Dynamic Subtitle Placement Considering the Region of Interest and Speaker Location

Wataru Akahori[1], Tatsunori Hirai[2] and Shigeo Morishima[3]
[1]*Waseda University / JST ACCEL, Tokyo, Japan*
[2]*Komazawa University, Tokyo, Japan*
[3]*Waseda Research Institute of Science and Engineering / JST ACCEL, Tokyo, Japan*
akahori@akane.waseda.jp, thirai@komazawa-u.ac.jp, shigeo@waseda.jp

Keywords: Dynamic Subtitles, Eye-tracking, Region of Interest, Speaker Detection, User Experience.

Abstract: This paper presents a subtitle placement method that reduces unnecessary eye movements. Although methods that vary the position of subtitles have been discussed in a previous study, subtitles may overlap the region of interest (ROI). Therefore, we propose a dynamic subtitling method that utilizes eye-tracking data to avoid the subtitles from overlapping with important regions. The proposed method calculates the ROI based on the eye-tracking data of multiple viewers. By positioning subtitles immediately under the ROI, the subtitles do not overlap the ROI. Furthermore, we detect speakers in a scene based on audio and visual information to help viewers recognize the speaker by positioning subtitles near the speaker. Experimental results show that the proposed method enables viewers to watch the ROI and the subtitle in longer duration than traditional subtitles, and is effective in terms of enhancing the comfort and utility of the viewing experience.

1 INTRODUCTION

Placing subtitles in a video has been widely used in various situations such as foreign language videos, noisy environments, or for people with hearing impairments. Conventionally, subtitles are rendered at a fixed position, i.e., the bottom-center of the screen. However, because the human perceptual span for reading is narrow (McConkie et al., 1989; Rayner, 1975), the viewer's attention is drawn from the main video content to subtitles. Thus, subtitles disturb the viewer's ability to concentrate on the visual content. Moreover, frequent changes in gaze point between the ROI and the subtitles can cause eyestrain. Accordingly, a method that enables users to view video content and subtitles efficiently is required.

To address these problems, previous studies vary the position of subtitles according to the speaker's location (Hong et al., 2011; Hu et al., 2015) and the viewer's gaze position (Akahori et al., 2016; Katti et al., 2014). Herein, we refer to subtitles that change position as dynamic subtitles. These dynamic subtitling methods enable the viewer to follow the active speaker easily while understanding the content of the spoken dialog. However, the methods based on the speaker detection (Hong et al., 2011; Hu et al., 2015) cannot place dynamic subtitles robustly when it is difficult to detect the speaker. Furthermore, the methods based on the estimated ROI using the viewer's gaze position (Akahori et al., 2016; Katti et al., 2014) did not tackle the problem of frequent subtitle position changes.

We propose a dynamic subtitling method based on both eye-tracking data and a speaker identification algorithm. The proposed method estimates the ROI (calculated using eye-tracking data of multiple viewers), detect the active speaker (identified by combining audio and visual information), and positions subtitles based on the ROI and the speaker's location. The proposed method positions subtitles in a manner that does not interfere with the ROI and enables viewers to recognize the speaker easily. We conducted an eye-tracking data analysis and a user study to verify the effectiveness of the proposed method.

2 RELATED WORK

Placing speaking dialog in the image has been studied to improve accessibility and understanding for applications such as a word balloon in comics (Cao et al., 2014; Chun et al., 2006; Kurlander et al., 1996). Although word balloon placement is helpful for optimiz-

Akahori W., Hirai T. and Morishima S.
Dynamic Subtitle Placement Considering the Region of Interest and Speaker Location.
DOI: 10.5220/0006262201020109
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 102-109
ISBN: 978-989-758-227-1

