IBC2023

AI IMAGE ANALYSIS IN ERA OF SHORT-TIME VIEWING

M. Maezawa, R. Endo, T. Mochizuki

NHK (Japan Broadcasting Corporation), Japan

ABSTRACT

In the era when short videos are preferred, broadcasting stations have been enhancing momentums to distribute summary videos of broadcast content on social networking services (SNS). Therefore, we have developed automatic generation systems for news and programme summary videos. Using a video summarisation artificial intelligence (AI) that has learned the image composition and camerawork typical of important scenes, it is possible to automatically generate summary videos with a high quality close to videos edited by actual programme production staff. These systems have been on trial/practical use in various NHK broadcasting stations. The generated summary videos are posted daily on SNS. Furthermore, considering a programme website is also important content that could boost viewer contact rates, we developed a support system for programme website creation using an AI to extract thumbnails automatically. Using thumbnail candidate images extracted automatically by AI that has learned the unique features of a programme's representative images, you can create programme websites with minimal effort. These technologies can streamline the production of Internet-content such as summary videos and programme websites. Moreover, they will greatly boost Internet deployments of various broadcasting programmes.

INTRODUCTION

With the consumption of short video clips on the Internet increasing, broadcasters are attempting to boost user contact with their content by producing summary videos and distributing them on the Internet. Editorial operations to create summary videos require certain specialities and high work costs. For content that is updated daily, such as news programmes, automation of the summary video production process is especially desirable. Therefore, we are developing automatic video summarisation technologies.

In this paper, we introduce a system that automates the production of news summary videos. Our automatic video summarisation technology has the following features. In news footage, the importance of shots is strongly correlated with image composition and camerawork, such as zooming in on important people, panning to show items of interest in detail, and special shooting angles for buildings involved in incidents. We trained an artificial intelligence (AI) system to learn picture-making such as the image composition and camerawork typical of summary videos produced by skilled editors. Moreover, our technology used the similarity of keywords with an anchor’s introduction speech to evaluate the importance of each video segment.

This system summarises 15–30-minute news programmes into about 1–2 minutes, with about 20 summary videos being produced and distributed each day. Summary videos that used to take a skilled editor more than one hour to produce can now be generated about 10–20 minutes after the end of the broadcast programme, making it possible to produce and distribute summary videos without losing the immediacy of the news.

With regard to common programmes not including an anchor's introduction, we developed a practical system to generate summary videos automatically. As with the news video summarisation system, the AI on this system learned picture-making typical of important scenes using summary videos created by professional video production staff. This system has also been on trial use in a number of NHK regional broadcast stations to post summary videos on social networking services (SNS).

Along with summary videos, attractive thumbnails on programme websites can boost the number of accesses to broadcast contents. An attractive thumbnail contains an eye-catching image that is representative of the video. We developed a support system for programme website creation using AI to select attractive thumbnails from videos. This AI was generated by learning the composition of good thumbnails and their suitability as representatives of the video.

The summary videos and programme websites could lead users to broadcast content, so such technologies are expected to provide a bridge between the Internet and broadcasting.

RELATED WORK

Video Summarisation

In this section, we describe practical examples of automatic summarisation technology in video production sites.

In sports videos, the important scenes to be used in summary videos, such as successful attempts and scores, are clear. Therefore, practical use of summarisation technology is being promoted in various sports events.

As a practical example for tennis, there is a summary video generation system developed by IBM that was used at the Wimbledon Championships 2018 (1). Using AI that has learned a large number of point-scoring scenes collected from past game footage, it analyzes the cheers of the audience, movement of the players, scores of the players, etc., and automatically extracts the important scenes in the game. As for golf, at the Masters Tournament 2019, IBM developed a technology to generate a summary video in a short time (2). The importance of each shot is determined on the basis of the presence or absence of specific actions (such as fist pumps), facial expressions, changes in the volume and tone of the cheers, and the position of the ball, among other factors. Each importance is assigned automatically, and a summary video of each player is automatically generated. Practical examples for soccer include uniform number recognition, player movement analysis, and automatic summarisation technology using techniques such as image trimming in accordance with the results. A summary video distribution service using this technology has started in Italy's Serie A (3). As for basketball, to popularise and improve the value of the 3-player basketball league “3x3.EXE PREMIER,” a summary video generation service is in operation that uses technology to detect objects such as scoreboards and recognise displayed characters (4).

In fields other than sports, it is difficult to define important scenes due to differences in editorial staff's points of view and viewers' tastes, and there are few technologies that have been put to practical use.

Hakuhodo DY Media Partners Inc., Tokyo University of Science, and M Data Co.,Ltd. have developed a trial version of a system that uses AI to automatically generate summary videos of dramas. They tested it for serial dramas broadcasted from July 2019 (5). However, this technology requires metadata about the utterances and telop contents of the performers, which is manually assigned to each scene of the TV drama video. Such data is not attached to most of the programme videos owned by broadcasting stations. Therefore, we developed a practical technique that can automatically generate a summary video only from a programme video.

