IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 June 2023, accepted 22 August 2023, date of publication 1 September 2023, date of current version 11 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311256

RESEARCH ARTICLE

Automatic Camera Selection, Shot Size, and Video Editing in Theater Multi-Camera Recordings

ECKHARD STOLL[1,2], STEPHAN BREIDE[2], STEVE GÖRING[1], AND ALEXANDER RAAKE[1], (Member, IEEE)

[1]Audiovisual Technology Group, Technische Universität Ilmenau, 98693 Ilmenau, Germany
[2]Audio Visual Media Center, South Westphalia University of Applied Sciences, 59872 Meschede, Germany

Corresponding author: Eckhard Stoll (eckhard.stoll@tu-ilmenau.de)

This work was supported in part by the Open Access Publication Fund of the Technische Universität Ilmenau, and in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant DFG-437543412.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics committee of Technische Universität Ilmenau for the Project SoPhoAppeal under Application No. DFG-437543412, and the I3TC Center.

ABSTRACT In a non-professional environment, multi-camera recordings of theater performances or other stage shows are difficult to realize, because amateurs are usually untrained in camera work and in using a vision mixing desk that mixes multiple cameras. This can be remedied by a production process with high-resolution cameras where recordings of image sections from long shots or medium-long shots are manually or automatically cropped in post-production. For this purpose, Gandhi et al. presented a single-camera system (referred to as Gandhi Recording System in the paper) that obtains close-ups from a high-resolution recording from the central perspective. The proposed system in this paper referred to as ''Proposed Recording System'' extends the method to four perspectives based on a Reference Recording System from professional TV theater recordings from the Ohnsorg Theater. Rules for camera selection, image cropping, and montage are derived from the Reference Recording System in this paper. For this purpose, body and pose recognition software is used and the stage action is reconstructed from the recordings into the stage set. Speakers are recognized by detecting lip movements and speaker changes are identified using audio diarization software. The Proposed Recording System proposed in this paper is practically instantiated on a school theater recording made by laymen using four 4K cameras. An automatic editing script is generated that outputs a montage of a scene. The principles can also be adapted for other recording situations with an audience, such as lectures, interviews, discussions, talk shows, gala events, award ceremonies, and the like. More than 70 % of test persons confirm in an online study the added value of the perspective diversity of four cameras of the Proposed Recording System versus the single-camera method of Gandhi et al.

INDEX TERMS Multi-camera theater recordings, cropping, automatic montage, 4K, automatic video editing.

I. INTRODUCTION

Professional recordings of live events are usually implemented with multiple cameras by qualified personnel. Camera work, directing, and a montage of theater recordings are artistic crafts that require training, experience, and skill. Therefore, it is difficult for amateurs to achieve acceptable results here. Without training in theory and practice, amateurs often do not know the design rules for aesthetically pleasing images and usually cannot follow the movements of the performers quickly and competently.

Figure 1 shows three examples of an amateur recording [1]. Here, in (a) the heads are cut off, in (b) the distances to the edge of the picture are not kept, or in (c) the actors are placed

The associate editor coordinating the review of this manuscript and approving it for publication was Francisco J. Garcia-Penalvo.

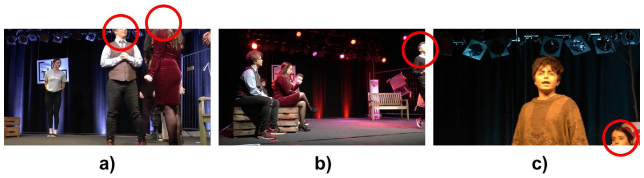

FIGURE 1. Examples of an amateur recording [1]: (a) heads are cut off, (b) wrong edge of the picture, (c) too low placement.

FIGURE 2. Cropping of image sections from (a) medium long shots: (b) medium close shots (orange) and (c) close shots (green) [2]. The shot above is from a camera standing at stage right. The shot below is from a camera standing at stage left.

too low in the picture. This can be remedied by a production process in which medium shots and close-ups are obtained subsequently in post-production instead of during shooting. To this aim, the recording is done with high-resolution cameras (4K, 6K, 8K,...), which are fixed or only slightly panned and zoomed. Only long shots or medium long shots are recorded, which capture the entire action. This can also easily be done by laymen on the cameras. Figure 2 a) shows images [2] from 4K/UHD cameras with a resolution of 3840 × 2160 pixels (blue). From these, cropped b) medium close (orange) and c) close shots (green) can be extracted. The purpose of this paper is to exemplify how multi-camera recordings of theater performances or other stage performances can be improved in a non-professional environment using a semi-automatic algorithmic approach. A production method is developed using high-resolution cameras from which image sections are automatically cropped from long shots or medium long shots.

A similar approach has already been used by Gandhi [3]. However, only one camera is used here, centered to record the whole stage. The cropped close-ups thus also have the perspective from the center. Figure 3 b) shows this arrangement, which is called the Gandhi Recording System (GRS) in this paper. Professional recordings use multiple cameras to increase perspective variety. Cameras placed at the front left and right of the stage allow the actors' faces to be shown more clearly during dialogues. Side cameras each show the full face and both eyes of the actors, as Figure 2 illustrates. The right camera shows the woman and the left camera the men.

The recording system presented in this paper referred to as "Proposed Recording System (PRS)" in the following is shown in Figure 3 c) and extends the Gandhi Recording System (GRS) to four perspectives. Figure 3 d) and e) show two variants of professional recording systems, which will be explained in more detail later. Variant e) represents the approach with the highest number of Degrees of Freedom, using four cameras that all can zoom, with three cameras on-stage that can be moved to follow actors. As this approach is infeasible for high-quality yet amateur recordings, it is not considered as reference for the PRS. Instead, variant d), which uses fixed cameras, is defined as the Reference Recording System (RRS). Based on professional recordings made with such an RRS approach, rules for camera selection, image cropping, and montage are derived, which are then used in the Proposed Recording System (PRS). Furthermore, Figure 3 a) shows a Simple Recording System (SRS) with only one camera without cropping, which will also be described in more detail later. In an online study, also presented in this paper, the three amateur systems a), b), and c) (SRS, GRS, and PRS) are evaluated by the test persons. It is shown that the participants prefer the PRS compared to the other systems.

The paper is structured as follows: Section II gives an overview of the state of the art. Afterward, Section III presents the setup of the Proposed Recording System, that generates image crops and cut sequences using recordings from four high-resolution cameras. This is done according to rules obtained from the analysis of professional theatrical recordings. Section IV deals with this analysis of professional recordings. It is determined which broadcast series is suitable as a Reference Recording System. The open-source analysis tool OpenPose is used as a multi-person recognition system for locating people and determining their movements. Section V shows how a script breakdown can be automatically created with image analysis by reconstructing the shot sizes based on the overall stage set. In Section VI the selected method for speaker detection using audio analysis is described. Scene analysis and flow chart development is presented in Section VII. Section VIII shows an example of how the results of the scene analysis can be implemented in the Proposed Recording System. For a multi-camera recording with non-professionals, an example is given of how a process for automatic editing can be designed and how it leads to an automatic editing script. Section IX reports on the online study.

II. RELATED WORK

The published paper "Modelling of an automatic vision mixer with human characteristics for multi-camera theater recordings" [4] investigates a subtask for an automated editing system and is thematically close to the present paper. Therefore, relevant related works are listed in both papers. In the following, a brief overview of commonly established approaches for automatic shot selection and image cropping is summarized and evaluated. Fully automatic techniques for editing videos are being developed for various fields. In Gandhi and Ronfard the focus is on the automatic detection

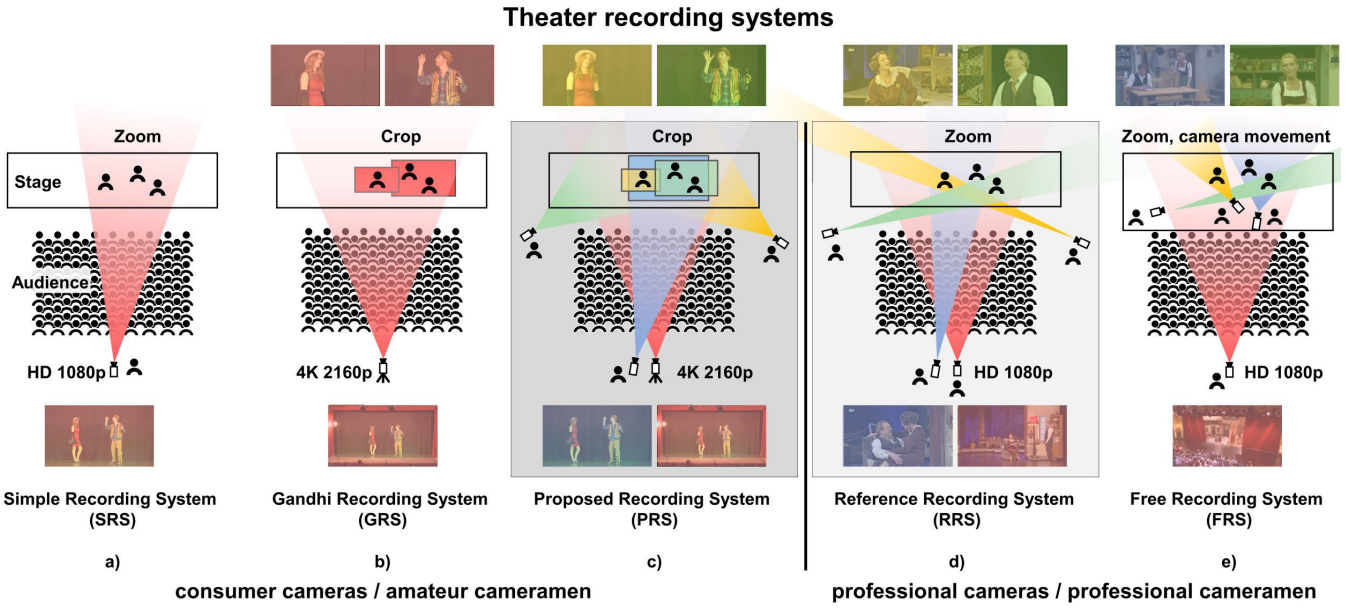

FIGURE 3. Theater recording systems. Amateur recording systems on the left side: a) Simple (without cuts), b) Gandhi cropping close-ups ([3]), c) Proposed Recording System with four cameras and cropping. Professional recording systems on the right side: d) Reference Recording System with four fixed-position cameras, e) Moving cameras on stage.

and naming of actors on a stage [5]. In that work, a method is developed to distinguish the external appearance of clothes, which is implemented based on color differentiation. Furthermore, Gandhi [3] apply person recognition to theater recordings for automatized generation of image crops. Only one camera is used and the different shots are cropped out but not edited together. The system is then further developed in [6] to improve the tracking of actors in motion. In [7], the method is applied to 4K footage of dancers, and multiple shots are output simultaneously on a split screen. Improvements, such as the use of a two-stage method (detection of timestamps for image cuts and optimization of crops for pans and zooms), were made by Rachavarapu et al. [8]. In this work, the eye movement of 5 viewers watching a wide-screen video is captured with an eye tracker. The video is to be cropped to a smaller aspect ratio and is optimized in x-position and zoom. The y-position is not changed and this restricts the algorithm so that faces or bodies may be cut off by the image boundary. Chen et al. [9] investigate the computational complexity of optimal rectangle search in attention-based automatic image cropping. Li and Zhang [10] generate image cropping using Collaborative Deep Reinforcement Learning trained by eye-tracking. Cropping is used to enhance the quality of experience (QoE) of 4K videos when played back on small screen devices such as smartphones [11]. Here the regions in the image that are frequently viewed are cropped and displayed in full format [12], [13].

Escobar and Parikesit perform an analysis of a theater video recording of a puppet show [14] and using difference frames of each two consecutive frames, the intensity of movement of the puppets is measured and assigned to narrative scene segments. Leake et al. are concerned with automated video editing for dialogues [15]. Using software such as OpenPose [16], [17], [18], [19] and OpenFace [20], face recognition and tracking are applied and the video clips are matched to the textual dialogues in the script. Cuts are performed only when speakers change. Moreover, automatic segmentation of videos is performed for tutorial videos [21], for example, and a semi-automatic video editing system is being developed to support the production of concise tutorials [22]. Automatic camera control of a single camera is used especially in amateur sports because production with many cameras and a camera crew is expensive. Quiroga et al. [23] film a basketball court with a fixed 4K camera and obtain a lower resolution automated virtual camera to follow the game action. A similar method has been developed for ice hockey [24]. Soccer fields have a very large width. Different approaches are used to obtain a high-resolution 180Â° image as a basis [25], [26], [27] and specialized tracking algorithms track the game action [28], [29]. These methods are also used in other sports such as table tennis [30], tennis [31], or field hockey [32].