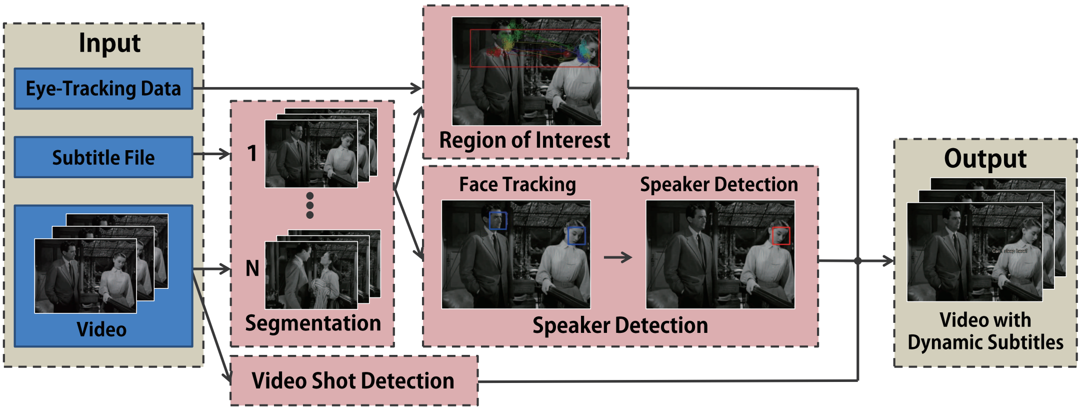

Figure 1: Overview of the proposed method.

ing script location to avoid interfering with the main visual content, it cannot be easily applied to time-varying images such as video content. To deal with the time-varying images, we determine the subtitle position based on temporal information of eye gaze data.

There are two approaches to vary the position of subtitles in a video: speaker-following subtitling and gaze-based subtitling. Speaker-following subtitling allows viewers to follow both the speaker and the subtitle. Hong *et al.* identified the active speaker based on lip motion and positioned subtitles in a low-salient region near the speaker (Hong et al., 2011). Hu *et al.* extended that speaker identification method to optimize subtitle placement (Hu et al., 2015). These methods can help viewers better recognize the speakers. However, speaker identification is challenging if the speaker's face is small or not frontal and if the characters are moving. Therefore, a more robust video subtitling method is preferable. Gaze-based subtitling aims to position subtitles robustly based on the viewer's ROI. Katti *et al.* proposed an interactive online subtitle positioning method that captures a user's eye-tracking data (Katti et al., 2014). This method detects the active speaker robustly based on the fact that the viewer can easily identify and track the speaker. However, because the viewer's future interest is estimated by buffering previous gaze locations, subtitles remaining on the screen can overlap the main video content when the content moves dynamically. In our previous work, we placed subtitles in response to averaged group eye-tracking data and reduced the risk of occluding the critical position of visual attention (Akahori et al., 2016). However, because this method does not detect the speaker, the subtitle positions may confuse the viewers match the subtitle with the corresponding character. Moreover, this method did not consider the subtitle position consistency, which makes the viewers feel uncomfortable.

Compared with previous studies, the proposed method combines the benefits of speaker-following subtitling and gaze-based subtitling methods. The proposed method can place dynamic subtitles near the active speaker robustly so that the subtitles enable the viewers to recognize the speaker. The proposed method also avoids the subtitles from overlapping with important regions and frequent position change.

3 PROPOSED METHOD

Given a video, a corresponding subtitle file, and multiple viewer gaze data as input, the proposed method outputs a video with subtitles that appear at different positions for each subtitle segment. Here, the subtitle file is a text file in SRT format that includes the timing and content of each subtitle. Timing information indicates the time when subtitles appear and disappear.

An overview of our proposed method is shown in Figure 1. First, we divide the input video into speaking segments based on the timing information in the subtitle file and detect individual scenes using a shot segmentation technique (Apostolidis and Mezaris, 2014). Then, we estimate the ROI and detect the active speaker. Finally, we position the subtitle based on the ROI, the speaker's location, and the shot timing information.

3.1 Estimating the ROI

To prevent subtitles from overlapping an important region in a scene, we estimate the ROI using eye-tracking data. Various methods to calculate salient regions based on low-level image features have been proposed (Harel et al., 2006; Hou and Zhang, 2007; Itti et al., 1998). However, as these methods do not

Table 1: The description about video clips.

