Proc. AI4TV Workshop @ ACM MM 2020. Author's accepted version. The final publication is at https://doi.org/10.1145/3422839.3423061

Video Analysis for Interactive Story Creation: The Sandmännchen Showcase

Miggi Zwicklbauer
miggi.zwicklbauer@rbb-online.de
Rundfunk Berlin-Brandenburg
Berlin, Germany

Willy Lamm
willy.lamm@rbb-online.de
Rundfunk Berlin-Brandenburg
Berlin, Germany

Martin Gordon
martin.gordon@rbb-online.de
Rundfunk Berlin-Brandenburg
Berlin, Germany

Konstantinos Apostolidis
kapost@iti.gr
CERTH-ITI
Thessaloniki, Greece

Basil Philipp
basil.philipp@genistat.ch
Genistat
Zürich, Switzerland

Vasileios Mezaris
bmezaris@iti.gr
CERTH-ITI
Thessaloniki, Greece

ABSTRACT

This paper presents a method to interactively create a new Sandmännchen story. We built an application which is deployed on a smart speaker, interacts with a user, selects appropriate segments from a database of Sandmännchen episodes and combines them to generate a new story that is compatible with the user requests. The underlying video analysis technologies are presented and evaluated. We additionally showcase example results from using the complete application, as a proof of concept.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → **Information retrieval**; **Multimedia information systems**; **Recommender systems**; • **Human-centered computing** → **Interaction design**; • **Applied computing** → **Publishing**.

KEYWORDS

Machine learning; Sandmännchen; video analysis; smart speaker

ACM Reference Format:
Miggi Zwicklbauer, Willy Lamm, Martin Gordon, Konstantinos Apostolidis, Basil Philipp, and Vasileios Mezaris. 2020. Video Analysis for Interactive Story Creation: The Sandmännchen Showcase. In *2nd International Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV'20), October 12, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3422839.3423061

1 INTRODUCTION

The days of a passive public depending upon a handful of selected broadcasters for their information and entertainment are long gone. Thanks to the internet, and thanks to smartphones in particular, users can access the content they want, where and when they want it. Content covering every topic and niche is today available in every format, whether audio, video, image or text. Content can now be individually tailored by broadcasters / professional content creators for a particular playout channel and a relevant target group. Professional content creators and owners can create new, or reinvent existing, broadcast channels to successfully find an audience for their content.

In the ReTV project we have intensively explored and researched[1] how end users can benefit from AI-based recommendation and user profiling systems. The result is the "4u2" use case[2], which aims to provide consumers with quick and easy access to personalised content from broadcasters and media archives via novel publication channels. To realize and test this use case, we have developed an application for use with smart speakers. This application, the Abendgruß, is based on the well-known children's programme "Unser Sandmännchen"[3], from Rundfunk Berlin-Brandenburg (rbb)[4], which is a seven-minute show broadcast daily. Targeted at pre-school children, "Unser Sandmännchen" accompanies children to bed with a bedtime story at 18:00.

Sandmännchen episodes are simply structured: there is always a framing story, consisting of an intro and an outro (the Sandmännchen arrives and leaves a selected setting in a specified way), which surrounds a main story, usually an adventure of one or more of his friends. There is a continuously growing amount of archive material from which these structural elements can be drawn.

Currently, Sandmännchen episodes are created primarily for TV broadcasting. This means that the selection of production elements (i.e., which framing story to use, which main story elements to combine with which other) depends strongly on the requirements of rbb's daily (live) TV programme. Live broadcasting requires that Sandmännchen elements are limited by available broadcast time, and rbb's on-demand platforms, such as the Mediathek, host only these TV versions of the programme.

Here the Abendgruß application for smart speakers comes into play. Thanks to the app, there are unrestricted possibilities for the creation of personalised Sandmännchen episodes. These can be created according to user preference, and are no longer defined by the requirements of live TV broadcast. Users create their own

AI4TV'20, October 12, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8146-8/20/10…$15.00
https://doi.org/10.1145/3422839.3423061

[1] https://retv-project.eu/mdocs-posts/d6-1-requirements-for-consumer-use-case/
[2] https://retv-project.eu/4u2-personalised-ai-driven-content/
[3] https://www.sandmann.de/elternseite/unser_sandmaennchen/
[4] rbb is the public service broadcaster for Germany's capital Berlin and the surrounding federal state of Brandenburg: https://www.rbb-online.de//unternehmen/der_rbb/

Proc. AI4TV Workshop @ ACM MM 2020. Author's accepted version. The final publication is at https://doi.org/10.1145/3422839.3423061

