Chop Chop: A TikTok Cooking Montage Editing Assistant

Andrea Dahl*
Stanford University
California, USA
ahdahl@stanford.edu

Christina Ding*
Stanford University
California, USA
christinading@stanford.edu

Alex Hurtado*
Stanford University
California, USA
hurtado@stanford.edu

Jacob LeBlanc*
Stanford University
California, USA
jleb304@stanford.edu

ABSTRACT

TikTok content creators who are interested in making sound-satisfying cooking montages have to spend long hours editing down long form videos into a short 1 min TikTok. We built a system that identifies key clips from long form videos to facilitate creators in the editing process. The TikTok creator can input their video into our system which automatically identifies action clips (clips with a spike in audio level) and transition clips (clips with little visual change) and saves them locally to the user's computer. In a user study, we found that the average amount of time spent editing a TikTok went down 36% when using our clips with raw footage rather than raw footage alone. We conclude that this work could be applied to other types of sound-satisfying video montages outside of TikTok and outside of the cooking context.

CCS CONCEPTS

• **Human-centered computing → Graphical user interfaces**.

KEYWORDS

content creation, video editing, TikTok

ACM Reference Format:
Andrea Dahl, Christina Ding, Alex Hurtado, and Jacob LeBlanc. 2018. Chop Chop: A TikTok Cooking Montage Editing Assistant. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the rising popularity of TikTok, cooking content creators have begun to focus on producing short-form montages featuring quick cuts between steps of food preparation[2]. While impressive, these sensorily satisfying videos are time-consuming to create as the creator must scrub through footage of their entire preparation to find and synchronize successive video cuts. This project aims to facilitate this process by automatically identifying and editing raw long-form footage down to a pleasant quick cut montage using audio cues. The biggest challenge of creating these works of video montage is the amount of time and attention required to edit them down from long-form footage. Currently, these edit creators must record the entirety of their meal preparation, load that footage onto a video editing software, and then painstakingly navigate the footage, clip out the moments of interest, and align the clips to a sonically pleasing rhythm. Our design goal is to reduce the time expense and cognitive burden taken on by creators to make these video montages by automating the process.

2 RELATED WORKS

This project builds on the contribution led by "UnderScore: Musical Underlays for Audio Stories" as well as other papers[5][6][4]. Namely, we aim to empower amateur producers that may otherwise lack the adequate time or technical resources to transform their raw footage into short cooking montages. Our application of audio cues based on background noise differentiates itself from other audio based applications since it does not rely at all on human speech or a transcript. Another study [3] proposes a new approach for accelerating video editing processes by using a contrastive learning framework to train a 3D ResNet model to predict good regions to cut from unedited videos. For our project, we extend this work by incorporating audio detection to further help segment video sections. Using these audio cues, we pull out interesting video segments based on the sounds occurring for users to use in the editing process. As such, our tool is not fully automatic since we give some degree of autonomy to the user.

3 SYSTEM OVERVIEW

With the goal to reduce the time required to edit quick cut cooking montages, we created an automated system that identifies *action clips* and *transition clips* within raw footage as suggestions for the creator to use during the editing process. At a high-level, action clips capture moments of audibly-significant action (e.g. chopping a carrot, cracking an egg, stirring a pan). Transition clips capture slices of visually-consistent footage (e.g. pouring oil or plating a meal). We found that this classification of clips offered creators a balanced toolkit to create their quick cut montages. Our technical system is available to view on Github[1].

*All authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

3.1 Action Clip Identification and Extraction

In order to identify action clips, we must process the footage's audio into a more useful form for our automated analysis. We first sample the audio signal into 10 millisecond chunks (converted at 44100 fps) into an array of audio energy. Then, we take the root-mean-square energy (RMSE) of that chunk's array to yield its average loudness. Thus, we can use spikes of RMSE in the audio as indicators of an audibly-significant action. Quantitatively, we can identify a slice of audio as audibly-significant if its RMSE is above a selected threshold – we found that the 97th percentile of the audio's RMSE worked well as an indicator threshold. This slice represents the peak of some RMSE spike in the audio. Using an audibly-significant slice of audio, we can identify the clip of the entire action by expanding the slice outwards to encapsulate adjacent audio signals above a selected baseline level of RMSE: the audio's median RMSE. This expanded slice captures the entirety of that particular RMSE spike. We then extract this slice of footage as an action clip and offer it to the user.

