi, 2018] or video summaries[Gygli et al., 2014; Wu et al., 2020]. AI technology is still not good enough to completely automate even simple tasks such as subtitling in video editing[Soe and Slavkovik, 2021] without human

[9] A caveat with semi-automated subtitling is that the video must contain speech

intelligence. Therefore, the solution to automation in video workflows must utilize both human intelligence and AI and thus the user interfaces and interactions with AI must be considered and prioritized.

One important thing we have learned from our work on semi-automated subtitling[Soe et al., 2021] is that adding automation changes users' interactions with the tool. Therefore, adding automation requires rethinking and redesigning existing user interfaces and interactions. This should encourage more design and user experience -focused research in semi-automating video workflows.

Another reason for working towards crafting better user experience instead of trying to push the boundary of AI performance is the diminishing returns of deep learning. Deep learning, the dominant method in AI-based audio and video processing, had been shown to offer diminishing returns with more data and computational power [Thompson et al., 2021]. To explain it differently, although deep learning methods perform better with increases in data and computational power, the improvements gets smaller and smaller. In addition, Thompson et al. [2021] pointed out the rising environmental costs of training deep learning models and suggested the AI community to explore other AI methods. As the performance improvements of deep learning methods will stagnate due to diminishing returns, I would recommend exploring efficient human–AI interactions and find ways of using AI to support video editing. We should explore designs and interactions where the mistakes of AI are expected and handled with clever designs.

7.7 Limitations

The main limitation in this thesis is that the participants' usage of the prototypes only happened during the evaluation. In such a short time, learning and adapting to work with AI assistance is limited. A long-term usage of AI embedded prototypes could improve users' performance, remove some short-term usability issues encountered, or reveal entirely different long-term problems. For instance, in the AI–assisted subtitling prototype, with long-term usage, the users might learn to predict and correct the kind of errors that the speech-to-text model is most likely to make. They might learn the fact that English speech-to-text model cannot detect out of dictionary characters such as names from other languages. With long-term usage, the users might get better at expecting and correcting errors in non-English names.

Since the amount of time participants have for an evaluation is a very limited resource, we were unable to explore the impact of individual design elements in our studies as an

experimental condition. In particular, in Soe et al. [2021], we did not evaluate the changes in effectiveness and efficiency from highlighting low confidence[10] words. If we were to use confidence highlighting as an additional experimental condition, the participants would have to subtitle three videos instead of two. Therefore, we decided to design our study comparing AI assistance with the baseline. We only implemented confidence highlighting as a part of the AI-assisted interface, and we simply made an observation that most of the highlighted errors were fixed.

Another methodological limitation is that measuring the quality of the work in video editing tasks quantitatively is difficult. This fact is true for even simple video tasks such as subtitling. For instance, we used Word Error Rates(WER) to measure how accurate are the subtitles [Soe et al., 2021]. Let's consider two different errors for the text, "I want an apple". Two different sentences with errors — "I want apple", and "I want an orange" will have the same WER score[11]. However, the first type of error does not change the semantics of the sentence. The reasons for using WER and alternative measures are discussed in details in [Soe et al., 2021]. For the cropping and panning work, like many other video editing tasks, there is no standard measure for the quality of the work.

When using AI to create easier to use workflows, comparisons with a baseline tool is not always practical. We were unable to compare with baseline (manual cropping and panning) in [Soe and Slavkovik, 2022] because the baseline is difficult to perform for the novices in a reasonable time. The baseline cropping and panning uses marking for crop areas and keyframes for controls, while our tool uses cinematic idioms.

In the survey results of the first study[Soe and Slavkovik, 2021], the participants answered the question without significant previous experience with AI in video editing. Therefore, the opinions from the survey are mostly based on how AI is portrayed in the media and the participants' imagination of what AI is. These opinions could change based on successful or failed interactions with AI-powered tools in their workflow. It is up to the designers to create better user experiences of using AI in video editing.

7.8 Future work

In this section, two types of future work will be presented. The first type is interesting topics that we have thought about but have not explored in this thesis. Afterwards, the consequences of this thesis on the future research of semi-automation in video editing

[10] Confidence is a score from the speech-to-text model predicting how certain it is about its prediction of a word

[11] WER only measures syntactic corrections

are presented to conclude this section.

The utility of keyframes, frames that represent each shot, should be explored in video editing workflows. Truong and Venkatesh [2007] summarized a review of keyframes and their potential applications in video skimming[12]. Keyframes can be used to represent a video in a few static images. For example, a 4 minutes video with 25 frames per second will have 6000 frames. With each keyframes being used to for each shot, the same video can be represented with around 10 to 20 static images. By using keyframes, some video editing tasks can be reformulated as easier image editing tasks. The user's editing decisions on the keyframes can be extrapolated to edit the entire video.

