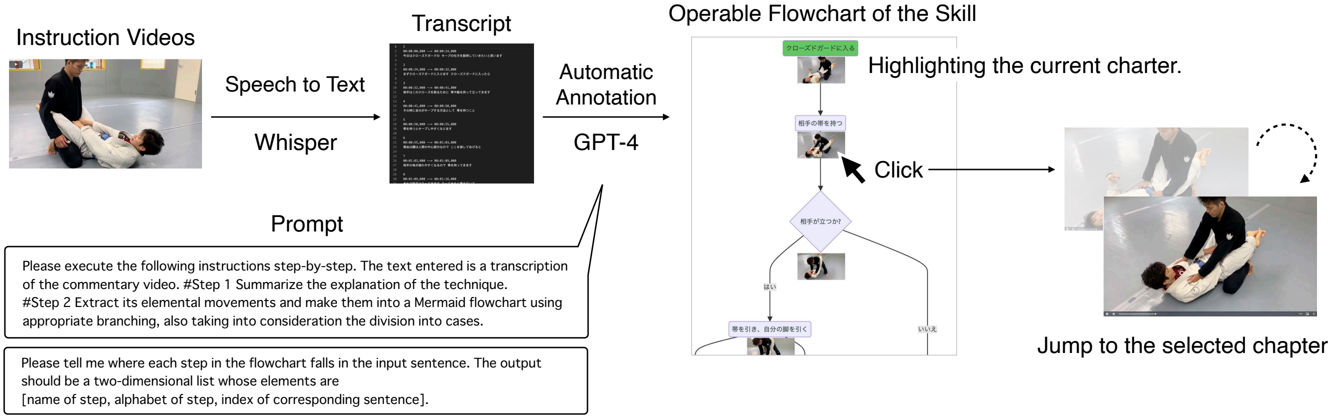
SkillsInterpreter: A Case Study of Automatic Annotation of Flowcharts to Support Browsing Instructional Videos in Modern Martial Arts using Large Language Models

Kotaro Oomori
oomorik@acm.org
The University of Tokyo
Bunkyo-ku, Tokyo, Japan

Yoshio Ishiguro
The University of Tokyo
Bunkyo-ku, Tokyo, Japan
ishiy@acm.org

Jun Rekimoto
The University of Tokyo
Bunkyo-ku, Tokyo, Japan
Sony CSL Kyoto
Bunkyo-ku, Kyoto, Japan
rekimoto@acm.org



Please execute the following instructions step-by-step. The text entered is a transcription of the commentary video. #Step 1 Summarize the explanation of the technique. #Step 2 Extract its elemental movements and make them into a Mermaid flowchart using appropriate branching, also taking into consideration the division into cases.

Please tell me where each step in the flowchart falls in the input sentence. The output should be a two-dimensional list whose elements are
[name of step, alphabet of step, index of corresponding sentence].

Figure 1: SkillsInterpreter automatically generates a flowchart of the skills recorded in a video by large language models from the speech contained in the video, and uses it as a user-operable tool for video browsing. The generated flowchart explores desired scenes, checks the current chapter, and reviews the skill structure while watching the video.

ABSTRACT

The use of video for learning physical skills such as modern martial arts is becoming popular. Physical skills such as modern martial arts require decisions depending on the situation. An example of these decisions is selecting an appropriate off-balance technique based on the position of the opponent's feet. However, the existing interface does not support video browsing based on the structure of the physical skills, including situations and the decisions that should be made at that time. We hypothesize browsing based on the structure can help the user's skill comprehension. In this paper, we propose a structure-based video browsing method, SkillsInterpreter, which automatically generates a flowchart of the speech-contained skill instruction video by large language models (LLMs). The generated flowchart explores desired scenes, checks the current chapter, and reviews the skill structure while watching the video. Our study included interviews with experts and evaluations with learners in modern martial arts. Based on our two studies, it was suggested that SkillsInterpreter can support video-based skill learning in modern martial arts, especially in Brazilian Jiu-Jitsu, which needs situation-specific decision making.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**.

KEYWORDS

Video Browsing Interface, Large Language Models, Flowchart, Martial Arts

ACM Reference Format:
Kotaro Oomori, Yoshio Ishiguro, and Jun Rekimoto. 2024. SkillsInterpreter: A Case Study of Automatic Annotation of Flowcharts to Support Browsing Instructional Videos in Modern Martial Arts using Large Language Models. In *The Augmented Humans International Conference (AHs 2024), April 04–06, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652920.3652942

AHs 2024, April 04–06, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0980-7/24/04
https://doi.org/10.1145/3652920.3652942

1 INTRODUCTION

Video-based learning is widely used in the acquisition of physical skills. For example, fitness is one of the hot topics in the social streaming service [16]. As another example, in modern martial arts, in addition to distributing instructional videos on social streaming services such as YouTube, multiple platforms sell instructional videos [5, 17]. Video-based learning has the advantage of being suitable for learning from a favorite instructor and for repetition.