III. OVERVIEW OF THE PROPOSED RECORDING SYSTEM (PRS)

In the following section, an overview of the Proposed Recording System (PRS) is provided, it consists of several steps, which are described in detail in the following.

Figure 4 shows the schematic setup of the PRS, a) in use for recording a theater performance with four 4K cameras positioned left and right of the stage (green and yellow) and behind the audience (blue and red). In the next step, b), the camera signals pass through a plot analysis, that detects plot elements and logs them in a plot script, such as appearances and departures of protagonists, the number of persons, positions, and movements, face recognition,

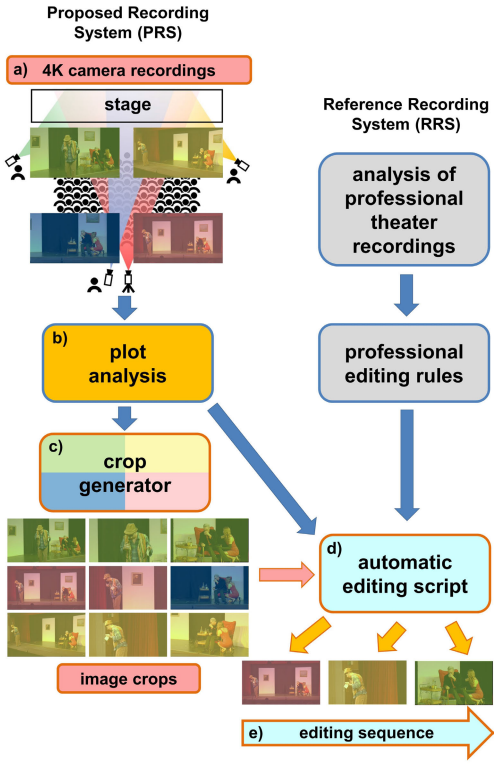

FIGURE 4. Schematic structure of the Proposed Recording System (PRS).

recognition of speaking persons, relationships of persons to each other (e.g., dialogues), etc. In the subsequent step, c), the plot script is used to search for and generate image crops from the 4K recordings, and in the last step, d), an automatic editing script is generated from which an editing sequence can be created. For this purpose, professional editing rules are applied, which are created from the analysis of professional theater recordings of the Reference Recording System (RRS). The audio-video software tools used for the plot analysis are presented below, and Figure 28 then shows the complete process.

In the previous work by the authors [33], an online study is presented to evaluate the quality and preference for three versions of the same scene with differently cropped image sections. Now in this paper, an analysis method for professional recordings is developed, rules for camera selection, image cropping, and montage are extracted and an exemplary application of automation is shown. Here, the automatic editing script in Figure 4 d) can be given an individual, human-like behavior. Analyses of professionally working vision mixers are performed in [4] and result in a model. The developed algorithm can be applied to the automatic editing script.

IV. THEATER VIDEO RECORDINGS ANALYSIS OF THE REFERENCE RECORDING SYSTEM (RRS)

A. SELECTION OF A TV THEATER SERIES

For the analysis of professional recordings, a TV broadcast series is used, where the procedures are similar for each production, such as stage size, number of cameras, camera

FIGURE 5. Tracking shots using cameras on stage for Komödienstadel (top row) [37] and Chiemgauer Volkstheater (bottom row) [38].

positions, etc. German television regularly broadcasts series of shows by three different specific theater ensembles. The Komödienstadel [34] and the Chiemgauer Volkstheater [35] are recordings by the Bayerischer Rundfunk (BR).[1] The Ohnsorg Theater in Hamburg [36] is recorded by the Norddeutscher Rundfunk (NDR).[2] In the Komödienstadel and Chiemgauer Volkstheater, however, cameras are on stage and moved around during the production, as Figure 5 shows, corresponding to recording system e) in Figure 3.

The local audience in the theater has to live with the possibly disturbing cameras on stage because the focus here is on the TV recording. In Figure 3 e) this recording system is shown with the designation Free Recording System (FRS). Since this practice is not similar to the recording process in non-professional recordings, where the cameras are not allowed to disturb the audience much or at all, Komödienstadel and Chiemgauer Volkstheater are not included in the analysis, and only recordings from the Ohnsorg Theater in Hamburg are examined and the underlying set-up is used as Referece Recording System (RRS). In Figure 3 the RRS is shown in d). Figure 6 shows the Ohnsorg Theater in Hamburg. Here two fixed cameras are used at the front left (green, camera 1) and right (yellow, camera 2) of the stage, as well as two fixed cameras (blue, camera 3 and red, camera 4) in the audience. Such a setup is also used for the Proposed Recording System (PRS) in Figure 3 c) presented in this paper. In order to even less disturb the audience, cameras 3 and 4 are then usually placed behind the audience.

B. VALUE ASSIGNMENT OF SHOT SIZES

The real-time multi-person recognition system OpenPose [16], [17], [18], [19] is an open-source software that recognizes the human body, hand, facial, and body points (135 key points in total) on single images. For example, OpenPose is used for attitude shot size analysis [39]. The OpenPose library is based on a neural network for human pose recognition and has been trained with 25,000 images of over 40,000 people with annotated body joints [40]. OpenPose analyzes individual images or, in the case of videos, image by image sequentially and outputs data sets (x,y values) for the recognized persons for each image. For the body,

[1] https://www.br.de
[2] https://www.ndr.de

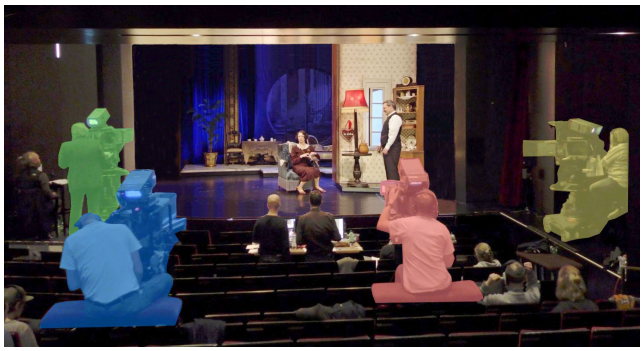

FIGURE 6. Camera arrangement in the Ohnsorg Theater. Two cameras (green and yellow) are at the front left and right of the stage and two cameras (blue and red) are in the auditorium.

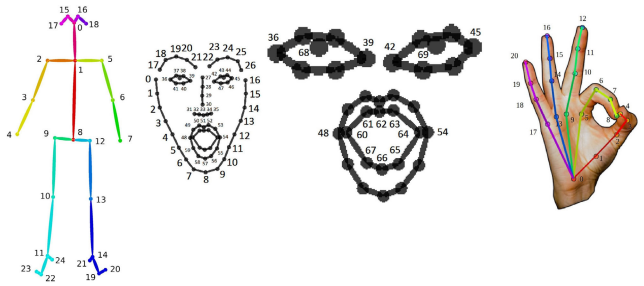

FIGURE 7. OpenPose key points (illustration courtesy of [16]).

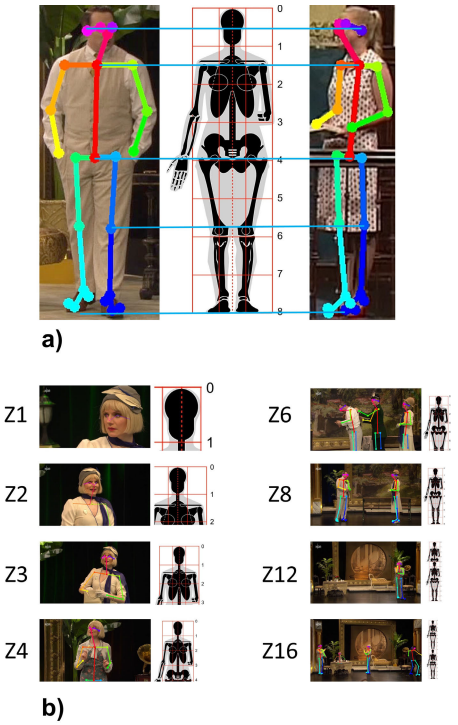

FIGURE 8. OpenPose key points in the 8 zones model and mathematical implementation of shot sizes (using an GNU Free Documentation License image from [41]).

25 body points are specified: Nose, neck, right shoulder, right elbow, etc. If the flag `--face` is set during image analysis, 70 points are assigned to recognized faces. By setting the flag `--hand`, 21 key points are specified per hand. For the analyses in this paper the version OpenPose v1.7.0 is used.

Figure 7 from [20] shows the assignment of key points in the datasets.

Furthermore, Figure 8 shows how the key points are detected and colored joints are drawn. For each frame, a data set with the x-y coordinates of all key points of all persons is estimated. In movie language, shot sizes are referred to as long shot, medium long shot, medium close-up, close-up, and so on. These are discrete divisions. Koga-Browes [42] gives an overview of how the shot sizes from different literature sources (books) can be assigned to a division of the human body into 8 zones. It can be seen that this division shows differences from author to author. For example, one author indicates that a close-up shows the head up to the middle of the chest, while another author shows the close-up up to below the chest. Also, for example, a slow zoom from a long shot to a close-up cannot be assigned an exact indication of shot size over time using the discrete labels. For analysis and automation purposes, it is therefore expedient to specify the shot sizes in a mathematically continuous manner. A division of the body proportions into 8 zones is made according to Bernhard [41]. How this division corresponds to the key points is shown in Figure 8 in a). The crown of the head is assigned the value 0 and the lower foot point the value 8. The hips of persons are then located at about 4. Neck points have about the value 1.5 and the ear points are located at about 2/5 between crown 0 and neck point 1.5 and thus at about 0.6. The knee points are located at about 3/4 between 5 and 6 and thus at about 5.75.

Since different people have different physiques and assume different postures, the assigned values can only serve as guide values for determining the shot sizes. Figure 8 shows in b) how shot sizes can be assigned. For identification purposes, the letter Z (zone) is used as the unit. Z1 (one zone) means the head of the person is visible. Z2 (two zones) corresponds to a close-up and the person can be seen up to the chest. Z4 corresponds to the medium close up to the hip. Z6 corresponds to the medium shot up to the knee, and Z8 is a medium long shot showing people from head to toe. Z16 corresponds to a long shot showing two people stacked on top of each other. Continuous intermediate values can also be determined, such as Z3.48 or Z11.2.

C. TRACKING

Since OpenPose's recognition is based on individual images, the information from the previous image is not included in videos. As a result, the order of the recognized persons varies from frame to frame. Figure 9 shows an analysis of a video sequence over 400 frames (16 seconds) from "Dream dancer", which was recorded with a fixed 4K camera during the main rehearsal. Eight people from a training course are sitting on chairs and talking. Shown in the graph are the

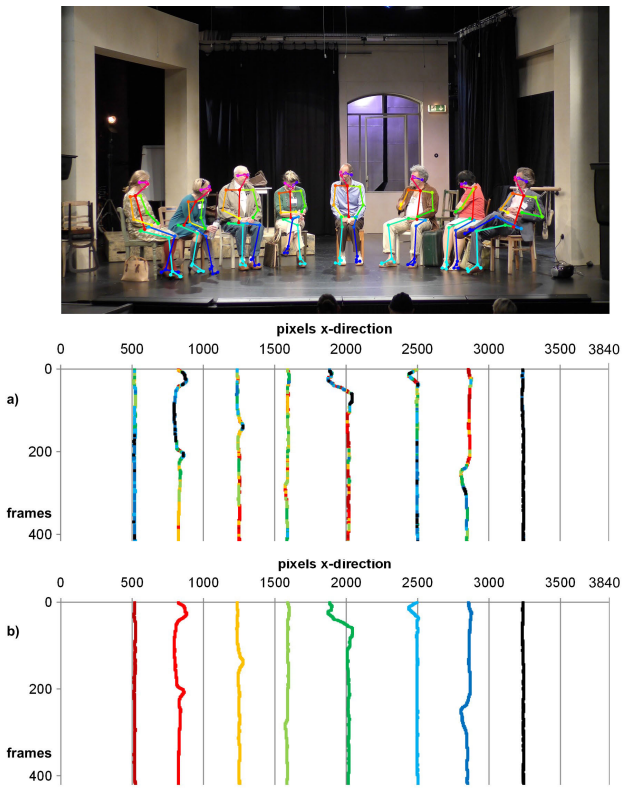

FIGURE 9. Unsorted data mapping of neckpoints a) and tracking b) over time (vertical axis).