Personalization is an important aspect of AI in supporting video editing. To put it in different words, AI should continually learn from the user and so that it can adapt to the user's need better. In the future work section of our assisted subtitling paper [Soe et al., 2021], we proposed how a speech-to-text model can be personalized with corrections made by the users. Learning from corrections made by the users has potential to solve the current limitations of speech-to-text methods. In particular, the limitation of only being able to detect the words in the dictionary can be circumvented by using user's corrections to update the dictionary.

The results from the survey of our paper [Soe and Slavkovik, 2021] highlighted under-explored areas of video editing such as logging and metadata creation, providing recommendations, suggestions, and personalization. Using AI to help users in video editing should also consider supporting tasks in video editing, such as using AI to log and create metadata for videos, using AI to help organize files and resources in video projects. Though these tasks do not involve manipulating videos, they are in line with our goal to support users with AI in video editing.

Another interesting area to explore from our paper [Soe and Slavkovik, 2021] is providing recommendations in video editing for both style and content. Generating recommendations for video consumption is an established research area with popular industry applications such as in Netflix and Spotify [Steck et al., 2015]. Can these recommender systems be adapted to provide suggestions in video editing? Suggestions can be used to assist video editors find relevant video segments, suggest background music that suits the video, suggest graphics and data, or suggest aesthetically pleasing transitions and video graphics.

I, as a researcher, am more inclined towards systems and technology rather than societal impact. As Marda and Narayan [2021] argued about the importance of ethnographic

[12]Video skimming is creating a shorter version of a video

studies in exploring societal impact of AI applications, I think ethnographic studies would also lead to better understanding of the broader context and societal impact of using AI in video editing. For example, how would the usage of AI changes the way video editors work, and the nature of the videos they produce in the long run? As more AI applications are introduced into the video editing workflows as well as other media creation, it will be more important to perform ethnographic studies of AI usage to get a more complete picture, to limit the potential harm that it can have through influencing how videos are created, and its consequences on the stories told through the videos.

Bibliography

V. Ambati. *Active learning and crowdsourcing for machine translation in low resource scenarios*. phd, Carnegie Mellon University, USA, 2012. AAI3528171 ISBN-13: 9781267582157.

S. Amershi, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, and P. N. Bennett. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–13, Glasgow, Scotland Uk, 2019. ACM Press. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300233. URL `http://dl.acm.org/citation.cfm?doid=3290605.3300233`.

E. R. Babbie. *Survey research methods*. Wadsworth Pub. Co, Belmont, Calif, 2nd ed edition, 1990. ISBN 978-0-534-12672-8.

X. Bai, J. Wang, D. Simons, and G. Sapiro. Video SnapCut: robust video object cutout using localized classifiers. *ACM Transactions on Graphics*, 28(3):1–11, July 2009. ISSN 0730-0301, 1557-7368. doi: 10.1145/1531326.1531376. URL `https://dl.acm.org/doi/10.1145/1531326.1531376`.

T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, Xi'an, May 2018. IEEE. ISBN 978-1-5386-2335-0. doi: 10.1109/FG.2018.00019. URL `https://ieeexplore.ieee.org/document/8373812/`.

L. Baraldi, C. Grana, and R. Cucchiara. Shot and Scene Detection via Hierarchical Clustering for Re-using Broadcast Video. In *Computer Analysis of Images and Patterns*, volume 9256, pages 801–811. Springer International Publishing, Cham, 2015. ISBN 978-3-319-23191-4 978-3-319-23192-1. doi: 10.1007/978-3-319-23192-1_67. URL `http://link.springer.com/10.1007/978-3-319-23192-1_67`. Series Title: Lecture Notes in Computer Science.

Y. Bengio, Y. Lecun, and G. Hinton. Deep learning for AI. *Communications of the*

ACM, 64(7):58–65, June 2021. ISSN 0001-0782. doi: 10.1145/3448250. URL https://doi.org/10.1145/3448250.

F. Berthouzoz, W. Li, and M. Agrawala. Tools for placing cuts and transitions in interview video. *ACM Transactions on Graphics*, 31(4):67:1–67:8, July 2012. ISSN 0730-0301. doi: 10.1145/2185520.2185563. URL https://doi.org/10.1145/2185520.2185563.

R. Borgo, M. Chen, B. Daubney, E. Grundy, G. Heidemann, B. Höferlin, M. Höferlin, H. Leitte, D. Weiskopf, and X. Xie. State of the Art Report on Video-Based Graphics and Video Visualization. *Computer Graphics Forum*, 31(8):2450–2477, Dec. 2012. ISSN 01677055. doi: 10.1111/j.1467-8659.2012.03158.x. URL http://doi.wiley.com/10.1111/j.1467-8659.2012.03158.x.

