AI-DRIVEN SMART PRODUCTION

Hiroyuki Kaneko, Jun Goto, Yoshihiko Kawai, Takahiro Mochizuki,
Shoei Sato, Atsushi Imai, Yuko Yamanouchi

NHK, Japan

ABSTRACT

NHK has developed a new AI-driven broadcasting technology called "Smart Production" to quickly and accurately gather and analyze diverse types of information from society and deliver information to a wide range of viewers. Smart Production uses artificial intelligence (AI) to analyze diverse types of information obtained from social media and open data as well as know-how related to program production possessed by broadcast stations. This approach makes it possible to extract events and incidents in society and present the results of analysis to producers. In particular, image analysis technology for recognizing objects in video and speech recognition technology for generating transcripts of interviews enable metadata to be automatically generated for video footage. Additionally, to convey information to a wide range of viewers including hearing/visually impaired persons, research and development is progressing on technology for automatically converting broadcast data into content that can be easily understood by viewers with special needs.

INTRODUCTION

In recent years, as program material transmission lines become capable of high speed and the capacity of recording media increases, broadcasting stations have become capable of obtaining large amounts of video and audio contents for program creation. Also, as the use of social media becomes widespread, first reports of accidents and incidents and information about social trends now appear on platforms such as Twitter. Furthermore, it is now possible to monitor open data such as sensor information released by municipalities and incorporate such data into news programs. The work of extracting materials needed for programs from vast amounts of video materials and finding information useful for news presentation from social media data has become extremely burdensome for program production members. Also, in order for the produced programs to reach all viewers, including foreign and hearing and visually-impaired persons, it is necessary to convert the format of the contents to suit the viewing and listening environment of the viewer.

Consequently, in collaboration with other departments at NHK, the NHK Science & Technology Research Laboratories is engaged in the research and development of AI-driven content production technologies and universal service that allows human-friendly broadcasting to reach all people, including viewers with hearing and visual impairments and foreign people (Figure 1).

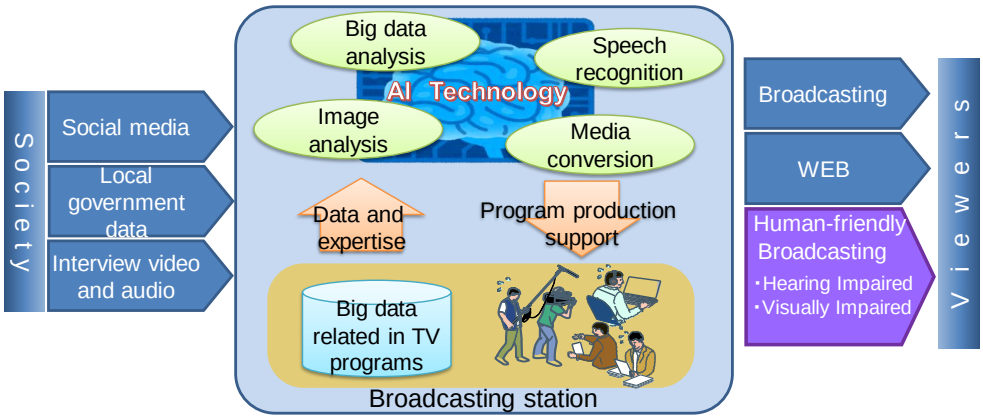

Figure 1 – Smart Production

INTELLIGENT PROGRAM PRODUCTION

Textual big data analysis technologies

We are conducting research and development of technologies to support TV program production by analyzing big data such as information related to programs that broadcasting stations possess and information posted on social networking service (SNS). We introduce here a social media analysis system that obtains information useful for new program production from social media, including Twitter posts (tweets), and categorizes the information, such as the occurrence of fire, traffic accidents, or natural disasters. We also introduce a system that automatically prepares news manuscript drafts about river conditions during heavy rains and typhoons by using broadcasting stations' past news manuscripts and information from river sensors.