In the research of interfaces for viewing videos, the browsing operation, in which the user moves the playback position to a desired scene, is a vital component. Using content summaries in video browsing tools is effective in facilitating video comprehension. VideoDigests presents a transcript-based interface that generates a summary of the sections in a video and allows the user to click through to the video for browsing [14]. Automatic summarization of data is a powerful tool in natural language processing. There have been attempts to extract and represent structure from text [8, 12]. Also, like VideoDigests, it shows the importance of content-based video browsing rather than just the usual time-based operation with a seek bar [3].

Videos, especially those that include speech instruction, such as instructional videos, have logical structures. This structure is essential for video comprehension. Especially in fields such as sports, where one needs to make instantaneous judgments according to the situation, it is necessary to understand the content in a structured manner. However, an attempt has yet to be made to support video browsing by extracting the structure in these videos and automatically annotating them to reflect this structure. Systematic understanding of how to deal with a situation requires organization. Still, the existing one-dimensional way of presenting chapters and captions doesn't reflect this on the user interface (UI). In this paper, we propose SkillsInterpreter, a video browsing support system that automatically extracts the logical structure of the speech contained in a video, creates a flowchart that summarizes the content, and automatically determines at which point each component of the diagram may correspond to a scene containing an utterance, allowing the user to look back on the scene with a single click. The flowchart generation and the linking of utterances to components are each interpreted using the Large Language Models (LLMs), which can solve general tasks in natural language processing [2, 11]. SkillsInterpreter must solve the dual tasks of properly extracting structure from the transcript and estimating the corresponding speech segment. For improving the estimation performance of LLMs, we leveraged several promt engineering techniques [18, 19]. Following the idea of Chain-of-Thought (CoT) prompting [18], LLMs summarize the video content, generate a flowchart, and estimate the correspondence between each step in the flowchart and the transcript.

As an example of a characteristic topic with decision-making depending on the situation, we selected modern martial arts as the topic of the video. An example is selecting an appropriate off-balance technique based on the position of the opponent's feet. Because of the interaction between players in modern martial arts and the complex interplay of bodies, it isn't easy to automatically tag what is being done from visual information. Therefore, SkillsInterpreter used transcript-based processing. It took advantage of the characteristic of instructional videos that verbal explanation while acting. Our study included interviews with experts in modern martial arts, particularly Brazilian jiu-jitsu (BJJ)[1], mixed martial arts (MMA)[2], and kickboxing[3]. Interviews were conducted with two professional instructors, providing insight into the potential of this system to support skill acquisition in martial arts.

In addition, four videos, two each explaining skills in BJJ and kickboxing, were tested on 12 participants, eight with BJJ experience and four with kickboxing experience. The experiment was compared to a browsing system with clickable subtitles. We hypothesize that the learner should be able to understand 100% of the video content as much as possible to use the technique effectively. The BJJ videos show the percentage of people who got a perfect score **87.5% for the proposed** and **37.5% for the baseline**. The results and user feedback suggested the effect of supporting instructional video comprehension of automatic annotation of flowcharts in martial arts, especially in BJJ.

Also, the framework of this study can be used for competitions in martial arts and other domains, and examples of applications other than skill explanation in martial arts are presented, including flowcharts generated for cooking and arguments with contrasting structures.

2 RELATED WORK

2.1 Neural Language Processing for Content Comprehension

Neural language processing has been studied to comprehend content, such as text and video, effectively. Pavel et al. analyzed video content on a transcript basis of written data to facilitate users' video comprehension [13, 14]. LLMs have attracted attention in natural language processing as a method that can solve a wide range of tasks with good performance [2, 11]. For improving the task-solving ability of the LLMs, prompt engineering techniques and methods of devising prompts that serve as input to LLMs are researched [18, 19]. The task-solving ability was focused on, which is increasingly used in interaction research mention catailyst, get assist, seascape, graphlogue. As the existing research focused on the automatic generation of flowcharts with LLMs, Graphologue is an interactive system that utilizes LLMs to facilitate information-seeking and question-answering tasks by representing the structure of long reply sentences from LLMs in a flowchart [8].

However, previous research has yet to attempt to assist users by applying automatic flowchart generation to video that represents the logical structure of video, such as case classification and its conditions (an example is in Fig. 2). In our study, we proposed an interface that enables video browsing by adapting flowchart generation from text data to video transcripts, reflecting the structure of the speech in the video.

[1] Brazilian Jiu-Jitsu (BJJ) is a grappling-based martial art whose central theme is the skill of controlling a resisting opponent in ways that force him to submit [1].

[2] MMA combines wrestling and striking martial arts into one complete discipline, including techniques from Thai-boxing, judo, Brazilian jiu jitsu and boxing [7].

[3] Kickboxing is a striking style that incorporates punches and kicks. The essence of kickboxing is that it is a stand-up fighting style. This is to say that it focuses on striking, and there is no ground fighting involved [20].