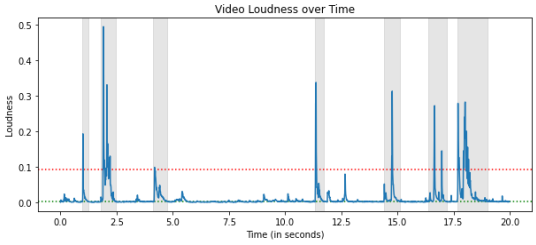

Figure 1: The blue line graph in this plot represents our audio's RMS energy over time. The dotted red line represents the threshold required to identify a clip as audibly-significant, and the dotted green line near the baseline represents the audio's median RMS energy. Action clips are identified via spans of gray background.

3.2 Transition Clip Identification and Extraction

Similarly to action clips, we must process the footage's video into a more useful form for our automated analysis. We first sample the video frames at 2 fps to yield our reference frames. For more efficient computation, we downsample each frame by a factor of 4 and convert them to grayscale. We then calculate the pixel difference between neighboring reference frames before taking the L0 norm of the frame-by-frame difference. This process yields the reference frames' normalized video difference score. Similarly, we can use inverted spikes of video difference as indicators of a visually-consistent action. We can identify the slice of video as visually-consistent if its frame-by-frame differences are below a selected threshold – we found that the 10th percentile of the frame-by-frame differences worked well as an indicator threshold. This slice represents the nadir of some video difference spike. Using a visually-consistent slice of video, we can identify the clip of the entire action by similarly expanding the slice outwards to encapsulate adjacent frame-by-frame differences below a selected baseline level of video difference: the video's median frame-by-frame difference. This expanded slice captures the entirety of that particular video difference spike. We can finally extract this slice of footage as a transition clip and offer it to the user.

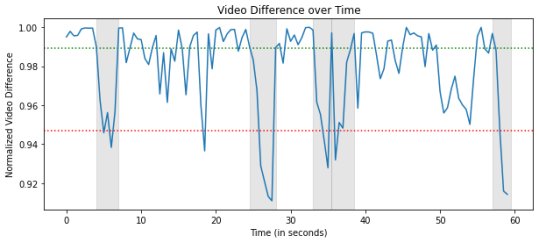

Figure 2: The blue line graph in this plot represents our video's frame-by-frame differences over time. The dotted red line represents the bottom threshold required to identify a clip as visually-consistent, and the dotted green line at the top represents the footage's median video difference. Transition clips are identified via spans of gray background.

4 EXPERIMENT OVERVIEW

For our experiment, we recorded two sample cooking videos for study participants to edit in iMovie. The raw footage of both videos was around 5 minutes in length, and participants were asked to edit the videos down to a 30-second TikTok-style montage. We had each participant first edit the raw footage of one video, then try editing the other video with the extracted clips. At the end of the study, they would create two TikToks - one using only the raw footage, and one using the clips. All of the participants were also given the raw footage for the extracted clip videos in case they wanted to reference the original footage. We timed each process, then asked them to share qualitative feedback and fill out a short feedback form afterwards. All studies were conducted through Zoom, and we had participants share their screens during the process so we could follow along easily. We split our study into two phases. In the first phase, our main goal was to test our idea with users and discover whether our tool provided value to the video editing process. We manually extracted action clips from both videos. Action clips were organized chronologically in their own folder. We recruited 6 participants for our first phase. 3 of them were given Video 1 action clips, and the other 3 were given Video 2 action clips. We noticed that users often referenced the raw footage to manually extract transition-style clips themselves to pad all the action clips, so we incorporated this feedback into the next phase. The second phase of our study was very similar to the first. However, this time we automatically extracted both action clips and transition clips so that users could create more balanced TikToks. Action clips and transition clips were ordered chronologically in their own separate folders. We recruited 7 participants, with 3 of them editing the raw footage of Video 1 and then the extracted clips of Video 2, and vice versa for the other 4 participants. Again, participants were tasked with creating two TikTok montages by the end of the study. We timed the processes and gathered feedback through post-study discussions and a form.

5 RESULTS

5.1 Quantitative

From our second experiment, we did find that our ChopChop tool's clips do decrease the time required to create TikTok cooking montage edits. Specifically, we found, as seen in Table 1, that the average amount of time spent editing a TikTok went down 36% when using clips instead of just raw footage. This does include an outlier in our dataset with Participant B who had an increase in time instead of a decrease. If we remove Participant B from the dataset, the average decrease in time instead goes to 50.9%, showing that ChopChop does successfully complete the goal of saving time.