(1) Social media analysis system

Broadcasting stations often manually search for information useful for program production from social media and use the information as initial reports after determining their veracity. If a person who happens to be at the scene of an incident or accident can post information about it to SNS, the occurrence of the event can be known more quickly than in the past. A large amount of labor is needed to uncover useful information from the vast amount of SNS postings, placing a large burden on the local production team. We therefore developed a system that learns tweets that had been determined to be useful for news broadcasts by local producers and searches for and presents useful tweets. We have begun field trials in collaboration with local news teams (1) (Figure 2).

This system uses a recurrent neural network to determine whether words appearing in tweets include information useful for news broadcasts. It has learned to sort information into 24 news categories, such as fires and traffic accidents. With this system, it is possible to automate a portion of the work that involves the confirmation of each piece of information by local broadcast producers. The system also accepts feedback from program production members to acquire new data for learning, which is used to maintain and improve the tweet extraction function. We are thus conducting research to improve the accuracy of categorizing newsworthy tweets through the use of image recognition technology to identify objects in images attached to tweets such as fires and fire engines.

Figure 2 – Social media analysis system

Figure 3 – Automated news manuscript creation system

(2) Automated news manuscript creation system

Broadcasting stations gather, analyze, and use for broadcasting content sensor information released by public agencies and local municipalities. Constantly monitoring such high-volume open data and quickly creating news manuscripts for broadcasting contents that include such data are heavy burdens for program production members. We therefore developed a news manuscript creation support system that automatically creates news manuscripts for initial reports of river conditions during heavy rains and other weather phenomena. This system uses information from water level sensors in rivers and previous broadcast news manuscripts (Figure 3). We conducted field trials of this system at local news stations during the 2017 rainy season.

The water level of rivers is obtained as numerical data from the Foundation of River & Basin Integrated Communications, which distributes the data every 10 minutes. The information includes the monitoring location, current water level, and four water level notification thresholds indicating conditions such as flood warning and flood danger.

Using previous broadcast news manuscripts accumulated at news stations, this system uses a neural network to automatically extract fixed expressions and identify river names and news expressions used during warnings of water levels. These expressions are used to create a template. A news manuscript draft is created on the basis of the template and the obtained water level data, which is matched to the warning water level used in the past broadcasts stored at the broadcasting station. By applying revisions, reporters can also create their own original news manuscripts about river conditions.

Video analysis technologies

To enable the creation of programs with high-quality and attractive contents, we are advancing research of video summarization technology and monochrome film colorization technology as video-analysis-driven program production technologies.

(1) Automated video summarization system

To support the production of program preview video and digest video, we are purusing research of technology to produce automated video summarization. We have developed a

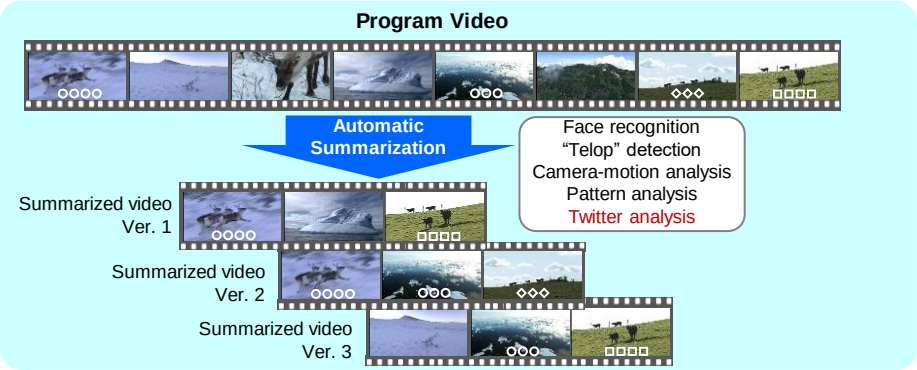

Figure 4 – Automated video summarization system

system that automatically generates a digest video that reflects the various intentions of program production members (2) (Figure 4).