2.2 Interface for Support instructional Video browsing

Video in instruction has been researched as a helpful learning tool. In the field of interaction research, Several research were conducted to support instructional video browsing [13, 14, 22]. Several focus on acquiring physical movements [3, 6, 9]. PoseAsQuery proposed a direct manipulation method that uses body movements directly as queries for video retrieval [6]. RubySlippers proposed a hands-free method of content-based video browsing using voice commands as input [3]. FlowAR proposed a method of presenting images while moving using a head-mounted display to achieve hands-free performance as well [9].

Although our work is the first study to apply flowchart generation from transcripts by LLMs to video browsing and typical inputs using hands (clicks, drags) as the input modality, future applications using other modalities than clicks as input are promising. In particular, hands-free is an essential factor in the acquisition of body movements, and the achievement of this is critical future work in this study as discussed in listerature[3, 6, 9].

As an example of how we have attempted to generate graphs from video content to assist in video browsing, VideoGraph [21] used visual diagrams of scene images to represent the relevance of scenes from visual information. While they dealt with visual information, our study focused on the characteristics of the instructional video, in which the actions are explained and demonstrated with speech. Therefore, we aimed to support video browsing more effectively by generating a flowchart from the transcript to reflect the logical structure, such as the division of cases shown in Fig.2.

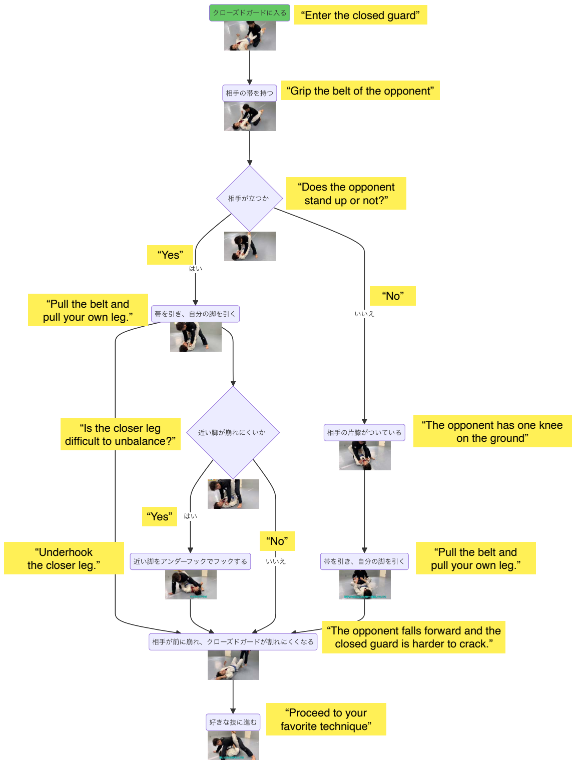

クローズドガードに入る
"Enter the closed guard"
相手の帯を持つ
"Grip the belt of the opponent"
相手が立つか
"Does the opponent stand up or not?"
"Yes"
はい
"No"
いいえ
"Pull the belt and pull your own leg."
帯を引き、自分の脚を引く
"Is the closer leg difficult to unbalance?"
近い脚が崩れにくいか
相手の片膝がついている
"The opponent has one knee on the ground"
"Yes"
はい
"No"
いいえ
"Underhook the closer leg."
近い脚をアンダーフックでフックする
帯を引き、自分の脚を引く
"Pull the belt and pull your own leg."
相手が前に崩れ、クローズドガードが割れにくくなる
"The opponent falls forward and the closed guard is harder to crack."
好きな技に進む
"Proceed to your favorite technique"

Figure 2: An example of a flowchart generated by the system. The video is https://youtu.be/ORwOPudfOG8?feature=shared (2024.1.9 accessed) As shown in the figure, the original flowchart was generated in Japanese, and the English part marked up in yellow is the translation by the author.

3 PROPOSED METHOD

3.1 Preprocess

Our flowchart auto-generation method works on a transcript basis, so the first step is to process the speech-to-text from the video. Whisper [15] is used for speech-to-text. Then, we get transcripts of the video in the format of a srt file.

GPT-4 [11] is the LLM for automatic flowchart generation from transcripts. GPT-4 outputs a flowchart described in Mermaid notation and a list of indices of the utterances in the SRT file corresponding to each flowchart component. In our prompts, we elicit the format of outputs, and the prompt is phrased respectfully. This is the technique of instruction in prompting [19]. Our prompts consist of two parts. In the first prompt, we also leveraged a prompt engineering technique of CoT for improving the estimation ability of LLM, which gives the LLM an explicit intermediate step [18]. Our prompts are zero-shot, so we leavaraged zero-shot CoT prompting [10]. The first prompt is *"Please execute the following instructions step-by-step. The text entered is a transcription of the commentary video. #Step 1 Summarize the explanation of the technique. #Step 2 Extract its elemental movements and make them into a Mermaid flowchart using appropriate branching, also taking into consideration the division into cases."* The entire transcript is summarized, and the summary and original utterances are given as input when the flowchart is generated. In the second prompt, GPT-4 estimates the correspondence of each flowchart step and the transcripts. The output format is the list of [name of step, alphabet of step, index of the corresponding sentence]. The second prompt is *"Please tell me where each step in the flowchart falls in the input sentence. The output should be a two-dimensional list whose elements are [name of step, alphabet of step, index of corresponding sentence]."* Note that this work's videos and prompts are in Japanese.