Thumbnail Extraction

Techniques for selecting representative images of video have been proposed using basic image information such as colour histograms, colour layouts, texture features related to luminance co-occurrence, motion vectors between frames, and face sizes (6)(7)(8). However, the various factors to be considered in the selection of images are not always accurately reflected. In addition, a method using a convolutional neural network (CNN), which is the mainstream for tasks such as image classification, has also been proposed (9). However, this technology is targeted at YouTube, and since the number of video playbacks is used as training data when learning CNN, it is difficult to apply it to broadcast video.

The AVA database (10) is a dataset in which images are divided into two classes, high quality and low quality, on the basis of aesthetic visual analysis by photographers. Jin et al. proposed a NN that classifies images into high/low quality using the AVA database as training data (11). First, the feature extraction network computes the image features, and then the classification network computes the 2-class probability distribution from the image features. The selection of thumbnails by programme production staff takes into consideration elements related to aesthetic preferences, such as image colour and composition. For this reason, we apply this method in our programme thumbnail extraction technology. Our technology can improve the efficiency of thumbnail selection work and drive the Internet deployments of broadcast contents.

NEWS VIDEO SUMMARISATION

News that is updated at short intervals is content that is highly compatible with the Internet community, which is sensitive to changes in topics. At NHK as well, expectations are rising for the Internet delivery of summarised news as a means of increasing the opportunities for viewers to experience broadcast contents. Therefore, we developed an automatic summarisation system for news broadcast videos (12). In this chapter, we describe the automatic summarisation technology for news videos and the functions of the system.

Automatic Split into News Subjects

A typical news programme video consists of a plurality of news items. Each news item consists of a part (lead video) in which the studio announcer reads the news, followed by a main story video. Since the content of each news item is completely different, to generate a summary video of the entire news programme, it is necessary to automatically divide the news programme video into news item units and summarise the main story video of each news item individually.

This section describes the flow of automatic segmentation into news items. First, a news programme video is automatically divided into shots, and frame images are sampled from each shot. Then, each frame image is input to the “lead image determination AI” created in advance, and the probability (score) of each image being included in the lead video is calculated. This process uses a support vector machine that has learned the image features of lead images in various news programmes. Next, in each shot, the average score of the images belonging to it is determined. Finally, shots whose average score is equal to or higher than a threshold are defined as lead videos, shot sequences between the lead videos are defined as main story videos, and intervals of each news item are determined.

News Video Summarisation NN

A video summary of each news item is generated by extracting important video segments from the main story video. We describe the News Video Summarisation Neural Network (N-VSNN) for estimating the importance score of each video segment. Figure 1 shows the structure of N-VSNN. In video production at a broadcasting station, various elements unique to television, such as the type and size of the subject, composition, and camera movement, are considered. Therefore, we made it possible to input multiple modal features to N-VSNN and adopted the following four types of image features (hereafter referred to as “feature datasets”) as feature data.

- Subject feature: Each image sampled at equal intervals from the video segment is input to the existing trained image classification CNN model (13), and the feature vector is obtained from the intermediate layer. The average feature vector (2048 dimensions) for all images is taken as the subject feature.
- Object class feature: In the object feature calculation process, the vector (1000 dimensions) obtained in the same way using the final layer instead of the intermediate layer is used as the object class feature.
- Face class feature: From each image sampled in the same way as the subject feature, the face region image detected by Kawai et al.'s method (14) is input to the existing trained face classification model (15) to obtain a feature vector. Then, for each feature vector, a similarity vector is generated using 1,000 face classes created by clustering a large number of face feature vectors. This similarity vector consists of the inner product of the feature vector and the centre vector of each class. A face class feature (1000 dimensions) is obtained by multiplying and summing the similarity vectors for all face regions by a weighting factor in accordance with the face size.
- Camera movement feature: The video segment is divided equally into three sub-segments, the first half, the middle part, and the second half. A vector connecting motion histograms (6 regions x 8 directions) in each sub-segments is taken as a camera motion feature (144 dimensions).