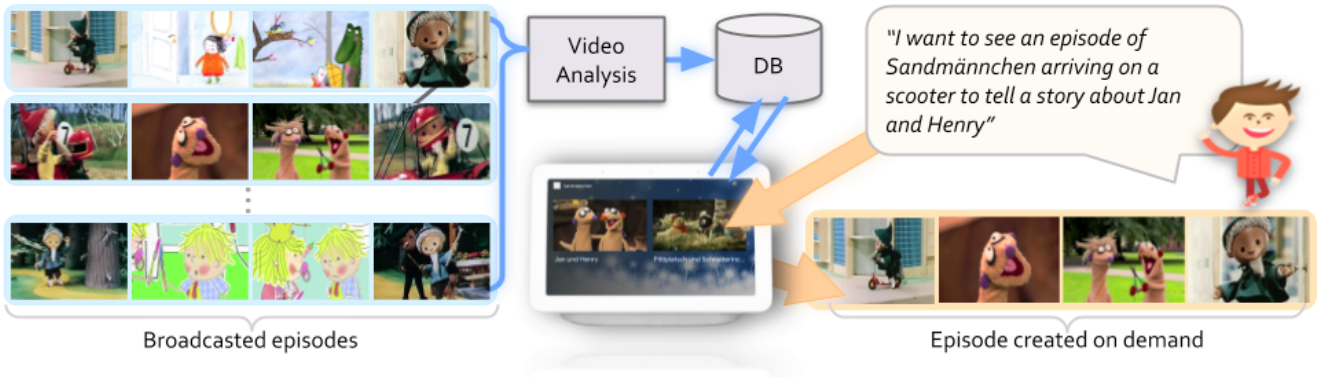

Figure 1: Our concept for a Sandmännchen custom story generation application.

Sandmännchen episode completely free from editorial restrictions, and their choices are driven only by their personal preferences. The Abendgruß application is supported by functions that we developed for video adaptation and re-purposing that are based on Artificial Intelligence (AI) techniques, most notably deep neural networks.

2 SANDMÄNNCHEN VIDEO ANALYSIS

2.1 Related Work

To be able to properly (and with minimal human intervention) select segments from a Sandmännchen episode, we constructed a video analysis framework. Such video analysis starts with temporally fragmenting the video to meaningful segments, and subsequently employing annotation methods for generating high-level metadata (e.g., concept / object detection results).

There exists a plethora of video temporal fragmentation methods. Most of them deal with segmenting a video to elementary structural units called shots, which are defined as groups of consecutive frames captured without interruption by a single camera [1]. Older methods employed handcrafted rules based on color features [17, 25, 39, 40], local descriptors [1, 4, 6, 8, 22] or a fusion of such features [10, 15]. Due to the success of deep convolutional neural networks (DCNN) on various computer vision tasks, most recent attempts are based on the use of DCNNs [9, 11, 18, 37, 38]. Beyond techniques that segment videos to shots, there is a substantial number of works that deal with the fragmentation at different granularities (coarser or finer), with some methods performing video decomposition into coarse and semantically-coherent temporal segments, known as scenes [5, 16, 29, 32], while others working on the finer side of fragmentation and further decomposing shots into visually coherent parts that correspond to individual video capturing activities, usually referred to as sub-shots [2]. However, none of the aforementioned methods have direct application in our scenario: we need to detect specific parts of the structure of Sandmännchen episodes. For this, we need to design a domain-specific method, like [43] where soccer games are segmented according to the specific semantics of this sport, or [21] where the individual stories (on different topics) within a TV news broadcast are determined.

Regarding concept-based annotation, and again as a result of the widespread use of DCNNs, the focus has moved from employing Support Vector Machines (SVM) [3, 41] or local descriptors [23, 24] to an explosion of DCNN model architectures for concept detection [13, 14, 30, 33, 34, 44] as well as object detection [12, 19, 20, 27] in image / video. Of particular interest are methods that can be used to "adjust" (i.e., retrain) DCNN models that were trained for a visual annotation task on one dataset, to a new, considerably different dataset, a task known as "finetuning" [26, 28, 35].

Our goal is to construct a method to fragment a Sandmännchen episode taking into consideration the peculiarities of the application domain as discussed in Section 1, i.e., the presence of an intro / outro and a main story part, and annotate the main story part with the main involved character. For this we will adjust and employ methods of the literature, combining them to construct a complete Sandmännchen story creation framework.

2.2 Temporal Segmentation

Each Sandmännchen episode has three parts. In chronological order, these are:

(1) The introductory part, where the Sandmännchen arrives in a different vehicle and setting every time, enters a room with children, and starts narrating his story.
(2) The main part of the episode is the story that the Sandmännchen narrates. The Sandmännchen is not visible - it is assumed we are fully immersed in his story. This part deals with a different character, or set of characters, each time.
(3) The closing part, when the Sandmännchen has finished his narration and he is leaving in the same way that he arrived.

