World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:17, No:1, 2023

The Influence of Audio on Perceived Quality of Segmentation

Silvio R. R. Sanches, Bianca C. Barbosa, Beatriz R. Brum, Cléber G. Corrêa

Abstract—In order to evaluate the quality of a segmentation algorithm, the researchers use subjective or objective metrics. Although subjective metrics are more accurate than objective ones, objective metrics do not require user feedback to test an algorithm. Objective metrics require subjective experiments only during their development. Subjective experiments typically display to users some videos (generated from frames with segmentation errors) that simulate the environment of an application domain. This user feedback is crucial information for metric definition. In the subjective experiments applied to develop some state-of-the-art metrics used to test segmentation algorithms, the videos displayed during the experiments did not contain audio. Audio is an essential component in applications such as videoconference and augmented reality. If the audio influences the user's perception, using only videos without audio in subjective experiments can compromise the efficiency of an objective metric generated using data from these experiments. This work aims to identify if the audio influences the user's perception of segmentation quality in background substitution applications with audio. The proposed approach used a subjective method based on formal video quality assessment methods. The results showed that audio influences the quality of segmentation perceived by a user.

Keywords—Background substitution, influence of audio, segmentation evaluation, segmentation quality.

I. Introduction

ONLINE meetings are ways of group work that enable people to collaborate to achieve a goal. These meetings usually are conferences that use the Internet as a means of communication, as the displacement of participants to a specific physical location can generate costs for companies and institutions. During the pandemic caused by the coronavirus (COVID-19), online meetings have become the safest form of communication between people worldwide.

Apps like Skype, Hangouts, and WhatsApp allow users to make group calls in commercial and academic areas [1], [2], [3]. A common concern for a participant in an online meeting is that his background is visible to all participants. Sometimes, a participant prefers to hide his background [4]. Also, for an immersive sensation, a uniform background would be more desirable than a mosaic of different background scenarios [5].

A chroma-key algorithm [6] can remove the background of a scene. However, this approach requires a constant color background that covers the entire viewing area of the camera providing a controlled environment is not desirable in practical terms. There are algorithms able to extract the element of interest in an uncontrolled environment (with an arbitrary background) so that the system can replace the original background [7], [8], [9], [3]. As different algorithms are available, selecting the most appropriate video conference application background is essential.

Several state-of-the-art subjective [10] and objective [11], [12] metrics evaluate the segmentation quality. Such metrics can evaluate segmentation algorithms in different application domains, such as surveillance systems, intelligent environments, and video retrieval [13]. According to certain domains, there are also specific metrics to evaluate segmentation algorithms when used in background replacement applications [12], [14], [15].

Some applications that perform background substitution have characteristics that distinguish them from applications that use the element of interest for different purposes (people tracking, vehicle tracking, and fall detection). One of the essential characteristics of communication systems is audio. In video conference systems with background substitution, for example, participants communicate through the application, and it is possible that the audio influences their perception of the quality of the segmentation. However, state-of-the-art metrics should have considered audio as a component that can influence the user's perception.

Several studies simulate sensory modalities (for example, audio and vision). McDonald et al. [16] showed that involuntary auditory attention affects the perception of visual stimuli, increasing the perception of visual stimuli. Driver and Noesselt [17] showed that specific sensory brain responses and perceptual judgments related to one sense could be affected by relationships with other senses. Psychology has shown that visual and auditory events analyzed together improve visual perception. In [18], the results showed that the hearing modality could affect the perception in the visual modality. In [19], the audio feature improved visual perception.

This work aims to identify if the audio influences the user's perception of segmentation quality in background substitution applications with audio, such as video conference systems. The proposed approach used a subjective method based on formal video quality assessment methods.

II. Quality of Segmentation

Assessing the quality of the segmentation made by an algorithm is a problem that authors investigate in different application domains [13], [12]. There are two types of assessments: objective and subjective [12]. Subjective assessments, which require volunteers and particular infrastructure, are more accurate [15]. However, objective assessments can test the quality of a segmentation algorithm with no users in the process [12], [15].

S. R. R. Sanches, B. C. Barbosa and C. G. Corrêa are with Universidade Tecnológica Federal do Paraná, 1640, Alberto Carazzai Avenue, Cornélio Procópio, PR, Brazil (e-mail: silviosanches@utfpr.edu.br, biancabarbosa@alunos.utfpr.edu.br, clebergimenez@utfpr.edu.br).

B. R. Brum is with Universidade de São Paulo, 400, Trabalhador São Carlense Avenue, São Carlos, SP, Brazil (e-mail beatrizbrum@usp.br).

This section presents the main metrics to assess segmentation quality, emphasizing those specific to evaluate segmentation algorithms in background replacement applications. This section also details subjective methods used to assess segmentation quality. Such methods are one of the necessary steps in developing perceptual objective metrics.

A. Objective Assessment

An objective assessment uses the result of the segmentation of an algorithm to test its quality. The algorithm must segment a set of videos (datasets) that simulate the scene of the application for which the algorithm was developed [20], [21].

An objective metric calculates values such as true positives (TP), false positives (FP), false negatives (FN), or true negatives (TN). TP corresponds to pixels correctly classified as foreground, FP are pixels incorrectly classified as foreground, FN are pixels incorrectly classified as background, and TN are pixels correctly classified as background [21].