V. Braun and V. Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*, APA handbooks in psychology®, pages 57–71. American Psychological Association, Washington, DC, US, 2012. ISBN 978-1-4338-1005-3. doi: 10.1037/13620-004.

C. Bregler, M. Covell, and M. Slaney. Video Rewrite: driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97*, pages 353–360, Not Known, 1997. ACM Press. ISBN 978-0-89791-896-1. doi: 10.1145/258734.258880. URL http://portal.acm.org/citation.cfm?doid=258734.258880.

J. Brousseau, J.-F. Beaumont, G. Boulianne, P. Cardinal, C. Chapdelaine, M. Comeau, F. Osterrath, and P. Ouellet. Automated Closed-Captioning of Live TV Broadcast News in French. *8th European Conference on Speech Communication and Technology*, page 5, 2003.

S. Butler and A. Parkes. Film sequence generation strategies for automatic intelligent video editing. *Applied Artificial Intelligence*, 11(4):367–388, June 1997. ISSN 0883-9514, 1087-6545. doi: 10.1080/088395197118190. URL http://www.tandfonline.com/doi/abs/10.1080/088395197118190.

S. Bødker. When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction changing roles - NordiCHI '06*, pages 1–8, Oslo, Norway, 2006. ACM Press. ISBN 978-1-59593-325-6. doi: 10.1145/1182475.1182476. URL http://portal.acm.org/citation.cfm?doid=1182475.1182476.

Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *arXiv:1710.08092 [cs]*, May 2018. URL http://arxiv.org/abs/1710.08092. arXiv: 1710.08092.

J. Casares, A. C. Long, B. A. Myers, R. Bhatnagar, S. M. Stevens, L. Dabbish, D. Yocum, and A. Corbett. Simplifying video editing using metadata. In *Proceedings of the conference on Designing interactive systems processes, practices, methods, and techniques - DIS '02*, page 157, London, England, 2002. ACM Press. ISBN 978-1-58113-515-2. doi: 10.1145/778712.778737. URL http://portal.acm.org/citation.cfm?doid=778712.778737.

S. M. Casner, E. L. Hutchins, and D. Norman. The challenges of partially automated driving. *Communications of the ACM*, 59(5):70–77, Apr. 2016. ISSN 0001-0782. doi: 10.1145/2830565. URL https://doi.org/10.1145/2830565.

P.-Y. Chi, J. Liu, J. Linder, M. Dontcheva, W. Li, and B. Hartmann. DemoCut: generating concise instructional videos for physical demonstrations. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, UIST '13, pages 141–150, New York, NY, USA, Oct. 2013. Association for Computing Machinery. ISBN 978-1-4503-2268-3. doi: 10.1145/2501988.2502052. URL https://doi.org/10.1145/2501988.2502052.

C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani. State-of-the-art Speech Recognition With Sequence-to-Sequence Models. *arXiv:1712.01769 [cs, eess, stat]*, Dec. 2017. URL http://arxiv.org/abs/1712.01769. arXiv: 1712.01769.

J.-H. Choi and J.-S. Lee. Automated Video Editing for Aesthetic Quality Improvement. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1003–1006, Brisbane Australia, Oct. 2015. ACM. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806386. URL https://dl.acm.org/doi/10.1145/2733373.2806386.

Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2023, 2020.

E. Clark, A. S. Ross, C. Tan, Y. Ji, and N. A. Smith. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18*, pages 329–340, Tokyo, Japan, 2018. ACM Press. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172983. URL http://dl.acm.org/citation.cfm?doid=3172944.3172983.

G. Cloud. Speech-to-Text: Automatic Speech Recognition, 2022. URL https://cloud.google.com/speech-to-text.

S. G. a. C. M. U. CMU. The CMU pronouncing dictionary, 2014. URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

L. F. Cranor. A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, UPSEC'08, pages 1–15, USA, Apr. 2008. USENIX Association.

A. del Pozo, C. Aliprandi, A. Álvarez, C. Mendes, J. P. Neto, S. Paulo, N. Piccinini, and M. Raffaelli. SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 432–436, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/183_Paper.pdf.

D. Deng, Y. Zhou, J. Pi, and B. E. Shi. Multimodal Utterance-level Affect Analysis using Visual, Audio and Text Features. *arXiv:1805.00625 [cs, eess]*, May 2018. URL http://arxiv.org/abs/1805.00625. arXiv: 1805.00625.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

G. Dove, K. Halskov, J. Forlizzi, and J. Zimmerman. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 278–288, New York, NY, USA, May 2017. Association for Computing Machinery. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025739. URL https://doi.org/10.1145/3025453.3025739.