For the temporal segmentation of Sandmännchen episodes, following visual inspection of a large set of episodes, we decided to detect the intro transition (i.e., transition from the introductory part to the main story) and the outro transition of an episode (i.e, transition from the main story to the closing part of the episode). In most cases, the frames around the intro and outro transitions contain a characteristic camera zooming in and out from a screen, respectively. The screen is different every time, sometimes being a TV screen, other times being just a projection on wall. The zooming is accompanied with a fading transition, where in most cases the camera zooming fades out to a white frame.

We performed a statistical analysis of the (temporal) position and the duration of the intro and outro transitions on a set of randomly selected 80 episodes. We calculated the mean position of the intro

transition to be at second 84 from the start of the video, with a standard deviation of 39 seconds. Similarly, the mean position of the outro was found to be at second 346, with a standard deviation of 151 seconds. The duration of the transitions also varied among episodes; they were on average 1.4 seconds long, with a standard deviation of little over 1 second. From this analysis it becomes clear that using a simple heuristic for segmenting a Sandmännchen episode to its three main parts based on, e.g., just time information, would fail.

We first implemented a DCNN-based method to segment the video to shots by adopting and extending the method of [9]. Our extension marginally improves the accuracy of the original method when applied to Sandmännchen videos, and concerns two directions:

(1) The inclusion of a post-processing stage similar to the technique used in [1] to analyse the computed similarity scores between consecutive frames. Specifically, the time series formed by the shot transition probabilities is first smoothed using a moving average filter with a temporal window of 5 frames. Then, the first order derivative of the smoothed time series is calculated to discover the local minima and maxima. Each discovered local maximum is considered a shot transition.
(2) The introduction of a trick to use a larger temporal window for the input to the network without affecting speed. Specifically, instead of analysing 10 consecutive frames, we choose to analyse frames using a quadratic incremental step, i.e., the inference of our model for frame with index x is based on the analysis of frames with indices $x-8$, $x-4$, $x-2$, $x-1$, x, $x+1$, $x+2$, $x+4$, $x+8$, $x+16$. This way, we allow the model to look at a larger temporal window while still using just 10 frames as input to our model (i.e., the time efficiency remains unaffected).

We also trained a Random Forest classifier on a set of simple (and cheap to compute) frame features that, based on our intuition, are able to capture the variations of the sought transitions. These features are the following:

- ECR: We compute the Edge Change Ratio (ECR), which represents the amplitude of edge changes between two consecutive frames [42].
- Homogeneity: We convert the video frame to grayscale and we compute the range of the pixel intensity values, as a means to quantify the visual information contained in the frame.
- Blackness: We convert the input frame to grayscale and we compute the average of all pixel intensity values, to quantify how black the frame appears to be.
- Whiteness: We convert the input frame to the HLS [31] colorspace and we compute the range of all pixel values in the L (i.e., "lightness") component, to quantify how white the frame appears to be.
- Blurriness: We calculate the variance of the Laplacian of each frame, a well-known practice in image analysis, to quantify how blurry the frame is.

To detect the three parts of a Sandmännchen episode (i.e., the intro, outro and main story) we first perform temporal segmentation of the input video to shots. In parallel, we employ our Random Forest classifier model trained on the above-mentioned image analysis features for classifying the video frames into two classes: "normal frame" and "transition frame". Then, taking a further step and not relying solely on this frame-level prediction (i.e., on whether a frame was correctly classified as belonging / not-belonging to a transition), we incorporate the results of shot segmentation for making a video-level prediction for the intro and outro transitions. This is accomplished by employing the following simple domain rules:

- For a frame to be considered a "transition frame", besides having a high inferred probability from the Random Forest classifier, it must also belong to either the first or the last 1/3 of the video. This rule was employed since the main story part on all analysed Sandmännchen episodes was the largest part of each episode, and of course is always in the middle of the video.
- For a frame to be considered a "transition frame", it must additionally have a temporal distance of no more than four seconds from a shot boundary. We employed this rule since the longest transition was observed to be four seconds long, and the intro and outro transitions are always marked as a shot change by the shot segmentation module.

After the application of these additional domain rules we select the shot that contains the highest ranked "transition frame" and belongs to the first 1/3 portion of the video as the last shot of the introductory part. Consequently, we select the shot that contains the highest ranked "transition frame" and belongs to the last 1/3 portion of the video as the last shot of the main story.