FIGURE 10. Occlusion: If one person is in front of another, only one person is detected.

FIGURE 11. Face extraction from the final applause in the play "A better gentleman", Ohnsorg Theater 2019 [43].

x-values of the neck points on the horizontal axis and the temporal sequence in frames on the vertical axis. Each person record that OpenPose recognizes in an image is assigned a color in the sequence of recognition. It can be clearly seen in a) that the sequence varies from frame to frame. For data evaluation, the data must be sorted so that tracking of the respective persons is possible and an assignment can be made, as in Figure 9 b).

D. OCCLUSION

Occlusions or partial occlusions must be considered separately, as shown for example in Figure 10.

At the beginning of the sequence (frame 1), two persons are on stage. They are completely detected by OpenPose. A woman joins them and is detected as the third person (frame 82). The woman then occludes the man in the center (frame 294) and only two persons are detected. The woman walks forward and to the side and the man in the middle is again detected as the third person (frame 501). Due to the coincidence of two tracking paths, a person assignment is no longer possible after the occlusion. If both persons are visible again, face recognition can be used for identification.

E. FACE RECOGNITION

Usually, in a typical Ohnsorg Theater play recording, during the final applause of the play, all actors are first seen in a long shot, as Figure 11 above left shows. A camera then pans over all actors in a close-up. This is shown in the other images in Figure 11. Such a sequence can easily be recognized with OpenPose and the faces of the individual actors can be extracted. For face recognition, the Betaface API version 2.0 from betaface.com [44] is used. A variety of information can be extracted with Betaface, such as basic face recognition (identification, verification or 1:1, 1:N matching), biometric measurements, face analysis, tracking of faces and facial features in videos, detection of age, gender, ethnicity, and emotions, detection of skin, hair, and clothing colors, analysis of the shape of hairstyles and description of the shape of facial features.

Figure 12 shows the application of 1:N matching to the play "A better gentleman". The left column "Face" shows faces from the play. These are compared with the 10 faces from the final applause, a matching probability is determined and listed sorted by this under "Matches". If both eyes of a face are visible for a given "Face" (a to g), the face is assigned to the correct person with at least 80 % match probability. The distance to the next most likely person is 10 % or more. In this case, the correct person is recognized.

If a face is turned to the side so that only one eye is visible as shown in Figure 12 h), face recognition no longer works reliably. Since all persons are tracked, face recognition can be done at a later time when the person is looking forward. With OpenPose the face rotation can be detected.

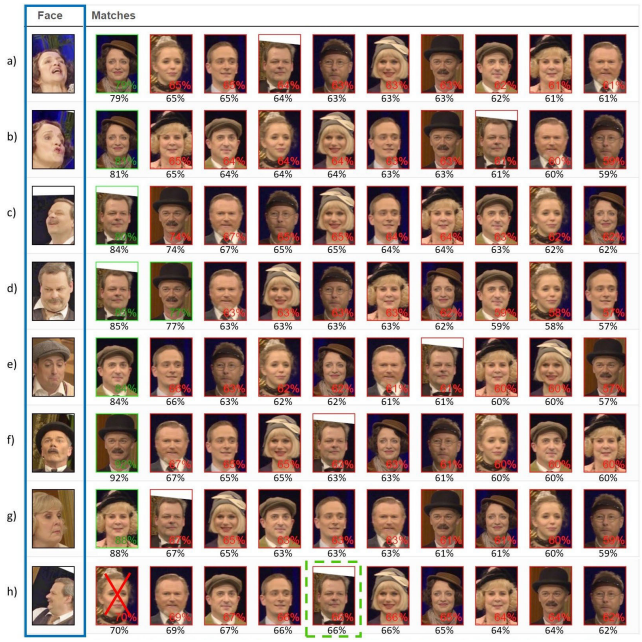

FIGURE 12. Face recognition with betaface.com. If both eyes are visible, the correct person is recognized with a match probability of at least 80 %.

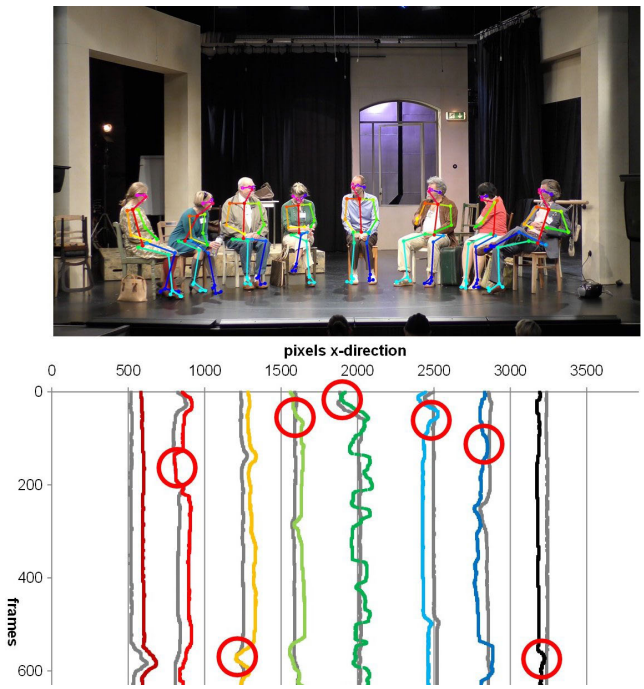

FIGURE 13. Detection of nose points and neck points as a measure of facial rotation. Red circles mark places where faces look forward.

Figure 13 shows the scene of Figure 9, where the x-values of the nose points (colored) and the x-values of the respective neck points (gray) are plotted. If the x-value of the nose point is to the right or left of the x-value of the neck point, the person looks to the right or left. If the x-values are close to each other, the person looks forward. The red circles mark

FIGURE 14. Examples of different recognition of head alignment according to the model of Kobayashi et al. [45].

positions where the respective tracked persons look forward for the first time in this sequence. These locations are suitable for facial recognition and establishing the identity of the persons. Tracking keeps the respective identity until occlusions or partial occlusions take place or a new person enters the stage. Then the identities of the persons involved have to be established anew.

F. HEAD AND BODY ALIGNMENT

A model that uses the body points of the head and body to determine the alignment of the person in the image is proposed in [45] and [46]. From the OpenPose data, an angle between 0° and 360° is specified separately, indicating a head in different angular positions (in terms of yaw angle, with 0Â° being the frontal position). Here an example is shown in Figure 14: 0° to the front, 90° to the right, 180° to the back, 270° to the left. Both eyes can be seen well in a range from 0° to 45° and 315° to 360° (≙ 0°). These ranges are suitable for face recognition with Betaface. Alternative tools for determining head positions include Face^{++} AI Open Platform [47] and Jeeliz Face Filters [48].

V. PLOT ANALYSIS OF THE REFERENCE RECORDING SYSTEM (RRS)

To identify professional editing rules (see Figure 4), the Reference Recording System RRS is analyzed (cf. Figure 3 d)). In order to determine the selection of the respective camera and shot size of professional theater recordings, it is necessary to reconstruct the entire stage action from the shots of the TV recording and thus evaluate the rules, according to which the script breakdown was made.

A. SCRIPT BREAKDOWN

Script breakdown in film terminology refers to deciding which camera distances, camera angles, shot sizes, and camera movements to use for a scene. Usually, a storyboard is created, consisting of drawn images, to visualize the script breakdown. ‘‘Script breakdown of a movie is a creative process subject to the cinematographer’s or director’s own taste and individual style. . . For camera positions, shot sizes, lens choices, and camera movements are always also chosen points of view and influence the effect of a film image in many ways.’’ [49]

Camera locations in film can be chosen at will. In theater recordings, however, the cameras are often at fixed locations (see Figure 6). Camera movement and tracking shots are not possible. The work of script breakdown, therefore, involves determining the best suitable camera (camera location) and shot size (focal length) in order to depict the individual

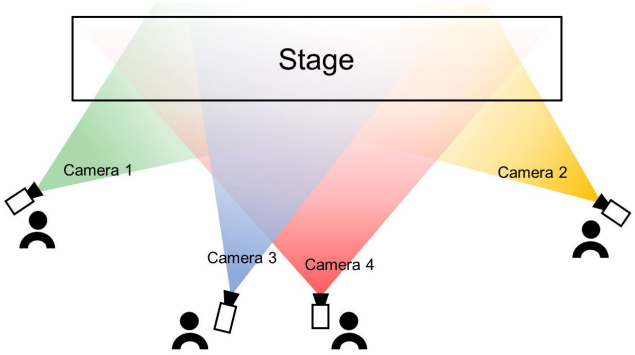

FIGURE 15. Camera positions in the Ohnsorg Theater.

actions and to accentuate the overall stage action accordingly, or to visually emphasize protagonists or groups.

B. RECONSTRUCTION OF SHOT SIZES INTO THE STAGE SET

Parameters for the process of script breakdown can be obtained by determining the individual shot sizes in relation to the action and the stage set. How many people are on stage? Where are the characters standing? How do they move? Who is speaking? Which camera is selected in the Reference Recording System? What shot size is used? In the case of existing TV recordings, however, only the broadcast version is available. A continuous long shot to analyze the spatial assignment of the characters is not available. Therefore, a reconstruction into an empty stage set is proposed and developed in this paper.

Figure 15 shows the camera positions in an Ohnsorg Theater recording with the numbering of cameras 1 to 4. Each camera is assigned to a color: camera 1 (green), camera 2 (yellow), camera 3 (blue), and camera 4 (red).

Similarly, Figure 16 shows the first 15 shots (1 to 15 above) from a scene from "A better gentleman" [43] and the reconstruction (1 to 15 below) into the long shot of the stage set, which is shown using bright highlighting. A long shot of the entire stage set can be obtained, for example, from the opening or closing long shots of a scene. The shots of the scene are fitted into the stage set according to their position and assigned to the corresponding cameras by color. The advantage of such a representation is that it reproduces the spatial positions of the characters and illustrates them well. The representation is two-dimensional, as if from a view from a back row of the audience.

C. PERSPECTIVE SHIFTS OF THE BACKGROUND

Due to the different camera angles, perspective shifts of the background occur during the fit. The stage image is obtained from camera 4 from a closing long shot. Therefore, all other shots of camera 4 like S1, S6, S13, S15, etc. (red in Figure 16) fit almost exactly into the stage set as shown in Figure 17, bottom right (camera 4). The green circles show how the lamp and cabinet fit almost exactly into the background.

FIGURE 16. The first 15 shots of a scene from "A better gentleman" [43] (top) and the reconstruction into the stage set (bottom).

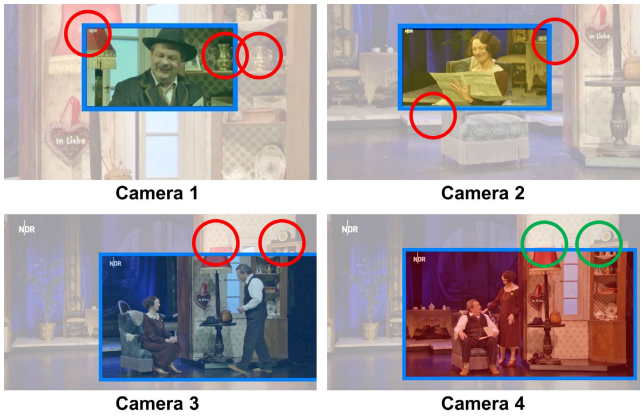

FIGURE 17. Perspective shifts (red circles) and perfect fit of camera 4 (green circles).

The images from the other three cameras, however, show perspective shifts (parallaxes). In the case of cameras 1 and 2, shown here in enlarged form in Figure 17, these deviations in the background are clearly visible because the cameras are located at the front left and right of the stage and record the stage action at a different camera angle. The man recorded with camera 1 is standing in the same position as before in a long shot, but the lamp and the shelf are shifted in the frame due to the different camera angles, as shown by the red circles. The woman recorded with camera 2 is also well-fitted to her position on the armchair. However, the background deviates greatly. For camera 3, which is close to camera 4, the parallax is small, as shown by the shifts of the lamp and cabinet in the red circles in Figure 17 bottom left. However, the shifts of the background are not relevant for the analysis, since person positions and person movements are the focus

FIGURE 18. Person removal, shots from camera 4.

