
Video Editing Tools
with Artificial Intelligence
Video Editor

Video Editing Tools with Artificial Intelligence

Hachik Yazadzhiyan

Technical University of Sofia,
Branch Plovdiv
Department of Computer Systems and Technologies

Abstract . **Video editing can be a very tedious task, so it's no surprise that artificial intelligence is increasingly being used to streamline the workflow or automate tedious tasks. However, it is very difficult to get an overview of what intelligent video editing tools exist in the research literature and the automation needs of video editors. So, we've identified the field of smart video editing tools in research and are surveying the opinions of professional video editors. We also summarized the current state of the art in artificial intelligence research with the intention of identifying what are the possibilities and current technical limitations towards truly intelligent video processing tools. The findings contribute to the understanding of the field of intelligent video editing tools, highlight unmet needs for automation, research, and provide general suggestions for further research on intelligent video editing tools.**

Keywords: AI, video editing, Adobe Premier, DaVinci Resolve

I. INTRODUCTION

Video is the most popular form of content on the Internet. According to the Visual Networks Index, 82.5% of Internet traffic in 2023 was video content. Mobile phones, video sharing and social media platforms make taking and posting videos easier and faster than ever before. However, editing these videos still takes too much time. Video remains a pain to edit because it requires working on individual frames, in addition to being a two-track medium with both audio and image. There are various attempts to make video editing easier. One approach is to automate the editing process using artificial intelligence (AI).

Here we are interested in the state of the art in video editing automation, focusing specifically on the mismatch between what is desired and what is achievable with current AI technology.

Okun et al. [Okun et al., 2015] define video editing as the act of cutting and joining parts of one or more sources together to make a single edited movie. Video editing tools can be roughly defined as (computer) programs that people can use to perform the task of video editing, ie. combining video segments. Video editing is one area where AI is being used to automate or augment the tasks of human video editors.

Smart video editing tools have been trying since the beginning of digital video editing to make video editing easier. One of the key topics in intelligent video editing tools is the problem of allowing video manipulation from high-level abstractions, such as frames and dialog, rather than frames. An early example of such a tool is Silver [Casares et al., 2002] from 2002, which provides intelligent selections of video clips as well as abstract views of video editing using video metadata. A more recent example of a smart video editing tool is Roughcut [Leake et al., 2017]. Roughcut enables computational editing for dialogue-driven scenes using user-entered scene dialogue, raw footage, and editing idioms. There is an open source tool autoEdit [Passarelli, 2019] and research [Berthouzoz, 2012] that allow text editing of video interviews by linking text transcripts to the videos.

Fully AI-controlled video production has recently attracted much research interest [Xue et al.; Hua et al., 2004]. Currently, AI-controlled video production is focused on creating automated video summaries or mixes. These fully automated video editing methods, as used to create video summaries and composite applications, are not considered intelligent video editing tools, as they are simple algorithms that perform a very narrow and specific query that does not require intelligence or user interaction.

Advances in artificial intelligence in image processing, vision, and natural language processing have made

video editing. But has the dream come true to break away from the drudgery of video editing? The answer, of course, depends on whose dream we're talking about. How can progress in automated video editing be evaluated? This question of assessing progress in automatic video editing can be approached from two angles. The first is to review the literature, while the second is to examine the expectations of human video editors and compare them to the state of the art in artificial intelligence. In this article, we do both.

The main challenge that smart video editing tools tried to solve was streamlining the video editing process for users. This is usually done by eliminating the tedious tasks of going through videos frame by frame. The proposed solutions and tools differ in a number of ways that stem from the approach to the problem, the intended goal, the underlying technology, the level of abstraction(s), the interactions offered, and the modalities for said interactions.

We reviewed the state-of-the-art video AI applications from two perspectives: i) general video AI technology; ii) specific AI video editing technology. Common AI technology for video includes a wide variety of video tasks such as object tracking, object detection, speech recognition, video reasoning, action detection, sentiment detection in videos. The specific AI technology for video editing is much narrower such as processing video scripts, frames and scenes, and video editing mining rules.

We surveyed 13 video editors whose video editing experience ranged from 1 year to 22 years. The survey covers their experience in video editing, thoughts on AI video editor, and video editing automation needs. The survey responses are then used to perform a thematic analysis to compile an overview of expectations, requirements and issues regarding automation in video editing tools. We compare the opinions and expectations revealed by the survey with current knowledge of machine learning for content creation/manipulation, automated video editing, and other AI tools. We discuss how the latest in AI can create an ideal AI video editing tool for video creators.