Some authors use these values to measure the performance of a segmentation algorithm. However, the most efficient metrics use these values as input to get more reliable measures. The Precision metric, for example, can be calculated according to the equation

$$\frac{TP}{TP+FP} \tag{1}$$

and the Recall metric can be defined as

$$\frac{TP}{TP+FN}. \tag{2}$$

According to (3), the F-score metric that is also often used to test segmentation quality [22], [21], is calculated based on the Precision and Recall metrics.

$$\text{F-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

The metrics presented in (1), (2) and (3) are widely used to test performance of segmentation algorithms used in surveillance systems [21]. These systems do not intend to replace the original background of each segmented frame.

Applications such as video conferences with background substitution [8] and augmented reality systems [23], for example, use the element of interest obtained in segmentation processes to compose a new scene with a new background. Precise identification of the edges of the element of interest influences the quality of the final scene when the application requires background replacement.

Many authors that present segmentation algorithms targeted for video conference applications [7], [24], [8] measure the quality of segmentation by calculating the percentage of pixels classified correctly (PCC) regarding the ground truth segmentation [25]. The PCC metric is defined as

$$\text{PCC} = \frac{TP+TN}{TP+TN+FP+FN}. \tag{4}$$

Like the metrics to test the performance of algorithms in surveillance systems, the PCC metric does not consider the user's perception to assess the quality of the segmentation.

Sanches et al. [14] presented an objective metric that considers the user's perception. However, it is only helpful to select segmentation algorithm parameters when used in augmented reality applications. The authors proposed a set of subjective experiments to find levels of discomfort caused by different segmentation errors.

Gelasca and Ebrahimi [12] presented an objective metric based on subjective assessments to analyze error by simulating spatial and temporal occurrences. The Perceptual Spatio-Temporal (PST) metric considers annoyance levels, and the authors focus on some background substitution applications (augmented reality, surveillance systems, and video compression). The PST metric classifies and penalizes the different misclassified pixels according to changes in the object's shape and, afterward, their size. The authors defined four types of errors, called "artifacts": added regions $\mathcal{A}_r$, added background $\mathcal{A}_b$, inside holes $\mathcal{H}_i$, and border holes $\mathcal{H}_b$.

Added region is the over-segmented part of the background that does not form any semantically meaningful region. This region is disjointed from the correctly segmented objects. Added background is the over-segmented part of the background attached to the correctly segmented object that makes the object larger. Inside holes are under-segmented parts contained inside the objects visible through the object parts of the background. Border holes are under-segmented parts directly attached to the object's border, making the object thinner [12].

PST metric takes into account temporal aspects such as sudden disappearance of artifacts, surprise effect, and expectation effect [12], resulting in four objective perceptual metrics $PST_{\mathcal{A}r}$, $PST_{\mathcal{A}b}$, $PST_{\mathcal{H}i}$ and $PST_{\mathcal{H}b}$. Last, the final metric linearly combines these metrics:

$$\text{PST} = a \times PST_{\mathcal{A}r} + b \times PST_{\mathcal{A}b} + c \times PST_{\mathcal{H}i} + d \times PST_{\mathcal{H}b}. \tag{5}$$

where the weights (a, b, c and d) are obtained by optimization processes [12].

Perceptual Application-Dependent Metric (PAD) presented in [15] also considered the user's perception regarding the quality of the segmentation. First, the authors defined a set of artifacts that annoys the users. Then, they applied an optimization strategy to choose a subset of these artifacts.

The artifacts that compose the PAD metric are $\mathcal{FN}$ (average of errors on the foreground relative to the total amount of foreground pixels), $\mathcal{EW}$ (average of errors on the foreground relative to the amount of pixels in the window), $\mathcal{PW}$ (average of errors on the background relative to the total amount of pixels in the window), $\mathcal{T}$ (average of errors in all frames related to ground truth) and $\mathcal{F}$ (average of errors on face region).

To consider temporal inconsistencies in consecutive frames, the fitting function *Weibull* transformed the artifacts according to equation $T_{art} = 1 - e^{-(x*art)^y}$ where art is each artifact that makes up the metric. The values x and y are the processing *Weibull* function inputs obtained from the

optimization strategy applied to the set of artifacts. Thus, $art = \mathcal{FN}$, $x = 0.051$ and $y = 0.7589$ for $T_{\mathcal{FN}}$; $art = \mathcal{EW}$, $x = 15.981$ and $y = 0.5416$ for $T_{\mathcal{EW}}$; $art = \mathcal{PW}$, $x = 1.2058$ and $y = 760$ for $T_{\mathcal{PW}}$; $art = \mathcal{T}$, $x = 0.0374$ and $y = 0.7583$ for $T_{\mathcal{T}}$; and $art = \mathcal{F}$, $x = 0.000198$ and $y = 0.6279$ for $T_{\mathcal{F}}$. PAD metric is defined according to (6)

$$\text{PAD} = a + b*T_{\mathcal{FN}} + c*T_{\mathcal{EW}} + d*T_{\mathcal{PW}} + e*T_{\mathcal{T}} + f*T_{\mathcal{F}} \quad (6)$$

where $a = 0.4835$, $b = 0.1990$, $c = 0.0873$, $d = 0.0581$, $e = 0.1790$ and $f = 0.7450$ are weights obtained from an optimization process [15].