A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*, 43(3):411–435, Dec. 2014. ISSN 0925-9902, 1573-7675. doi: 10.1007/s10844-014-0304-9. URL http://link.springer.com/10.1007/s10844-014-0304-9.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. ISSN 1573-1405. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

J. A. Fails and D. R. Olsen. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pages 39–45, New York, NY, USA, Jan. 2003. Association for Computing Machinery. ISBN 978-1-58113-586-2. doi: 10.1145/604045.604056. URL https://doi.org/10.1145/604045.604056.

Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian Active Learning with Image Data. *arXiv:1703.02910 [cs, stat]*, Mar. 2017. URL http://arxiv.org/abs/1703.02910. arXiv: 1703.02910.

A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox. A semi-automatic approach to home video editing. In *Proceedings of the 13th annual ACM symposium on User interface software and technology - UIST '00*, pages 81–89, San Diego, California, United States, 2000. ACM Press. ISBN 978-1-58113-212-0. doi: 10.1145/354401.354415. URL http://portal.acm.org/citation.cfm?doid=354401.354415.

A. Girgensohn, S. A. Bly, F. Shipman, J. S. Boreczky, and L. Wilcox. Home Video Editing Made Easy - Balancing Automation and User Control. In M. Hirose, editor, *Human-Computer Interaction INTERACT '01: IFIP TC13 International Conference on Human-Computer Interaction, Tokyo, Japan, July 9-13, 2001*, pages 464–471. IOS Press, 2001.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA, Nov. 2016. ISBN 978-0-262-03561-3.

A. Graves. Sequence Transduction with Recurrent Neural Networks. *arXiv:1211.3711 [cs, stat]*, Nov. 2012. URL http://arxiv.org/abs/1211.3711. arXiv: 1211.3711.

D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. XAI-Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, Dec. 2019. doi: 10.1126/scirobotics.aay7120. URL https://openaccess.city.ac.uk/id/eprint/23405/. Publisher: American Association for the Advancement of Science.

M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating Summaries from User Videos. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 505–520, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10584-0. doi: 10.1007/978-3-319-10584-0_33.

S. HAI. Introducing Stanford's Human-Centered AI Initiative, 2022. URL https://hai.stanford.edu/news/introducing-stanfords-human-centered-ai-initiative.

J. Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, Feb. 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1807184115. URL https://www.pnas.org/content/116/6/1844. Publisher: National Academy of Sciences Section: Colloquium Paper.

T. T. Hewett, R. Baecker, S. Card, T. Carey, J. Gasen, M. Mantei, G. Perlman, G. Strong, and W. Verplank. *ACM SIGCHI curricula for human-computer interaction*. ACM, 1992.

L. E. Holmquist. Intelligence on tap: artificial intelligence as a new design material. *Interactions*, 24(4):28–33, June 2017. ISSN 1072-5520. doi: 10.1145/3085571. URL https://doi.org/10.1145/3085571.

A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G. C. Crişan, C.-M. Pintea, and V. Palade. Interactive machine learning: experimental evidence for the human in the algorithmic loop: A case study on Ant Colony Optimization. *Applied Intelligence*, Dec. 2018. ISSN 0924-669X, 1573-7497. doi: 10.1007/s10489-018-1361-5. URL http://link.springer.com/10.1007/s10489-018-1361-5.

E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*, pages 159–166, Pittsburgh, Pennsylvania, United States, 1999. ACM Press. ISBN 978-0-201-48559-2. doi: 10.1145/302979.303030. URL http://portal.acm.org/citation.cfm?doid=302979.303030.

X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-Based Automated Home Video Editing System. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5): 572–583, May 2004. ISSN 1051-8215. doi: 10.1109/TCSVT.2004.826750. URL http://ieeexplore.ieee.org/document/1294950/.

C.-w. Huang. Automatic Closed Caption Alignment Based on Speech Recognition Transcripts, 2003.

G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *International Workshop on Faces in Real-Life Images*, Oct. 2008. URL https://hal.inria.fr/inria-00321923.

X. D. Huang, H. W. Hon, and K. F. Lee. Large-vocabulary speaker-independent continuous speech recognition with semi-continuous hidden Markov models. In *Proceedings of the workshop on Speech and Natural Language - HLT '89*, page 276, Cape Cod, Massachusetts, 1989. Association for Computational Linguistics. ISBN 978-1-55860-112-3.

doi: 10.3115/1075434.1075480. URL http://portal.acm.org/citation.cfm?doid=1075434.1075480.