Finally, based on the GPT-4 estimation results, representative frame images in each step are automatically extracted. This makes it easier for the user to get an overview of the scene corresponding to each step. In the automatic frame extraction process, the frame located in the middle of the speech segment of the first transcript corresponding to each step of flowchart is extracted. For example, when speech segments 2, 3, and 4 correspond to chapter A, the frame in the mid-position of speech segment 2 is selected.

The preprocess results in a flowchart of the skill, the correspondence between each step and its utterance, and a video frame image that serves as a thumbnail of the step. An example of the finally generated flowchart is in Fig.2. Fig. 1 shows a summary of the prompts and data processing.

3.2 Interface

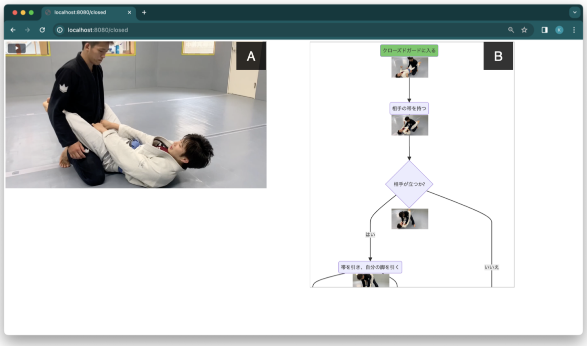


Figure 3: Interface Overview. (A) Web video player, which plays the input video. Users can control the video with the sequence toolbar. (B) Operable flowchart of the skill in the input video. The chapter will be played when the user clicks the image below each chapter of the flowchart. The user can freely move and zoom in on the flowchart by pan and zoom operations.

3.2.1 *Interface Design.*

The design of our interface is intended to allow the user to grasp the flowchart and the input video on a single screen. The design was based on this concept, discussed among the co-authors. The positioning of the video and flowchart portions was inspired by the positioning of the video and the transcribed list of the YouTube interface. The size of the generated flowchart varies depending on the input, but in order not to spoil the concept of being able to check the video and flowchart on a single screen" for any input, the position of the flowchart portion on the interface is fixed. The zoom and pan movements on the flowchart allow it to be freely expanded, reduced, and re-positioned.

3.2.2 *Implementation.*

Our proposed interface mainly consists of two parts. On the left side of the interface is a video player capable of playing input video (Fig. 3 (A)). A user typically uses a mouse or touch input to the seek bar for video browsing. With the part of an operatable flowchart of the skill (Fig. 3 (B)), by clicking on the images placed under the components, the user can navigate to the point in time when the utterance at the youngest index in the list of corresponding sentences was started. The currently playing part is highlighted in green, allowing the user to know in real-time which part is currently playing and where it is located within the whole. The size of the generated flowchart varies from video to video. To accommodate this, the user can freely zoom in and out and move around the flowchart by performing zoom and pan operations on the flowchart.

The interface was implemented using p5.js [4] and video.js [5], and the server using flask-socketio [6]. Socket communication records the browsing log, and the current playback position is sent to the server in real-time and output to a log file.

4 STUDY I: INTERVIEWS WITH PROFESSIONAL MODERN MARTIAL ARTS INSTRUCTORS

We conducted semi-structured interviews about the system with two professional instructors. One interviewee has two years of experience teaching Brazilian Jiu-Jitsu (BJJ), a modern martial art. The other has ten years of experience teaching BJJ, kickboxing, and Mixed Martial Arts. The interviewee was asked about his experience with the system and his experiences with it for four videos. The interview was one hour long. Coding was performed on the interview results to create a topic-based evaluation of the system. As a gratuity, a payment of 80.12 dollars was made to each interviewee.

4.1 Findings

Findings from the interviews are presented below, along with chapters. The original interview is in Japanese, and the following comments are translations.

The flowchart represents every single skill. Both of the instructors mentioned that the flowchart of this method represents every single skill: "The one I know is different from this one in that it is not a structure with a single technique, which means this(proposed) system is finer than the one I know"; "Of all the things that have chapters for one technique, it's never been this easy to watch at all."

The possibility to facilitate user browsing. Both of the instructors mentioned the possibility of facilitating user browsing: "When I'm sparring, for example, and I'm having trouble with something, instead of having to search for it all over again, I can just jump to it, so it's quick and easy, and I feel like I can solve my problems right away"; "With this system, it's easy to find it because it's already written in a sentence."

Low cost to make the annotation. Instructor2 mentioned that the cost to make the flowchart is low: "If you can make a chart with no cost, fine, you can go on and on."

The user can watch the flowchart and the movie on one screen. Instructor2 mentioned that the proposed interface feels good because the user can watch the flowchart and the video on one screen: "The good thing is that users can see it on one screen, and it's easy to understand what is being explained."

Future direction make users understand the skill without watching videos. Instructor1 mentions the possibility that the flowchart of the proposed method could serve as a teaching tool for the technique without directly viewing the video: "It may be difficult, but if the explanation I'm talking about is lightly written next to the video, I may be able to complete the chart alone without watching the video, well, it's better to watch it, but I think I can explain it with just the chart."