B. Subjective Evaluation and Application Typical Scenario

The best way to consider the user's perception to test the quality of a segmentation algorithm is by applying subjective methods in which users test videos with segmentation errors [15]. The methods used for this purpose are the same used in Image Quality Assessment (IQA) [26], Video Quality Assessment (VQA) [27] or Service Quality Evaluation (with accompanying audio) [28].

Applying a subjective method to each new segmentation algorithm requires much effort because it is necessary to recruit users and configure the environment to apply the experiments. The objective metrics, such as those discussed in Section II-A, are more appropriate because of their practicality.

Objective metrics that consider the user's perception – such as PST and PAD – apply subjective experiments as one stage of their development process. The main objective of these experiments is to identify the levels of discomfort caused by the different segmentation errors presented in the video frame.

Subjective assessments usually are performed according to recommendations that suggest, for example, how to configure the physical environment and what is the ideal profile for the users [28], [29]. The whole process comprises to: *(i)* define a set of types of errors, called artifacts; *(ii)* generate videos containing these artifacts, which simulate the environment of an application domain; *(iii)* recruit a group of users (participants or volunteers) who should test these videos and *(iv)* analyze the results to identify the level of discomfort caused to users by each artifact. The artifacts and the levels of discomfort caused by each are the basic information that comprises objective perceptual metrics, such as PST and PAD.

One of the essential steps in conducting subjective experiments is to generate videos that simulate an application domain's environment. In video conference applications with background substitution, for example, the typical usage scenario is a person with the visible head and torso in the foreground [15]. Volunteers in the subjective experiments should test videos that have: *(i)* the most significant similarity possible with the application presented and *(ii)* simulate errors or present absolute segmentation errors (artifacts) to the volunteers.

In the subjective experiments used to develop the PST and PAD metrics, the authors used videos during the subjective experiments that simulated the scenario of specific applications (augmented reality and video conference). These are potential applications for the algorithm under evaluation, however, augmented reality and video conference have at least one crucial feature that needs to be considered by PST and PAD metrics. In these systems, communication between users can occur. Some studies show that audio in the environment can affect the visual quality of a video [30], [31]. The subjective experiments that generated the data to define the PST and PAD metrics did not contain any videos with audio. Therefore, the environment of a video conference system and augmented reality applications were not simulated, which may compromise the efficiency of these metrics.

Subjective experiments that contain videos with audio may be necessary to define objective metrics that test algorithms whose potential applications are systems that use audio.

III. Method to Identify the Audio Influence

This section presents details of the method to identify the influence of audio on the perceived quality of segmentation. Fig. 1 shows the main steps of this method.

A. Foreground Layers Generation

Initially, the approach proposed captures videos that simulate a videoconferencing environment (*Source Videos* Block of Fig. 1), in which a typical usage scenario is a person with a visible head and torso in the foreground [15]. Although these applications run in environments with an arbitrary background, a constant color background (blue) was the choice for segmentation to occur more accurately (later segmentation errors were simulated). A Motorola Moto Z Play Smartphone device (model xt1635) recorded 160 frames with High Definition (HD) resolution (1280×720 pixels) at a rate of 30 frames per second (fps). The person in the foreground says, "4K resolution is higher than full HD resolution," while the device captures audio and video.

The chroma-key algorithm developed by Bergh and Lalioti [6] removed the constant background of the captured video. It generated layers with only the element of interest (a person). Although segmentation occurs with quality when the scene's background is constant, the generated layers may contain pixel sorting errors, particularly in the regions near the edge of the element of interest. Segmentation errors are common in these areas because the color of the element of interest and the color of the background [32] can influence these pixels. We edited these layers to correct these errors to set apart the elements of interest (see "Foreground Layers" Block of Fig. 1).

B. Videos Composition

The set of layers obtained by the segmentation combined with the video source generated new videos with the original background replaced by a new constant color (see *Videos Composition* Block of Fig. 1). In video conferences, the new background can be any image with the exact resolution as the frames of the original captured video. However, the gray (**R**ed=127, **G**reen=127, **B**lue=127) is the color that less affects the human viewer according to opinions of psychophysical experts [12]. The new backgrounds of the new videos used