Movie name	Clip ID	Activity level	Subtitle segments	Detected shots
"Roman Holiday"	C1	High	22	5
	C2	Low	31	3
	C3	Middle	32	9
"Charade"	C4	Middle	31	26
	C5	Middle	27	25
	C6	High	30	14

accurately predict human eye gaze in target-oriented situations, methods that use object localization, detection, and segmentation have also been proposed to consider human intrinsic attention due to anticipation and intention (Cerf et al., 2008; Kanan et al., 2009; Yang and Yang, 2012). Moreover, because predicting eye gaze is challenging, methods that use gaze data as direct input have been proposed (Akahori et al., 2016; Jain et al., 2015; Katti et al., 2014). Following these methods, we use eye-tracking data to estimate the ROI and avoid positioning subtitles that are overlapping with visually important regions.

3.1.1 Eye-tracking Data Collection

First, we collected eye-tracking data of multiple viewers to determine the ROI. 6 2-minute video scenes with English audio tracks were selected from two movies in the public domain, i.e., "Roman Holiday" and "Charade" (Table1). 5 participants (4 males, 1 female) were recruited from graduate students aged 23-28 years ($\mu = 25.0, \sigma = 1.87$). All participants were native Japanese speakers with normal or corrected eyesight, no hearing-impairments, and having a basic knowledge of English. The participants were asked to sit approximately 1.3 m from a 42-inch display and watch the six video clips. The order of the clips was randomized for each participant. Gaze points were recorded using a Tobii X3-120 eye-tracker at 120 Hz. During the measurement, the participants could move their heads freely.

3.1.2 Estimating the ROI

Katti *et al.* initially positioned subtitles near the active speaker and made the subtitles to track the position of the active speaker (Katti et al., 2014). However, this strategy was less effective than initializing subtitles near the speaker and leaving the subtitles at that position. Therefore, after calculating the ROI from all eye gaze positions for each subtitle segment, we initialize subtitles based on the ROI and leave them there.

For each subtitle segment, the ROI is computed from the gaze position $\boldsymbol{r}_i(t) = (r_i^1(t), r_i^2(t)), i = 1, 2, \ldots, N$, where N is the number of viewers. With the N gaze dataset, the mean value $\boldsymbol{\mu} = (\mu^1, \mu^2)$ and the standard deviation $\boldsymbol{\sigma} = (\sigma^1, \sigma^2)$ are computed as follows:

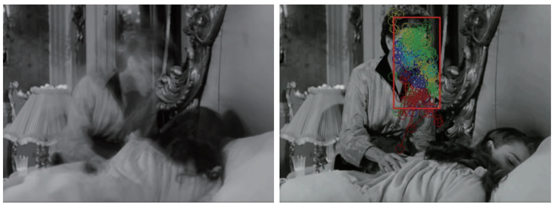

Figure 2: (Left) Averaged image in a subtitle segment; (right) the last frame with the plots of eye-tracking data and the estimated ROI (red rectangle) in the subtitle segment.

$$\mu^k = \frac{1}{N(t_e - t_s)} \sum_{i=1}^{N} \sum_{t=t_s}^{t_e} r_i^k(t), \quad (1)$$

$$\sigma^k = \sqrt{\frac{1}{N(t_e - t_s)} \sum_{i=1}^{N} \sum_{t=t_s}^{t_e} (r_i^k(t) - \mu^k)^2}, \quad (2)$$

where $k \in \{1, 2\}$ represents the x and y axes, and t_s and t_e represent the eye gaze sample number wherein the subtitles appear and disappear respectively. Finally, using the mean value $\boldsymbol{\mu}$ and the standard deviation $\boldsymbol{\sigma}$, a rectangular area $\boldsymbol{\mu} \pm 2\boldsymbol{\sigma}$ is created to surround the ROI, as shown in Figure 2. Each color plot represents the eye-tracking data of five viewers respectively. If the viewers' eye gaze are assumed to follow normal distribution for each subtitle segment, approximately 95% of the eye gaze positions are included in this rectangular area.

3.2 Speaker Detection

For each subtitle segment, we detect an active speaker to enable viewers to recognize the speaking character. The lip motion feature is widely used in previous speaker detection work (Everingham et al., 2006; Hong et al., 2011). Hu *et al.* combined audio-visual features and detected the active speaker more precisely than the methods based on the lip motion (Hu et al., 2015). Therefore, we detect the active speaker based on their algorithm.