This paper is structured as follows: Section 2 provides an overview of intelligent video editing tools and AI techniques in video. Intelligent video editing tools in the literature are compared and summarized in Section 3. The survey of (human) video editors, including the procedure and summarized results are in Section 4. In Section 5, previous works on solutions for intelligent video editing tools and that of user expectations from the survey are contrasted and some AI techniques are proposed as a potential solution to meet user expectations. Finally, the paper summary, conclusions and future work are presented in Section 6.

II. Ease of Use

Creating better tools to facilitate video editing has always been a research agenda since the introduction of digital video.

First, we reviewed different approaches to creating intelligent and automatic video publishing tools. Next, we turn to AI methods that have been applied directly or indirectly in video editing.

One of the first projects that tried to simplify video editing was Silver [Myers et al., 2001; Long et al., 2004]. In the first version of the Silver project, there are different types of views, namely Transcript, Timeline, Preview and Script views.

This editing tool also explores smart editing with smart selection, cut, delete, copy, paste and re-attach using photos and scene detection. Smart editing is built using the video's metadata layer in the form of text transcription, short boundary detection, and optical character recognition (OCR) to generate metadata. In the second iteration of the Silver tool [Long et al., 2004], the tool implemented lenses (clip, photos, frames) and semantic scaling to make easier tasks such as visualizing the correct frames to cut when joining two video segments. Most smart video tools are designed to create only one specific type of video. For example, QuickCut [Truong et al., 2016] was created for composing narrated videos, and Video digests [Pavel et al., 2014] was created for summarizing video lectures.

Next, we look at methods for fully automated video editing. Automatic video editing is the computational processing and compositing of video segments without any input from a human editor. Automatic video editing can be performed on recorded videos or, on a much larger scale, on video archives. Early work on automated video editing focused on rule-based strategies for generating video sequences [Butler and Parkes, 1997] or a semantic-based method for selecting and automatically editing user requests in the domain of video documentaries [Bocconi, 2004] . Mashups, a combination of multiple videos for a single event, is another type of automated video editing. Work called Virtual Director [Shrestha et al., 2010] created a method for generating mixes for concert recordings

that maximizes what makes a good concert video based on rules derived from interviewing video editors and film grammar literature. Automated video editing can also be used to broadcast live events. The work of [Radut et al., 2020] discussed not only the prototype AI video editor for live events, but also evaluation and discussions of evaluation methods to measure the quality of AI-edited live events. AI can also be used to automatically create edited videos. Made by Machine When AI encountered the archive from the BBC, it created 150 short complications from the BBC archive [R&D, 2018]. [Taskir et al., 2006] presented a summation method for skimming video programs using speech transcript from speech recognition systems. [Truong et al., 2016] provides a summary of video abstraction or summation methods, such as generating a series of keyframes or moving images in order to provide information about a video in the shortest possible time. Work on automated video editing of corporate meetings by [Wu et al., 2020] uses learned editing decisions from human-edited video and uses two attention models on both audio and video. EDL-Edit Decision List is a text-based language that encodes compositional decisions with an ordered list of clips and timecode data. It is used as the output of many automated video editing systems [Taskir et al., 2006; Wu et al., 2020; Passarelli, 2019], which video editors can then use the automated editing solutions encoded with EDL to continue their editing in software such as Adobe Premier Pro or Davinci Resolve.

How AI techniques can be used to extract information from videos is a very diverse area of research. We are particularly interested in face recognition, object detection, object tracking, scene detection, sentiment analysis, video reasoning, and video captioning. Face recognition refers to the problem of identifying whether a human face is present in an image and possibly whose, while object detection is the problem of identifying a specific object in an image. Object tracking is the problem of identifying and locating a specific object and tracking its movement through the frames of a video. Scene detection or video segmentation is the identification of segments that are semantically or visually related in a video. Sentiment analysis is the problem of matching the mood that would be conveyed by a piece of content: whether it is happy, sad, ironic, etc. Video captioning [Wu et al., 2016] is an AI technique that generates natural descriptions that capture the dynamics of video.

There are video AI techniques that are more specific to video editing. [Matsuo et al.] presented a data mining technique to detect editing patterns (composed of loose, medium-tight frames and rules) from videos in order to create reproducible editing patterns. Earlier work by [Butler and Parkes, 1997] presented a rule and query-based approach to automating video editing. Automated video editing by modeling the editing process and using semantics is presented in [Nack and Parkes, 1997].