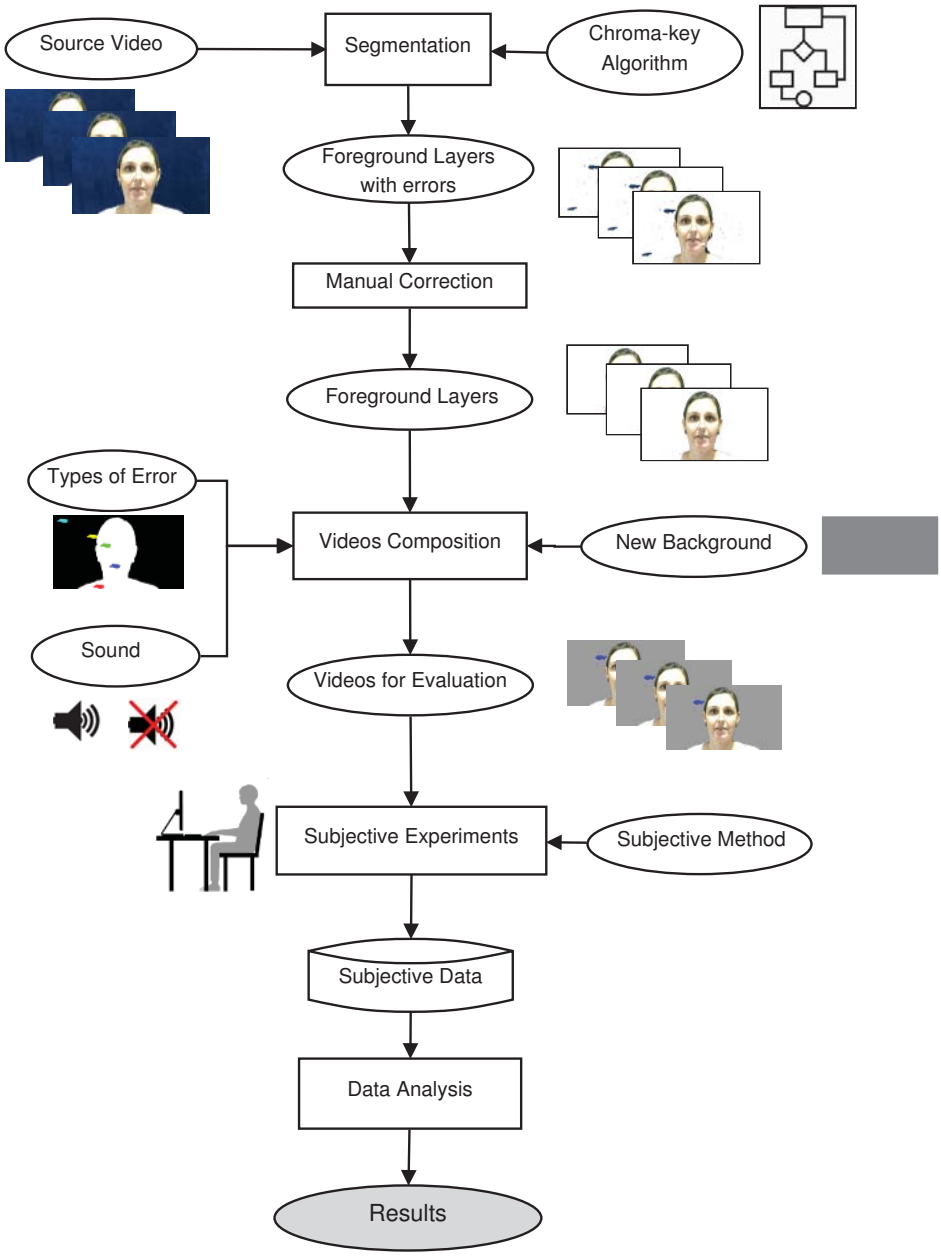

Fig. 1 The method to identify the influence of audio in the perceived quality of segmentation

in the experiments are gray (see *New Background* Block of Fig. 1) so that users better perceive the segmentation errors in this region.

Five regions of the video frame were defined, and inserted segmentation errors were in these regions (see *Types of Error* block of Fig. 1). In each of these regions, segmentation errors cause a different discomfort to the user [15]. Each new video produced has segmentation errors only in one of these regions. These new videos present all simulated segmentation errors in the pattern of asymmetric blobs with 6247 connected pixels. Preliminary experiments in this study showed that the blob-like errors are more noticeable than the errors in isolated pixels scattered around the frame. In addition, the most recent segmentation algorithms found in the literature [21] have errors in the asymmetric form. The regions of a frame that contain error simulations are:

- *Mouth*: errors in the person's mouth, in the element of interest;
- *Corner*: errors in the background, near the left corner of the window;

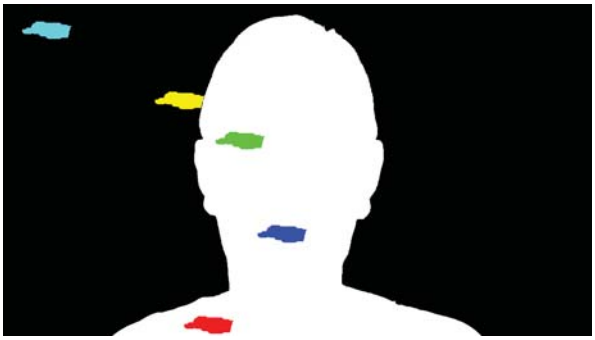

Fig. 2 Regions of the frame where errors were simulated: *Mouth* (dark blue color), *Corner* (light blue color), *Eye* (green color), *Torso* (red color) e *Near the head* (yellow color)

- *Eye*: errors in the element of interest, in the person's right eye;
- *Torso*: errors in the element of interest, in the region on the right side of the person's body;
- *Near the head*: errors in the background, next to the element of interest, on the right side of the person's head.