The system allows program production members to freely set weight distributions for various information contents as guides for automated summarization. These contents include "viewer response based on Twitter-analysis," "persons appearing in the scene based on image analysis," and "telop and camerawork." On the basis of these cues, the systems can automatically generate a digest video using particular aspects of materials, such as "display of large telop," "close-up of performers," and "rapid zoom-in." It is also possible to generate a digest video using mass feedback from the viewing audience by analyzing program-related comments posted on SNS.

(2) Automatic colorization technology for monochrome video

As an AI-driven technology to support the efficient production of programs, we have developed a system that automatically converts monochrome film video to color video (3) (Figure 5). By colorizing monochrome films with this technology, conditions during the time of filming can be conveyed with greater freshness.

DNN was trained by using data of video from about 20,000 programs gathered from past TV program videos and color films stored in the NHK archives. Three DNNs for color estimation, color correction, and propagation of color information to adjacent frames were used to automatically convert monochrome video to color video.

Since color correction based on historical factuality is necessary during program production, we have also developed a system that takes into account user-instructed colors when colorizing video. The necessary operations consist of simply clicking several target regions on the image and specifying the color that should be applied or the boundary of the color. In this way, the user can correct colors easily.

Until now, specialists could only color frames by hand one at a time. Several days were required to colorize a video of just several seconds. Using our developed system, it is now possible to shorten the task of colorizing a five-second monochrome film from about 30 minutes to 30 seconds.

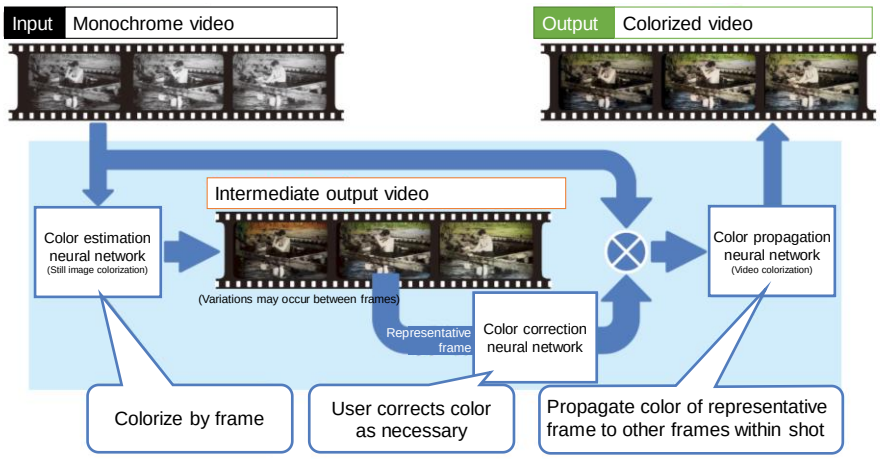

Figure 5 – Overview of automatic colorization technology

Speech recognition technologies

Programs are produced by sifting through a large amount of gathered video materials to find needed information. For this task, transcription of the audio contents of video materials is indispensable so that production members can browse the list of contents easily and view the content itself. A system that produces such transcription quickly and efficiently is thus demanded. We therefore developed a transcription production system that uses speech recognition technology and a user interface for easy correction of the recognition results when viewing (4) (Figure 6).

In order to reduce the labor of the operating procedures, the developed system allows users to quickly access the area he or she wishes to view by displaying thumbnails and major keywords. By synchronizing the display of speech and text on a per-word basis, text revision of the recognition results can be carried out with just a few operations. Also, by creating a web-application-based interface, we allow the system to be accessed anywhere within the broadcasting station.

At present, the departments of several broadcasting stations are conducting trials on revising the results of recognized speech from gathered news materials and recorded meetings. Going forward, we plan to improve the system on basis of their feedback.

The speech recognition technology used in this system was developed for the production of closed captions. At present, it can recognize clear, articulate speech, such as broadcast speech. However, gathered video materials including interviews contain unclear speech. As a result, many portions of the materials effectively cannot be used for broadcasting. To allow producers to confirm facts and increase accuracy, transcripts of these unclear portions are necessary. We are therefore continuing to pursue research and development of technologies for recognizing speech with low intelligibility.