III. SMART VIDEO EDITING TOOLS

In this chapter, we review the literature and summarize how previous intelligent video editing tools formulate and solve the problem of making video editing an easier task. Literature searches for intelligent video editing tools were performed using the keywords (intelligent OR intelligent OR automated OR AI) AND (video editor OR video editing) in computer science literature databases, which are DBLP, ACM Digital Library, and Google Scholar. The titles and abstracts were then read and filtered based on the inclusion criteria, which stipulated that the included papers should describe an intelligent approach to creating a video editing tool for users. Included documents should also contain a description and/or implementation of the user interface. The references of the included papers were also scanned to find more related literature. The resulting papers were then summarized and grouped into three topics, namely video editing tasks, interaction with an automated editor (human-computer interaction), and AI technology. The list of papers included in this section is in Table 1

A. Video editing tasks - of smart video editors

This subsection introduces the field of intelligent video editing tools in terms of various tasks addressed in the video editing workflow and summarizes the approach for each task. Video segmentation is the most common task that smart video editors try to solve. All previous work reviewed in this section uses some form of video segmentation method, but different approaches are used to perform the segmentation. From now on, a video is defined as a continuous segment of video from a single source file. The first segmentation method identified was the use of shot detection.

A frame is a continuous continuous sequence of images Okun et al. [2015]. Shot detection segmentation is performed with the image analysis methods in [Casares et al., 2002; Long et al., 2004] and shot detection is performed using features that are camera motion, brightness, and duration in [Shipman, 2008]. [Wu et al., 2015] uses frame and subframes to segment user-generated videos (a subframe is defined as a basic unit of video that contains consistent camera movement and self-contained semantics). [Casares et al., 2002] also consider the segmentation of video and audio at different locations in the case of L-cuts.

The second segmentation method uses synchronization with a text transcript to create cuts that depend on the content or meanings of the video audio. [Leake et al., 2017] used pre-written lines from the scene dialog script to segment the video, while [Pavel et al., 2014] ran the text transcription through Bayesian topic segmentation [Eisenstein and Barzilay, 2008] BSec to identifying sections and subsections of informational videos.

Other segmentation approaches include approaches such as using gaze [Kimura et al., 2005], starting from user-marked points, similarity measures of frames to those points [Chi et al., 2013], and fitness assessment for reduction in talking head interview videos [Bertuso, 2012]. User-generated video summaries by [Cattelan et al., 2008] created segmentation using user viewing actions and comments. Video segmentation is also essential to discuss the next task, which is to compose video segments.

Composing video segments is the second most common task covered in our list of AI video editing tools. The most common approach to streamlining the composition of video segments is the use of in-scene dialogues [Leake et al., 2017] or text transcriptions [Berthouzoz, 2012; Wu et al., 2015] as the starting point of the composition. Video dialogue must be written and provided as input, but a text transcript can be generated using speech recognition technology.

For example, [Truong et al., 2016] used a text transcript converted from narrated audio or voice-over instead of manually created text transcripts.

In Roughcut[Leake et al., 2017], the video segments created for each respective dialogue and speaker using automation. The dialogue order follows the script provided. However, the creative composition can be changed by the user who chooses a combination of video editing idioms. Similarly, the story schema created by the editor was used to compose segments [Truong et al., 2016]. Segmentation is done by cutting unwanted parts, such as certain phrases from an interview or repeated words, by selecting the corresponding text from the video recording [Berthouzoz, 2012].

[Chi et al., 2013] uses user-supplied tags in video clips as a way to organize and compose segments, and editors can change the composition of the overall segment by arranging the tags that correspond to steps in a demo video. Composition for the purpose of creating user-generated video summaries is done using modeled viewer intent from gazes [Kimura et al., 2005] and with viewing activities and user comments in [Cattelan et al., 2008].

Visualizing the timeline and videos comes in the form of viewing the timeline and videos at different levels of abstraction and different representation of the video timeline in an alternative way, such as text. Video rendering abstractions can be in the form of frame, frame, and clip.

A frame is a still image of a video, while a frame is a continuous sequence of images [Okun et al., 2015]. An overview of various forms of video abstractions is available in [Truong and Venkatesh]. Visualization of clips using a representative frame is discussed as a method to allow rapid judgment of video content on the timeline [Long et al., 2004].