Influence of the number of chapters in the flowchart. Instructor2 mentioned the number of chapters in the flowchart: "In the video, the amount of charting is not too detailed or too general. I think the amount of charts is also quite important, too much detail can be a bit annoying"; "If it is too detailed in a long movie, it can be a pain in the ass to find it. I wonder if there are controls for that."

[4] https://p5js.org/
[5] https://videojs.com/
[6] https://flask-socketio.readthedocs.io/en/latest/

Table 1: The list of the experiences of instructors. BJJ skill level of each instructor is based on their own belt color [4]

Number	Martial Arts Experience	Teaching Experience	BJJ Skill Level
Instructor1	6 years of BJJ	2 years of BJJ	Advanced
Instructor2	10 years of Judo, 15 years of BJJ, and 20 years of kickboxing and MMA	10 years of BJJ, kickboxing, and MMA	Expert

Future direction of applying multiple videos. Instructor2 mentions the possibility of recommending the following video to watch by using the proposed method to multiple videos at once: "For each technique, for example, there is a back-escape situation, if we can see what we should look at next, it would be good."

Revise the word of the chart by hands. Both of the instructors mentioned revising the chart by hand: "I think it's OK to make these mistakes because I know what they are. If I think there is a mistake, I can correct it by hand"; "Oh, maybe the AI is wrong. In any case, I'm talking about the left arm...well, watch and correct these details."

Points to be improved in the interface. Instructor1 mentioned the point to enhance the interface: "If the chart section could be expanded, it might be easier to look at it again since you can see it all at once"; "It's a bit of both, but there are people who read text-based information, so it's hard to find the right balance"; "It might be hard to see on a smartphone, though, on a computer, it's nice."

5 STUDY II: SYTEM EVALUATION WITH MODERN MARTIAL ARTS PRACTITIONERS

We conducted a comparative experiment to evaluate the system's performance for browsing and comprehending instructional videos.

The total number of participants was 12, including 10 men and two women (mean 33.92 and SD 10.85.) Participants were recruited through an online talk group at a modern martial arts gym. All participants were Japanese speakers, and the experiment was conducted using Japanese-language videos. Participants were asked to have previous experience with BJJ or kickboxing as a requirement for participation. Each participant could select their favorite video category in which they had experience, either BJJ or kickboxing. The experience of each participant is shown in the table. As a gratuity, 7.03 dollars was made to each participant.

5.1 Procedure

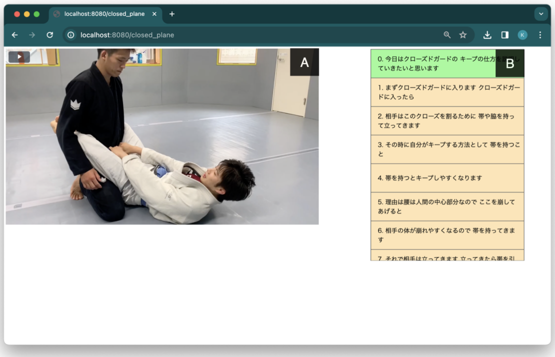
Each participant experienced the proposed method and the baseline in the experiment, respectively. The main task of the experiment was to have the participants view a martial arts instructional video in the proposed/baseline interface at their leisure until they felt they "understood" the technique (i.e., they could answer a quiz that asked about the content). When the participant signals to the experimenter that they "understand" the method, they are asked to stop watching the video and to answer a three-choice quiz with four questions, each asking about their understanding of the content. The three-choice quiz was quality-checked by a professional martial arts instructor. This was repeated twice per person, once as a suggestion and once as a baseline, and concluded by collecting Likert scales and comments on the system's experience. The experiment lasted approximately 30-45 minutes.

5.2 Baseline

As a baseline for the experiment, we employed a scrolling selection interface with clickable subtitles, such as those implemented in YouTube (Fig. 4). As in the proposed method, the currently playing part is highlighted in green, and by clicking on each part, the user can navigate to the position where the utterance began.

The main differences between the proposed method and the baseline are as follows:

- While the proposed method extracts the structure from the transcript and makes a flowchart, the baseline presents the transcript as is.
- While the proposed method moves to the part of the flowchart corresponding to a chapter by clicking on it, the baseline moves to the part of the clicked utterance.
- In the proposed method, the flowchart is manipulated by zoom and pan movements, whereas in the baseline, the subtitles are manipulated by moving up and down by scrolling.



0. 今日はクローズドガードの キープの仕方を説明し ていきたいと思います

1. まずクローズドガードに入ります クローズドガードに入ったら

2. 相手はこのクローズを割るために 帯や脇を持って立ってきます

3. その時に自分がキープする方法として 帯を持つこと

4. 帯を持つとキープしやすくなります

5. 理由は腰は人間の中心部分なので ここを崩してあげると

6. 相手の体が崩れやすくなるので 帯を持ってきます

7. それで相手は立ってきます 立ってきたら帯を引

Figure 4: Baseline interface. (A) Web video player, which plays the input video. Users can control the video with the sequence toolbar. (B) Operable subtitles of the skill in the input video. When the user clicks the image to blow each subtitle, the part of it will be played.