M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in Context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, June 2015. doi: 10.1109/CVPR.2015.7298710. ISSN: 1063-6919.

X. Jin, M. Evans, H. Dong, and A. Yao. Design Heuristics for Artificial Intelligence: Inspirational Design Stimuli for Supporting UX Designers in Generating AI-Powered Ideas. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–8, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8095-9. doi: 10.1145/3411763.3451727. URL https://dl.acm.org/doi/10.1145/3411763.3451727.

D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny. ICDAR 2015 competition on Robust Reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160, Aug. 2015. doi: 10.1109/ICDAR.2015.7333942.

W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. *arXiv:1705.06950 [cs]*, May 2017. URL http://arxiv.org/abs/1705.06950. arXiv: 1705.06950.

B. A. Kitchenham and S. Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University, July 2007. URL https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf. Backup Publisher: Keele University and Durham University Joint Report.

S. Kopf, T. Haenselmann, J. Kiess, B. Guthier, and W. Effelsberg. Algorithms for video retargeting. *Multimedia Tools and Applications*, 51(2):819–861, Jan. 2011. ISSN 1380-7501, 1573-7721. doi: 10.1007/s11042-010-0717-6. URL http://link.springer.com/10.1007/s11042-010-0717-6.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge. Understanding Low- and High-Level Contributions to Fixation Prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808, Venice, Oct. 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.513. URL http://ieeexplore.ieee.org/document/8237775/.

V. Këpuska. Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). *International Journal of Engineering Research and Applications*, 07 (03):20–24, Mar. 2017. ISSN 22489622, 22489622. doi: 10.9790/9622-0703022024. URL http://www.ijera.com/papers/Vol7_issue3/Part-2/D0703022024.pdf.

J. Lazar, J. H. Feng, and H. Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

M. Leake, A. Davis, A. Truong, and M. Agrawala. Computational video editing for dialogue-driven scenes. *ACM Transactions on Graphics*, 36(4):1–14, July 2017. ISSN 07300301. doi: 10.1145/3072959.3073653. URL http://dl.acm.org/citation.cfm?doid=3072959.3073653.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7553 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Computer science;Mathematics and computing Subject_term_id: computer-science;mathematics-and-computing.

C.-Y. Lee and S. Osindero. Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2231–2239, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.245. URL http://ieeexplore.ieee.org/document/7780614/.

D. C.-E. Lin, A. Germanidis, C. Valenzuela, Y. Shi, and N. Martelaro. Soundify: Matching Sound Effects to Video. *CoRR*, abs/2112.09726, 2021. URL https://arxiv.org/abs/2112.09726. arXiv: 2112.09726.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

J. Lindley, H. A. Akmal, F. Pilling, and P. Coulton. Researching AI Legibility through Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, New York, NY, USA, Apr. 2020. ISBN 978-1-4503-6708-0. URL https://doi.org/10.1145/3313831.3376792.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

I. S. MacKenzie. *Human-Computer Interaction: An Empirical Research Perspective*. Morgan Kaufmann, Amsterdam, 2013. ISBN 978-0-12-405865-1. URL http://www.sciencedirect.com/science/book/9780124058651.

V. Marda and S. Narayan. On the importance of ethnographic methods in AI research. *Nature Machine Intelligence*, 3(3):187–189, Mar. 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00323-0. URL https://www.nature.com/articles/s42256-021-00323-0. Number: 3 Publisher: Nature Publishing Group.

Y. Matsuo, M. Amano, and K. Uehara. Mining Video Editing Rules in Video Streams. In *Proceedings of the Tenth ACM International Conference on Multimedia*, MULTIMEDIA '02, pages 255–258, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1-58113-620-X. doi: 10.1145/641007.641058. URL https://doi.org/10.1145/641007.641058. event-place: Juan-les-Pins, France.

N. McDonald, S. Schoenebeck, and A. Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):72:1–72:23, Nov. 2019. doi: 10.1145/3359174. URL https://doi.org/10.1145/3359174.

Y. Mirsky and W. Lee. The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, 54(1):1–41, Jan. 2021. ISSN 0360-0300, 1557-7341. doi: 10.1145/3425780. URL http://arxiv.org/abs/2004.11138. arXiv: 2004.11138.

A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny. Deep Belief Networks using discriminative features for phone recognition. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5060–5063, May 2011. doi: 10.1109/ICASSP.2011.5947494. ISSN: 2379-190X.

L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 169–176, New York, NY, USA,

Nov. 2011. Association for Computing Machinery. ISBN 978-1-4503-0641-6. doi: 10.1145/2070481.2070509. URL https://doi.org/10.1145/2070481.2070509.