[Casares et al., 2002] offers timeline visualization in different levels of abstractions, which are script, editable transcript, and timeline views. A second approach to timeline visualization is by representing the timeline in terms of textual transcripts [Casares et al., 2002; Truong et al., 2016; Bertuso, 2012; Pavel et al., 2014]. As noted in the previous paragraphs, the textual representation of the timeline can sometimes be manipulated at the word level to make changes to the actual composition of the video frames.

Intelligent video manipulation is discussed in only two of the intelligent video editors. The first work [Casares et al., 2002] used smart selection, clipping, clipping, pasting and re-pinning. All these actions are performed using photo borders with image analysis. The second work simply provides intelligent selection or intelligent cutting of video segments using the video transcript [Berthouzoz, 2012]. The lack of many examples of this task and the tasks mentioned below may be due to the fact that all smart video editing tools are proof of concepts, thus lacking these very important but non-essential features.

Create transitions. The simpler method of creating transitions is covered in two separate works. [Truong et al., 2016] created a method to automate an aesthetically pleasing transition by formulating transit tasks as dynamic programming where bad transition points, such as jumps, are penalized. [Berthouzoz, 2012] uses a different approach to create hidden transitions by using hierarchical clustering of frames and finding the shortest path between frames as transaction points.

Record videos. [Truong et al., 2016] presented a new approach for registering video clips with audio annotations by registering video during the capture process. In their work, logging can be done by audio, in addition to logging with tags during the recording review.

B. Interaction with automation

In this section, the mode of interaction as well as the level of video abstractions [Truong and Venkatesh] will be summarized. The primary mode of interaction used in most of the smart video editing tools we examined is through a graphical user interface (GUI) with a keyboard and mouse. An exception is the one case from [Kimura et al., 2005], which investigates gaze-based interaction. However, the level of abstraction and granularity of control that users have varies across tools. In video editing tools without any abstraction, editing must be done at the individual frame level, which is very labor-intensive. However, some of the smart video editing tools offer video manipulation at multiple levels of abstraction. Two examples of tools offering multiple abstractions are Silver [Casares et al., 2002] and Quickcut [Truong et al., 2016]. Silver which offers three abstractions which are: clip, shot and frames. Quickcut offers abstractions in terms of spoken words and frames.

Some video editing tools operate at a very high level of abstraction. In DemoCut [Chi et al., 2013], for example, video editing is done by abstracting steps and markers for those steps. Similarly, in RoughCut[Leake et al., 2017], the user can manipulate the timeline by using dialogue lines in a dialogue script and editing decisions in the form of idioms. Manipulation at higher levels, however, comes at the cost of the ability to make finer adjustments at the frame level. However, in three of the intelligent video editing tools [Leake et al., 2017; Truong et al., 2016; Passarelli, 2019] video editing work can be exported as EDL (Edit Description Language), which can be used with other commercial video editing software to make frame-level corrections and complete the editing process on video.

C. AI technology is used

Video segmentation. Earlier work on intelligent video editing tools relied on image analysis to detect frame boundaries and find representative frames for each frame [Casares et al., 2002] or using a combination of image analysis, make background knowledge and model matching in [Shipman, 2008].

Frame detection rules are hand-crafted for targeted video types. In [Casares et al., 2002], hand-crafted transcripts were aligned to videos using speech recognition. Segmentation of video lectures into thematically coherent units is performed by topic segmentation on the textual transcript of the video in [Pavel et al., 2014]. Another form of segmentation with audio annotations is explored in [Truong et al., 2016]. It works by using motion-based segmentation and refinement through audio annotation into semantically relevant segments. Motion-based segmentation is performed by detecting continuous motion in the video, while semantic segmentation corresponds to actions or topics in the video.

An introduction to the principles of imaging using computational techniques can be found in [Wu et al., 2015] and [Leake et al., 2017]. Domain-specific principles for detecting video clipping points, selected videos and selected audio fragments, are selected through interviews and presented as optimization problems [Wu et al., 2015]. In [Leake et al., 2017], 12

basic film editing idioms (jump avoidance, emotion enhancement, etc.) are represented in terms of feature parameters that enter as inputs to a hidden Markov model to generate editing solutions.
A Hidden Markov Model (HMM) is a statistical approach to modeling sequences in which the series of internal states are hidden. The features used in HMM include labels generated using speech-to-text, face recognition, and structural information from the clips.