Fig. 2 shows the regions with simulated errors. The dark blue color pixels are the *Mouth* errors. The light blue color pixels are the *Corner* errors, the green color pixels are the *Eye* errors, the red color pixels are the *Torso* errors, and the yellow color pixels are the *Near the head* errors.

After determining the shape and regions with simulated segmentation errors, the proposed experiment created new videos combining the set of layers that contains the element of interest, the new background, and a single region of the frame (see Fig. 2). Furthermore, the experiment used a new video that combined each type of simulated error with each frame region. In addition, the approach proposed generates a video without segmentation errors to a reference in the subjective experiments. Fig. 3 shows a frame of each new video created.

Note that Fig. 3 shows false positives (pixels belonging to the background classified as belonging to the element of interest) with the same color as the original background and false negatives (pixels belonging to the element of interest classified as background) with the same color as the new background. All generated videos have a version preserving the captured audio and a different variation without the audio, determining the audio status (*On*/*Off*). Twelve videos were obtained (*Videos for Evaluation* block of Fig. 1) for conducting the experiments.

C. Performing Subjective Experiments

The next step in identifying the influence of audio on the perceived quality of segmentation is to apply subjective experiments in which volunteers give their opinions regarding the quality of videos with segmentation errors displayed to them. These subjective experiments were conducted using the Subjective Assessment Method for Video Quality (SAMVIQ) [33] (*Subjective Method* in Fig. 1). The TV industry uses the SAMVIQ to assess the quality of videos in multimedia applications since it is more precise than other methods directed to the same purpose [34], [15].

Organizations such as ITU [29] (International Telecommunications Union) and EBU [35] (European Broadcasting Union) recommend these methods, which suggest how to perform each step of then must and how to configure the physical environment [28]. These recommendations include details followed in this work: the number of volunteers and the distance from these volunteers to the display; the size, type, and intensity of light emitted by the display, which must be adequate for the application assessed; and the color of the image background when the system works on images of reduced size.

During an experiment, a volunteer views a reference video (without segmentation errors), as shown in Fig. 3a. This volunteer tests the other videos with different segmentation errors displayed in the sequence by giving a grade on the scale between 0 (poor) and 100 (same quality as reference). The SAMVIQ method also requires that the volunteer evaluates the reference video shown among the videos with segmentation errors (hidden references).

In the experiments, we did not ask the volunteers about the level of annoyance caused by segmentation errors. We instructed each volunteer to give his/her opinion on the quality of the displayed video. The videos kept all rated characteristics except the segmentation error and the audio status. So, the volunteers tested the annoyance of this combination error/audio status. The volunteers tested videos generated from the same source video combined with the same background. Each volunteer tested the segmentation error (or region where the error occurs) and the audio status because only these characteristics vary.

The volunteers for the experiments are students, faculty, staff, and external volunteers. 60 volunteers participated in four experiment batches at this research step (15 volunteers in each batch), and each volunteer gave their opinions in a single batch.

The experiments show videos with two audio statuses. In the first status, the audio was disabled (*Off*), and in the second status, the audio was enabled (*On*).

Table I shows the configuration of each batch in the experiments. Column 1 shows the audio status, column 2 shows an error display sequence, and column 3 shows the number of volunteers in the batch. Each batch displayed the errors in one of the two sequences: Sequence A (*corner*, *eye*, *torso* and *near the head*) and Sequence B (*near the head*, *torso*, *eye*, *corner* and *mouth*).

TABLE I
CONFIGURATION OF EACH BATCH OF THE SUBJECTIVE EXPERIMENTS

Audio status	Error Sequence	Number of Volunteers
Off	Sequence A	15
On	Sequence A	15
Off	Sequence B	15
On	Sequence B	15

IV. Results and Discussion

This section presents the results of the analysis of the data generated by the subjective experiments. These data are values