3.2.1 Face Tracking and Landmark Localization

First, face tracks are obtained using following tracking-by-detection procedure. For each subtitle segment, the face detector (King, 2015) is executed for each frame[1]. The detecting results are used to establish correspondence between pairs of detected faces within the subtitle segment. For a given pair of faces in different frames, the overlap region between the former detected face and the latter detected face is calculated, and a match is declared if the overlap region is 40% or more to the latter detected face region. However, it is difficult to detect a face in every frame if the face is dynamically moving, which causes a labeling failure. Accordingly, using the face detection results as input, we perform face tracking using an object-tracking technique (Danelljan et al., 2014) for each detected face to complement the frame where the face is not detected[1]. Finally, we detect facial feature points in the tracked face region using the method proposed by Uricar *et al.* (Uřičář et al., 2012).

3.2.2 Speaker Detection Algorithm

The main idea of the speaker detection algorithm proposed by (Hu et al., 2015) is the cascade classifier, which comprises four features: (i) mean squared distance (MSD), i.e., the distance between consecutive frames in the mouth region; (ii) center contribution (CC), i.e., the distance between a candidate's face position and the center of the screen; (iii) length consistency (LC), i.e., the consistency between the length of a candidate's face tracking time and the length of the speaking time; and (iv) audio-visual (AV) synchrony, i.e., the synchrony score between the audio features and lip motion features. Speaker detection is performed by chaining these four features in a cascade in the following order: MSD, CC, LC, and AV. The design of the cascade structure is based on the observation that only speakers pass to the next step. Details of the algorithm are described in (Hu et al., 2015).

The speaker detection accuracy is shown in Table 2[2]. Here, precision is the proportion of correctly detected speakers and recall is the proportion of the number of detected speaker segments to the number of all subtitle segments, except when the active speaker is not visible on the screen.

Table 2: Precision and recall of the speaker detection algorithm to the input video clips.

Clip ID	Precision (%)	Recall (%)
C1	61.9	59.1
C2	88.9	51.6
C3	66.7	40.0
C4	77.8	67.8
C5	81.3	48.1
C6	63.0	56.7

3.3 Subtitle Placement

Subtitles are positioned based on the estimated ROI, speaker detection results, and shot-change timing information. We consider the following points to improve user experience: (1) ease of speaker recognition (it is easy to recognize a speaker in a subtitle segment); (2) aesthetics (subtitles should not overlap visually important content, *e.g.,* a face, in the video); and (3) suppression of varying the subtitle position (the distance between continuously appearing subtitles should be small to avoid interfering with the viewer's cognitive process). In the following, we describe how to select the candidate subtitle region and placement of subtitles.

3.3.1 Candidate Subtitle Region

We determine the candidate subtitle region to enable the viewers to follow the important visual content and avoid overlapping with the content. In previous work (Hong et al., 2011; Hu et al., 2015), candidate subtitle positions were close to the speaker (*e.g.,* above left, above, above right, below left, below, below right, left and right). However, when the ROI is large in a subtitle segment, there is not enough space to place a horizontal subtitle at the positions to the left or right of the ROI. Moreover, most viewers are accustomed to subtitles positioned at the bottom of the screen, and the distance between consecutive subtitles should be small. Thus, we place subtitles just below the ROI.

3.3.2 Subtitle Placement

To enable viewers to easily recognize the active speaker, the *x*-coordinate of the center of the subtitle S_x is calculated as follows:

$$S_x = \begin{cases} \frac{(Sp_x + R_x)}{2} & if\ speaker\ is\ detected, \\ R_x & if\ speaker\ is\ not\ detected, \end{cases} \tag{3}$$

where Sp_x is the *x*-coordinate of the center of the speaker's face and R_x is the *x*-coordinate of the center of the ROI. When the active speaker is not detected, we can place dynamic subtitles robustly according to the video content by positioning the subtitle based on

[1] The DLib C++ library provides open-source implementations of (Danelljan et al., 2014; King, 2015) in http://dlib.net/.

[2] We apply $\theta_1 = 5$ for C1, C2, and C3 and $\theta_1 = 6.5$ for C4, C5, and C6. We also apply $\theta_2 = 2$, $\theta_3 = 2$, $\theta_4 = 0.1$, $\theta_5 = 2$ throughout.