IV. WHAT EDITORS WANT FROM A SMART VIDEO EDITING ASSISTANT

In this section, we report on the survey we conducted to explore the opinions of (human) video editors on what constitutes an ideal AI video editor.

Study procedure
General information.
The average video editing experience among our survey participants was 9.75 years, with the shortest ranging from 1 year to the longest being 22 years. In terms of the type of video participants work with, each participant listed about 3 types of video. The most common types of videos are advertising, documentary, presentation, sports, social media and news videos.
In terms of software programs, participants used an average of 5 video editing programs. The most commonly mentioned editing programs are Adobe Premier Pro and DaVinci resolve. In addition, lesser-known programs such as Avid, Flimora 9 and VizStory are also mentioned, as well as video services such as Rev.com and Descript.com. When asked what AI technology they have heard of in the context of video editing, 8 out of 13 respondents answered with AI technology in branded and commercial offerings such as Adobe Premier Pro CC and Magisto. For the remaining 5 respondents, the answers covered AI techniques which are automatic correction, noise reduction, background removal, video stabilization, object detection/image annotation, segmentation, zooming, deep falsification, automated video signals, face recognition and speech to text.

Perfect AI editor. The answer to the question “What would you like the perfect AI video editing tool to be?” answers vary greatly from one another. However, we identified five themes in the responses. They are AI as a tool for video editing tasks, AI as a tool for project management tasks, automatic tools for improving aesthetic quality, human AI problems, and AI for content discovery.
AI as a tool for video editing tasks is the largest category that most responses fall into. This contains keywords which are shot recognition, video compositing, filtering out bad videos based on dialogue lines, syncing songs and subtitles, translation and language understanding. The project management AI topic includes terms that are video metadata creation, data management, and adoption. The next topic is AI to improve aesthetic quality, which covers automatic color grading and automatic audio equalization. In addition to this, there are human-AI concerns regarding AI, such as the balance of control and automation, user-centric AI, and personalization. Finally, terms in the content discovery AI topic include suggesting stock videos and stock music videos based on existing content on the timeline.
When asked how to interact with the AI video editor, most of the responses mentioned that they would like to interact via voice, followed by those who would like to interact via a keyboard and mouse GUI. In addition to these two main modes of interaction, various other modes such as touch interface, gestures, brain computer interface are mentioned several times. Some answers mentioned contextual commands based on project status as essential for communication, as well as an AI OFF button that allows automation to be easily turned off.
The last question on the topic of the perfect AI editor asks about the level of abstractions (human) editors would like to work with. Most of the respondents said that they would like to manipulate the video at the keyframe level in their vision of a perfect AI video editing tool. The second and third most popular levels of abstraction are clips and frames. Other types of abstractions that are mentioned once each are sequences, history, and frame. Two respondents mentioned that they would like a flexible type of abstraction where they can adjust the level of abstractions for the underlying principles of interaction.
AI and Workflow. The second part of the study explores the following topics: tasks in the video editing workflow that participants want to automate, and related issues of level of autonomy and modes of interaction for these tasks. The answers to the questions in the workflow section consist of four themed tasks: Video Editing Tasks, Aesthetic Enhancements, Video Pre-Editing Tasks, and Suggestion Tasks.

The most popular keywords used to describe video editing tasks were segmentation and subtitling, which were mentioned in three responses. The second most popular keywords for video editing tasks are video segmentation and bad frame filtering (which were mentioned twice). In addition, the following video editing tasks are mentioned once: content analysis, video archiving, face detection, crop placement, transition frame selection, audio-to-video synchronization, and two-channel audio and video selection.

The next frequently mentioned topic task in the answers that we would like to automate is aesthetic quality improvements. The most popular terms for these are color correction and audio equalization. In addition, tasks such as visual improvements, background removal, and stutter removal were mentioned once each. For pre-editing tasks, video recording is mentioned twice and automated timecode creation is mentioned once. In terms of suggestive tasks, answers include a clip suggestion with editing styles, general help, and music suggestions.

The final question is how well editors expect the editor to be AI compared to the tools they've used. In this question , four respondents said they wanted it to be very similar or familiar. Two responses said they wanted there to be a basic level of similarity. The keyword "easy to use" found in two answers is another word that can have a similar meaning to "familiar". One respondent mentions a plug-in approach to integrate AI video editing tools into existing tools. Only two respondents said they expect AI video editors in the future to be very different or not at all.