5.3 Results

Within the current experiment's scope, we could not find a statistical difference between the baseline subtitle-lined interface and the baseline subtitle-lined interface. However, the difference in the score between the proposed and baseline was marginally significant ($p < 0.1$) with Wilcoxon signed rank test.

Table 2: The list of the experiences of participants selected BJJ. The BJJ skill level is based on their belt color [4].

Number	Martial Arts Experience	BJJ Skill Level
P1	10 years of Karate, 6 years of Judo, 3 years of Shorinji Kempo, 1.5 years of BJJ, and 1 years of boxing	Beginner
P2	0.5 years of MMA, kickboxing and BJJ	Beginner
P3	1 years of MMA, boxing and BJJ	Beginner
P4	5 years of Judo and 2 years of BJJ	Intermediate
P5	7 years of kickboxing and 2 years of BJJ	Intermediate
P6	10 years of Judo and 2 years of BJJ and MMA	Intermediate
P7	5 years of boxing, 2.5 years of BJJ, and 1 years of grappling	Intermediate
P8	2.5 years of kickboing and 2 years of BJJ	Beginner

Table 3: The list of the experiences of participants selected kickboxing.

Number	Martial Arts Experience
P9	10 years of kickboxing and BJJ, and 4 years of Sambo
P10	9 years of kickboxing and BJJ
P11	3 years of kickboxing
P12	0.5 years of kickboxing and BJJ, and 2 years of wrestling

We hypothesize that the learner should be able to understand 100% of the video content as much as possible to use the technique effectively. Regarding the percentage of participants who achieved all correct answers in the quiz, the overall percentage was 75% for the proposed method and 50% for the baseline. In the BJJ videos, the percentage of people who got a perfect score was 87.5% for the proposed and 37.5% for the baseline.

The NASA-TLX subjective ratings showed significant trends in temporal demand and frustration measures ($p < 0.1$), but no significant differences were detected in all measures (Table 5) with Wilcoxon signed rank test.

5.4 User Feedback

The following is a summary of the feedback findings by comment, proposed method, and baseline condition.

Potential of the Proposed Method. Several participants mentioned the potential of the proposed method to assist in understanding martial arts instructional videos: P2 "It was easy to understand that the choices were divided by charts, so I could read what was going to happen in the video"; P3 "Because it was different from the usual instructional videos, it was hard to get used to it for a moment, but once I got used to it, it was comfortable. I felt it was easy to understand the contents of the videos systematically"; P4 "Jiu-jitsu has a lot of two-choice elements of what to do when certain events occur, so it was good to see that visualized"; P5 "I liked the charts because they were easier to understand."; P9 "The load is a little heavier than the subtitled ones, but I felt that the charts were better if the emphasis is on learning"; "It was good that the units were organized, and it was also good to have a bird's-eye view of the overall technology"; "The flow just seemed like a lot of work for the creator to make"; P12 "The flow was easier to understand, but I felt the mental demand of trying to understand the meaning of the technology was greater."

About the Baseline. Several participants mentioned the baseline for clicking on the subtitle: P6 "Subtitle is easier to understand"; P2 "Too many subtitles, too many small pieces of information. I felt that typos reduced my comprehension"; P12 "The latter was a physical burden to scroll."; P4 "While subtitle support would make the video easier to understand, we felt that having subtitles side by side might make it easier to search, but not to follow while watching"; P9 "It was refreshing to learn the techniques while looking at the captions. It was easy to look back on it later, and it would be better if it scrolled automatically."

6 DISCUSSIONS

The feedback from comments and the difference in the percentage of people who got a perfect score (87.5% and 37.5%) in the BJJ videos suggest that this study may have a positive effect. Especially since, as P4 comments, BJJ has a vital element of choice, which is addressed on a scene-by-scene basis, automatic generation of flowcharts may provide support. There was feedback that the flowchart supported the understanding of the technology, as in the case of P2 and P3.

On the other hand, the interface design of SkillsInterpreter is preliminary, it was also indicated that issues still need to be addressed as an interface. One instructor pointed out that the flowchart highlights the current chapter but that it sometimes suddenly moves to a distant location on the flowchart without any navigation. This is related to the issue that the more complex the flowchart becomes, the more difficult it becomes to display the entire flowchart in a large enough size alongside the video, and we plan to explore solutions to this issue in the future. One simple solution is to automatically move the flowchart so that the currently playing part is displayed. However, ensuring that the user's free viewing is not compromised is also necessary. Besides directly referencing the interface design, P12 rated the proposed method as easy to understand and mentioned that he felt its mental load was high. There is no statistical difference between the two methods. Still, mental demand is the

Table 4: Results of the score of the quiz. The score is one point per question, with a maximum score of 4 points. Values are rounded to two decimal places.