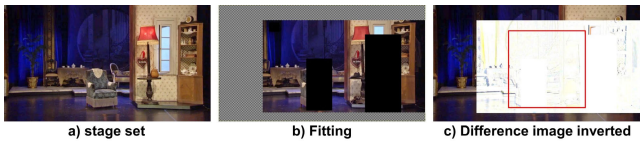

FIGURE 19. Fitting into the stage set by adjusting the inverted differential image.

of the investigation. The fits are made so that the size of the persons (head size, shoulder area, etc.) and the neck positions fit as well as possible from shot to shot.

D. MATCHING BACKGROUND OF CAMERA 4

The shots of camera 4 such as S1, S6, S13, S15 etc. (red in Figure 16) are used as keyframes since they can be fitted exactly into the stage set. Fine adjustment is done by matching the background.

For this purpose, the persons are covered with a black rectangle in the shots of camera 4, as shown in example E6 in Figure 18 a), as can be seen in b). The persons and their skeletal coordinates are obtained with OpenPose and the rectangular coordinates are determined from them.

Now, the image is fitted into the stage set, which can be seen in Figure 19 a). If the position and image size are optimally adjusted, as shown in b), the gray values of the individual pixels almost erase each other in a difference image. In c) the difference image is shown inverted for better illustration. The more exact the fit, the brighter the inverted difference image.

The representation with the inverted image is chosen here for better illustration. The fitting can also be done with common methods of motion estimation. For example, the routines from MathWorks implemented in MATLAB and Simulink can be used for this.

The fitted image can be finely adjusted in three degrees of freedom x, y and size. Figure 20 shows the fine adjustment in x-direction, y-direction and size. It can be seen how the inverted difference image change when the fit is shifted from the optimum in the x or y direction or the image size are too small or too large relative to the optimum. The curve plots show the relative white values of the inverted difference images. The average of the gray values is formed from all pixels (mean value) and set in relation to a white image (100 % white).

As the measurement curves in Figure 20 show, only a value around 98 % is achieved at the optimum. A hundred

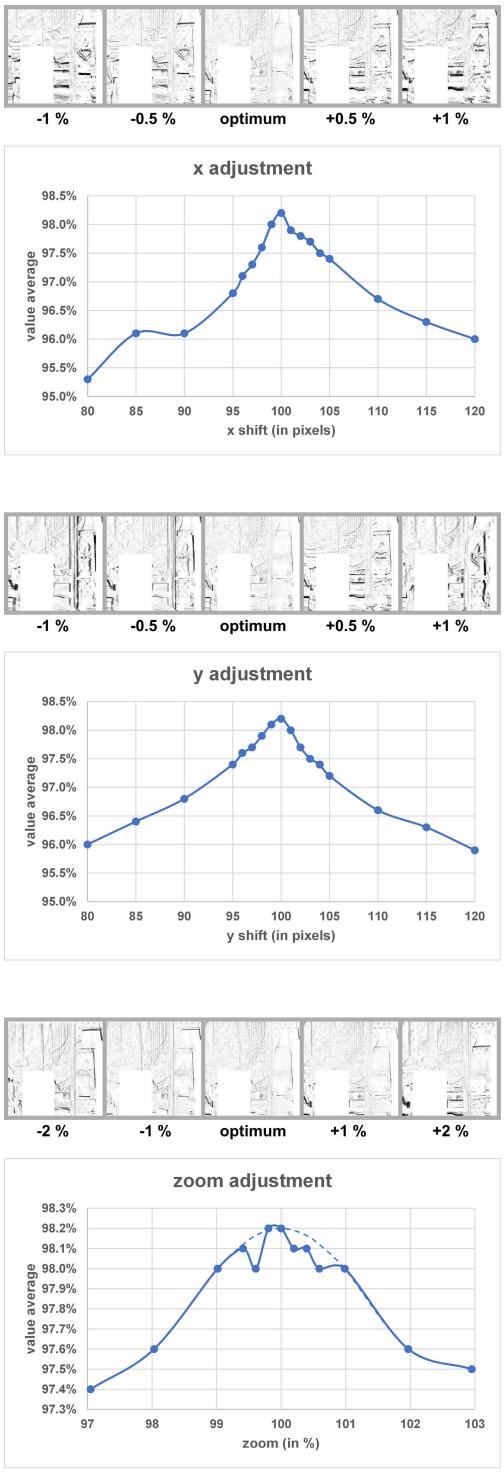

FIGURE 20. Fine adjustment in x-direction, y-direction and size.

percent erasure of the two images does not take place for various reasons. On the one hand, an optical lens system like that of camera 4 shows geometric aberrations. These have a different effect on a long shot than on a zoomed-in shot such as E6. Particularly at the edges of the image, barrel distortion

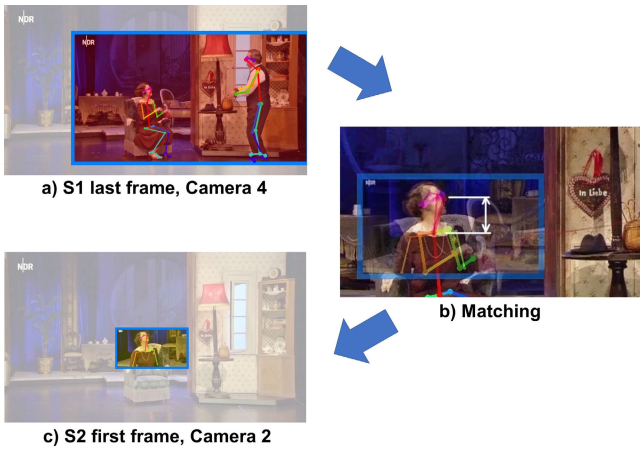

FIGURE 21. Matching with OpenPose of last frame S1 with first frame S2.

or pincushion distortion can occur [50], [51]. Furthermore, camera 4 can also pan left or right from the center when zoomed in. In doing so, the image plane twists with respect to the image plane that the long shot has. Further influences are the digitization process, the grid structure of the image sensor, interpolation in the scanning process, and interference ín images with fine details (moiré effects). The influences can be reduced during fine adjustment by not including the edges of the images during the matching process, but only a rectangle from the respective center of the image. In Figure 19 c), the rectangle is drawn in red and the calculations in Figure 20 were performed only within this rectangle.

E. MATCHING BODY POINTS

Figure 21 b) shows enlarged semitransparent superimposed matching of the last frame from shot S1 (from Figure 16) of camera 4, shown in a), to the first frame of shot S2 of camera 2, shown in c). In the matching process, shot S2 is first resized so that the vertical distance from neck point to nose point is the same, because perspective shifts have little effect on the vertical distances of the body points, since all cameras are at approximately the actors' eye level. The distance from neck point to nose point is shown in white in b). The neck point (red) between the shoulder points serves as a fixed point during matching. It is assigned the same coordinates in the first frame of S2 in c) as in the last frame of S1 in a). In the semi-transparent overlay b) it is easy to see that the neck points are exactly on top of each other, while the other body points are shifted horizontally. Shot S3 in Figure 16 shows the man alone. In shot S2, however, he is not visible, so the first frame of S3 is also fitted to the last frame of the long shot S1, where the man was still visible. For S4 (woman), the fit is taken from S2, and for S5 (man), the fit is taken from S3.

VI. AUDIO ANALYSIS

For the analysis of theater recordings and the automation of editing, it is necessary to detect the respective speaking person in the image. This can be done using audio analysis in combination with OpenPose, yielding audiovisual analysis.

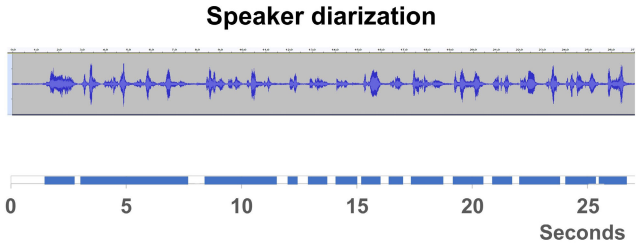

FIGURE 22. Speaker diarization. Recognition of speech segments and pauses in the audio file.

TABLE 1. Diarization output of the in- and end-points (in seconds) of the individual speech segments in a text file.

	NR	IN	OUT
SPEAKER <NA>	1	1,46 s	2,764 s
SPEAKER <NA>	1	3,011 s	7,687 s
SPEAKER <NA>	1	8,397 s	11,526 s
usw.			

FIGURE 23. Detection of lip movements with OpenPose based on the difference of the y-values of the upper lip and lower lip.

A. SPEAKER DETECTION

Speaker diarization is the process of partitioning audio into homogeneous segments according to speaker identity. The open-source toolkit used in this paper, pyannote-audio 2.1.1 [52], is based on PyTorch, an open-source program library focused on machine learning [53].

The diarization of an excerpt of an audio track of a play is shown in Figure 22. The waveform of the audio file is shown above and the result of the diarization is shown below. The in- and out-points of the individual speech segments are output line by line in seconds in a text file, as Table 1 shows.

Speaker diarization can differentiate individual voices. However, the mapping is too inaccurate when differentiating between multiple people who have the same gender and similar voices. Therefore, the assignment to the respective speaking person is done with OpenPose.

Figure 23 shows on the right the medium long shot of a scene from the play "A Midsummer Night's Dream" [54] and on the left the detected key points of the faces. The mouth movement is detected by the difference of the y-values of

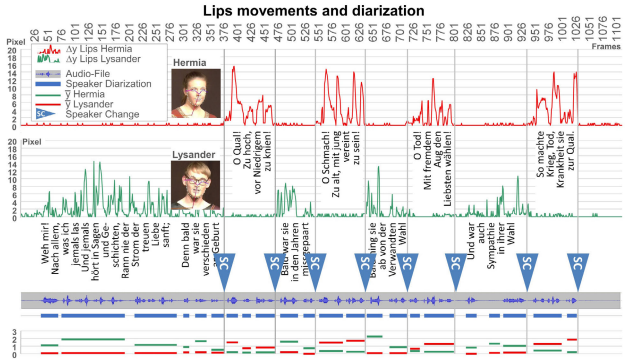

FIGURE 24. Assignment of the speaking persons. Audio speech segments are matched to the correct person based on lip movements.

the upper lip and lower lip. According to Figure 7, these are the key points (KP) 62 and 66. Since the zero point of the pixel values in image processing programs is in the upper left corner of the image, the y-value of the lower lip is the larger one and the mouth opening is determined by Equation 1.

$$\Delta y = y_{lower\ lip} - y_{upper\ lip} = y_{KP\ l66} - y_{KP\ l62} \tag{1}$$

Figure 24 shows how the diarization segments are assigned to the respective speaking persons. At the top, the mouth openings and the spoken text of the two persons are shown over time. The lip movements of the woman (Hermia) are shown in red and the lip movement of the man (Lysander) in green. Below that, the waveform of the audio file and the result of the diarization are plotted in blue. For the assignment of a diarization segment to the speaking person, the change of the mouth opening to the respective previous image is measured. If only the mouth opening was taken into account, a person with an open mouth throughout could achieve higher values than the person whose mouth is moving. Within a segment, therefore, the magnitudes of the changes in mouth opening are formed frame by frame so that both opening or closing of the mouth results in a positive contribution. The values are added and divided by the number of images in the segment to form the mean value. This mean mouth movement $\bar{y}$ (in pixels) is calculated for a segment consisting of n images from Equation 2.

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} \mid \Delta y_k - \Delta y_{(k-1)} \mid \tag{2}$$

The calculated mean values for Hermia in red and Lysander in green are listed below in Figure 24, each parallel to the blue segment. The values of the speaking person are between 0.5 and 2.5 pixels and are each higher than those of the non-speaking person, whose values are between 0 and 0.5 pixels. For the first six segments (between frame 1 to frame 375 in Figure 24), Lysander has the higher value $\bar{y}$. He is the speaking person. At the seventh segment (from frame 380), $\bar{y}$ of Hermia is greater than $\bar{y}$ of Lysander. Here, a speaker change takes place and Hermia is now the speaking person. The detected speaker change and all following ones are marked with blue arrows "SC" in Figure 24.