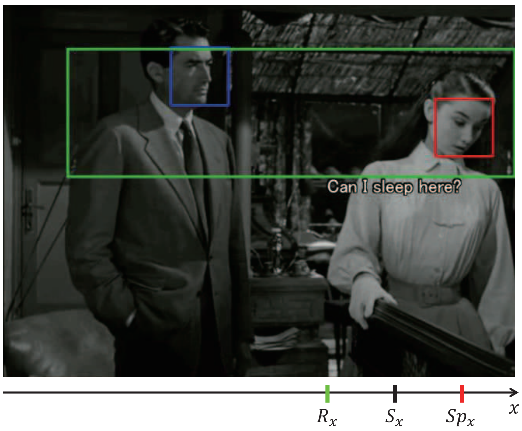

Figure 3: Example subtitle placement result (calculated ROI (green rectangle), detected speaker (red rectangle), and detected non-speaker (blue rectangle)).

the region of interest. To avoid overlapping subtitles with the ROI, we equalize the *y*-coordinate of the upper edge of the subtitle with the *y*-coordinate of the lower edge of the ROI. Figure 3 shows the placement result.

Since frequent position change of subtitles disturbs viewers, we suppress changes in the *y*-coordinates of the subtitles when the time interval between continuously appearing subtitles is small. First, we create clusters in which the time interval between consecutive subtitles is within 0.3s. Then, we unify the *y*-coordinates of the subtitles to the maximum value of the *y*-coordinate within each cluster. Thereby, the subtitles do not overlap the ROI and the position changes less frequently.

If a shot change is included in a subtitle segment, the dynamic subtitles might cause cognitive burden on the viewers. Although the subtitles should change its position according to the current scene, frequent changes are not preferable in terms of viewer's cognitive burden. Therefore, we place subtitles at the bottom-center of the screen when a shot change is included in the subtitle segment so that the subtitles do not interrupt the viewer's concentration.

4 EXPERIMENT

To assess the effectiveness of the proposed method, we conducted eye-tracking data analysis and a user study.

4.1 Participants

19 participants (17 males, 2 females) were recruited from graduate and undergraduate students aged 21-26 years ($\mu = 23.2, \sigma = 1.36$). Note that these participants are not the same as participants in Section 3.1.1. All participants were native Japanese speakers with a basic knowledge of English, normal or corrected eyesight, and no hearing impairments.

4.2 Video Clips and Subtitles Setup

Subtitles play an important role, especially when watching a foreign language video. We assume such a situation by placing Japanese subtitles in six video clips with English audio (Table 1). For each video clip, we produced three modes (Table 3). Although Hu *et al.* displayed subtitles with a blurb (Hu et al., 2015), Katti *et al.* pointed out that blurbs distract the viewer's understanding (Katti et al., 2014). Therefore, in our method, all subtitles were displayed as white text with a thin black outline without a blurb.

4.3 Experience Design

The participants were shown 18 video clips in total, 3 subtitle modes (Table 3) for each of the 6 clips (Table 1), while capturing their eye-tracking data in the same environment described in Section 3.1.1. A short break was taken between successive clips. Note that the clips were shown in random order. To evaluate comfort and utility, after watching each of the video clips, the participants were asked the following questions,

1). Did you feel uncomfortable with the position of the subtitles? (7-point Likert scale, 7: not uncomfortable at all; 1: quite uncomfortable)

2). Did you feel that the subtitle placement method was useful? (7-point Likert scale, 7: quite useful; 1: not useful at all)

3). When did you feel comfortable or uncomfortable when watching the video? (open ended)

4.4 Eye-Tracking Data Analysis

We computed the percentage of the duration that visual fixations fall into the rectangle ROI that proposed in Section 3.1.2 or the subtitle region in all subtitle segments. Figure 4 shows these percentages that are averaged over viewers. A two-tailed t-test was conducted to determine whether the difference between the average points was statistically significant.

The visual fixations of the participants when watching *Dynamic Subtitles2* (the proposed method) seemed to be included within the ROI and the subtitle region in longer duration than *Static Subtitles*. Thereby, the proposed method reduced unnecessary

Table 3: The description of subtitle mode.