2.3 Character Annotation

As discussed in Section 2.2, the Sandmännchen appears in the introduction and closing part. The main story deals with a different protagonist each time. The protagonist can be a single character (e.g., Kalli - a blonde boy) or a character set, which will always appear together throughout the whole main story (e.g., Rita und das Krokodil - Rita and her very hungry friend, Crocodile, who lives in the bathtub). For the sake of brevity, in the sequel the term character may refer to a single character or a character set. There are 30 characters in total. We decide to employ a DCNN model of the EfficientNet state-of-the-art architecture [36]. We utilized the weights of an EfficientNet instance trained on the 1000 classes of the ImageNet challenge [7] as the initial weights of our model and then fine-tuned it to be able to detect the character of the main story with a similar technique to [26]. We decided to select a subset of 11 characters out of the total 30 characters since we consider this as a good starting point for getting our application to life. The 11 detectable characters are: 1) Herr Fuchs und Frau Elster, 2) Jan und Henry, 3) Kalli, 4) Der kleine König, 5) Der kleine Rabe Socke, 6) Die Moffels, 7) Meine Schmusedecke, 8) Pittiplatsch, Schnatterinchen und Moppi, 9) Plumps, 10) Pondorondo, and 11) Rita und das Krokodil.

Our model annotates each frame of an input video with the detection score for each one of the 11 characters. However, we do not rely solely on this frame-level character predictions but we also calculate a video-level prediction. For this we perform majority voting

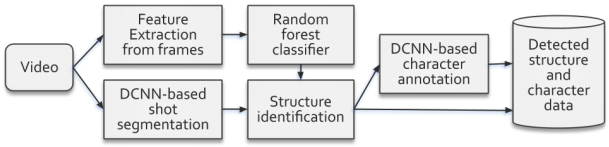

Figure 2: The framework of video analysis methods used in our application.

over the frame-level predictions, since a Sandmännchen episode deals with a single character in its main story part. Although the audio stream could also have been used for performing this classification, our results indicate that this is not needed, as perfect results are observed at the video level (following the majority voting) by using just the visual classifiers. The framework of all utilized video analysis methods described in Section 2.2 as well as in the current Section, is summarized in Fig. 2.

Regarding the selection of a framing story, our application (Section 3) will initially rely on a finite set of previously-annotated episodes. Therefore, it was deemed that for the identification of the vehicle that Sandmännchen uses in the intro and outro sections of an episode there is no need for developing a video analysis method to automate it, at least for this first phase of the Abendgruß application.

2.4 Video analysis service

The video analysis techniques discussed in previous sections have all been incorporated into a video analysis component. This component is deployed as a REST service that: a) retrieves a video file, b) performs the temporal segmentation of a Sandmännchen episode, c) analyzes the main part to identify the main character, and d) stores the results in a JSON-structured file which can be downloaded using a specific type of call.

The REST service works in an asynchronous way, i.e., through a 3-step process. The first step relates to an HTTP POST call that enables the submission of a video for analysis and the initiation of a relevant session in the REST service. The second step is associated to an HTTP GET call that queries the status of the initialized session and the progress of the analysis. Finally, the third step is performed by another HTTP GET call that enables the retrieval of the results of a successfully completed session.

3 SANDMÄNNCHEN APPLICATION

3.1 Application overview

Since our focus is on video content, Abendgruß is designed primarily for the use with smart speakers with display. The first prototype was developed as an action for Google Assistant, focusing on the Google Nest Hub. Here's how it works:

- To start Abendgruß, the user has to say "OK, Google, mit Abendgruß sprechen (OK, Google, speak to Abendgruß.)".
- The Nest Hub answers: "In Ordnung, ich starte die Testversion von Abendgruß (All right, I'm starting the test version of Abendgruß.)."
- The application opens.
- The user sees the start screen and gets a welcome combined with a call to action: "Hallo! Um deinen eigenen Abendgruß zu sehen, sage das Wort Abendgruß" (Hello! To watch your own Abendgruß, say the word "Abendgruß".).
- After saying "Abendgruß", two options are shown. First, the user can choose how the Sandmännchen should arrive. For example, "Zu Fuß oder auf dem Elefanten? (By foot or on the elephant?)". In other words, the framing story (see Section 1) is defined in this step.
- Secondly, the user determines her/his main story by answering the question "Und welche Geschichte möchtest Du heute sehen?" (And what story do you want to see today?). Again two options are presented, e.g., "Rita und das Krokodil oder Die Moffels? (Rita and the crocodile or The Moffels?)".
- The Abendgruß application finally shows an automatically-generated Sandmännchen video which consists of the respective framing and main story elements.

3.2 Communication with audio and video analysis APIs

The Abendgruß application aims to be conversational, which means that it needs to deal with users speaking a command in many different ways. A user might just say "Jan" instead of "Jan und Henry". Mapping all possible inputs to clearly defined API calls is usually done with a chatbot framework. We use Google's Dialogflow[5], since it is tightly integrated with the Google Assistant.

When a user speaks to the Abendgruß application on the Google Assistant, their commands are sent to Dialogflow and mapped to API calls. Those calls are then sent to the Abendgruß API, which either returns options for the user to choose from, or the customized video in the final step. If no mapping is possible, Dialogflow will tell the user that their command could not be understood.

