Automation in Video Editing: Assisted workflows in video editing

THAN HTUT SOE, University of Bergen, Norway

Capturing, publishing and distribution of video content have become accessible and efficient. However, video editing task remains very time consuming. Video being a time-based and dual-tracked (audio and video) medium meant each video clips must be inspected and editing decisions has to be performed on individual frames. Introducing automation into video editing has been attempted since the beginning of digital video with little or no success so far. We, hereby, present an breakdown of tasks involved in video editing and argue for an approach to introduce automation these smaller tasks instead of entire editing workflow. By working on tasks, the impact of introduction of automation can be measured and user experience evaluated. In addition, we laid out the challenges in our approach to introducing automation into video editing and presented some AI techniques that can be applied to video editing workflows and AI concerns relevant to the topics.

Additional Key Words and Phrases: video, video editing, automation, assisted workflow

1 INTRODUCTION

Video is the most popular form of content on the Internet measured in internet traffic. According to the Cisco visual networking index [6], 75% of the Internet traffic in 2017 has been video content. With mobile phones, video sharing platforms and social media it is easier than ever to capture and publish videos. However, editing video is very time consuming. Video is a tedious medium to work with as it requires inspecting and manipulating videos at individual frames and it is a dual-track medium with both audio and video. There are various attempts to automate video editing and creating easier video editing workflows with semi-automation. In this position paper, we will focus on the latter and laid out the overview and challenges in introducing automation into video editing workflows.

In the whole video production workflow, video editing is a part of the post-production process [13] which took place after the media assets has been created via filming or acquisition from other sources. Video editing is defined as the process of assembling shots and scenes into a final product, making decisions about their length and ordering [15]. Nonlinear editing is defined as editing that does not require that the sequence be worked on sequentially [15]. All modern digital video editing software are non-linear editors in which the original video, audio or images are not modified but a new edit is specified based on the cuts and modifications of the existing media assets. The video editing software such as Adobe Premier or Final Cut Pro has decades of development behind them. Being mature software, the user interfaces and interactions of video editing tools are very similar or common across many different video editing programs. However, these established interfaces and interactions are created and evolved to edit videos without automation. The editing made by non-linear editors consists of an ordered list of media assets used in the edit and time codes of used media assets and it is usually stored in a file format called edit decision list (EDL).

Entirely automated video editing has received a lot of research interest. At present, automated video production is aimed at creating *video summaries* or *mashups*. Video summaries are highlights from video clips which fulfill some

Workshop proceedings *Automation Experience at the Workplace*
In conjunction with CHI'21, May 7th, 2021, Yokohama, Japan

Website: http://everyday-automation.tech-experience.at

selection criteria such as importance, being aesthetically pleasing, or having some level of interests [5, 18]. Mashups are videos that are produced by concatenating video segments from video clips which are usually recorded at the same event by different cameras [14, 20]. These completely automated video editing methods, as used to create video summaries and mashups, are intended to create a simple highlight compilation of the videos and there is no user interaction involved. It is clear that though fully automated video editing is useful for some cases, the usage of them is very limited in the workplace as they are simply highlights generators that are not configurable.

Intelligent video editing tools have been attempted since the beginning of digital video editing with the goal of making video editing easier. These tools make video editing easier by offering semi-automation or allowing manipulation of videos at a higher level of abstraction than just manipulating frames (e.g. spoken words, shots and dialogue). One early example of such a tool is *Silver* [4] from 2002, which provides smart selections of video clips, as well as abstract views of video editing, by using metadata on from the videos. A more recent example of an intelligent video editing tool is *Roughcut* [11]. *Roughcut* allows the computational editing for dialog driven scenes using user input of dialog for the scene, raw recordings, and editing idioms. There is an open source tool*autoEdit* [16] and a research [2], which enable text-based editing of video interviews by linking text transcripts to the videos.

2 ASSISTED WORKFLOWS IN VIDEO EDITING

Typical tasks involved in video editing are described in Figure 1. This set of tasks has not been verified with industry practices but rather created from consultation with just a single video editing product manager. The main purpose of the task is to introduce the tasks involved in video editing. The presence, emphasis and order of each of the tasks will have different permutations depending on video editing contexts. The tasks required for video editing can depends on both the type of the video being edited as well as the organization context where the video is being produced.