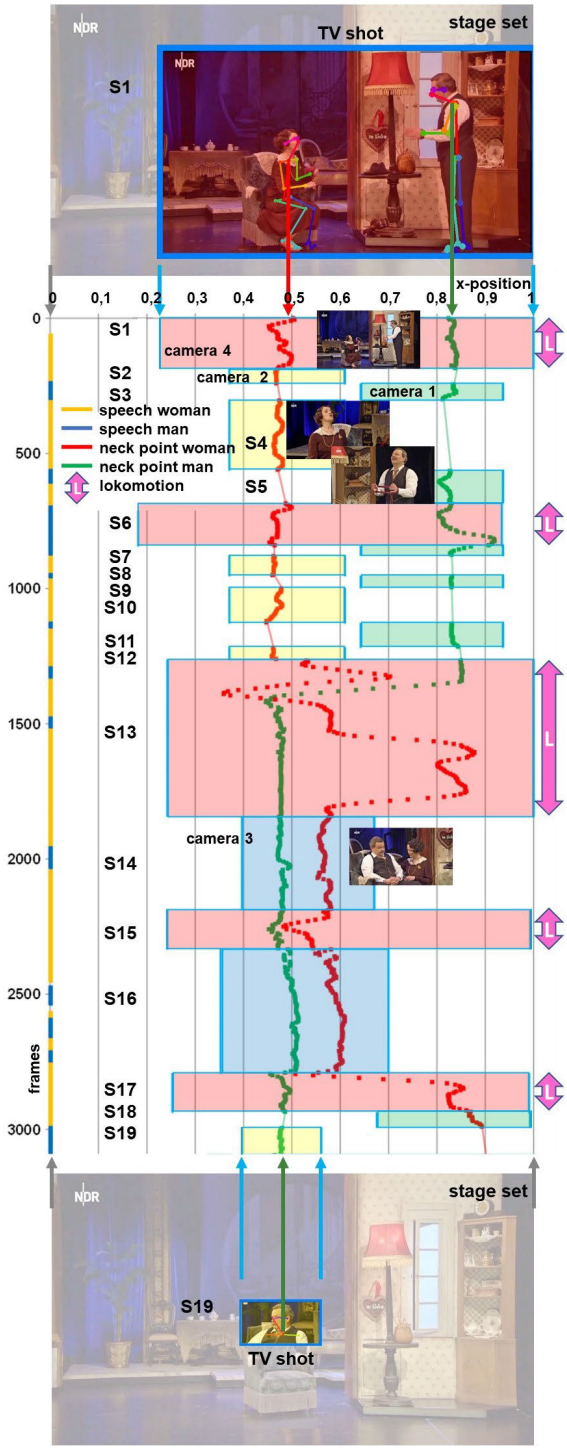

FIGURE 25. Scene analysis from "A better gentleman" [43]. Movement of the neck points over the time axis (vertical) and adjustment of the shots in the color mapping of the cameras.

VII. SCENE ANALYSIS

Step by step, parameters, and rules of script breakdown for the Reference Recording System are now obtained. For this

purpose, different scenarios are investigated, which differ with respect to the number of actors and their stage action. In this paper, the investigations are presented for scenes with two persons, limiting the presentation for economic usage of paper space. If two persons are on stage, they are usually spatially arranged on stage in such a way that both persons can be easily seen from all seats in the audience and do not cover each other. The characters can walk around, switch sides, get close to each other, sit together at a table or sofa, have dialogues, etc., to name just a few examples of possible actions.

A. EXAMPLE

As an example, the analysis of a scene from the play ''A better gentleman'' [43] is presented here. Figure 25 shows the scene from Figure 16 analyzed over time. The time axis runs from top to bottom. Above, it is shown with arrows how the neck points and fitting edges of shot S1 map to the diagram. The fit of shot S19 is shown below.

Shown in red and green on the horizontal axis are the x-positions of the neck points of the woman and the man in the stage image, relative to the stage image width. The value 0 is the left image edge and the value 1 is the right image edge of the stage set. The neck points are chosen for the position indication because they are independent of head movements, in contrast to the points of mouth, nose, and ears. The y-axis oriented downwards shows the temporal progression in frames. Each shot is fitted into the stage set as explained before. Dialogues such as S2 to S5 and S7 to S12 are shown in close-ups of camera 1 (green) and camera 2 (yellow). This is a shot/reverse shot situation typical in film [55]. The long shot from camera 4 (red) is taken when a locomotion of the persons takes place, as in S1, S6, S13, S15, etc. For example, in shot 12, the woman is sitting in an armchair and is shown in close-up. Just before she walks over to the man, shot 13 cuts to camera 4 and shows the long shot. The woman fetches the man (frame 1300), pulls him over to the armchair so he can sit down (frame 1400). The woman walks around the armchair and stops there briefly. After frame 1500, in frame 1600 she goes to the table by the lamp and puts the ball of wool there in the needlework bag. After frame 1700 she goes to the armchair and squats down to the man at frame 1800. Only then the locomotion is over and camera 3 is used.

In this case, the two people in S14 are now so close that they are in Intimate Distance according to Hall [56]. The persons can now no longer be captured individually by the cameras and are shown in a medium shot by camera 3. Also, after the locomotion in S15, in S16 the persons are again in Intimate Distance and are shown by camera 3. Only after the locomotion of the woman in S17, which goes to the right side of the stage, the distances are large enough that in S18 and S19 the shot/ reverse shot of camera 1 and 2 is used again.

The shot changes can be mapped and classified as a graph as shown in Figure 26:

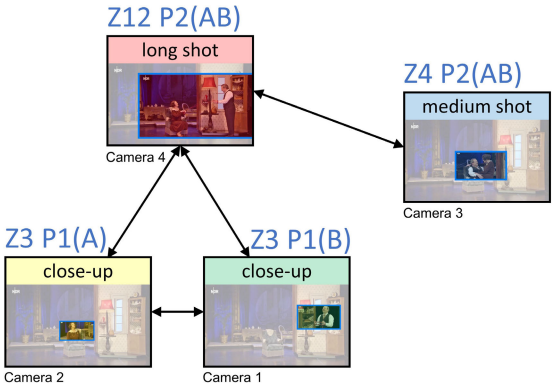

FIGURE 26. Graph of the shot changes.

- Type Z12 P2(AB): Medium long shot or long shot; shows both people (camera 4).
- Type Z4 P2(AB): Medium shot; shows both people when they are close together (camera 3)
- Type Z3 P1(A): Medium close-up; shows the person who is to the left of the other person (camera 2)
- Type Z3 P1(B): Medium close-up; shows the person who is to the right of the other person (camera 1).

The shot sizes are indicated with Z in the system of body proportions introduced in Subsection IV-B. Since several shots are summarized in one type, the specification e.g. Z12 is to be understood as an order of magnitude which summarizes shot sizes from e.g. Z8 to Z16 in one type. The letter P indicates the number of persons. P1 shows one person, P2 two persons. The letters A and B indicate the positions of the persons from left to right. This scheme for position indication remains, even if the actors change places. If the woman is to the left of the man, the woman is labeled A and the man is labeled B. If the man is to the left of the woman, the man is labeled A and the woman is labeled B. The shot changes in the scene are shown in Figure 26. Starting points are long shots Z12 P2(AB). If the woman and the man move close together, there is a change to a medium shot Z4 P2(AB) showing both people, and from there back to the starting point Z12 P2(AB), when the people move apart again. If the woman and man are far enough apart that they can be shown alone in close-ups, sequences from the shots Z3 P1(A) and Z3 P1(B) follow in alternation.

B. DEVELOPMENT OF FLOW CHART REPRESENTATION

The flowchart in Figure 27 can be developed for automatic editing from the analyses of scenes with two people.

The curtain opens and the scene usually begins with a long shot (a). The end of the scene is also concluded with a final long shot. If the end of scene (b) has not yet been reached, it is analyzed where locomotion takes place (c). In these cases, the long shot is shown. In the parts of the scene where no locomotion takes place, the distance of the actors is analyzed (d).

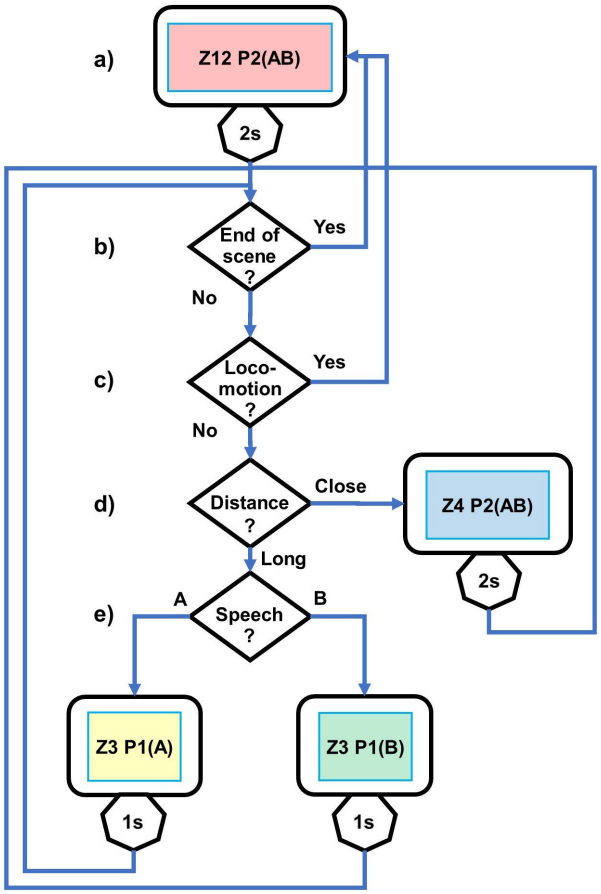

FIGURE 27. Flow chart for 2 persons generated from the analysis of the Reference Recording Systems (RRS).

The flowcharts generated by the authors for 3, 4, or more people on stage have the same structure. Close distances are shown with camera 3 in medium shot. If the actors are far enough apart, the dialog is shown in shot/reverse shot (e). Close shots should be at least 1 second long, medium shots, and long shots at least 2 seconds, because the viewers have to take in more image information.

VIII. PROPOSED RECORDING SYSTEM (PRS): MULTI-CAMERA RECORDING BY NON-PROFESSIONALS

The rules obtained by the analysis of the Reference Recording System (RRS) are now applied in a practical example carried out with the Proposed Recording System (PRS). In the auditorium of the Gymnasium of Benedictiner in Meschede, the play "Astoria" is performed by the theater group Theatiner [57]. The play is recorded with four 4K cameras. Panasonic HC-WX979 4K camcorders are used on the left and right, and a 4K Sony Alpha 6500 APS-C E-mount and a fixed Blackmagic Pocket Cinema Camera 4K are used at the back. The fixed camera (camera 4) recorded the long shot throughout. It is aligned once at the beginning of the performance and the camera positioning not changed.

It is specified that close-ups should have at least Z3 (see Figure 8) and the presented video format HD 1280 × 720. With this presented video format, four fixed cameras cannot yet be used. Medium-length shots of Z9 are necessary if Z3 is to be cropped. The (cropping) factor 3 results from the recording format 4K and the presented video format (3840/1280 = 3).

Therefore, camera 1 to 3 are operated by inexperienced students. All are instructed to take medium-length shots (Z9) and not to zoom or pan within a scene. The front left camera should always capture the person or persons on the right side of the scene and the right camera should capture the person or persons on the left side of the scene. If there is significant movement, for example, to the other side of the stage, the camera should be panned to the new position if necessary and then not moved again. Because the cameras take medium long shots Z9 and shows the people from head to toe, little correction is needed. During the correction, the camera is not used for the presented video anyway. The fixed camera 4 then takes the shot, because locomotion is taking place (see flow chart Figure 27). In the future, the process will use four fixed cameras without a camera crew at all if 6K or 8K cameras are used. With 6K cameras (6144 × 3456) and cropped Z3 close-ups, the cropping factor is 4.8 (6144/1280 = 4.8). The cameras can then capture the persons with Z14.4. This is a long shot that shows the entire stage if it is a small stage. With 8K cameras (7680 × 4320), even larger stages are fully captured. The cropping factor is 6 (7680/1280) and thus a Z3 close-up can be cropped from a Z18 long shot.

A. PROCESS

Figure 28 shows the process flow leading to the automatic editing script. The same tools are used as for the analysis of the Reference Recording System described in Section IV. The 4K camera recordings (step a) are analyzed with OpenPose (step b), and face recognition and motion paths are created. With speaker diarization (step c) the timecodes are also recorded, identifying which person speaks at which time. The motion paths are examined (step d), indicating where locomotion takes place (step e), and then in the automatic editing script (step h) the long shot of camera 4 is selected. The speaker changes from c) are also registered in the script and control the cuts f) as trigger points. Instead of exact cuts, which can appear robotic, the vision mixer model (vm-model) from [4] can be applied. A parameter can be used to specify whether the cuts tend to be placed before the speech pause or delayed after the speech pause. Professional vision mixers have their own individual style, which was analyzed and modelled in [4].