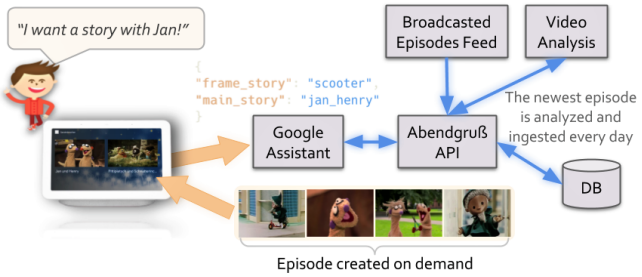

Figure 3: Overview of the Abendgruß architecture.

The Abendgruß API periodically checks if new Sandmännchen episodes have been published, by monitoring selected Web sources. If this is the case, they are sent through the Video analysis service, and the results are stored in the Abendgruß database, ready to be integrated into future stories. See Fig. 3 for an overview on how the different software services work together.

	Training dataset	*Test dataset*
Videos	N/A	35
Shots	N/A	1453
Frames analyzed	14375	238535
Total duration (seconds)	N/A	9541

Table 1: Dataset specifications for training and evaluating the Sandmännchen character identification.

4 EXPERIMENTS AND RESULTS

4.1 Video Analysis Results

Our models for the identification of the main story character were trained and evaluated on a dataset we manually curated. The specifications of this dataset are reported in Table 1. The training dataset is in the form of a set of selected frames, while the testing dataset consists of videos in order to be able to evaluate the whole character identification process, i.e., including the video-level character inference.

	Normal frame	*Transition frame*
Normal frame	0.94	0.06
Transition frame	0.01	0.99

Table 2: Confusion matrix for Sandmännchen episode structure segmentation (frame-level identification results).

For the Sandmännchen episode structure segmentation, we compiled a training dataset of 50 Sandmännchen episodes, by manually annotating the structure of each episode. We also annotated 30 Sandmännchen videos to create a testing dataset. In Table 2 we report the frame-level identification results in the form of a confusion matrix. Overall, for the detection of the transitions, using the Random Forest classifier at frame-level we achieve 88.54% F-score. Employing the additional domain rules, as discussed in Section 2.2, for the video-level prediction of transitions we reach a 91.67% F-score.

Regarding the Sandmännchen character identification frame-level predictions, in Tables 3 and 4 we report the evaluation results and confusion matrix, respectively. We observe that by just using the DCNN model for the frame-level predictions, there are classes that perform very well (e.g., Herr Fuchs und Frau Elster with 92.8% accuracy)) but also classes with noticeably bad performance (e.g., Meine Schmusedecke with 51.3% accuracy). Our intuition for explaining this sometimes low accuracy, besides the varying difficulty of detecting each character due to its specific characteristics, is that the video frames that are analysed do not always depict the main character (see Fig. 4 for indicative examples). However, we should highlight that after employing majority voting to infer video-level predictions, as discussed in Section 2.3, we achieve a perfect score on our test dataset, i.e., 100% accuracy for all classes.

[5]https://cloud.google.com/dialogflow/

Character	*F1-score*	*Accuracy*
HerrFuchs & ...	0.68	0.93
Jan & Henry	0.68	0.60
Kalli	0.77	0.84
KleineKonig	0.86	0.89
KleineRabeSocke	0.73	0.57
Luzi & Moffels	0.68	0.58
MeineSchmusedecke	0.49	0.51
Pittiplatsch & ...	0.64	0.56
Plumps	0.79	0.91
Pondorondo	0.83	0.99
Rita & Krokodil	0.71	0.98

Table 3: Sandmännchen character frame-level identification results per class.

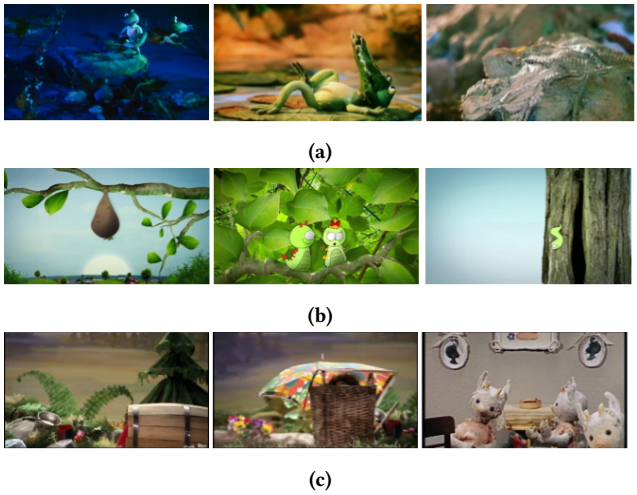

Figure 4: Sample frames from the main story part of Sandmännchen episodes where the main character is not depicted. The frames in (a) are from a Herr Fuchs und Frau Elster episode, in (b) from a Kalli episode, and in (c) from a Pittiplatsch, Schnatterinchen und Moppi episode.