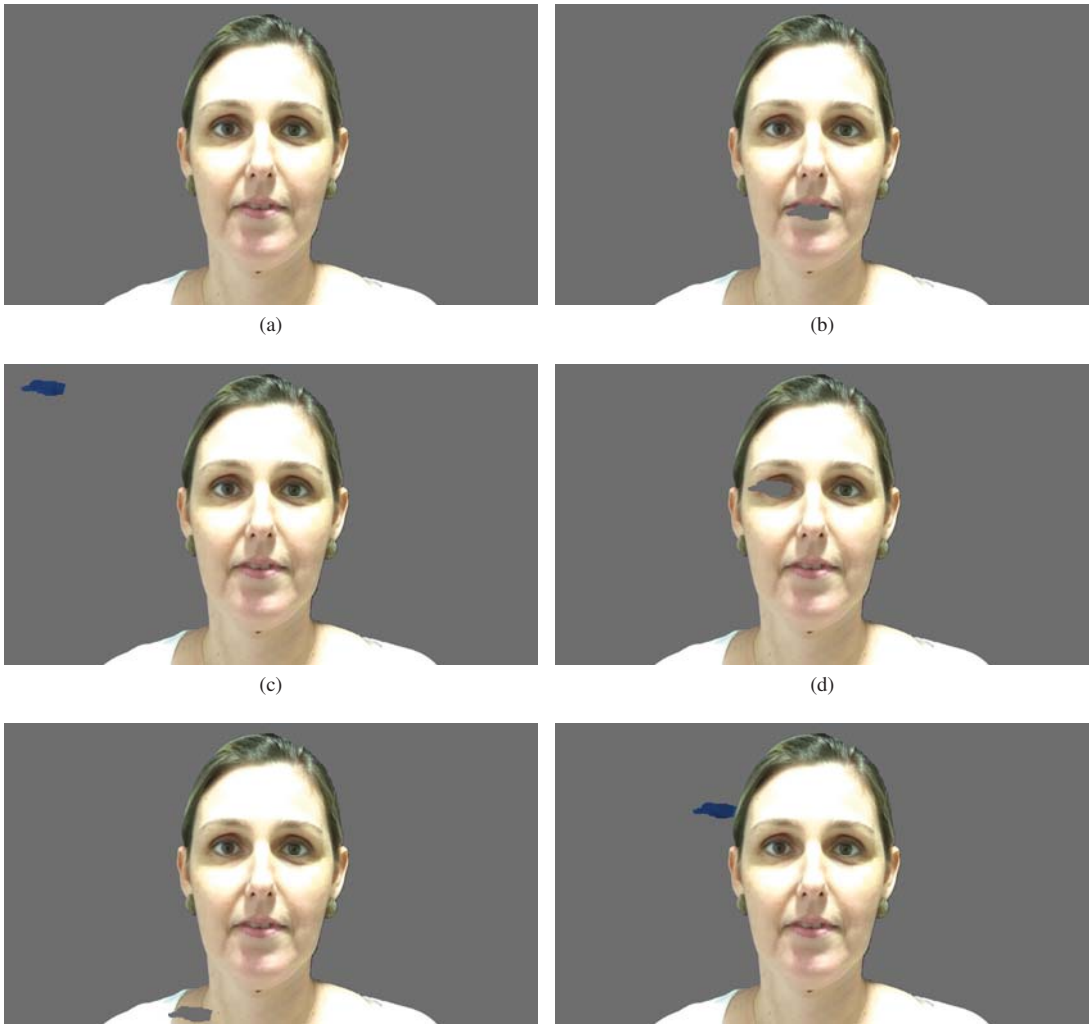

Fig. 3 Examples of video frames used in the experiments that have simulated segmentation errors; note that (a) shows a reference video frame, which does not have segmentation errors

that represent the subjective (perceived) overall quality of the segmentation (Opinion Score (OS)) [12]. This analysis aims to identify if the audio influences the quality of segmentation perceived by users.

Firstly, the Analysis of Variance (ANOVA) was applied to analyze the differences among the OSs from the analysis of the videos shown in Fig. 3. This analysis considered the following factors with their respective levels: type of error (corner, mouth, eye, torso, and near the head), audio status (*Off* and *On*), and error sequence (Sequence A and Sequence B of Table I).

Fig. 4 shows the Q-Q Plot, which tests the appropriateness of the analysis. The chart shows that the residuals have a normal distribution and there are outliers in the *Mouth* and *Eye* errors.

Fig. 5 shows another analysis of residual distribution. Residuals are distributed around zero and do not have a specific shape, which means that the use of the ANOVA was appropriate.

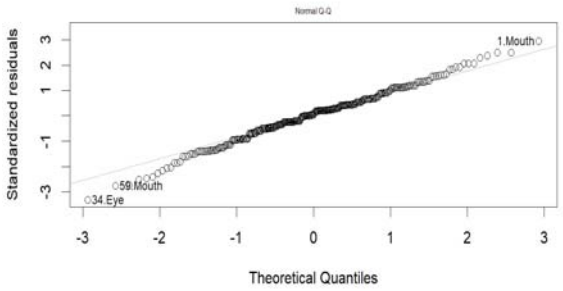

Fig. 4 Q-Q Plot of residuals from the analysis; the residuals are normally distributed

The analysis of the interaction effect between the type of error and audio status using ANOVA resulted in a test statistic with $F = 7.101$ and $p - value = 1.84e - 05$, showing

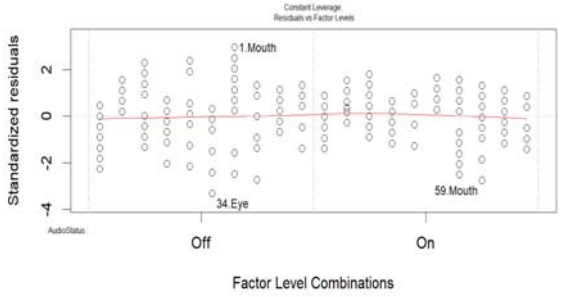

Fig. 5 Analysis of residual distribution, considering the levels of the factor audio status (*On* and *Off*), and showing that residuals are distributed around zero and there are some outliers

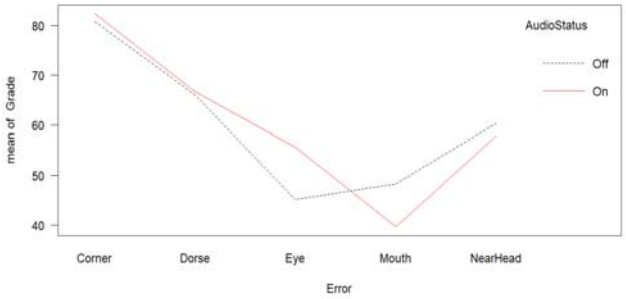

Fig. 6 OS means comparison of the audio status for each type of error; the *Mouth* error is more perceptible for the *On* audio status

a significant difference, considering a confidence level or $\alpha = 0.05$. ANOVA showed a coefficient of determination (R^2) equal to 0.6826, which indicates how much the approach explained the total variability of the data. This total variation of the response variable (segmentation quality OS) is reduced by the factors (type of error, audio status, and error sequence), with the coefficient approximating 1. Thus, the coefficient is a reasonable value for ANOVA, which shows its suitability to describe the perceived quality of segmentation in videos with and without audio and residual analysis.