The distance of the actors is examined in step g. If the distance is close, a medium shot of camera 3 is selected, which shows both persons. If the distance is long, close-ups are cropped in the images of camera 1 and 2 and entered into the script as shot/reverse shot.

B. AUTOMATIC EDITING SCRIPT

Figure 29 shows the result of the automatic editing script. The video input of the four cameras and the selected crops are

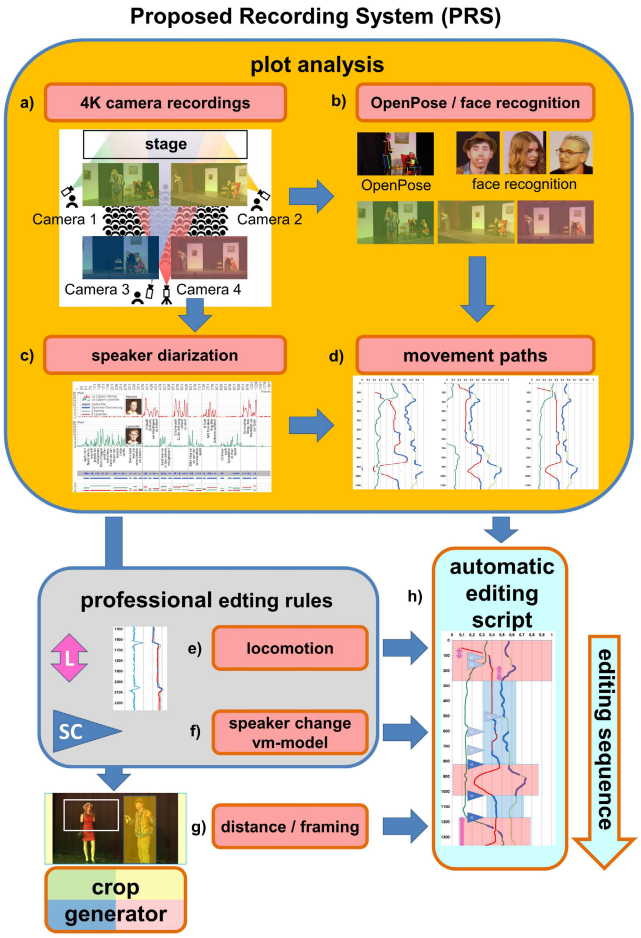

FIGURE 28. Process flow of the proposed recording system (PRS).

shown in a), the steps followed by the automatic editing script in b), and the main output in c). After the entry long shot S1, the man is shown in close-up until the woman appears. If new persons appear, the cut is brought forward to the long shot with a time lead of about 1 second, so that the viewer can see the entire entry of the person and not only when it is detected by OpenPose.

Figure 29 b) shows how each speaker change (triangle SC) is cut to a new shot. Only the SCs in light blue marked triangles are within a locomotion and there is no cut. What is special about this scene is that the man in the close-ups mostly speaks to the audience (S2, S4, S7, S13, S17, S19) and occasionally to the woman (S11 and S15). Usually, close-ups of the man are shown by camera 1 (green). But when he is speaking to the audience, camera 3 is chosen, which is behind the audience. Where the man is looking is detected by face rotation analysis (subsection IV-F). In Figure 29 b), the nose point is drawn in gray in addition to the neck point. It is easy to see in S11 and S13 how the nose point is located to the left of the neck point when the man speaks to the woman. The close-ups are cropped from the medium long shots and a Z-value of 4 was set for the close-ups in this scene. In Figure 29 a) the image sections of the close-ups are drawn as white frames in the camera images from camera 1 to 3. It is also possible to crop in the long shot as in S10, S16, and S18 when the actors are closer together. This is measured by distance analysis with OpenPose and adjusted accordingly. The automated edit script then outputs a text-based EDL list (Edit Decision List) listing timecode, camera, and frame. Video editing programs such as the open-source software ShotCut can read and process the lists and generate the main output.

IX. ONLINE STUDY "PERSPECTIVE DIVERSITY"

The effort required to film with the Proposed Recording System (PRS) using four cameras is higher than for the Gandhi Recording System (GRS) [3] using only one camera. To assess how test persons evaluate the use of the PRS with its four camera perspectives, as shown in Figure 3 c), an online study was carried out. For this purpose, a comparison is made with the GRS, as shown in Figure 3 b), which uses only one camera in central perspective and generates image crops from the captured footage. In addition, the test participants compare the two recording methods with a Simple Recording System (SRS), as shown in Figure 3 a), which records with one camera, no cropping, and no cuts. It is noted that an online test is not a subjective test under laboratory conditions as defined in Recommendation ITU-R BT.500-14 [58]. The results merely represent a set of subjective opinions that can only provide evidence of a tendency. However, in e.g. [59], [60], it has been shown that such online tests can be used to gather similarly reliable results as in the case of a lab test for the quality evaluation of high-resolution images and videos.

A. CREATING DIFFERENT VERSIONS

Three different versions are created, excerpts of which can be seen in Figure 30. Four scenes with 27 shots are shown from two theater recordings with an approximate total playing time of 2:40 minutes. Version 1, the Simple Recording System (SRS), shows the scenes only in medium long shot. This is a common recording practice among non-professionals. The recording is done with only one camera. Here, the camera is set up in the center behind the audience. Either a continuous long shot is recorded or the image is adjusted to the action at the beginning of a scene by zooming in so that all the people in the image can always be seen from head to toe (medium long shot). Then the camera is not moved for the rest of the scene. In version 1 there are no zooms but only fixed camera shots. There are no cuts to close-ups in contrast to Gandhi. Each scene is shown in only one shot without cuts. Therefore, in version 1, shot E2 and shot E3 in Figure 30 show the same image as E1. Version 1 actually has only four shots (one shot for each scene) in contrast to version 2 and version 3, which have 27 shots.

Version 2, the Proposed Recording System (PRS), shows camera shots as they can be generated by an automatic editing script according to Section VIII. Version 3, the Gandhi Recording System (GRS) shows shots as they can be obtained

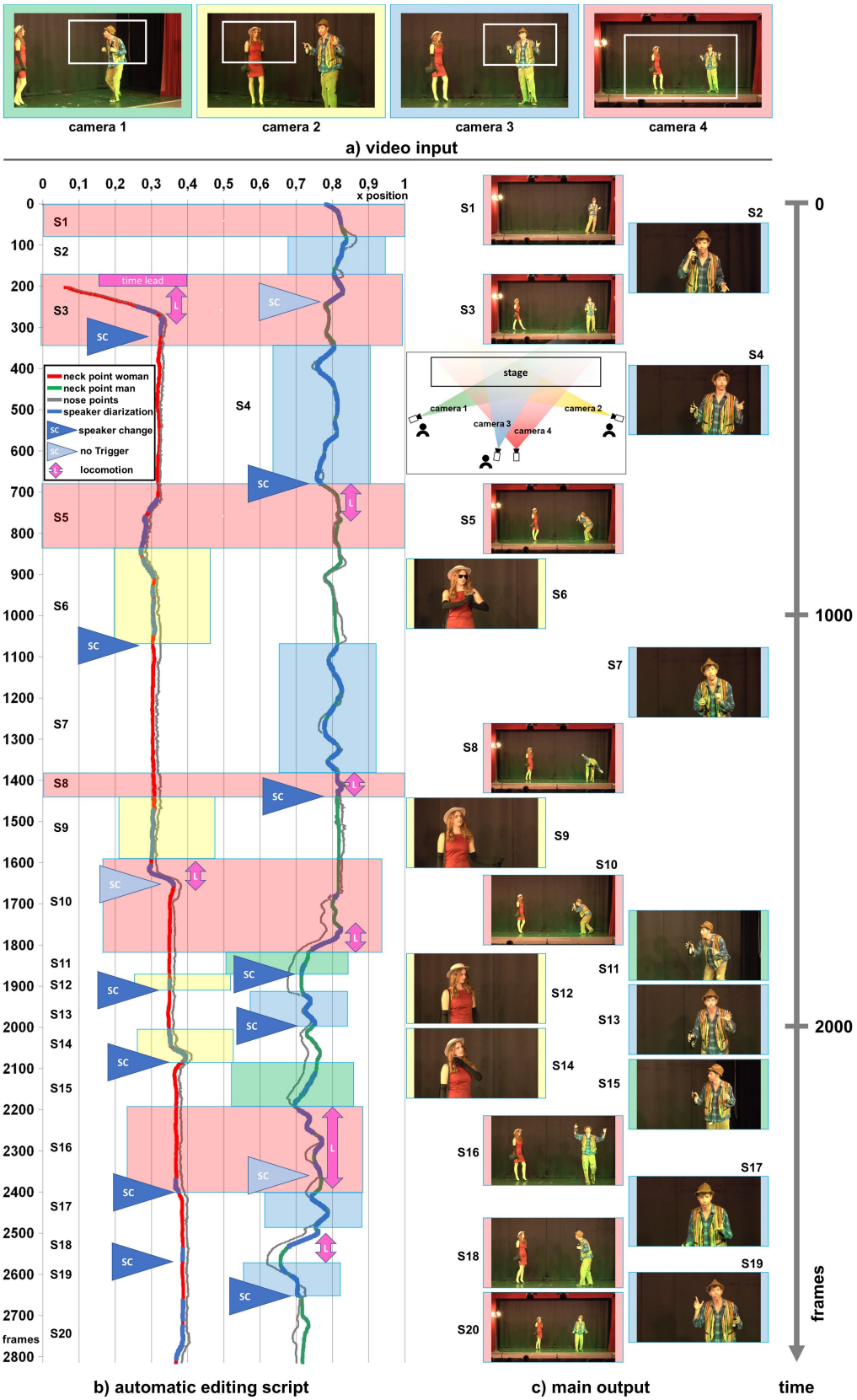

FIGURE 29. Generated automatic editor script and main output. Scene from "Astoria" [57].

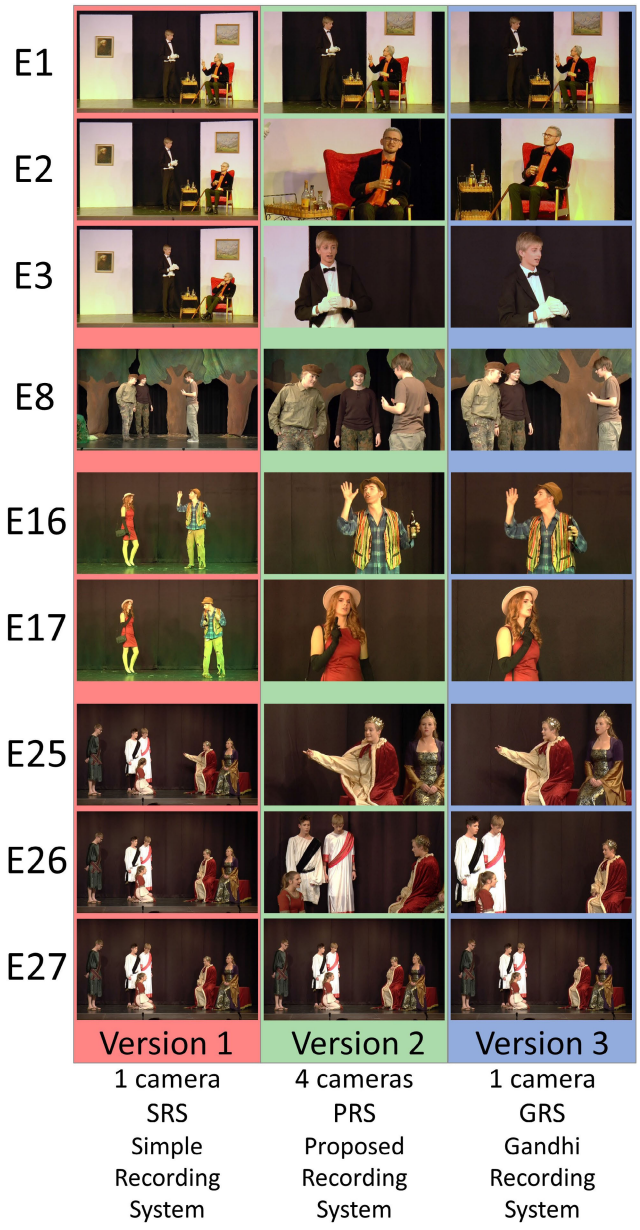

FIGURE 30. Excerpt from the 27 shots of the three versions.