4.2 Sandmännchen Application Results and Examples

We have already presented our Abendgruß application prototype to project stakeholders and a wider audience at various fairs such as the IFA[6]. A small user study with a questionnaire for parents regarding the concept of the application, was conducted. The aim was to find out to what extent the application meets the expectations of parents and their children, in the context of a smart speaker application for children, and where adjustments may still be necessary. The feedback was very positive throughout. In particular, the idea of enabling the end user to interact directly with the content of a broadcaster in order to personalise it was met with great approval. The fact that the smart speaker was the device of choice was considered reasonable and relevant to the times. With regard to the target group of the Abendgruß, i.e., pre-school children or parents with

[6]https://www.ifa-berlin.com/en/

	HerrFuchs & ...	*Jan & Henry*	*Kalli*	*KleineKonig*	*KleineRabeSocke*	*Luzi & Moffels*	*MeineSchmus.*	*Pittiplatsch&...*	*Plumps*	*Pondorondo*	*Rita & Krokodil*
HerrFuchs & ...	92.8	0.0	0.0	0.0	0.0	0.0	0.0	7.1	0.0	0.0	0.0
Jan & Henry	0.58	59.5	9.25	1.7	0.0	0.5	6.9	4.0	2.3	3.4	11.5
Kalli	1.1	0.3	83.7	1.9	0.0	1.5	5.0	0.3	1.1	0.3	4.2
KleineKonig	0.0	0.0	0.0	88.5	0.0	0.0	4.1	0.0	0.0	1.0	6.2
KleineRabeSocke	0.0	0.0	5.1	1.7	57.2	0.0	14.5	0.0	0.8	0.8	19.6
Luzi & Moffels	3.11	2.33	11.2	1.9	0.0	57.9	5.8	2.7	4.2	6.2	4.3
MeineSchmus.	1.8	5.4	12.6	0.0	0.0	14.4	51.3	6.3	1.8	2.7	3.6
Pittiplatsch & ...	5.8	7.7	12.3	0.6	0.0	6.4	1.3	55.8	6.4	1.3	1.9
Plumps	0.00	1.2	0.0	0.0	0.0	0.0	0.0	6.2	91.2	1.2	0.0
Pondorondo	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	98.7	0.0
Rita & Krokodil	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	97.9

Table 4: Confusion matrix for Sandmännchen characters frame-level identification.

Figure 5: Screenshot of the smart speaker screen for the available options on how the Sandmännchen arrives.

Figure 6: Screenshot of the smart speaker screen for the available options on the main story to include in the constructed video.

pre-school children, the application also convinced the audience with its simple, child-friendly operation. Moreover, the approach of using AI techniques to realise the concept generated great interest and was considered highly innovative. In addition, rbb confirmed that the idea of the Abendgruß for smart speakers fits perfectly into the broadcaster's plans to 1) open up new distribution channels and 2) expand its digital offer for the Sandmännchen.

Additionally, we should clarify that the application always allows the user to select between two options for each of the into/outro and main story, and these two options are selected randomly each time the application is used; thus, the user cannot select the exact same characters/stories again and again. Offering a variety of episodes, we ensure that the educational effect of the Sandmännchen series is not lost.

A typical way to use our developed application starts with the question about how you want the Sandmännchen to arrive. We provide two randomly-selected options, as can be seen in Fig. 5. After the user answers, the next screen will provide another two randomly-selected options for the main story, as illustrated in Fig. 6. We provide below three indicative usage examples of the developed application, as a proof of concept.

Figure 7: Interactive creation of a Sandmännchen story - Example #1.

Example #1.

APPLICATION: Wie soll das Sandmännchen heute ankommen - zu Fuß oder auf einem fliegenden Pferd (How shall the Sandmännchen arrive today - by foot or on a flying horse)?

USER: Zu Fuß (By foot).

APPLICATION: Und welche Geschichte möchtest Du heute sehen - Der kleine König oder Die Moffels (And what story do you want to see today - The little King or The Moffels)?

USER: Die Moffels (The Moffels).

The application will select the appropriate videos and segments from the database, as seen in Fig. 7, and will start playing the video of the constructed story after saying:

APPLICATION: Hier ist Dein Video (Here is your video).

Example #2.

APPLICATION: Wie soll das Sandmännchen heute ankommen - auf dem Elefanten oder zu Fuß (How shall the Sandmännchen arrive today - on the elephant or by foot)?

Figure 8: Interactive creation of a Sandmännchen story - Example #2.

USER: Auf dem Elefanten (On the elephant).