S. Mori, C. Suen, and K. Yamamoto. Historical review of OCR research and development. *Proceedings of the IEEE*, 80(7):1029–1058, July 1992. ISSN 00189219. doi: 10.1109/5.156468. URL http://ieeexplore.ieee.org/document/156468/.

A. C. Morris, V. Maier, and P. D. Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *INTERSPEECH*, 2004.

B. A. Myers, J. P. Casares, S. Stevens, L. Dabbish, D. Yocum, and A. Corbett. A multi-view intelligent editor for digital video libraries. In *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries - JCDL '01*, pages 106–115, Roanoke, Virginia, United States, 2001. ACM Press. ISBN 978-1-58113-345-5. doi: 10.1145/379437.379461. URL http://portal.acm.org/citation.cfm?doid=379437.379461.

H. Müller, A. Sedley, and E. Ferrall-Nunge. Survey Research in HCI. In J. S. Olson and W. A. Kellogg, editors, *Ways of Knowing in HCI*, pages 229–266. Springer, New York, NY, 2014. ISBN 978-1-4939-0378-8. doi: 10.1007/978-1-4939-0378-8_10. URL https://doi.org/10.1007/978-1-4939-0378-8_10.

F. Nack and A. Parkes. The Application of Video Semantics and Theme Representation in Automated Video Editing. In *Representation and Retrieval of Video Data in Multimedia Systems*, pages 57–83. Springer US, Boston, MA, 1997. ISBN 978-0-7923-9863-9. doi: 10.1007/978-0-585-31786-1_4. URL http://link.springer.com/10.1007/978-0-585-31786-1_4.

D. T. D. Nguyen, A. Carlier, W. T. Ooi, and V. Charvillat. Jiku director 2.0: a mobile video mashup system with zoom and pan using motion maps. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 765–766, Orlando Florida USA, Nov. 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868.2654884. URL https://dl.acm.org/doi/10.1145/2647868.2654884.

NHTSA. Automated Vehicles for Safety | NHTSA, 2022. URL https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.

N. J. Nilsson. *Artificial intelligence: a new synthesis*. Kaufmann, San Francisco, Calif, 5th print edition, 1998. ISBN 978-1-55860-467-4 978-1-55860-535-0.

Y. Niu and F. Liu. What Makes a Professional Video? A Computational Aesthetics Approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 22 (7):1037–1049, July 2012. ISSN 1051-8215, 1558-2205. doi: 10.1109/TCSVT.2012.2189689. URL http://ieeexplore.ieee.org/document/6162974/.

M. Obach, M. Lehr, and A. Arruti. Automatic Speech Recognition for Live TV Subtitling for Hearing-Impaired People. In *Volume 20: Challenges for Assistive Technology*. Assistive Technology Research Series, 2007. URL `http://ebooks.iospress.nl/publication/641`.

J. A. Okun, S. Zwerman, K. Rafferty, and S. Squires, editors. *The VES handbook of visual effects: industry standard VFX practices and procedures*. Focal Press, Taylor & Francis Group, New York, 2015. ISBN 978-0-240-82518-2 978-1-138-01289-9.

L. Onnasch, C. D. Wickens, H. Li, and D. Manzey. Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(3):476–488, May 2014. ISSN 0018-7208, 1547-8181. doi: 10.1177/0018720813501549. URL `http://journals.sagepub.com/doi/10.1177/0018720813501549`.

J. Owens and G. Millerson. Overview of Video Production. In *Video Production Handbook*. Routledge, 5 edition, 2011. ISBN 978-0-240-52221-0. Num Pages: 11.

P. Passarelli. autoEdit Fast Text Based Video Editing, 2019. URL `http://www.autoedit.io/`.

A. Pavel, C. Reed, B. Hartmann, and M. Agrawala. Video digests: a browsable, skimmable format for informational lecture videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 573–582, Honolulu Hawaii USA, Oct. 2014. ACM. ISBN 978-1-4503-3069-5. doi: 10.1145/2642918.2647400. URL `https://dl.acm.org/doi/10.1145/2642918.2647400`.

S. Poria, E. Cambria, and A. Gelbukh. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1303. URL `http://aclweb.org/anthology/D15-1303`.

M. Radut, M. Evans, K. To, T. Nooney, and G. Phillipson. How Good is Good Enough? The Challenge of Evaluating Subjective Quality of AI-Edited Video Coverage of Live Events. *Workshop on Intelligent Cinematography and Editing*, page 8 pages, 2020. ISSN 2411-9733. doi: 10.2312/WICED.20201127. URL `https://diglib.eg.org/handle/10.2312/wiced20201127`. Artwork Size: 8 pages ISBN: 9783038681274 Publisher: The Eurographics Association Version Number: 017-024.