Subtitle Mode	Description
Static Subtitles (SS)	Traditional subtitles positioned at the bottom-center of the screen.
Dynamic Subtitles1 (DS1)	Speaker-following subtitles (Hu et al., 2015).
Dynamic Subtitles2 (DS2)	Gaze-based and speaker-following subtitles (proposed method).

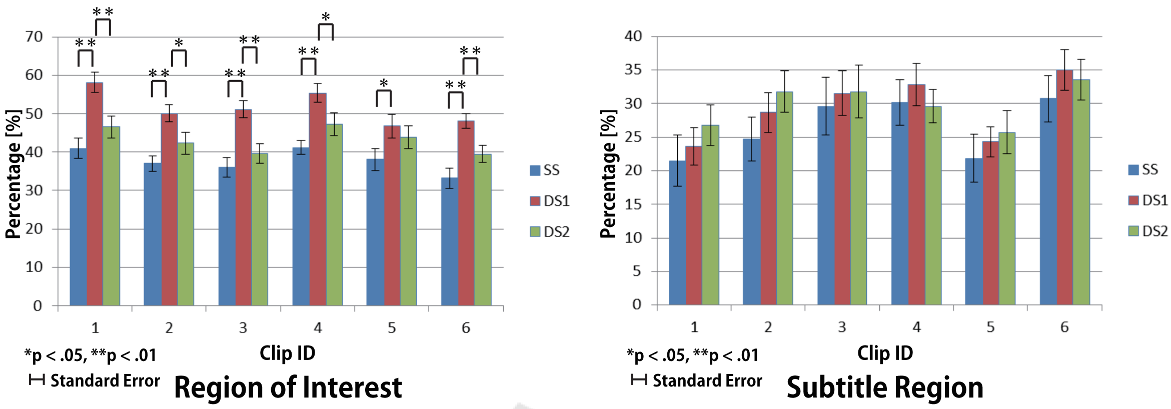

Figure 4: The percentage of eye fixation included in the ROI (left) and subtitle region (right).

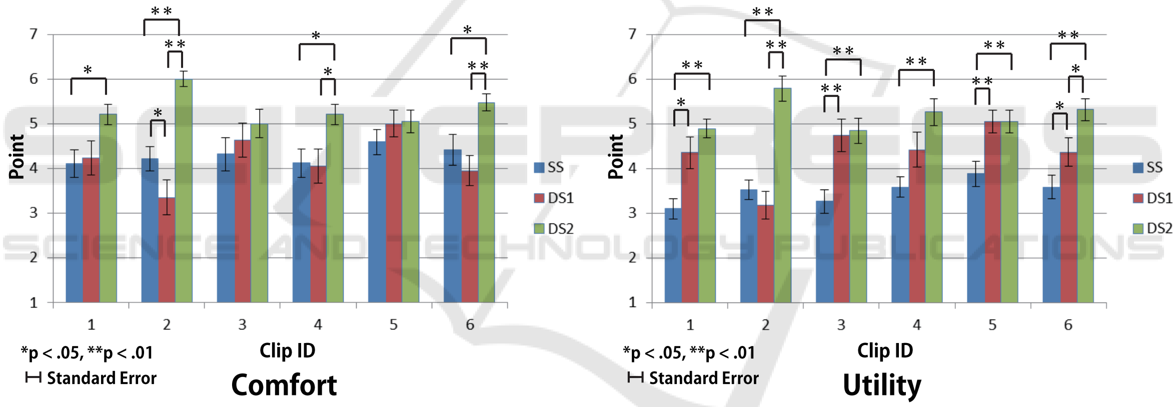

Figure 5: Results of comfort (left) and utility (right) for each subtitling method. (7-point Likert scale, 1: bad; 7: good).

eye movements between the video content and the subtitle. Since frequent eye movements interfere with understanding the video content, the proposed method enables to avoid such disruption. As for the ROI, although the percentage of *Dynamic Subtitles1* (Hu et al., 2015) shows particularly high values, since the subtitles were included in the ROI in most cases, these may overlap the important video contents.

4.5 User Study

Figure 5 shows the averaged results of the user study. A two-tailed Wilcoxon signed-rank test was conducted to determine whether the difference between the average score was statistically significant.