APPLICATION: Und welche Geschichte möchtest Du heute sehen - Kalli oder Herr Fuchs und Frau Elster (And what story do you want to see today - Kalli or Mr Fox and Ms Magpie)?

USER: Herr Fuchs und Frau Elster (Mr Fox and Ms Magpie).

The application will again select the appropriate videos and segments from the database, as seen in Fig. 8, and will start playing the video of the constructed story after saying "Hier ist Dein Video (Here is your video)".

Figure 9: Interactive creation of a Sandmännchen story - Example #3.

Example #3.

APPLICATION: Wie soll das Sandmännchen heute ankommen - auf dem Motorrad oder mit dem Ballon (How shall the Sandmännchen arrive today - on the motorcycle or by balloon)?

USER: Auf dem Motorrad (On the motorcycle).

APPLICATION: Und welche Geschichte möchtest Du heute sehen - Meine Schmusedecke oder Pittiplatsch und Schnatterinchen (And what story do you want to see today - My Cuddle Blanket or Pittiplatsch and Schnatterinchen)?

USER: Pittiplatsch und Schnatterinchen (Pittiplatsch and Schnatterinchen).

Similarly to the previous examples, the application will select the appropriate videos and segments from the database, as seen in Fig. 9, and will start playing the video of the constructed story after saying "Hier ist Dein Video (Here is your video)".

5 CONCLUSIONS AND NEXT STEPS

We presented an application for smart speaker devices equipped with a display to interactively create custom videos of Sandmännchen episodes. We presented the underlying video analysis technologies that enable the automatic generation of such videos, and performed experimental tests to evaluate their effectiveness on our specific usage scenario. The application as a whole was evaluated in a qualitative way and examples of use were reported as a proof of concept.

With respect to future work, our goal for the next version of the application, aimed at Amazon Alexa, is to adjust / expand all the video analysis components so that the creation of more diverse personalised episodes using voice commands can be achieved. Specifically, we plan to expand the set of identifiable characters so as to include all of the Sandmännchen friends, and introduce a method to also automate the annotation of the intro and outro sections of an episode with information on the vehicle that the Sandmännchen uses.

ACKNOWLEDGMENTS

This work was supported by the EU's Horizon 2020 research and innovation programme under grant agreement H2020-780656 ReTV.

REFERENCES

[1] Evlampios Apostolidis and Vasileios Mezaris. 2014. Fast shot segmentation combining global and local visual descriptors. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6583–6587.

[2] Konstantinos Apostolidis, Evlampios Apostolidis, and Vasileios Mezaris. 2018. A motion-driven approach for fine-grained temporal segmentation of user-generated videos. In *International Conference on Multimedia Modeling*. Springer, 29–41.

[3] Yusuf Aytar, O Bilal Orhan, and Mubarak Shah. 2007. Improving semantic concept detection and retrieval using contextual estimates. In *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 536–539.

[4] Junaid Baber, Nitin Afzulpurkar, Matthew N Dailey, and Maheen Bakhtyar. 2011. Shot boundary detection from videos using entropy and local descriptor. In *2011 17th International conference on digital signal processing (DSP)*. IEEE, 1–6.

[5] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2015. A Deep Siamese Network for Scene Detection in Broadcast Videos. *CoRR* abs/1510.08893 (2015). arXiv:1510.08893 http://arxiv.org/abs/1510.08893

[6] Edward JY Cayllahua Cahuina and Guillermo Camara Chavez. 2013. A new method for static video summarization using local descriptors and video temporal segmentation. In *2013 XXVI Conference on Graphics, Patterns and Images*. IEEE, 226–233.

[7] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. 2009. Imagenet: A large-scale hierarchical image database. In *In CVPR*.

[8] Anderson Carlos Sousa e Santos and Helio Pedrini. 2017. Shot boundary detection for video temporal segmentation based on the weber local descriptor. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 1310–1315.

[9] Michael Gygli. 2018. Ridiculously fast shot boundary detection with fully convolutional neural networks. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–4.

[10] Rachida Hannane, Abdessamad Elboushaki, Karim Afdel, P Naghabhushan, and Mohammed Javed. 2016. An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram. *International Journal of Multimedia Information Retrieval* 5, 2 (2016), 89–104.

[11] Ahmed Hassanien, Mohamed Elgharib, Ahmed Selim, Sung-Ho Bae, Mohamed Hefeeda, and Wojciech Matusik. 2017. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks. *arXiv preprint arXiv:1705.03281* (2017).

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 630–645.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition. 4700–4708.

[15] Kai Jin, Hong Cai Feng, Qi Feng, and Chi Zhang. 2013. Shot Boundary Detection Algorithm Based on Multi-Feature Fusion. In *Applied Mechanics and Materials*, Vol. 347. Trans Tech Publ, 3866–3871.

