Dynamic Subtitles in Cinematic Virtual Reality

Sylvia Rothe
LMU Munich
Munich, Germany
sylvia.rothe@ifi.lmu.de

Kim Tran
LMU Munich
Munich, Germany.
ki.tran@campus.lmu.de

Heinrich Hußmann
LMU Munich
Munich, Germany
hussmann@ifi.lmu.de

TVX '18, June 26–28, 2018, SEOUL, Republic of Korea
ACM 978-1-4503-5115-7/18/06.
https://doi.org/10.1145/3210825.3213556

Abstract

Cinematic Virtual Reality has been increasing in popularity in recent years. Watching $360°$ movies with a Head Mounted Display, the viewer can freely choose the direction of view, and thus the visible section of the movie. Therefore, a new approach for the placements of subtitles is needed. There are three main issues which have to be considered: the position of the subtitles, the speaker identification and the influence for the VR experience. In our study we compared a static method, where the subtitles are placed at the bottom of the field of view, with dynamic subtitles, where the position of the subtitles depends on the scene and is close to the speaking person. This work-in-progress describes first results of the study which point out that dynamic subtitles can lead to a higher score of presence, less sickness and lower workload.

Author Keywords

Cinematic Virtual Reality, 360-degree video, subtitles, dynamic subtitles, static subtitles, speaker identification, presence, sickness, task workload, screen-referenced subtitles

CCS Concepts

•**Human-centered computing** → **User interface programming;** *Virtual reality;*

Introduction

$360°$ movies are attracting widespread interest and have many possible applications, e.g. telling stories about exciting locations in the world or ancient places of interest in history. In Cinematic Virtual Reality (**Cinematic VR**), the viewer watches a $360°$ movie using a Head Mounted Display (**HMD**). Therefore, the viewer is inside the movie and has the possibility to look around. For watching movies in foreign languages, but also for supporting hearing-impaired viewers, subtitles are needed. Not all rules of subtitling can be transformed from traditional movies to Cinematic VR. The freedom of the viewer to choose the viewing direction requires new approaches for subtitling.

In traditional movies, usually **static subtitles** are used. These subtitles are mostly at the bottom of the movie and do not change their position. This method is also called center-bottom subtitles [7]. For reducing head and eye movements during watching movies with subtitles, there are attempts to use **dynamic subtitles** placed near the speaker. The position of these subtitles is dynamically changing and depends on the scene. Other names for these subtitles are speaker following subtitles [7] or positioned subtitles [1].

Regarding subtitles in Cinematic VR there are three main issues. The first issue is the **position** of the subtitle. The viewer can move the head, thereby the field of view (**FoV**) is changing. There is no bottom in a $360°$ image, so the standard location for static subtitles is missing. Using the bottom of the display is one approach for static subtitles in Cinematic VR. Dynamic subtitles can benefit from more space between the speakers in Cinematic VR. In traditional movies usually there is only little room between dialog partners, if they are in the same shot. In other cases, only one person can be seen in one shot - the dialog partner in the next one. In Cinematic VR all talking persons are on the image at the same time with some distance to each other - so the eye movements between speaking persons and bottom-based subtitles are mostly greater than for subtitles placed between the persons.

The second issue is **speaker identification**. The problem of speaker identification is more relevant in Cinematic VR than in traditional videos, as all persons in the room are visible in the $360°$ image at the same time, even if the viewer sees just a part of it. Placing the subtitles near the speaker, helps to identify the speaker, however the viewer is restricted in the choice of the viewing direction when reading the subtitles. In our experiments we used speaker names for the static method and placements close to the speaker for the dynamic method to indicate the speaker.

This leads to the third issue - the **VR experience** - which includes topics such as presence, sickness and workload. Watching the movie using a HMD, the viewer is inside the scene - part of the surrounding scenery. Since subtitles do not belong to this scenery, the presence could be reduced and additional workload or sickness could be caused.

Searching for a subtitling method in Cinematic VR, the following issues have to be taken into account:

- The subtitles have to be easily readable, and should support the viewer's understanding of the story.
- The subtitles have to be understandable with an easy way for speaker identification.
- The subtitles should not destroy the VR experience - with as little eye strain as possible, less sickness and high presence.

Since speaker identification is an important issue for subtitling, especially in Cinematic VR, we chose scenes with more than one speaker: one dialog scene with two people and a meeting room scene with several people. We compared different subtitle methods for these scenes.

As a first approach to this topic, we started studies for viewers with normal hearing abilities watching movies in foreign languages. We are aware of the fact that not all of our findings can be adapted to subtitles for **hearing-impaired** viewers. For parts of our user study we had one deaf participant, who gave us valuable hints for our further research. We did not include this data in our analysis, as we decided to work out subtitles methods for this specific user group in the near future.

Related Work

Placement of Subtitles in Traditional Videos

Kurzhals et al. [7] compared center-bottom subtitles with dynamic (speaker-following) subtitles in traditional videos. Dynamic subtitles led to higher fixation counts on points of interest and reduced saccade lengths. The participants had the subjective impression of understanding the content better with dynamic subtitles. In their experiments the audio was muted. Since in Cinematic VR, audio is an important cue for hearing people to recognize something new in the scene, even outside the FoV, we did not adapt this approach. Instead, we manipulated the audio.