	Raw Footage	Footage + Clips	Percent Change
Participant A	14:42	6:16	-57.37%
Participant B	8:28	12:58	+53.15%
Participant C	12:49	5:05	-60.34%
Participant D	20:57	10:21	-50.6%
Participant E	9:01	4:22	-51.57%
Participant F	26:34	15:34	-41.41%
Participant G	15:00	8:22	-44.22%
			-36.0%

Table 1: The table of participant time results from our second experiment. It shows the raw footage edit time, the clip edit time, and the percent change between the two, where a green/negative percent change shows that editing with the clips was faster than editing just the raw footage.

We can also break down these statistics to analyze the individual videos. Participants A, C, E, and G edited Video 1 from just the raw footage with an average time of 14:22.25. We can then compare this to the time it took for Participants B, D, and F to edit the clips of Video 1, which resulted in an average time of 12:57.67, which is a decrease in about 10%. However it is important to note that this does not take into account the skill level of the participants. We can analyze Video 2 in the same manner. Participants B, D, and F completed the editing process using only Video 2's raw footage with an average time of 18:39.67. We can then compare this to Participants A, C, E, and G who used Video 2's clips and had an average edit time of 6:01.25, which shows a 67% decrease in the time required to create a TikTok edit, also not taking into account the skill of the participants.

Time was not the only metric we were interested in with ChopChop. One of the other metrics we were interested in was self-reported ease of use, as demonstrated by the bar chart in Figure 3. This chart shows that our participants thought our tool was very easy to use, with an average 6.28 out of 7 on a 1-7 Likert scale. For us, this showed that the concept of ChopChop was not only a time saver, but would also be easy for users to actually get into using without a high barrier to entry. Another self-reported metric we were interested in was how enjoyable the experience was because this implied to us how likely people would be to actually use the tool despite easiness and time savings. This data is shown in Figure 4, and this also had an average 6.28 out of 7 on a Likert scale.

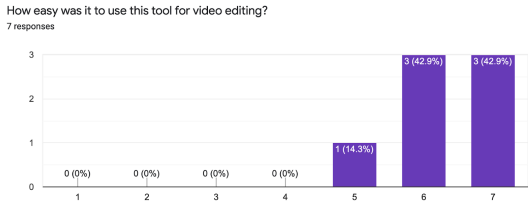

Figure 3: Our Likert scale for ease of use from our post-study survey. It shows 1 score of 5, 3 scores of 6, and 3 scores of 7 for an average score of 6.28.

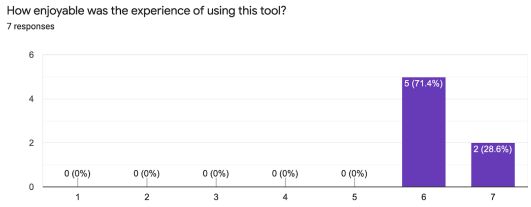

Figure 4: Our Likert scale for enjoyability from our post-study survey. It shows 5 scores of 6 and 2 scores of 7 for an average score of 6.28.

Outside of metrics that relate to people's likelihood to use the tool, we were also interested in the perceived effectiveness of the tool in pulling clips that editors would want to use in their cooking montages.

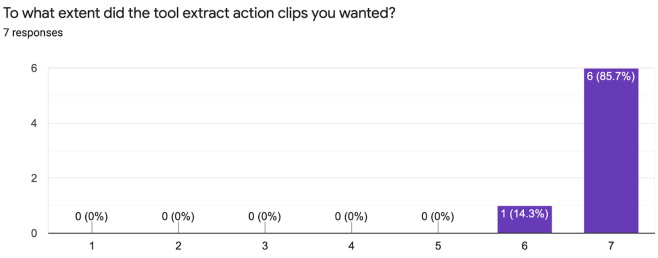

Figure 5: Our Likert scale for quality of action clips from our post-study survey. It shows 1 score of 6 and 6 scores of 7 for an average score of 6.85.

As such, we have two more Likert scales where we asked users the extent to which ChopChop pulled action and transition clips that they would have liked to use. First, the action clip results (as shown by the bar chart in Figure 5) show that users were very satisfied with the action clips pulled from our videos with an average score of 6.85 out of 7. We believe that this is because of the nature of the TikTok montages themselves that rely heavily upon clean crisp sounds that the ChopChop action clips represent. In Figure 6,

we can see how satisfied users were with the transition clips used. In this case, users were less satisfied, with an average Likert score of only 6.14 out of 7, but they were still very pleased with the clips included and most users only wished that one or two extra clips had been included.