Video	mean(SD)		p-value	Achieved Perfect scores	
	proposed	baseline		proposed	baseline
BJJ	3.88(0.35)	3.13(0.83)	0.06	87.5%	37.5%
Kickboxing	3.5(0.58)	3.75(0.50)	0.32	50%	75%
Summary	3.75(0.45)	3.33(0.78)	0.12	75%	50%

Table 5: Results of qualitative analysis by NASA-TLX. Ratings are discrete integer values from 0 to 10, with a value closer to 0 indicating a smaller load and better performance. Values are rounded to two decimal places.

scale	mean(SD)		p-value
	proposed	baseline	
Mental Demand	4.25(2.26)	3.92(3.09)	0.91
Physical Demand	2.42(1.44)	3.17(2.82)	0.28
Temporal Demand	2.33(1.83)	3.42(2.94)	0.06
Performance	1.92(1.62)	2.75(2.67)	0.39
Effort	3.58(2.47)	4.83(3.89)	0.20
Frustration	1.92(1.62)	3.5(3.29)	0.06
Overall load	2.84(1.64)	3.75(2.93)	0.24

only item for which the proposed method has a higher value, so this point should be carefully considered.

One limitation of our work is the limited number of professional instructors. We recruited only 2 participants. For deeper insight, a larger number of participants is needed.

although this study used prompts based on the CoT concept, the prompt design needs further examination and improvement. The lack of strict prompting can negatively impact the learning process. How the prompt design affects the quality of the generated flowcharts and what the optimal prompt is is a topic for future work.

Other feedback from professional instructors provided several future directions. We plan to investigate these possibilities, such as recommending the following instructional video based on the system of techniques and adjusting the number of chapters. This study applied the SkillsInterpreter mechanism to every single video independently. In the future, we plan to expand it to be used for multiple videos simultaneously and visualize the structural connections between videos, which could be applied to video recommendation. We will conduct this in the future. In addition, as the instructor commented, the amount of chapters could affect user understanding. It would be an exciting research topic in the future to adjust the number of chapters and investigate how the number of chapters makes a difference in experience.

7 APPLICATIONS IN OTHER DOMAINS

Our proposed method applies not only to martial arts skills but also to videos in other domains. Here are some examples of how it can be used. In addition to skill explanations, using this feature for lecture videos makes it possible to browse the video according to the logical structure of the lecture, e.g., the structure of contrasts (Fig. 5). This video discusses the differences between jiu-jitsu and MMA and generates a flowchart from which a contrasting structure is extracted, such as what is with or without each rule. Even outside of skill acquisition, the SkillsInterpreter framework can be practical in areas where it is necessary to compare and understand two or more things, as in this contrast. On the other hand, in a video like this, where people are sitting and talking all the time, it may not be necessary because there is little change in the thumbnail image.

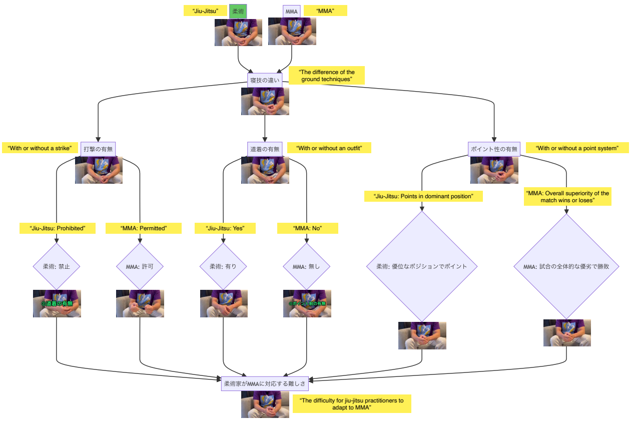

"Jiu-Jitsu"
柔術
MMA
"MMA"
寝技の違い
"The difference of the ground techniques"
"With or without a strike"
打撃の有無
道着の有無
"With or without an outfit"
ポイント性の有無
"With or without a point system"
"Jiu-Jitsu: Points in dominant position"
"MMA: Overall superiority of the match wins or loses"
"Jiu-Jitsu: Prohibited"
"MMA: Permitted"
"Jiu-Jitsu: Yes"
"MMA: No"
柔術: 禁止
MMA: 許可
柔術: 有り
MMA: 無し
柔術: 優位なポジションでポイント
MMA: 試合の全体的な優劣で勝敗
②道着の有無
③ポイント制の有無
柔術家がMMAに対応する難しさ
"The difficulty for jiu-jitsu practitioners to adapt to MMA"

Figure 5: An example of a flowchart generated by the system. A is made from https://youtu.be/99hK8-bbnhA?feature=shared (2024.1.16 accessed.) As shown in the figure, the original flowchart was generated in Japanese, and the English part marked up in yellow is the translation by the author.

Also, it can be applied to skills without case separation such as cooking a curry. As mentioned in the P4's comment of Study II, Brazilian jiu-jitsu has many selective elements in terms of which response to take in each situation, making it a domain that is well suited for visualization of structure through flowcharts. In areas where multiple options do not exist and where the choice factor is weak, the SkillsInterpreter framework may be limited to summarizing and chapter creation. It may be less effective than it could be. For example, in the case of cooking, another type of structuring other than flowcharts may be appropriate, such as dividing the process into different cooking utensils.