[16] Rodrigo Mitsuo Kishi, Tiago Henrique Trojahn, and Rudinei Goularte. 2019. Correlation based feature fusion for the temporal video scene segmentation task. *Multimedia Tools and Applications* 78, 11 (2019), 15623–15646.

[17] Onur Küçüktunç, Uğur Güdükbay, and Özgür Ulusoy. 2010. Fuzzy color histogram-based video segmentation. *Computer Vision and Image Understanding* 114, 1 (2010), 125–134.

[18] Rui Liang, Qingxin Zhu, Honglei Wei, and Shujiao Liao. 2017. A video shot boundary detection approach based on CNN feature. In *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, 489–494.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[21] Zhu Liu and Yuan Wang. 2018. TV news story segmentation using deep neural network. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–4.

[22] Jharna Majumdar, Dhanush M Adiga, and MP Ashray. 2019. Comparison of video shot detection methods using higher order local descriptor. In *Proceedings of the Third International Conference on Advanced Informatics for Computing Research*. 1–5.

[23] Foteini Markatopoulou, Vasileios Mezaris, and Ioannis Patras. 2016. Ordering of visual descriptors in a classifier cascade towards improved video concept detection. In *International Conference on Multimedia Modeling*. Springer, 874–885.

[24] Foteini Markatopoulou, Nikiforos Pittaras, Olga Papadopoulou, Vasileios Mezaris, and Ioannis Patras. 2015. A study on the use of a binary local descriptor and color extensions of local descriptors for video concept detection. In *International Conference on Multimedia Modeling*. Springer, 282–293.

[25] Zhenxing Niu, Xinbo Gao, Dacheng Tao, and Xuelong Li. 2008. Semantic video shot segmentation based on color ratio feature and SVM. In *2008 International Conference on Cyberworlds*. IEEE, 157–162.

[26] Nikiforos Pittaras, Foteini Markatopoulou, Vasileios Mezaris, and Ioannis Patras. 2017. Comparison of Fine-Tuning and Extension Strategies for Deep Convolutional Neural Networks. In *MultiMedia Modeling*, Laurent Amsaleg, Gylfi Þór Guðmundsson, Cathal Gurrin, Björn Þór Jónsson, and Shin'ichi Satoh (Eds.). Springer International Publishing, Cham, 102–114.

[27] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[28] Angie K Reyes, Juan C Caicedo, and Jorge E Camargo. 2015. Fine-tuning Deep Convolutional Networks for Plant Recognition. *CLEF (Working Notes)* 1391 (2015), 467–475.

[29] Daniel Rotman, Dror Porat, Gal Ashour, and Udi Barzelay. 2018. Optimally grouped deep features using normalized cost for video scene detection. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 187–195.

[30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[31] G Saravanan, G Yamuna, and S Nandhini. 2016. Real time implementation of RGB to HSV/HSI/HSL and its reverse color space models. In *2016 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 0462–0466.

[32] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. 2011. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 8 (2011), 1163–1177.

[33] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[35] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* 35, 5 (2016), 1299–1312.

[36] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *CoRR* abs/1905.11946 (2019). arXiv:1905.11946 http://arxiv.org/abs/1905.11946

[37] Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. 2018. Fast video shot transition localization with deep structured models. In *Asian Conference on Computer Vision*. Springer, 577–592.

[38] Dalton Meitei Thounaojam, Thongam Khelchandra, Kh Singh, Sudipta Roy, et al. 2016. A genetic algorithm and fuzzy logic approach for video shot boundary detection. *Computational intelligence and neuroscience* 2016 (2016).

[39] Efthymia Tsamoura, Vasileios Mezaris, and Ioannis Kompatsiaris. 2008. Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *2008 15th IEEE International Conference on Image Processing*. IEEE, 45–48.

[40] Zhi-min Xiao, Kun-hui Lin, Chang-le Zhou, and Qiang Lin. 2008. Shot Segmentation Based on HSV Color Model [J]. *Journal of Xiamen University (Natural Science)* 5 (2008).

[41] Jun Yang, Rong Yan, and Alexander G Hauptmann. 2007. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th ACM international conference on Multimedia*. 188–197.

[42] Ramin Zabih, Justin Miller, and Kevin Mai. 1995. A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. In *Proceedings of the Third ACM International Conference on Multimedia* (San Francisco, California, USA) *(MULTIMEDIA '95)*. Association for Computing Machinery, New York, NY, USA, 189–200. https://doi.org/10.1145/217279.215266

[43] YZ Zhang, JY Wang, and YW Dai. 2009. Soccer video shot segmentation based on self-adapting dual threshold and dominant color percentage. *J Nanjing Univ Sci Technol (Nat Sci)* 33, 4 (2009), 432–437.

[44] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8697–8710.