V. CHALLENGES FOR AI VIDEO EDITING TOOLS

Video Editing Tasks There is considerable overlap between the video editing tasks identified in the literature and mentioned in the results of our study. Here we focus on the unexplored parts of the video editing workflow. The first is to sync audio and video from different tracks. This task has already been investigated in a different but related context of automated video mixture generation [Wu et al., 2015; Shrestha et al., 2010]. Filtering out bad frames or bad segments in the video has not been studied, but it should be done with informed research to understand what the video editors meant when they said they were bad frames. Finally, we have language issues such as automatic translation, subtitles and language understanding when editing video. Subtitling and automated translation system can be automated using machine automation. However, understanding language in the context of video editing requires both advances in natural language processing and an understanding of the use of video editing terms and language in the context of video editing.

Video Logging and Metadata Creation Video logging is watching a video recording and tagging its content using timecodes. It has been identified both in the intelligent video editing literature and in our research as one of the tasks that users would like to automate. Current video recording techniques in the literature are very limited to specific applications; namely demo videos and dialogue based videos. Speech recognition to convert video to text and apply text processing techniques has many potential use cases in video recording and metadata creation. Another interesting area of AI to watch out for is video reasoning and understanding combinations of patterns from both visual and linguistic input [Wu et al., 2016].

Voice-based interactions for video editing are the most common mode of interaction that people say they would like to use to communicate with an AI video editing tool. The potential of voice user interfaces in video editing tools has not been explored. [Chang et al., 2019] presented the design space exploration of voice-based interactions for navigating instructional videos. Since voice interaction in video editing is completely uncharted territory, the starting point should be design research. Another possibility is to study context-free single voice commands instead of interactions. Tasks that the user would like to fully automate, such as aesthetic quality enhancements, file management or pre-editing tasks, should be ideal for voice command design.

Personalization in this context is the ability of a smart video editing tool to adapt to the user by learning from processed videos, produced videos and usage patterns in the software. This topic is absent from the literature but found in our study in the form of understanding the context of video editing and personalization. Rule-based learning of video editing rules is discussed in [Matsuo et al.].

It is important to note that the smart video editing tools discussed in Section 3 are designed to solve the editing of a specific type of video: in fact, eight of them target only one type of video. However, our survey of video editors listed three types of videos processed by each on average. Since most of the smart video editing tools are built for one type of video, there are concerns about the general applicability of the techniques mentioned in the results.

The video editing tasks described in the Smart Video Editing Tool literature are focused on the interactions during the basic video editing task. However, research participants need automation for additional tasks such as file/media organization, aesthetic quality improvements, pre-editing tasks and content suggestions. A video editing content suggestion suggests good videos or music segments to add to an existing video or story.

In the survey results, voice interaction was the most common mode of interaction desired by participants. However, only one work on an intelligent video editor includes voice [Chi et al., 2013], where voice annotations are used to tag videos. This popularity of voice interaction can be attributed to the popularity of AI-based voice assistance in mobile phones and smart home speakers, as well as the representation of AI as a voice in science fiction.

The AI techniques used in smart video editing tools are heuristics-based systems. Other approaches such as neural networks and machine learning-based approaches are less studied. The study by [Dove et al., 2017] concluded that machine learning is a difficult design material to work with to create user experiences, as it is difficult to create machine learning-based prototypes and requires machine learning collaborators training to perform it.

VI. CONCLUSION

In this paper, we have defined intelligent video editing tools and presented an overview of the existing literature on (intelligent) video editing, user interaction, and AI technology.

We also surveyed video editors about their needs for automation in their video editing workflow. Doing research in this field of study requires knowledge of video editing, human computer interaction, and AI or machine learning. Intelligent video editing tools requiring these three very different expertises is one of the reasons that the literature on them is very limited compared to each of the fields individually.

In this study, there is a large amount of crossover between literature and research findings in the area of video editing tasks. However, there are areas such as video recording or organizing video editing projects, aesthetic quality correction, and content suggestions that need to be explored further to fulfill the research needs. Our conclusion is that with greater involvement of the machine learning community, the ideal AI editor can be achieved. In future work, we intend to contribute to this goal.

REFERENCES

[1] F. Berthouzoz. Tools for Placing Cuts and Transitions in Interview Video. page 8, 2012. S. Bocconi. Semantic-aware automatic video editing. In Proceedings of the 12th annual ACM international conference on Multimedia - MULTIMEDIA '04, page 971, 2004.