Observing the statistical results, significant differences between *Off* and *On* levels of the audio status were identified for *Mouth* and *Eye* levels of the factor type of error. Thus, in these configurations, the evidence suggested by the test indicates that audio influenced the quality of segmentation perceived by the volunteers.

The experiments used descriptive statistics to analyze the audio status OS for each type of error and the OS of the errors regarding the audio status. Figs. 6 and 7 show the results.

Comparing audio status (*Off* and *On*), the *Torso* error did not change significantly between the two statuses. The *Corner* and *Eye* errors were less perceived when audio status was *On*; and the *Mouth* and *Near the head* errors were less perceived when audio status was *Off* (Fig. 7).

Mouth error is most noticeable on videos with audio. For this error, 100% of the volunteers gave scores lower than

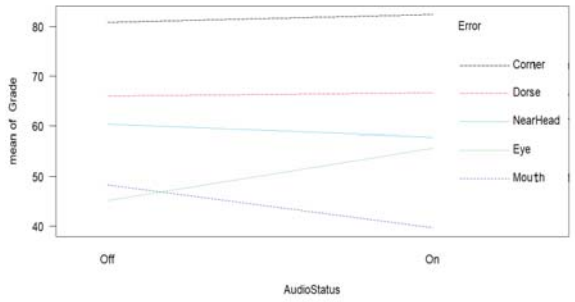

Fig. 7 OS means comparison of the error for each audio status; *Torso* and *Corner* errors are relatively equal

65 (between 0 and 100) to the videos with level *On* – the Mean Opinion Score (MOS) was 39.67. The *Eye* error is most noticeable on videos with audio level *Off*. Here, the MOS was 45 (75% of the volunteers gave scores less than 50). *Mouth* and *Eye* errors were the most noticeable when the analysis considered all videos. These results indicate that users focus on these regions (mouth and eyes) of the face during a conversation.

V. Conclusion

This paper identified that audio influences the user's perception of segmentation quality in background substitution applications, considering specific configurations. For this, the proposed approach applied a subjective method based on formal video quality assessment methods.

Most objective metrics that consider the user's perception apply subjective experiments as one stage of their development. In this experiment, a set of volunteers tests videos with and without segmentation errors to identify the levels of discomfort caused by the different errors.

The subjective experiments that generated the data used to define the state-of-the-art PST and PAD metrics did not contain audio in the videos. They consider only the errors of segmentation in the videos. Thus, the experiments do not simulate the complete application environment once the audio is a common component in the video conference and augmented reality systems (potential applications of the PST and PAD metrics).

The results obtained after applying the proposed subjective method, which uses videos with and without audio, showed that audio influences the quality of segmentation perceived by a user in certain situations. Therefore, this research concluded that the PST and PAD metrics could be more efficient if the subjective experiments used to define them were applied using audio-able videos.

References

[1] N. Austin, R. Hampel, and A. Kukulska-Hulme, "Video conferencing and multimodal expression of voice: Children's conversations using Skype for second language development in a telecollaborative setting," *System*, vol. 64, no. 47, pp. 87–103, 2017.

[2] T. Shimizu and H. Onaga, "Study on acoustic improvements by sound-absorbing panels and acoustical quality assessment of teleconference systems," *Applied Acoustics*, vol. 139, no. November 2017, pp. 101–112, 2018.

[3] Personify Inc, "ChromaCam," 2019, https://www.chromacam.me Accessed 26 Jun 2019.

[4] S. R. R. Sanches, R. Nakamura, V. F. da Silva, R. Tori, V. F. Silva, and R. Tori, "Bilayer Segmentation of Live Video in Uncontrolled Environments for Background Substitution: An Overview and Main Challenges," *IEEE Latin America Transactions*, vol. 10, no. 5, pp. 2138–2149, sep 2012.

[5] A. Parolin, G. P. Fickel, C. R. Jung, T. Malzbender, and R. Samadani, "Bilayer video segmentation for videoconferencing applications," in *Proceedings of the IEEE International Conference on Multimedia and Expo – ICME 2011*. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1–6.

[6] V. D. Bergh and V. Lalioti, "Software chroma keying in an immersive virtual environment," *South African Computer Journal*, vol. 24, pp. 155–162, 11 1999.

[7] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - CVPR '06*, vol. 1. Washington, DC, USA: IEEE Computer Society, Jun 2006, pp. 53–60.