Several studies investigated the placement of dynamic subtitles in traditional videos for reducing the switching rate and distance between regions of interests and subtitle [1, 2, 5]. In our work we investigate if dynamic subtitles are applicable in Cinematic VR environments.

Brown et al. [2] analyzed the eye tracking data for subtitles in regular videos. They found out that gaze patterns of people watching dynamic subtitles were more similar to the baseline, than watching with traditional subtitles. Most of the participants were more immersed and missed less of the content. However, a few people preferred traditional subtitles, because they found dynamic subtitles more distracting. Another mentioned disadvantage was that for viewers who do not need subtitles, dynamic subtitles are more disruptive. This weakness is not relevant for Cinematic VR, as every viewer can choose if subtitles are desired, in contrast to traditional videos, where several people can look at the same display.

Speaker Identification in Traditional Videos

Another problem besides placement of subtitles is the identification of speakers in cases where there are more than one speaker. To place the subtitles near the speaker is one of the methods which can help to solve the problem. Vy and Fels [8] compared subtitles including speaker names with subtitles next to the speaker. In their experiments the participants felt distracted by subtitles following the speakers who change the place. Speaker names were helpful for most participants, but not for deaf viewers, who are not aware of the voices and do not usually identify people by names, but rather by visual characteristics. A conclusion of the paper is that hearing-impaired persons need different methods of subtitling than hearing persons. Since our participants were hearing people we used names for speaker identification in the static method.

Static Subtitles in 360° Videos

In their work-in-progress Brown et al. [3] suggested four static methods of subtitling. We implemented these methods and compared them in a prestudy. All participants chose the Static-Follow method - where the subtitles are moving with the head of the viewer - as the most comfortable and best working. Hence, in our main study we compared this method with dynamic subtitling. For this presence, simulator sickness and task workload was measured and a semi-structured interview was carried out.

User Study

In our study we compared static and dynamic subtitling. For the **static subtitles**, the texts are fixed in front of the viewer and statically connected to the head movements. In our experiments they are 12.5° below eye level. For speaker identification, the name of the speaker was added at the beginning of the text.

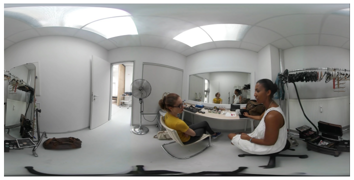

Figure 1: The scene of talk video

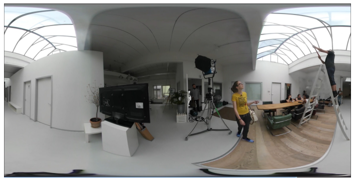

Figure 2: The scene of the meeting video

The position of the **dynamic subtitles** is near the speaker. It depends on the scenario where the subtitles are placed. Thus, the viewer has to look in the direction of the speaker to read the text.

Participants and Material

34 paid participants (26 men, 8 women, average age 22.9, 11 VR beginners) watched the videos using an Oculus Rift. They saw two short scenes recorded in a TV studio (3min length overall). In the first scene (Figure 1) two people talk to each other - we call this the "talk" video. In the second scene (Figure 2), there are several people in a meeting room, others are coming and leaving. This video is called the "meeting" video. We wanted to make sure, the participants did not understand the spoken text, therefore the audio was manipulated.

Study Procedure

After a general questionnaire part, every participant saw the same two short videos, each of them with one of the two methods. The order of videos and methods was counterbalanced.

All the head movements were tracked. After each video the task workload, sickness and presence parts of the questionnaire were answered.

Task workload: The workload was studied using the NASA-TLX questionnaire [4], where all six sub-scales were used: (1) Mental Demand, (2) Physical Demand, (3) Temporal Demand, (4) Performance, (5) Effort, (6) Frustration. In addition to the overall load, the subscale rates of each single item were compared for finding possible reasons for increased workload.

Simulator sickness: For measuring simulator sickness a reduced questionnaire of the Simulator Sickness Questionnaire (SSQ) of Kennedy et al.[6] was used. Since not all questions are relevant for Cinematic VR, six items were selected: (1) general discomfort, (2) fatigue, (3) headache, (4) eye strain, (5) difficulty focusing, (6) nausea, (7) difficulty concentrating.

Presence: To investigate the presence, we used parts of the presence questionnaire (PQ) of Witmer and Singer [9]. Since the PQ was developed for general Virtual Environments with interactivity and movement, we chose some of the questions which are relevant for Cinematic VR.

The questionnaire ended with some questions comparing the two methods. After each video a semi-structured interview was held and recorded.

First Results

Asking the participants directly about the preferences for the two subtitling methods, the results were well-balanced. However analyzing the scores of the NASA-TLX, SSQ and PQ questionnaires we could find some differences, which need a closer inspection. For some items dynamic subtitles led to a higher score of presence, less sickness and lower workload. Additionally, we got important hints for problems which we will investigate in the future.