A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin. A Local-to-Global

Approach to Multi-modal Movie Scene Segmentation. *arXiv:2004.02678 [cs]*, Apr. 2020. URL http://arxiv.org/abs/2004.02678. arXiv: 2004.02678.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *arXiv:1506.02640 [cs]*, May 2016. URL http://arxiv.org/abs/1506.02640. arXiv: 1506.02640.

O. Russakovsky, L.-J. Li, and L. Fei-Fei. Best of both worlds: Human-machine collaboration for object annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2121–2131, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298824. URL http://ieeexplore.ieee.org/document/7298824/.

M. K. Saini and W. T. Ooi. Automated Video Mashups: Research and Challenges. In M. Montagud, P. Cesar, F. Boronat, and J. Jansen, editors, *MediaSync: Handbook on Multimedia Synchronization*, pages 167–190. Springer International Publishing, Cham, 2018. ISBN 978-3-319-65840-7. doi: 10.1007/978-3-319-65840-7_6. URL https://doi.org/10.1007/978-3-319-65840-7_6.

M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. MoViMash: online mobile video mashup. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, page 139, Nara, Japan, 2012. ACM Press. ISBN 978-1-4503-1089-5. doi: 10.1145/2393347.2393373. URL http://dl.acm.org/citation.cfm?doid=2393347.2393373.

B. Settles. Active Learning Literature Survey. Computer Sciences Technical Report, University of Wisconsin–Madison, 2009. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

P. Shrestha, P. H. de With, H. Weda, M. Barbieri, and E. H. Aarts. Automatic mashup generation from multiple-camera concert recordings. In *Proceedings of the international conference on Multimedia - MM '10*, page 541, Firenze, Italy, 2010. ACM Press. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874023. URL http://dl.acm.org/citation.cfm?doid=1873951.1874023.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv:1712.01815 [cs]*, Dec. 2017. URL http://arxiv.org/abs/1712.01815. arXiv: 1712.01815.

A. Smith, V. Kumar, J. Boyd-Graber, K. Seppi, and L. Findlater. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *Proceedings of the 2018 Conference on Human Information Interac-*

tion&Retrieval - IUI '18, pages 293–304, Tokyo, Japan, 2018. ACM Press. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172965. URL http://dl.acm.org/citation.cfm?doid=3172944.3172965.

T. H. Soe and M. Slavkovik. AI video editing tools. What do editors want and how far is AI from delivering them? *CoRR*, abs/2109.07809, 2021. URL https://arxiv.org/abs/2109.07809. arXiv: 2109.07809.

T. H. Soe and M. Slavkovik. A content aware tool for converting videos to narrower aspect ratios. In *ACM International Conference on Interactive Media Experiences*, Alvaro, Portugal, 2022. ACM.

T. H. Soe, O. E. Nordberg, F. Guribye, and M. Slavkovik. Circumvention by design - dark patterns in cookie consent for online news outlets. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, NordiCHI '20, pages 1–12, New York, NY, USA, Oct. 2020. Association for Computing Machinery. ISBN 978-1-4503-7579-5. doi: 10.1145/3419249.3420132. URL https://doi.org/10.1145/3419249.3420132.

T. H. Soe, F. Guribye, and M. Slavkovik. Evaluating AI assisted subtitling. In *ACM International Conference on Interactive Media Experiences*, pages 96–107, Virtual Event USA, June 2021. ACM. ISBN 978-1-4503-8389-9. doi: 10.1145/3452918.3458792. URL https://dl.acm.org/doi/10.1145/3452918.3458792.

T. H. Soe, C. T. Santos, and M. Slavkovik. Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way, 2022. URL https://arxiv.org/abs/2204.11836.

H. Steck, R. van Zwol, and C. Johnson. Interactive Recommender Systems: Tutorial. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 359–360, New York, NY, USA, Sept. 2015. Association for Computing Machinery. ISBN 978-1-4503-3692-5. doi: 10.1145/2792838.2792840. URL https://doi.org/10.1145/2792838.2792840.

B. Suhm, B. Myers, and A. Waibel. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction*, 8(1):60–98, Mar. 2001. ISSN 10730516. doi: 10.1145/371127.371166. URL http://portal.acm.org/citation.cfm?doid=371127.371166.

N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. *IEEE Spectrum*, 58(10):50–55, Oct. 2021. ISSN 1939-9340. doi: 10.1109/MSPEC.2021.9563954. Conference Name: IEEE Spectrum.