[8] P. Yin, A. Criminisi, J. Winn, and I. Essa, "Bilayer segmentation of webcam videos using tree-based classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 30–42, 2011.

[9] H. Huang, X. Fang, Y. Ye, S. Zhang, and P. L. Rosin, "Practical automatic background substitution for live video," *Computational Visual Media*, vol. 3, no. 3, pp. 273–284, 2017.

[10] S. R. R. Sanches, D. M. Tokunaga, V. F. Silva, and R. Tori, "Subjective video quality assessment in segmentation for augmented reality applications," in *2012 14th Symposium on Virtual and Augmented Reality*, May 2012, pp. 46–55.

[11] E. D. Gelasca, T. Ebrahimi, M. C. Q. Farias, M. Carli, and S. K. Mitra, "Annoyance of spatio-temporal artifacts in segmentation quality assessment [video sequences]," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, vol. 1, Oct 2004, pp. 345–348.

[12] E. Gelasca and T. Ebrahimi, "On evaluating video object segmentation quality: A perceptually driven objective metric," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 319–335, april 2009.

[13] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An Expanded Change Detection Benchmark Dataset," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2014, pp. 393–400.

[14] S. R. R. Sanches, V. F. Silva, R. Nakamura, and R. Tori, "Objective assessment of video segmentation quality for augmented reality," in *Proceedings of IEEE International Conference on Multimedia and Expo – ICME 2013*. Washington, DC, USA: IEEE Computer Society, 2013, pp. 1–6.

[15] S. R. R. Sanches, A. C. Sementille, R. Tori, R. Nakamura, and V. Freire, "PAD: a perceptual application-dependent metric for quality assessment of segmentation algorithms," *Multimedia Tools and Applications*, Aug 2019.

[16] J. J. McDonald, W. A. Teder-SaÈlejaÈrvi, and S. A. Hillyard, "Involuntary orienting to sound improves visual perception," *Nature*, vol. 407, no. 6806, pp. 906–908, 2000.

[17] J. Driver and T. Noesselt, "Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments," *Neuron*, vol. 57, no. 1, pp. 11–23, 2008.

[18] J. Vroomen and B. d. Gelder, "Sound enhances visual perception: cross-modal effects of auditory organization on vision." *Journal of experimental psychology: Human perception and performance*, vol. 26, no. 5, p. 1583, 2000.

[19] P. Dalton and C. Spence, "Attentional capture in serial audiovisual search tasks," *Perception & Psychophysics*, vol. 69, no. 3, pp. 422–438, 2007.

[20] S. R. R. Sanches, C. Oliveira, A. C. Sementille, and V. Freire, "Challenging situations for background subtraction algorithms," *Applied Intelligence*, vol. 49, no. 5, pp. 1771–1784, May 2019.

[21] Université de Sherbrooke, "ChangeDetection.NET – a video database for testing change detection algorithms," 2019, http://www.changedetection.net. Accessed 20 Jun 2019.

[22] N. Goyette, P. M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 1–8.

[23] S. R. R. Sanches, V. F. Silva, R. Nakamura, and R. Tori, "Objective assessment of video segmentation quality for augmented reality," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*, July 2013, pp. 1–6.

[24] S. R. R. Sanches, V. F. da Silva, and R. Tori, "Bilayer segmentation augmented with future evidence," in *Proceedings of the 12th International Conference on Computational Science and Its Applications - Volume Part II*, ser. ICCSA'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 699–711.

[25] P. L. Rosin and E. Ioannidis, "Evaluation of global image thresholding for change detection," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2345–2356, 2003.

[26] Q. Li, Y.-M. Fang, and J.-T. Xu, "A novel spatial pooling strategy for image quality assessment," *Journal of Computer Science and Technology*, vol. 31, no. 2, pp. 225–234, Mar 2016.

[27] X. Huang, J. Søgaard, and S. Forchhammer, "No-reference pixel based video quality assessment for hevc decoded video," *Journal of Visual Communication and Image Representation*, vol. 43, no. C, pp. 173–184, 2017.

[28] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Geneva, Switzerland, 2009, https://www.itu.int/rec/r-rec-bt.500. Accessed 1 March 2019.

[29] ITU-R, "International telecommunications union – committed to connecting the world," 2019, http://www.itu.int. Accessed 21 February 2019.

[30] J. G. Beerends and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, 1999.

[31] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman, "Experienced quality factors: qualitative evaluation approach to audiovisual quality," *Multimedia on Mobile Devices 2007*, vol. 6507, no. February, p. 65070M, 2007.

[32] J. Wang and M. F. Cohen, "Image and video matting: A survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 2, pp. 97–175, Jan. 2007.

[33] F. Kozamernik, V. Steinmann, P. Sunna, and E. Wyckens, "Samviq – a new ebu methodology for video quality evaluations in multimedia," *SMPTE Motion Imaging Journal*, vol. 114, no. 4, pp. 152–160, april 2005.

[34] S. Péchard, R. Pépion, and P. L. Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *Proceedings of the International Workshop on Image Media Quality and its Applications – IMQA2008*, 2008, pp. 1–6.

[35] European Broadcasting Union, "EBU – european broadcasting union," 2019, http://www.ebu.ch. Accessed 22 February 2019.