In the comment part of the questionnaire and the semi-structured interview the following statements were mentioned :

Static Method, positive:

- “I can decide where to look.”
- “This method is similar to the method in TV.”
- “The subtitles are always visible.”

Static Method, negative:

- “It is difficult to assign the speaker.”

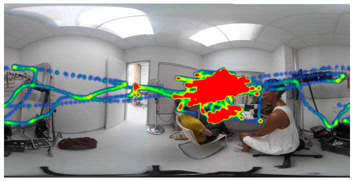

Figure 3: Heatmap of the head tracking data for the talk video with dynamic subtitles. There is a cluster in the area of the subtitles.

Dynamic Method, positive:

- “Subtitles can be assigned more easily to the speaker.”
- “Speakers and subtitles can be seen simultaneously.”
- “It is a more natural experience".
- “It is easier to absorb the content.”

Dynamic Method, negative:

- “I am forced to look at the speaker.”
- “It is sometimes difficult to discover the speaker.”
- “I did not know where the next subtitle will appear.”

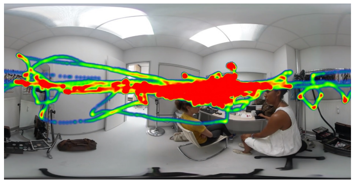

Figure 4: Heatmap of the head tracking data for the talk video with static subtitles. The viewers are looking more around than in the dynamic case

Inspecting the heatmaps of the head tracking data, we found differences for the talk scene (Figure 3 and Figure 4). In time intervals where people were speaking, the data of the dynamic methods are more concentrated around the speakers, which means less head movements.

Discussion

Static subtitles make it easier to look around but more difficult to absorb the content. For speaker identification the dynamic method is preferred. The viewer can see the speaking person and read the subtitles simultaneously without extensive eye movements. From there, it is easier to capture the content. However, if the speaking person is changing and the following person is not in the FoV, it needs some effort to find the new speaker and subtitle. If there is more than one speaking person in the movie, it is difficult to match the subtitles to the speakers using the static method. So it is more difficult to understand the story.

Even if the participants did not prefer one of the methods in the comparison part of the questionnaire, the questions about the VR experience result in better scores for the dynamic method. One reason could be that dynamic subtitles are integrated in the movie and static subtitles are part of the display. Participants noted, that dynamic subtitles are ”more natural , it coincides more with the real life”. Comparing the data regarding task workload, sickness and presence the dynamic subtitling method was more comfortable in several cases. There is less eye strain because the subtitles are placed near the speaking persons and the viewer is not forced to switch to the bottom of the FoV.

Conclusion and Future Work

We explored two types of scenes: a dialog of two persons and a group of speaking people. The protagonists did not change their positions during the conversation. For moving protagonist, who are speaking, dynamic subtitles need to move accordingly. Such scenarios require further testing.

The participants of this study were hearing people. So, the results can be helpful for finding subtitle methods for foreign languages. Because we want to continue our work with subtitles for hearing-impaired people, we had one deaf person at the end of our user study, who tried out the investigated methods. As we expected, the problem of speaker identification needs much more effort than for hearing people. For hearing people the voices of the protagonists are an aid which is not available for deaf people. Different colors, fonts or signs are already used in subtitling of traditional movies and could be adapted. However, the problem of speaker identification in Cinematic VR is harder than in traditional movies and needs more research.

For logging the viewing direction we used head tracking. The additional usage of an eye tracker could lead to more detailed results in the analysis of the viewing direction.

Both methods - static and dynamic subtitling - are helpful for understanding movies in foreign languages. Even if our work is just a first approach and we investigated just two special scenes, the result of this study encourages further studies in this field. We think there is much potential in dynamic subtitles which are not used in Cinematic VR at the moment. However, none of the investigated methods meet all requirements for each scenario in Cinematic Virtual Reality. A combination of the methods depending on the requirements could be a new approach. Additionally, the subtitling methods could be expanded with techniques of attention guiding to facilitate speaker identification.

Even if we are just at the beginning of finding useful subtitle methods for Cinematic VR, these techniques are also important in other areas such as Augmented Reality and other fields of Virtual Reality.

REFERENCES

1. M Brooks and M Armstrong. 2014. Enhancing Subtitles. *TVX2014 Conference, Brussels* (2014), 25–27.
2. Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic subtitles: the user experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*. ACM, 103–112.
3. Andy Brown, Jayson Turner, Jake Patterson, Anastasia Schmitz, Mike Armstrong, and Maxine Glancy. 2017. Subtitles in 360-degree Video. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*. ACM, 3–8.
4. Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.
5. Yongtao Hu, Jan Kautz, Yizhou Yu, and Wenping Wang. 2015. Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11, 2 (2015), 32.
6. Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.
7. Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6559–6568.
8. Quoc V Vy and Deborah I Fels. 2010. Using placement and name for speaker identification in captioning. In *International Conference on Computers for Handicapped Persons*. Springer, 247–254.
9. Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and virtual environments* 7, 3 (1998), 225–240.