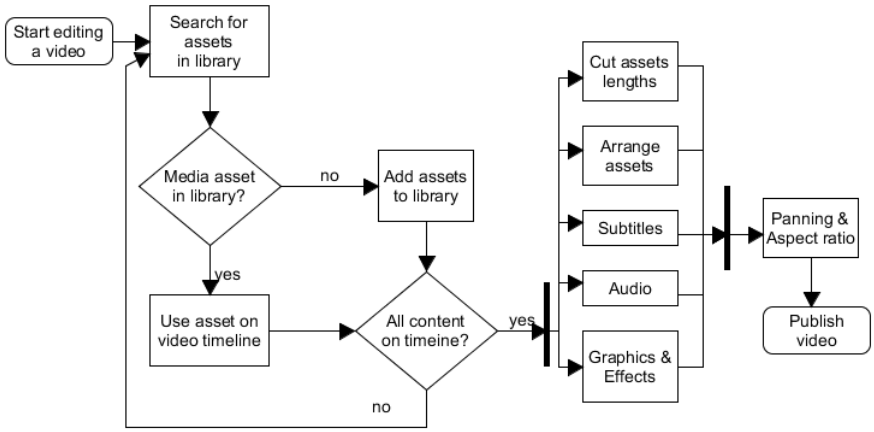

Fig. 1. Tasks involved in video editing

Semi-automated video editing tools that address the whole video editing process usually targets only a particular type of video editing need such as an interview view [2], an instructional video or a dialogue scene[11]. This is because discussing the entire video editing workflow is very context dependent and complex. In addition, the video workflows

might be personalized as well. Therefore, it might be more suitable to introduce automation for each individual task or a few combinations of tasks at first. Then we can provide a smaller automation blocks that the editors can personalized to fit their needs.

Another argument for introducing automation at task level is that it is each individual task can be measured and the changes user experience introduced by automation in each task can be studied. On top of that 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 task to create a dataset for the correct way of editing a whole video. We have studied the semi-automation in assisted subtitling in our work accepted for publication in Interactive Media Experiences 2021 as discussed in next paragraph [17].

How addition of semi-automation or having to work with automation in a subtitling changes performance, behavior, and experiences of novice users in subtitling is explored in the authors' paper [17]. In the paper the assisted subtitling prototype with speech to text allows the novice users to create slightly more accurate subtitles and much more efficiently. However, the users rate the experience with assisted subtitling more difficult than starting from scratch. The users experiences with introduction of automation into the subtitling workflow is summarized in the paper and it laid out usability problems that needs to be addressed for efficient human-machine collaboration in subtitling. In addition, the possible collaboration with state-of-the-art machine learning based speech to text systems and users in subtitling is laid out.

3 CHALLENGES IN INTRODUCING AUTOMATION INTO VIDEO WORKFLOWS

Efficiency or introduction of using automation or semi-automation in the workplace depends on crating efficient human-machine communication and well designed user interfaces that enable the communication. Introducing automation into a product itself is a challenge and some of the research highlights the challenges and general guidelines involved in building AI infused products [1]. The eighteen guidelines provided covers both building user experience, clarifying user expectations, matching social norms and learning from users' behaviour [1]. However, applying these guidelines for any specific scenario would require reevaluating them in that specific context.

AI techniques have been developed to manipulate or create video. Earliest work Video Rewrite [3] uses existing footage of a person to automatically create a video of said person speaking to a different audio track. This work was done with the intention of facilitating movie dubbing. AI synthesizing videos from existing footage became popular once again after deep neural networks were trained to synthesize fake videos - a process known as creating *deep fakes*. According to [12], a deep fake is a content generated by AI that is authentic in the eyes of a human observer. Deep fakes inspire negative concerns, however, they do also have potential applications in generating or adapting video content.

How AI techniques can be used to extract information from videos is a very diverse area of research. We are particularly interested in facial recognition, object detection, object tracking, scene detection, sentiment analysis, video reasoning and video captioning. Facial 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 across the frames in a video. Scene detection, or video segmentation, is identifying segments which are semantically or visually related in a video. Sentiment analysis is the problem of matching the sentiment that would be conveyed by a given content: is it happy, sad, ironic etc. Video captioning [20] is an AI technique that generates natural descriptions that capture the dynamics of the video.

Video editing workflows are complicated processes and depends on the context of the video production. In this position paper we try to laid out some of the challenges with video editing which is the part of post-production process of video workflows. Based our assisted workflow in subtitling evaluation, there are changes in performance characteristics as well as in user experience when automation automation in comparison with existing workflows. Similarly, the challenges in automation in video workflows can be summarized into

- Understanding existing workflows
- Deciding on which part of the workflows to automate and to which extent
- Working with the automation/AI technology
- Understanding the impact of automation by evaluation the tool

Upon solving these challenges, we can use the knowledge we have gained to create better tools with automation for the users. In addition, automation methods that can learn or adjust to user feedbacks can be explored.

Simply plugging ML into an video editing tool does not help user understand and utilize what ML could and could not do. Study of human factors and UI designs for ML is necessary to explore how users understand ML and affordances provided by introducing ML in the process. Dove et al. [8] stated in their study that ML is both underexplored opportunity for HCI researchers and has unknown potential as design material. The authors also pointed out the challenges which are the lack of understanding of ML in UX community, the data dependent nature of ML blackboxes and the difficulty of making interactive prototypes with ML.