Dynamic Subtitles2 (the proposed method) outperformed *Static Subtitles* in terms of utility. Several participants gave a favorable response for the proposed method.

> “ I felt it was easy to watch when the positions of the subtitle and the position of the speaker’s face were close.” (P11)

Similar comments were also given for *Dynamic Subtitles1* (Hu et al., 2015). From this result, placing subtitles near the active speaker is important for dynamic subtitle placement. Furthermore, some participants stated as follow.

> “ I felt comfortable when subtitles did not cover the speaker’s face during a conversation.” (P2)

Comparing with previous related work (Hu et al., 2015), we consider temporal gaze positions to avoid

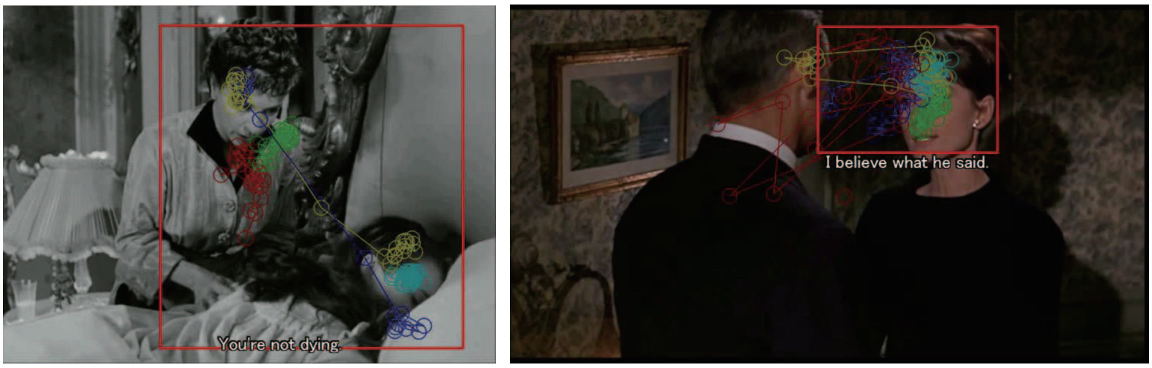

Figure 6: Examples of subtitles which are placed far away from the speaker's face (left) and placed on the speaker's chin (right). The red rectangle represents the ROI and each color plot represents the eye-tracking data of five viewers respectively.

the subtitles overlapping with the important video contents which move around. Moreover, some participants pointed out another interesting aspect.

> “ It was easy to watch when the subtitles were positioned over the speaker's chest.” (P15)

> “ It was strange when subtitles were displayed above the speaker's face.” (P18)

Therefore, This results show that the participants seem to prefer subtitles positioned below the speaker's face.

On the other hand, although *Dynamic Subtitles2* (the proposed method) outperformed *Dynamic Subtitles1* (Hu et al., 2015) and *Static Subtitles* in terms of comfort for four clips (C1, C2, C4, and C6), *Dynamic Subtitles2* (the proposed method) did not differ significantly from *Dynamic Subtitles1* (Hu et al., 2015) and *Static Subtitles* for two clips (C3 and C5). Note that some participants provided negative comments about the proposed method for C3 and C5, respectively as follows.

> “ It was difficult to watch when subtitles were positioned away from the speaker's face.” (P9)

> “ I felt uncomfortable when subtitles covered on the speaker's chin.” (P1)

As the *y*-coordinates of the subtitles depend on the standard deviation in y axis of multiple viewers' gaze positions, subtitles are positioned far away from or too close to the active speaker when the ROI is large or small by comparison with the speaker's face size as shown in Figure 6. In these cases, we can avoid the viewers with feeling uncomfortable by positioning below at a distance from the speaker's face. In addition, some participants stated the following aspect.