Based on the results of this interview and experiment, a possible case-by-case genre for which SkillsInterpreter has been suggested to be effective, such as BJJ, might be its troubleshooting application.

What other domains would our proposed method or any other structuring method be more effective in supporting learning? This is an interesting topic for future research.

8 CONCLUSIONS

In this study, we proposed SkillsInterpreter, a structure-based video browsing method that leverages LLMs. The system automatically generates a transcript-based flowchart of the techniques included in the instructional video from the video content and provides it to the user. As a domain that fits situation-specific decisions, modern martial arts was chosen as the theme of the video for this study, and a case study was conducted on the effects of SkillsInterpreter. The system was evaluated through interviews with two professional instructors and user experiments with 12 learners. The results, based on feedback through qualitative comments and the percentage of learners who received a perfect score on the confirmation quiz in the BJJ video (87.5% for the proposed and 37.5% for the baseline), suggest that the system can support the learning of the technique. Future research will be conducted to improve the system based on feedback obtained from professionals and learners and to validate the effectiveness of SkillsInterpreter in domains other than modern martial arts.

ACKNOWLEDGMENTS

This work was supported by JST Moonshot R&D Grant JPMJMS2012, JST CREST Grant JPMJCR17A3, and the commissioned research by NICT Japan Grant JPJ012368C02901.

REFERENCES

[1] Renzo Gracie Academy. 2024. *What is Brazilian Jiu-Jitsu (BJJ)?* Retrieved January 15, 2024 from https://renzogracieacademy.com/about/what-is-brazilian-jiu-jitsu-bjj/

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Minsuk Chang, Mina Huh, and Juho Kim. 2021. RubySlippers: Supporting Content-Based Voice Navigation for How-to Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 97, 14 pages. https://doi.org/10.1145/3411764.3445131

[4] BJJ Fanatics. 2023. *A Look At The BJJ Ranking System.* Retrieved December 30, 2023 from https://bjjfanatics.com/blogs/news/a-look-at-the-bjj-ranking-system

[5] BJJ Fanatics. 2024. *BJJ Fanatics - Brazilian Jiu-Jitsu Instructional Videos.* Retrieved January 16, 2024 from https://bjjfanatics.com/

[6] Natsuki Hamanishi and Jun Rekimoto. 2020. PoseAsQuery: Full-Body Interface for Repeated Observation of a Person in a Video with Ambiguous Pose Indexes and Performed Poses. In *Proceedings of the Augmented Humans International Conference* (Kaiserslautern, Germany) *(AHs '20)*. Association for Computing Machinery, New York, NY, USA, Article 13, 11 pages. https://doi.org/10.1145/3384657.3384658

[7] IMMAF. 2024. *WHAT IS MMA?* Retrieved January 15, 2024 from https://immaf.org/about/what-is-mma/

[8] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 3, 20 pages. https://doi.org/10.1145/3586183.3606737

[9] Hye-Young Jo, Laurenz Seidel, Michel Pahud, Mike Sinclair, and Andrea Bianchi. 2023. FlowAR: How Different Augmented Reality Visualizations of Online Fitness Videos Support Flow for At-Home Yoga Exercises. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 469, 17 pages. https://doi.org/10.1145/3544548.3580897

[10] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916 [cs.CL]

[11] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[12] Amy Pavel. 2019. Navigating Video Using Structured Text. https://api.semanticscholar.org/CorpusID:198186570

[13] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/2807442.2807502

[14] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video Digests: A Browsable, Skimmable Format for Informational Lecture Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 573–582. https://doi.org/10.1145/2642918.2647400

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[16] Karina Sokolova and Charles Perez. 2021. You follow fitness influencers on YouTube. But do you actually exercise? How parasocial relationships, and watching fitness influencers, relate to intentions to exercise. *Journal of Retailing and Consumer Services* 58 (2021), 102276. https://doi.org/10.1016/j.jretconser.2020.102276

[17] Dynamic Striking. 2024. *Dynamic Striking | Instructional Videos From The Biggest Names and Best Teachers in The Sport.* Retrieved January 16, 2024 from https://dynamicstriking.com/pages/about-us

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]

[19] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382 [cs.SE]

[20] yokkao. 2024. *What is Kickboxing | Techniques, Benefits and Rules.* Retrieved January 15, 2024 from https://asia.yokkao.com/pages/kickboxing

[21] Lei Zhang, Qian-Kun Xu, Lei-Zheng Nie, and Hua Huang. 2014. VideoGraph: a non-linear video representation for efficient exploration. *The Visual Computer* 30 (2014), 1123–1132.

[22] Yaxi Zhao, Razan Jaber, Donald McMillan, and Cosmin Munteanu. 2022. "Rewind to the Jiggling Meat Part": Understanding Voice Control of Instructional Videos in Everyday Tasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 58, 11 pages. https://doi.org/10.1145/3491102.3502036