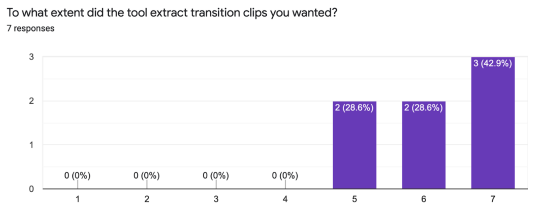

Figure 6: Our Likert scale for transition clip quality from our post-study survey. It shows 2 scores of 5, 2 scores of 6, and 3 scores of 6 for an average score of 6.14.

5.2 Qualitative

After the experiment, we also asked each participant about their experiences, and there were a couple of common responses. First, many of the participants wished that the transition and action clips had been included into one chronologically ordered folder rather than separate folders. This way, when imported into iMovie, all of the clips would be in the correct order automatically without users having to switch between action and transition clips to figure out where they fit into the entire video. As a result, we believe that adding this feature would allow an even greater time decrease using clips compared to only editing raw videos. Secondly, many participants commented on the change in creative process that resulted from using the clips as opposed to editing the entire video. Multiple participants noted that the clips seemed to offer less flexibility in the editing process since it was always just easier to use the pre-pulled clips rather than finding their own. Participant G particularly noted that this tool feels very corporate to him and takes away from the creative process, so those who create content professionally may be more inclined to use the tool than those who simply want a creative outlet, which may have implications for the other uses of this tool.

5.3 Limitations

Although our results are very promising, there are several holes in our work. Our results are based on a very small sample size, so there are many demographics who we think that our tool would be useful for that we were not able to recruit to participate in the study. Particularly, we were not able to test our tool with any professionals, whether they are profession TikTok creators or simply professional editors. We would have loved to be able to see if this editing style would be helpful for professionals who are very used to a specific style of editing and what the learning curve would be like for this kind of power user, especially since several of our participants did note that they saw the tool as something more useful for professionals. With more users and a wider range of users, we feel that we could make more general claims about the usefulness of this work.

6 CONCLUSION AND FUTURE WORK

Despite our relatively small sample size, it seems promising that our system lowered the editing time required for most participants. Our goal was to create a tool to make editing these TikToks faster and our system was able to achieve that without sacrificing on the quality of the video. We also heard from our participants that the process of editing felt easy and enjoyable when they had access to our provided clips. According to our participants, it was difficult to use the action clips and transition clips because they were organized into separate folders. They reported that it would have been easier and faster to use the tools if all the clips were grouped together and chronologically organized so that concatenating them chronologically would be easy. In the future, we would improve our system by organizing the resulting clips all in one place. Sound-satisfying montages are not unique to the cooking genre. We would be interested to see how our model would need to be tweaked, if at all, to be used to create other types of montages such as construction, crafting, cleaning, and more. The possibilities are vast for potential genres our tool could be used for and commercial TikTok editing tools are very popular among creators. We believe that there would be much user demand for our ChopChop system in its current and future states.

Outside of TikTok, we also feel that it would be interesting to dive deeper into the combination of visual and audio analysis to create single clips rather than separate clips. For example, we believe that this work could be used to find clips where there is little to no visual change but a large spike in volume, which may be useful in some cases for editors. By combining these techniques, users would have much more control over what type of clips they would like to pull to create videos as specific to their goal as possible. In this way, the technology could be applied to spaces outside of TikTok. On that note, in additional, more general iterations of the tool, we would like to add a way for users to specify the types of clips they would like in terms of volume and visual changes in order to pull clips that are more applicable to a wider array of videos rather than the specific cooking montage TikTok videos that we based this work on.

REFERENCES

[1] Alexanderjhurtado. [n.d.]. https://github.com/alexanderjhurtado/cs347_chopchop
[2] Mangiamoh on TikTok. [n.d.]. https://vm.tiktok.com/ZTd9PvSpQ/
[3] Alejandro Pardo, Fabian Heilbron, Juan Alcázar, Ali Thabet, and Bernard Ghanem. 2021. Learning to Cut by Watching Movies. (08 2021).
[4] Steve Rubin, Floraine Berthouzoz, Gautham Mysore, Wilmot Li, and Maneesh Agrawala. 2012. UnderScore: Musical Underlays for Audio Stories. (2012), 359–366. https://doi.org/10.1145/2380116.2380163
[5] Than Htut Soe. 2021. AI video editing tools. What editors want and how far is AI from delivering? *CoRR* abs/2109.07809 (2021). arXiv:2109.07809 https://arxiv.org/abs/2109.07809
[6] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. QuickCut: An Interactive Tool for Editing Narrated Video. (2016), 497–507. https://doi.org/10.1145/2984511.2984569