> “ I felt uncomfortable when the subtitles were displayed near a person who was not speaking.” (P19)

The dynamic subtitles can confuse the viewer if they are positioned near a non-speaker's face.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a dynamic subtitling method based on eye-tracking data and a speaker detection algorithm. Our goal was to reduce unnecessary eye movements and improve the viewing experience. The proposed method estimates the ROI to avoid positioning subtitles that interfere with important content. In addition, we detect the active speaker, which allows the viewer to recognize the speaker easily. We position subtitles below the ROI and near the active speaker. The results of eye-tracking data analysis demonstrated that the proposed method enabled to watch the ROI and the subtitle region in longer duration than the traditional subtitles. Moreover, the results of a user study demonstrated that participants generally preferred the proposed method over traditional and previous subtitle placement methods in terms of comfort and utility. Since the number of subtitled videos has increased for both the movie industry and video-sharing services, the proposed method can be applied to various videos in the future.

The bottleneck of the proposed method is that eye-tracking data of multiple viewers are required as input. However, methods to capture eye-tracking data at low cost (San Agustin et al., 2010) and in large quantities (Rudoy et al., 2012) have been proposed, and such methods may improve the usability of the proposed method.

As mentioned in Section 4.5, the dynamically positioned subtitles may confuse the viewer when a non-speaker is detected as the speaker or a speaker is not visible on the screen. Therefore, in the future work, we would like to improve the precision of the speaker detection by combining the content-aware analysis and eye-tracking data for better presentation of subtitles.

ACKNOWLEDGEMENTS

We thank S. Kawamura, T. Kato and T. Fukusato (Waseda University, Japan) for their advisory. This research was supported by JST ACCEL and CREST.

REFERENCES

Akahori, W., Hirai, T., Kawamura, S., and Morishima, S. (2016). Region-of-interest-based subtitle placement using eye-tracking data of multiple viewers. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pages 123–128. ACM.

Apostolidis, E. and Mezaris, V. (2014). Fast shot segmentation combining global and local visual descriptors. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6583–6587. IEEE.

Cao, Y., Lau, R. W., and Chan, A. B. (2014). Look over here: Attention-directing composition of manga elements. *ACM Transactions on Graphics (TOG)*, 33(4):94.

Cerf, M., Harel, J., Einhäuser, W., and Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*, pages 241–248.

Chun, B.-K., Ryu, D.-S., Hwang, W.-I., and Cho, H.-G. (2006). An automated procedure for word balloon placement in cinema comics. In *International Symposium on Visual Computing*, pages 576–585. Springer.

Danelljan, M., Häger, G., Khan, F., and Felsberg, M. (2014). Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press.

Everingham, M., Sivic, J., and Zisserman, A. (2006). Hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, volume 2, page 6.

Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552.

Hong, R., Wang, M., Yuan, X.-T., Xu, M., Jiang, J., Yan, S., and Chua, T.-S. (2011). Video accessibility enhancement for hearing-impaired users. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 7(1):24.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Hu, Y., Kautz, J., Yu, Y., and Wang, W. (2015). Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(2):32.

Itti, L., Koch, C., Niebur, E., et al. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Jain, E., Sheikh, Y., Shamir, A., and Hodgins, J. (2015). Gaze-driven video re-editing. *ACM Transactions on Graphics (TOG)*, 34(2):21.

Kanan, C., Tong, M. H., Zhang, L., and Cottrell, G. W. (2009). Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003.

Katti, H., Rajagopal, A. K., Kankanhalli, M., and Kalpathi, R. (2014). Online estimation of evolving human visual interest. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1):8.

King, D. E. (2015). Max-margin object detection. *arXiv preprint arXiv:1502.00046*.

Kurlander, D., Skelly, T., and Salesin, D. (1996). Comic chat. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 225–236. ACM.

McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., and Jacobs, A. M. (1989). Eye movement control during reading: Ii. frequency of refixating a word. *Perception & Psychophysics*, 46(3):245–253.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology*, 7(1):65–81.

Rudoy, D., Goldman, D. B., Shechtman, E., and Zelnik-Manor, L. (2012). Crowdsourcing gaze data collection. *arXiv preprint arXiv:1204.3367*.

San Agustin, J., Skovsgaard, H., Mollenbach, E., Barret, M., Tall, M., Hansen, D. W., and Hansen, J. P. (2010). Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 77–80. ACM.

Uřičář, M., Franc, V., and Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output svm. *VIsAPP*, 12:547–556.

Yang, J. and Yang, M.-H. (2012). Top-down visual saliency via joint crf and dictionary learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2296–2303. IEEE.