There are emerging fields of study in AI that aims to put human at the center of control. An example from creative work is writing with machines in the loop [7]. Clark et al. [7] performed an experiment with two machine-in-the-loop systems for story writing and slogan writing tasks and the participants enjoyed collaborating with machine even though third-party evaluations rated of stories written with machine-generated suggestions are not as good as stories written by humans alone. Visual story telling models generates descriptions of a series of pictures that describes an event. Hus et.al [10] analyzed how humans edits those machine-generated text. Explainable Artificial Intelligence (XAI) is an emerging field in machine learning to come up with techniques that are more explainable to human users[9].

4 CONCLUSION

When introducing automation into video editing work, there are three important factors to consider: understanding existing video editing tasks and workflows, deciding where and how to introduce automation and finally user experience of automated workflows. We have provided a sample of a video editing workflow; however it has to be elaborated and verified in a study with professionals in the industry. Deciding where and how to introduce automation could benefit from studying what are the needs and expectations with automation in the industry. The user experience of automated workflows depends a lot on the interactions and interfaces between human and automation. The results from our evaluation of automated subtitling tool workflow suggests that just adding automation onto existing interfaces is not sufficient. Since introduction of automation changes the way the tool is used, new interactions have to be crafted informed by users' experience and needs.

We argue for tasks-based automation in video editing workflows because in the context of introducing automation in video editing which is a creative and subjective task, it is better to automate away the tedious and time consuming tasks and leave the creative or the task of telling the story to human editors. However, there is a lot of potential in other types of automation in video editing such as writing a video with text [19] or computational highlight generation from video archives. Since, automation or AI technology in video is rapidly evolving, there should be more attempts to

introduce automation into not only video editing but the entire video production workflow. The success of introduction of automation in our opinion depends on crafting a good user experience and adaptable automation.

REFERENCES

[1] Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, and Paul N. Bennett. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Floraine Berthouzoz. [n.d.]. Tools for Placing Cuts and Transitions in Interview Video. ([n. d.]), 8.

[3] Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97*. ACM Press, Not Known, 353–360. https://doi.org/10.1145/258734.258880

[4] Juan Casares, A Chris Long, Brad A Myers, Rishi Bhatnagar, Scott M Stevens, Laura Dabbish, Dan Yocum, and Albert Corbett. [n.d.]. Simplifying Video Editing Using Metadata. ([n. d.]), 10.

[5] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. 2005. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 2 (Feb. 2005), 296–305. https://doi.org/10.1109/TCSVT.2004.841694

[6] V Cisco. 2018. Cisco visual networking index: Forecast and trends, 2017–2022. *White Paper* 1 (2018), 1.

[7] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. 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*. ACM Press, Tokyo, Japan, 329–340. https://doi.org/10.1145/3172944.3172983

[8] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. ACM Press, 278–288. https://doi.org/10.1145/3025453.3025739

[9] David Gunning. 2017. Explainable Artificial Intelligence (XAI). (Nov. 2017), 38.

[10] Ting-Yao Hsu, Yen-Chia Hsu, and Ting-Hao 'Kenneth' Huang. 2019. On How Users Edit Computer-Generated Visual Stories. *arXiv:1902.08327 [cs]* (Feb. 2019). http://arxiv.org/abs/1902.08327 arXiv: 1902.08327.

[11] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. 2017. Computational video editing for dialogue-driven scenes. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–14. https://doi.org/10.1145/3072959.3073653

[12] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *Comput. Surveys* 54, 1 (Jan. 2021), 1–41. https://doi.org/10.1145/3425780 arXiv: 2004.11138.

[13] Frank Nack. 2005. Capture and transfer of metadata during video production. In *Proceedings of the ACM workshop on Multimedia for human communication from capture to convey - MHC '05*. ACM Press, Hilton, Singapore, 17. https://doi.org/10.1145/1099376.1099382

[14] Duong Trung Dung Nguyen, Axel Carlier, Wei Tsang Ooi, and Vincent Charvillat. 2014. 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*. ACM, Orlando Florida USA, 765–766. https://doi.org/10.1145/2647868.2654884

[15] Jeffrey A. Okun, Susan Zwerman, Kevin Rafferty, and Scott Squires (Eds.). 2015. *The VES handbook of visual effects: industry standard VFX practices and procedures*. Focal Press, Taylor & Francis Group, New York.

[16] Pietro Passarelli. 2019. autoEdit Fast Text Based Video Editing. http://www.autoedit.io/

[17] Than Htut Soe, Frode Guribye, and Marija Slavkovik. 2021. Evaluating AI Assisted Subtitling. *Accepted for publication at ACM International Conference on Interactive Media Experiences (IMX 2021) 2021* (2021).

[18] C.M. Taskir, Z. Pizlo, A. Amir, D. Ponceleon, and E.J. Delp. 2006. Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia* 8, 4 (Aug. 2006), 775–791. https://doi.org/10.1109/TMM.2006.876282

[19] Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, and Ariel Shamir. 2019. Write-a-video: computational video montage from themed text. *ACM Transactions on Graphics* 38, 6 (Nov. 2019), 1–13. https://doi.org/10.1145/3355089.3356520

[20] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. 2016. Deep Learning for Video Classification and Captioning. *arXiv preprint arXiv:1609.06782* (2016).