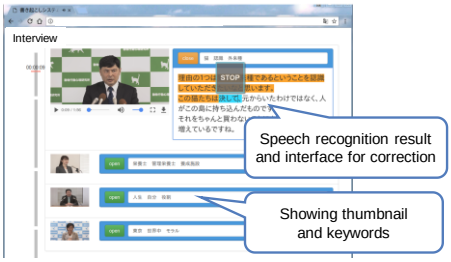

Figure 6 – User interface for speech transcription system

UNIVERSAL SERVICE

Automated audio description

Broadcasting stations provide on the secondary audio channel commentary of visual content that cannot be conveyed by the primary audio content alone. Understanding of the broadcasting content is improved for visually-impaired persons by supplementing visual information with audio commentary. However, such audio commentary is currently provided only for prerecorded programs in limited genres, such as dramas; audio commentary for live programs such as live sports broadcasts is not supported. We are therefore advancing research and development of automated audio description technology, which features automated synthesized speech (5) (Figure 7). By applying the automated audio description technology, we have developed a speech synthesis system that uses an “AI announcer” to read the news automatically.

(1) Automated audio description during live sports broadcasts

We sought to advance research with the aim of realizing synthesized speech during live sports broadcasts, with the goal of applying the technology to the 2020 Tokyo Olympics. In recent years, real-time sports-related data during live sports events, such as “who,” “when,” and “what happened,” is transmitted by sports production companies. Information such as the score, goals made, and penalties can be obtained in sequence. Automated audio description is a fully automated service that generates a script for explaining the game in progress from such data and delivering the script as audio that overlaps with the broadcast audio to an acceptable degree. The technology can create audio description instantaneously in situations where it is difficult to do so manually. It can also provide audio description for multiple sporting events taking place simultaneously. Automatic audio description can also be generated to be presented like an announcer’s play-by-play. The script read by speech synthesizer can also be used as-is for real-time closed captioning. We are proceeding with research to improve the attractiveness of methods for presenting automatic audio description in broadcast audio and to enrich play-by-play content.

(2) “AI announcer”

To realize full-fledged use of speech synthesis technology in broadcast programs, we are making preparations such as conducting speech synthesis technology research using DNN for reading news and organizing learning data. In April 2018, we implemented the technology in practical form as the AI announcer “Yomiko” on the program NEWSCHECK 11. Compared with the concatenative synthesis method, which uses a database of collected text and utterances on a large scale, we realized a natural voice tone for reading

news from an extremely small number of speech samples by deploying a DNN. Going forward, with an eye toward supporting the work of announcers in regional broadcasting stations, we seek to improve the technology through additional speech learning so that synthesized speech can sound more natural for a variety of contents.

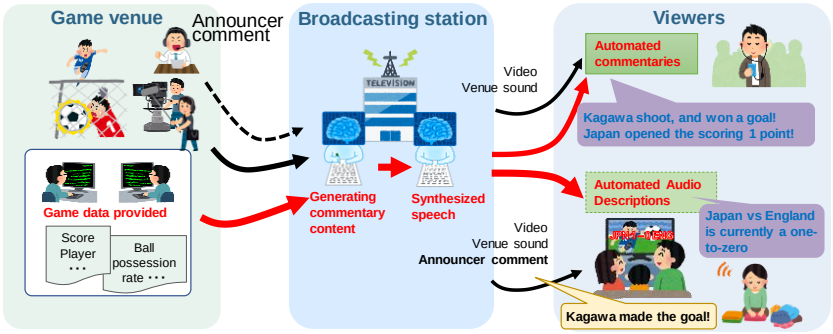

Figure 7 – User interface for speech transcription system

Sign language CG

Among hearing-impaired viewers, there are those who desire information to be presented by sign language because broadcast closed captions alone are insufficient for complete comprehension. However, the number of sign language casters who can express highly reliable sign language at broadcasting stations is limited, and each of them cannot be stationed at just one broadcasting station. We are therefore conducting research on sign language computer graphics (CG) generation technology in order to present emergency weather and disaster information using sign language in each region as the first report of an incident.