[2] S. Butler and A. Parkes. Film sequence generation strategies for automatic intelligent video editing. Applied Artificial Intelligence, 11(4):367–388, June 1997.

[3] J. Casares, AC Long, BA Myers, R. Bhatnagar, SM Stevens, L. Dabbish, D. Yocum, and A. Corbett. Simplifying Video Editing Using Metadata. page 10, 2002.

[4] RG Cattelan, C. Teixeira, R. Goularte, and MDGC Pimentel. Watch-and-comment as a paradigm towards ubiquitous interactive video editing. ACM Transactions on Multimedia Computing, Communications, and Applications, 4 (4):1–24, Oct. 2008.

[5] M. Chang, A. Truong, O. Wang, M. Agrawala, and J. Kim. How to Design Voice Based Navigation for How-To Videos. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–11, May 2019.

[6] P.-Y. Chi, J. Liu, J. Linder, M. Dontcheva, W. Li, and B. Hartmann. DemoCut: generating concise instructional videos for physical demonstrations. page 10, 2013. V. Cisco. Cisco visual networking index: Forecast and trends, 2017–2022. White Paper, 1:1, 2018.

[7] G. Dove, K. Halskov, J. Forlizzi, and J. Zimmerman. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. pages 278–288, 2017. J. Eisenstein and R. Barzilay. Bayesian unsupervised topic segmentation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08, page 334, 2008.

[8] 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.

[9] T. Kimura, K. Sumiya, and H. Tanaka. A video editing support system using user gazes. In

PACRIM. 2005 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 2005., pages 149–152, 2005.

[10]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.

[11]AC Long, BA Myers, J. Casares, SM Stevens, and A. Corbett. Video Editing Using Lenses and Semantic Zooming. page 10, 2004.

[12]Y. Matsuo, M. Amano, and K. Uehara. Mining Video Editing Rules in Video Streams. page 4. BA Myers, JP 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/IEEECS joint conference on Digital libraries - JCDL '01, pages 106–115, 2001.

[13]F. Nack and A. Parkes. The Application of Video Semantics and Theme Representation in Automated Video Editing. In HJ Zhang, P. Aigrain, and D. Petkovic, editors, Representation and Retrieval of Video Data in Multimedia Systems, pages 57–83. 1997.

[14]JA Okun, S. Zwerman, K. Rafferty, and S. Squires, editors. The VES handbook of visual effects: industry standard VFX practices and procedures. 2015.

[15]P. Passarelli. autoEdit Fast Text Based Video Editing, 2019.

[16]A. Pavel, C. Reed, B. Hartmann, and M. Agrawala. Video digests: a browseable, skimmable format for informational lecture videos. In Proceedings of the 27th annual ACM symposium on User interface software and technology, pages 573–582, Oct. 2014.

[17]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.

[18]Artwork Size: 8 pages ISBN: 9783038681274 Publisher: The Eurographics Association Version Number: 017-024.

[19]B. R&D. AI and the Archive - the making of Made by Machine, 2018. F. Shipman. Authoring, Viewing, and Generating Hypervideo: An Overview of Hyper-Hitchcock. 5(2):19, 2008.

[20]P. Shrestha, PH de With, H. Weda, M. Barbieri, and EH Aarts. Automatic mashup generation from multiple camera concert recordings. In Proceedings of the international conference on Multimedia - MM '10, page 541, 2010.

[21]C. Taskir, Z. Pizlo, A. Amir, D. Ponceleon, and E. Delp. Automated video program summarization using speech transcripts. IEEE Transactions on Multimedia, 8(4):775–791, Aug. 2006.

[22]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, Oct. 2016.

[23]T. Truong and S. Venkatesh. Video Abstraction: A Systematic Review and Classification. 3(1):37. 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, June 2020.

[24]Y. Wu, T. Mei, Y.-Q. Xu, N. Yu, and S. Li. MoVieUp: Automatic Mobile Video Mashup. IEEE Transactions on Circuits and Systems for Video Technology, 25(12):1941– 1954, Dec. 2015.

[25]Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang. Deep Learning for Video Classification and Captioning. arXiv preprint arXiv:1609.06782, 2016. C. Xue, L. Li, F. Yang, P. Wang, T. Wang, and Y. Zhang. Automated Home Video Editing: a Multi-Core Solution. page 2.

View publication stats