FIGURE 31. Comparison of the resolution of close-ups from the four cameras. The camera positions can be seen in Figure 28 a).

according to Gandhi. Versions 2 and 3 show close-ups of the person speaking cropped from medium long shots, in addition to medium long shots as people move on stage. If the person speaking cannot be shown alone because of the close distance to other persons, the group is shown as for example in E8, E25 or E26.

Version 2 and 3 have exactly the same cuts. They differ only in the camera angles of the close-ups. Version 2 shows the close-ups in shot/counter-shot with camera 1 and 2 (see Figure 28 a)). Version 3 shows the close-ups from the position behind the audience. Gandhi records in full stage long shot like Z18 and crops close-ups from them. The resolution is then much lower than for version 2 (PRS), as Figure 31 shows in an example.

Camera 4 is the 4K camera as installed on the Gandhi Recording System. Close-ups here have a resolution of only 853 × 480 pixels. Only an 8K camera would double the horizontal and vertical resolution of the close-ups, which would then roughly correspond to the resolutions of the close-ups from cameras 1, 2, and 3. These reach 1600 × 900 pixels in the example. In order to ensure that the low resolution of Gandhi (with a 4K camera) does not negatively influence the test persons in their evaluation of the perspective diversity, Gandhi is simulated as if the scene is recorded with an 8K camera. The close-ups of version 3 are obtained from camera 3 and thus have the same resolution as version 2. The differences in camera angles can be seen in Figure 30 at shots E2, E3, E8, E16, E17, E25, and E26. The front cameras in version 2 show more of the faces than in version 3. The study will show whether and how much this is an improvement for the test persons.

MAGIX Video Pro X14 (Version 20.0.3.181) is used to create all versions. All sequences are loaded into the editing software and downscaled to 1280 × 720 after editing and re-encoded using H.264 at 6 Mbit/s to enable the crowd-type online test. Filters such as ''sharpening'' or ''noise filter'' are deliberately not used.

B. ONLINE QUESTIONNAIRE

The development of the questionnaire was guided by different previous work [61], [62], [63], [64]. An online questionnaire is designed for the study, in which the three versions are evaluated. It was left to the respondents to decide on which end device they would carry out the evaluation. The three versions are arranged in three players. The entire questionnaire is integrated into a single website to enable the test participants to watch the films more than once and to enable comparative evaluations. Absolute ratings cannot be expected from this study, as in an online study the exact viewing conditions cannot be controlled [65] (e.g. display device, viewing distance, illumination, lighting conditions, reflections). The participants are asked to rate the liking of each sequence,

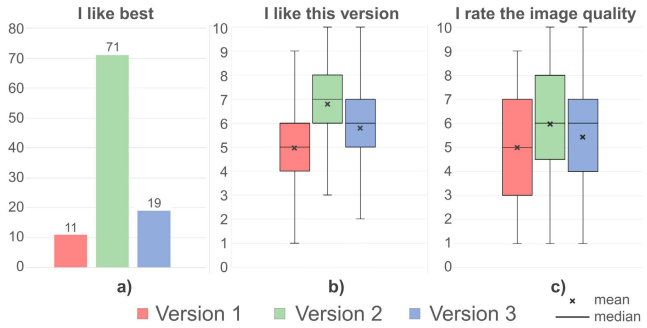

FIGURE 32. Results of the study "perspective diversity": a) distribution of the chosen favorite, b) evaluation of the preference, and c) evaluation of the image quality.

as well as the image quality. Age and gender are also asked.

C. SCALES

For each version, a direct liking rating is asked from the participants, using the phrasing "I like version X:" followed by the rating scale. Similarly, the image quality is to be rated, preceded by the instruction "The image quality for me is:" again followed by an interactive rating scale. The liking and image quality of the three versions are each evaluated on the 11-point scale (from 0 to 10) according to ITU-T P.910 [66] resp. [67] annotated with 5 attributes (bad - poor - fair - good - excellent). The 11-point scale presented in Appendix B of [67] is supposed to enable a finer-grained differentiation by intermediate values and additional identification of the endpoints than the 5-point Absolute Category Rating (ACR) scale often used for video quality evaluation.

D. PREFERENCE

Although a respondent's rating on a given scale might indicate a preference, a direct preference rating is asked, too, answering the question "Which version do you think is best?" followed by the question, "Why do you think this version is best?" For the latter question, in addition to knowledge of the preferred version, qualitative verbalized assessments of the respondents are also obtained, especially when two videos have yielded the same likability rating.

E. ACQUISITION

The following measures are taken to recruit the test subjects:

- Already during the recording of the plays, the audience is informed about the study via handout and asked to enter their name and email on the handout and to hand it in.
- Information was provided via various email distribution lists and multipliers such as institutions and associations are asked to also spread the call for the study in the social networks. Participation was also requested in the forum of the South Westphalia University of Applied Science.
- An appeal is made via the website of the local radio station Radio Sauerland.

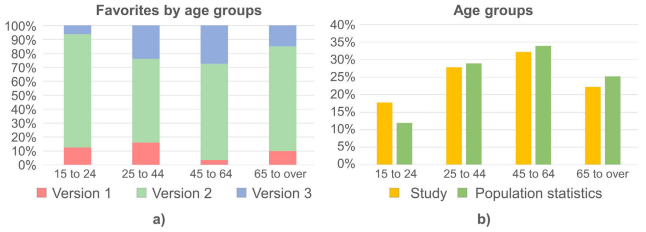

FIGURE 33. a) Favorites by age group and b) study compared with population statistics [68] (source: German Federal Statistical Office, cutoff date Dec. 31, 2019).

F. RESULTS

101 test persons aged 18 and over participate in the online study. Figure 32 a) shows the selection of the chosen preferred versions. Version 2 is chosen by over two-thirds of the test persons (71 votes = approx. 70.3 %). Version 1 receives 11 votes (= approx. 10.9 %) and version 3 receives 19 votes (= approx. 18.8 %). Results for the liking and image quality ratings are shown in Figure 32 b) and 32 c). For the Figure 32 b) and c) we performed statistical testing using the Kruskal-Wallis test with a p-value of 0.05. For the liking in Figure 32 b) all paired tests indicated statistically significant differences. For the image quality in Figure 32 c) only version 1 compared to version 2 was statistically significantly different, the other comparisons were not significant.

Version 2 and version 3 are similar. They have the same cuts, the same shot lengths, and the same shot sizes. Version 2 with its four perspectives, is clearly preferred as the favorite. This is assumed to be due to the fact that more of the faces can be seen through the shot-counter-shot position than in version 3 (Gandhi). Version 2, with a mean of about 6.78, is rated about one scale point higher in liking than version 3, with a mean of about 5.79.

Version 1 shows the entire stage in medium long shot throughout and corresponds the least to a professional theatrical recording. This is also reflected in the liking rating, which is the worst of the three versions, with a mean of about 4.97.

The test persons who chose version 1 as their favorite state that they feel like audience members and like to see the whole action and all reactions of the actors. For version 2, it is stated that facial expressions and gestures can be recognized better. Looking into the faces of the actors allows to experience the story and the emotions more closely. For version 3, it is stated that the slightly slanted perspective is more pleasant.

Because the close-ups of version 3 are obtained from camera 3, they have the same resolution as version 2. All shots in versions 1, 2, and 3 are only encoded by downscaling to the presented video format. Upscaling and thus quality degradation does not take place. Therefore, it is expected that the image quality of all three versions will be rated similarly.

In Figure 32 c), however, it can be seen that the image quality of version 2 is rated somewhat higher. Analysis of individual data sets shows that some test persons have adjusted the quality rating to the liking of the respective version. Physically, version 2 and 3 have the same quality in

the close-ups and the 10 medium long shots of the 27 shots are even identical.

It is checked whether there are noticeable differences in the choice of favorite depending on age. 90 test persons indicated their age. Figure 33 a) shows the favorite choice by age group. It can be seen that there are slight differences in the distribution of favorite choice in the age groups. In all age groups, version 2 is chosen as the favorite by a large margin over the other versions. Figure 33 b) shows the distribution of subjects among age groups compared to population statistics [68]. The distribution of the age groups of the test persons roughly corresponds to the distribution of the population statistics. The age group ''15 to 24'' is slightly more represented in relation to the population statistics. This may be due to the fact that many students responded to the call for participation in this study.

X. CONCLUSION

This paper exemplifies how multi-camera recordings of theater performances or other stage performances in a non-professional environment can be improved. A production method ''Proposed Record System (PRS)'' is presented using high-resolution cameras from which image sections are automatically cropped from long shots or medium long shots. To extract a set of rules and train algorithmic components, a novel method is presented on how professional theater recordings can be analyzed and the script breakdown of the director can be reconstructed by fitting it into the stage set. Rules for camera selection and framing are derived for scenes and presented in flow charts. A process is developed for applying the results to a multi-camera recording by non-professionals and for creating an automatic edit script that generates an automatic montage. An online study confirms the added value of the perspective diversity of four cameras of the Proposed Recording System (PRS) versus the single camera method of Gandhi et al. (Gandhi Recording System GRS). The PRS was preferred by over two-thirds of the test persons.

Editing is an artistic process. Different directors will create a different script breakdown for the same play. Nevertheless, basic principles can be registered in an automatic edit script and a proposal for an automatic montage can be generated.

The process presented in this paper consists of combinations of individual steps, as the process flow Figure 28 shows. As further development, the components can be implemented in a single program or included as plug-ins in video software. Shotcut (www.shotcut.org) as an open-source project is a good choice here, or Magix Video Pro X, for example, whose development department has also incorporated suggestions from users in the past [69].

Future work will focus on refining the method and analyzing further script breakdowns of scenes with several people. The principles can also be adapted for other recording situations with an audience, such as lectures, interviews, discussions, talk shows, gala events, award ceremonies, and the like.

REFERENCES

[1] Junges Theater Düren, *Der Besuch Der Alten Dame (The Visit), Friedrich Dürrenmatt. Ro-Ka Wirtz (Video Director)*, Junges Theater Düren, Düren, Germany, 2018.
[2] Theatiner, *Lysistrata—Der Krieg Muss Weg. Sophokles. Gymnasium of Benedictine. Schlomberg, P(Director)*, Gymnasium Benedictine, Meschede, Germany, 2019.
[3] V. Gandhi, ''Automatic rush generation with application to theatre performances,'' Ph.D. dissertation, Dept. Math.-Inform., Grenoble Univ., Grenoble, France, 2014.
[4] E. Stoll, S. Breide, S. Göring, and A. Raake, ''Modeling of an automatic vision mixer with human characteristics for multi-camera theater recordings,'' *IEEE Access*, vol. 11, pp. 18714–18726, 2023.
[5] V. Gandhi and R. Ronfard, ''Detecting and naming actors in movies using generative appearance models,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3706–3713.
[6] V. Gandhi, R. Ronfard, and M. Gleicher, ''Multi-clip video editing from a single viewpoint,'' in *Proc. 11th Eur. Conf. Vis. Media Prod.*, Nov. 2014, pp. 1–10.
[7] M. Kumar, V. Gandhi, R. Ronfard, and M. Gleicher, ''Zooming on all actors: Automatic focus+context split screen video generation,'' *Comput. Graph. Forum*, vol. 36, no. 2, pp. 455–465, May 2017.
[8] K. K. Rachavarapu, M. Kumar, V. Gandhi, and R. Subramanian, ''Watch to edit: Video retargeting using gaze,'' *Comput. Graph. Forum*, vol. 37, no. 2, pp. 205–215, May 2018.
[9] J. Chen, G. Bai, S. Liang, and Z. Li, ''Automatic image cropping: A computational complexity study,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 507–515.
[10] Z. Li and X. Zhang, ''Collaborative deep reinforcement learning for image cropping,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 254–259.
[11] R. Kumar, P. Assuncao, L. Ferreira, and A. Navarro, ''Retargeting UHD 4 K video for smartphones,'' in *Proc. IEEE 8th Int. Conf. Consum. Electron.*, Sep. 2018, pp. 1–5.
[12] T. Deselaers, P. Dreuw, and H. Ney, ''Pan, zoom, scan—Time-coherent, trained automatic video cropping,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[13] R. Kumar, L. Ferreira, P. A. Amado Assuncao, and A. Navarro, ''Retargeting 4K video for mobile access using visual attention and temporal stabilization,'' in *Proc. 9th Int. Symp. Signal, Image, Video Commun. (ISIVC)*, Nov. 2018, pp. 48–53.
[14] M. E. Varela and G. O. F. Parikesit, ''A quantitative close analysis of a theatre video recording,'' *Digit. Scholarship Humanities*, vol. 32, no. 2, pp. 276–283, 2017.
[15] M. Leake, A. Davis, A. Truong, and M. Agrawala, ''Computational video editing for dialogue-driven scenes,'' *ACM Trans. Graph.*, vol. 36, no. 4, p. 130, 2017.
[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, ''OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[17] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, ''Realtime multi-person 2D pose estimation using part affinity fields,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.
[18] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, ''Hand keypoint detection in single images using multiview bootstrapping,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4645–4653.
[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, ''Convolutional pose machines,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
[20] GitHub. (2020). *OpenPose: The First Real-Time Multi-Person System to Jointly Detect Human Body, Hand, Facial, and Foot Keypoints*. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose
[21] A. Truong, P. Chi, D. Salesin, I. Essa, and M. Agrawala, ''Automatic generation of two-level hierarchical tutorials from instructional makeup videos,'' in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–6.
[22] P.-Y. Chi, J. Liu, J. Linder, M. Dontcheva, W. Li, and B. Hartmann, ''DemoCut: Generating concise instructional videos for physical demonstrations,'' in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2013, pp. 141–150.
[23] J. Quiroga, H. Carrillo, E. Maldonado, J. Ruiz, and L. M. Zapata, ''As seen on TV: Automatic basketball video production using Gaussian-based actionness and game states recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3911–3920.