On the basis of XML data sent regularly by the Japan Meteorological Agency, the weather information uses a prepared sign language template for weather forecasts for which content such as “weather,” “temperature,” and “chance of rain” are filled with numerical data. The information is then presented by an automatically generated sign language CG animated character. A correct answer rate of 96% was obtained as the result of experiments to confirm whether deaf persons were able to understand the generated sign language expressions, confirming the effectiveness of sign language presentation by this method. At present, a weather information sign language CG evaluation web page has been established on the NHK Online website. It provides weather information in sign language CG form, and is automatically updated three times a day.

Furthermore, similar to the research of automated audio description technology that generates content from data, we are pursuing research on applying sign language CG to sports programs. To date, we have created a prototype system that presents sports video and sign language CG on a Web browser. This system automatically generates sign language CG of game conditions and rules using live sports-related data transmitted during the sports event (6) (Figure 8). We have also devised methods of visually presenting excitement at the sports venue. The responses to questionnaires given to hearing-impaired persons who participated in experiments showed that they had favorable opinions about information that could not be obtained from live play-by-play; this information was given while the game transmission was suspended. Going forward, we will conduct further evaluations of our developed systems with hearing-impaired participants,

determine functions of sign language CG needed for sports programs, and improve the system with an eye toward practical utilization in 2020.

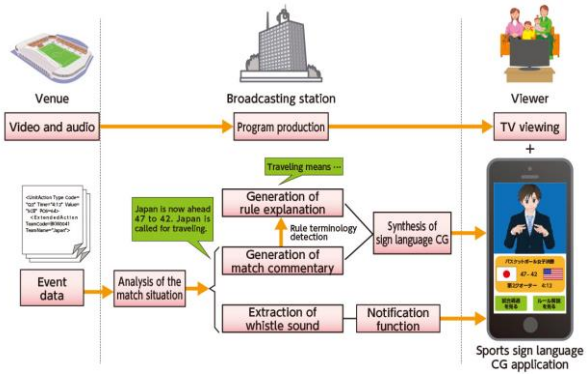

Figure 8 – Operation flow of sports sign language CG application

CONCLUSION

When utilizing a variety of information available in society and past TV program archives, the textual big data analysis, video analysis, and speech recognition technologies introduced here can enable broadcasting stations to quickly and efficiently obtain information needed for programs and allow production members to smoothly create programs. Also, realizing universal service that can present information accurately to all viewers, including persons with hearing and visual impairments, is a critical role of public broadcasting. Priority efforts for achieving this goal were introduced in this article.

Going forward, we seek to exploit the advantages of being close to local broadcasters where the research results are introduced, to incorporate agile development methods, and to continue to advance research and development so that the highest standards for broadcasting and services can be realized by 2020.

REFERENCES

1. Goto, J., et al., 2018. Automatic Tweet Detection based on Data Specified through News Production, Proc. of IUI2018 Companion, No.1.
2. Matsui, A., et al., 2017. Broadcast Video Summarization using Multimodal Contents Analysis, IEICE technical report, PRMU (in Japanese).
3. Endo, R., et al., 2017. Study of Multi-Scale Residual Network for Image-to-Image Translation, ITE technical report, ME (in Japanese).
4. Ito, H., et al., 2017. End-to-end Speech Recognition for Languages with Ideographic Characters, APSIPA ASC, Paper ID 118.
5. Kurihara, K., et al., 2017. Automatic Generation of Audio Descriptions for Sports Program, Proceedings of 2017 International Broadcasting Convention.
6. Uchida, T., et al., 2017. Sign Language Support System for Viewing Sports Programs, Proc. of ACM ASSETS 2017, p.339-340.