A. Truong, F. Berthouzoz, W. Li, and M. Agrawala. QuickCut: An Interactive Tool for Editing Narrated Video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 497–507, Tokyo Japan, Oct. 2016. ACM. ISBN 978-1-4503-4189-9. doi: 10.1145/2984511.2984569. URL https://dl.acm.org/doi/10.1145/2984511.2984569.

B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):3–es, Feb. 2007. ISSN 1551-6857. doi: 10.1145/1198302.1198305. URL https://doi.org/10.1145/1198302.1198305.

P. Turner. *A Psychology of User Experience: Involvement, Affect and Aesthetics*. Springer, Dec. 2017. ISBN 978-3-319-70653-5. Google-Books-ID: L5xBDwAAQBAJ.

N. van Berkel, M. B. Skov, and J. Kjeldskov. Human-AI interaction: intermittent, continuous, and proactive. *Interactions*, 28(6):67–71, Nov. 2021. ISSN 1072-5520. doi: 10.1145/3486941. URL https://doi.org/10.1145/3486941.

K. Vertanen and P. O. Kristensson. On the benefits of confidence visualization in speech recognition. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, page 1497, Florence, Italy, 2008. ACM Press. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357288. URL http://portal.acm.org/citation.cfm?doid=1357054.1357288.

J. Wang and A. Moulden. AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8095-9. doi: 10.1145/3411763.3443452. URL https://dl.acm.org/doi/10.1145/3411763.3443452.

M. Wang, G.-W. Yang, S.-M. Hu, S.-T. Yau, and A. Shamir. Write-a-video: computational video montage from themed text. *ACM Transactions on Graphics*, 38(6): 1–13, Nov. 2019. ISSN 0730-0301, 1557-7368. doi: 10.1145/3355089.3356520. URL https://dl.acm.org/doi/10.1145/3355089.3356520.

P. Wang, Y. Yang, Z. Huang, J. Cao, and H. T. Shen. WeMash: An Online System for Web Video Mashup. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 753–754, Orlando Florida USA, Nov. 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868.2654868. URL https://dl.acm.org/doi/10.1145/2647868.2654868.

S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai. ESPnet: End-

to-End Speech Processing Toolkit. *arXiv:1804.00015 [cs]*, Mar. 2018. URL http://arxiv.org/abs/1804.00015. arXiv: 1804.00015.

A. G. Wilson, C. Dann, C. G. Lucas, and E. P. Xing. The Human Kernel. *arXiv:1510.07389 [cs, stat]*, Oct. 2015. URL http://arxiv.org/abs/1510.07389. arXiv: 1510.07389.

H.-Y. Wu, T. Santarra, M. Leece, R. Vargas, and A. Jhala. Joint Attention for Automated Video Editing. In *ACM International Conference on Interactive Media Experiences*, pages 55–64, Cornella, Barcelona Spain, June 2020. ACM. ISBN 978-1-4503-7976-2. doi: 10.1145/3391614.3393656. URL https://dl.acm.org/doi/10.1145/3391614.3393656.

Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang. Deep Learning for Video Classification and Captioning. *arXiv preprint arXiv:1609.06782*, 2016.

J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

W. Xu. Toward human-centered AI: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, June 2019. ISSN 1072-5520, 1558-3449. doi: 10.1145/3328485. URL https://dl.acm.org/doi/10.1145/3328485.

Yale Song, J. Vallmitjana, A. Stent, and A. Jaimes. TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7299154. URL http://ieeexplore.ieee.org/document/7299154/.

Q. Yang, A. Steinfeld, C. Rosé, and J. Zimmerman. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, Apr. 2020. ACM. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376301. URL https://dl.acm.org/doi/10.1145/3313831.3376301.

K. Yi*, C. Gan*, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. CLEVRER: Collision Events for Video Representation and Reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik. Patch-VQ: 'Patching Up' the Video Quality Problem. *arXiv:2011.13544 [cs]*, Nov. 2020. URL http://arxiv.org/abs/2011.13544. arXiv: 2011.13544.

W.-Y. Yo, J.-J. Leou, and H.-H. Hsiao. Video retargeting using non-homogeneous scaling and cropping. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–5, Kaohsiung, Taiwan, Oct. 2013. IEEE. ISBN 978-986-90006-0-4. doi: 10.1109/APSIPA.2013.6694167. URL `http://ieeexplore.ieee.org/document/6694167/`.

J. Zimmerman, J. Forlizzi, and S. Evenson. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, page 493, San Jose, California, USA, 2007. ACM Press. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240704. URL `http://portal.acm.org/citation.cfm?doid=1240624.1240704`.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230866092 (print)
9788230852156 (PDF)