[24] H. Pidaparthy and J. Elder, "Keep your eye on the puck: Automatic hockey videography," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1636–1644.
[25] F. Perazzi, A. Sorkine-Hornung, H. Zimmer, P. Kaufmann, O. Wang, S. Watson, and M. Gross, "Panoramic video from unstructured camera arrays," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 57–68, May 2015.
[26] V. Popovic, K. Seyid, Ö. Cogal, A. Akin, and Y. Leblebici, "State-of-the-art multi-camera systems,' in *Design and Implementation of Real-Time Multi-Sensor Vision Systems*. Berlin, Germany: Springer, 2017, pp. 13–31.
[27] O. Schreer, I. Feldmann, C. Weissig, P. Kauff, and R. Schafer, "Ultrahigh-resolution panoramic imaging for format-agnostic video production," *Proc. IEEE*, vol. 101, no. 1, pp. 99–114, Jan. 2013.
[28] P. Ong, T. K. Chong, K. M. Ong, and E. S. Low, "Tracking of moving athlete from video sequences using flower pollination algorithm," *Vis. Comput.*, vol. 38, pp. 939–962, Jan. 2021.
[29] M. Tiwari and R. Singhai, "A review of detection and tracking of object from image and video sequences," *Int. J. Comput. Intell. Res.*, vol. 13, no. 5, pp. 745–765, Mar. 2017.
[30] H. Morimitsu, I. Bloch, and R. M. Cesar-Jr., "Exploring structure for long-term tracking of multiple objects in sports videos," *Comput. Vis. Image Understand.*, vol. 159, pp. 89–104, Jun. 2017.
[31] M.-Y. Fang, C. K. Chang, N. C. Yang, and C. M. Kuo, "Robust player tracking for broadcast tennis videos with adaptive Kalman filtering," *J. Inf. Hiding Multimedia Signal Process.*, vol. 5, no. 2, pp. 242–262, 2014.
[32] Z. Musa, M. Z. Salleh, R. A. Bakar, and J. Watada, "GbLN-PSO and model-based particle filter approach for tracking human movements in large view cases," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 8, pp. 1433–1446, Aug. 2016.
[33] E. Stoll, S. Breide, and A. Raake, "Towards analysing the interaction between quality and storytelling for event video recording," in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2020, pp. 1–4.
[34] (2021). *Der Komödienstadel: Episodenguide*. [Online]. Available: https://www.fernsehserien.de/der-komoedienstadel/episodenguide
[35] (2021). *Chiemgauer Volkstheater: Episodenguide*. [Online]. Available: https://www.fernsehserien.de/chiemgauer-volkstheater/episodenguide
[36] (2021). *Ohnsorg Theater: Episodenguide*. [Online]. Available: https://www.fernsehserien.de/ohnsorg-theater/episodenguide
[37] Bayerisches Fernsehen, *Obandlt is (Komödienstadel), Thomas Stammberger (Video Director)*, Bayerisches Fernsehen, M üchen, Germany, 2012.
[38] Bayerisches Fernsehen, *Der Kartlbauer (Chiemgauer Volkstheater), Thomas Kornmayer (Video Director)*, Bayerisches Fernsehen, München, Germany, 2017.
[39] V. Gupta. (2018). *Deep Learning Based Human Pose Estimation Using OpenCV (C++/Python)*. [Online]. Available: https://www.learnopencv.com/deep-learning-based-human-pose-estimation-using-opencv-cpp-python/
[40] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 246–260.
[41] I. Bernhard. (2010). *Study of Human Body's Proportions (Gnu Free Documentation License)*. [Online]. Available: https://commons.wikimedia.org/wiki/File: Human_body_proportions2.svg
[42] S. Koga-Browes, "Social distance portrayed: Television news in Japan and the UK," *Vis. Commun.*, vol. 12, no. 1, pp. 71–96, Feb. 2013.
[43] NDR, *A Better Gentleman (Ohnsorg Theater), Director: Henning Kasten*, Norddeutscher Rundfunk, Hamburg, Germany, 2019.
[44] Betaface.Com. *Facial Recognition Online (Demo)*. Accessed: Mar. 2, 2021. [Online]. Available: https://www.betaface
[45] Y. Kobayashi, S. Yamazaki, H. Takahashi, H. Fukuda, and Y. Kuno, "Robotic shopping trolley for supporting the elderly," in *Proc. Int. Conf. Appl. Human Factors Ergonom.* Cham, Switzerland: Springer, 2018, pp. 344–353.
[46] M. M. Islam, A. Lam, H. Fukuda, Y. Kobayashi, and Y. Kuno, "An intelligent shopping support robot: Understanding shopping behavior from 2D skeleton data using GRU network," *ROBOMECH J.*, vol. 6, no. 1, pp. 1–10, Dec. 2019.
[47] Megvii. (2022). *Face^{++} AI Open Platform*. [Online]. Available: https://www.faceplusplus.com/face-detection/
[48] WebAR.rocks. (2022). *Jeeliz Face Filters*. [Online]. Available: https://jeeliz.com/demos/faceFilter/demos/babylonjs/cube/
[49] H. Raschke, "Szenische auflösung," in *Wie Man Sich Eine Filmszene Erarbeitet (Praxis Film)*, vol. 2. Köln, Germany: Herbert von Halem Verlag, 2018.
[50] D. C. Brown, "Decentering distortion of lenses," *Photogrammetric Eng. Remote Sens.*, vol. 32, no. 6, 1966.
[51] J. G. Fryer, "Lens distortion for close-range photogrammetry," *Photogramm. Eng. Remote Sens.*, vol. 52, no. 1, pp. 51–58, 1986.
[52] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.Audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7124–7128.
[53] S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "PyTorch," in *Programming With TensorFlow*. Berlin, Germany: Springer, 2021, pp. 87–104.
[54] Theatiner, *A Midsummer Night's Dream, William Shakespeare, Gymnasium of Benedictine, Schlomberg, P(Director)*, Gymnasium Benedictine, Meschede, Germany, 2016.
[55] K. Thompson and D. Bordwell, *Film Art: An Introduction*, 8th ed. New York, NY, USA: McGraw-Hill, 2006.
[56] E. T. Hall, *The Hidden Dimension*. New York, NY, USA: Anchor Books, 1969.
[57] J. Soyfer, P. Schlomberg, and T. Krajewski, "Gymnasium of benedictine," Gymnasium Benedictine, Meschede, Germany, Tech. Rep., 2017.
[58] *Methodologies for the Subjective Assessment of the Quality of Television Images, Document Recommendation*, document ITU-R BT 500-14, ITU, Geneva, Switzerland, 2020.
[59] S. Göring, R. R. R. Rao, and A. Raake, "Quality assessment of higher resolution images and videos with remote testing," *Qual. User Exper.*, vol. 8, no. 1, p. 2, Dec. 2023.
[60] R. R. R. Rao, S. Göring, and A. Raake, "Towards high resolution video quality assessment in the crowd," in *Proc. 13th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9465425
[61] H. Kromrey, *Empirische Sozialforschung: Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung*, vol. 1040. Berlin, Germany: Springer, 2013.
[62] J. Friedrichs, *Methoden Empirischer Sozialforschung*. Berlin, Germany: Springer, 1990.
[63] S. Kirchhoff, S. Kuhnt, P. Lipp, and S. Schlawin, *Der Fragebogen*. Berlin, Germany: Springer, 2010.
[64] R. Porst, *Fragebogen: Ein Arbeitsbuch*. Berlin, Germany: Springer, 2013.
[65] Y. Zhu, G. Zhai, K. Gu, and Z. Che, "Closing the gap: Visual quality assessment considering viewing conditions," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.
[66] International Telecommunication Union, *Subjective Video Quality Assessment Methods For Multimedia Applications*, Recommendation document ITU-T P.910, Apr. 2008.
[67] International Telecommunication Union, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500-12, 2009.
[68] German Federal Statistical Office. (2019). *Population Statistics, Genesis-Table: 12411–0005*. [Online]. Available: https://www-genesis.destatis.de/
[69] MAGIX Magazine. (2018). *The Man Behind the Multicam: Eckhard Stoll*. Berlin, Germany. [Online]. Available: https://www.magix.com/ca/magazine/video/eckhard-stoll/

ECKHARD STOLL received the degree in electrical engineering from the University of Karlsruhe. He carried out his first "quasi"-multi-camera productions in the 1980s. He is currently the Artistic Director of the Audio Visual Media Center, University of Applied Sciences Südwestfalen, teaches in the field of media production. He is also the Production Manager of Multi-Camera Productions. In cooperation with TU Ilmenau, he researches in camera-based production technology. With one camera, four different performances of a play were recorded from four different camera positions, resulting in a multi-camera edit in post-production.

STEPHAN BREIDE received the Ph.D. (Dr.-Ing.) degree in electrical engineering with a focus on communications engineering, communications systems engineering, and television engineering from the Technical University of Braunschweig. He teaches as a Full Professor with the South Westphalia University of Applied Sciences in the field of communication services and applications. He is currently the Head of the Audio Visual Media Center. His focus is on multimedia applications and digital communication networks and the improvement of internet coverage in fixed wired broadband.

STEVE GÖRING received the B.Sc. and M.Sc. degrees in computer science from TU Ilmenau and the Ph.D. degree in visual quality prediction using machine learning, in 2022. His focus is also on data analysis problems for video quality models and video streams. In 2016, he was with the Audiovisual Technology Group. He was with the Big Data Analytics Group, Bauhaus University Weimar. He is currently working as a Computer Scientist with the Audiovisual Technology Group, TU Ilmenau. His specializations are data analytics/machine learning, video quality, and distributed communication/information systems.

ALEXANDER RAAKE (Member, IEEE) received the Ph.D. (Dr.-Ing.) degree from the Electrical Engineering and Information Technology Faculty, Ruhr-Universitat Bochum, in 2005, with the book Speech Quality of VoIP. He has joined TU Ilmenau, in 2015, as a Full Professor, where he heads the Audiovisual Technology Group. From 2005 to 2015, he held a Senior Researcher, an Assistant Professor, and a later Associate Professor positions with the An-Institut T-Laboratories, TU Berlin, a joint venture between Deutsche Telekom AG and TU Berlin, heading the Assessment of IP-Based Applications Group. From 2004 to 2005, he was a Postdoctoral Researcher with LIMSI-CNRS, Orsay, France. His research interests include audiovisual and multimedia technology, speech, audio, and video signals, human audiovisual perception, and quality of experience. Since 1999, he has been involved with the ITU-T Study Group 12's Standardization work on QoS and QoE assessment methods. He is a member of the Acoustical Society of America, the AES, VDE/ITG, and DEGA.