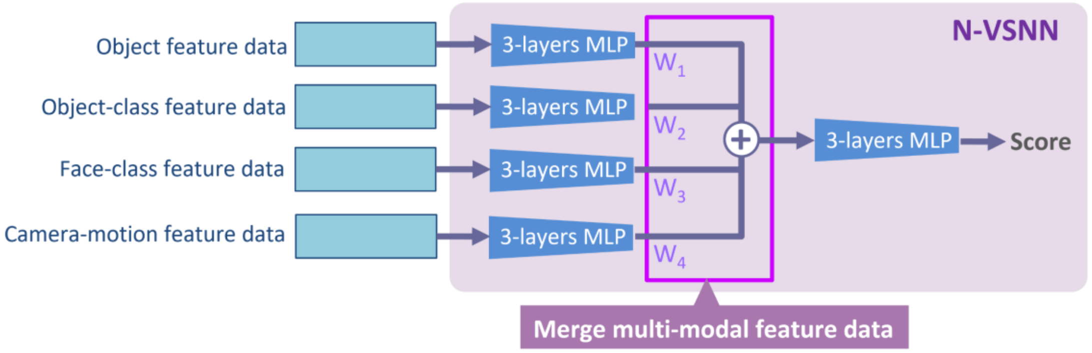

Figure 1 – News Video Summarisation NN

In the previous network, for each modal, 256-dimensional intermediate data is generated through a 3-layer multi-layer perceptron (MLP), and the four intermediate data are multiplied by weighting factors W_i (i = 1, …, 4) and added. When N-VSNN is trained, W_i is updated along with the MLP parameters. Finally, the importance of video segments is output through a post-network consisting of 3-layer MLPs.

As training data, we used about 200 summary videos distributed in the past on NHK's news website (16) and news main story videos as sources for each. N-VSNN was generated by training so that a high score would be output for sections of the programme video included in the summary video, and a low score would be output for other sections. By using N-VSNN, is it possible to extract video segments with picture-making unique to important scenes in news programmes, such as zooming in on important people, panning to show the subject in detail, and special shooting angles for buildings involved in the incident.

Speech Content Score

In the lead video, the announcer gives an overview of the news. Therefore, we consider that the interval where the lead video and the utterance content are similar is important, and introduced the “utterance content score” that expresses the similarity to the summarisation process.

In this section, we describe the calculation method of the utterance content score. Keywords are extracted by recognising the voice of the lead video, and each keyword is vectorised by Word2Vec (hereinafter referred to as a keyword vector). Keyword groups are extracted from each utterance period of the main story video by utterance period detection processing and speech recognition processing, and keyword vector groups are obtained in the same manner as the lead video. For each utterance segment, the degree of similarity between keyword vector groups with the lead video is used as the utterance content score.

News Video Auto-Summarisation System

In this section, we describe the News Video Auto-Summarisation System developed using the aforementioned processes and N-VSNN. Figure 2 shows the flow of generating a summary video using this system. The user uploads a news programme video to be summarised on the video input page. After uploading, a summary video is automatically generated by the following process.

1. Using the method of Automatic Split into News Subject sections, the news video is divided into news items, and steps 2–7 are performed for each news item.
2. A main story video is divided into shot units, and each shot is divided into fixed-length video segments.

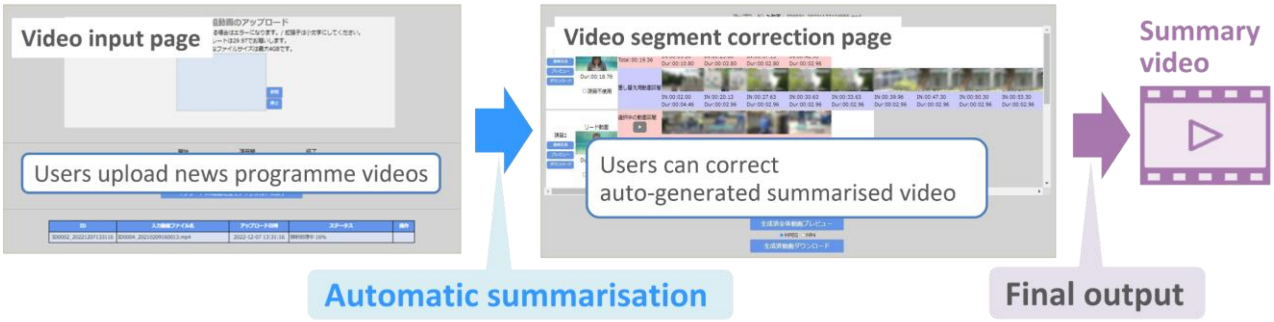
Video input page
Users upload news programme videos
Video segment correction page
Users can correct auto-generated summarised video
Summary video
Automatic summarisation
Final output

Figure 2 – News Video Summarisation System

3. For each video segment n, a feature dataset is computed and input to N-VSNN to compute the importance $S_{NN}(n)$.
4. The utterance content score $S_{SP}(m)$ is calculated for each utterance section m using the method in the Speech Content Score section.
5. A total score S(n) for each video segment is calculated using Equation (1).

   $$S(n) = S_{NN}(n) + S_{SP}(m^*)\ R(n, m^*) \qquad (1)$$

   Here, m^* is the index of the utterance segment with the highest overlapping rate with the video segment n, and $R(n, m^*)$ represents the overlapping rate.
6. The moving image segments are shot out in descending order of S(n). When the total duration of the clipped sections exceeds the length of the lead video, the process ends, and the video segments are connected to generate a summary video.
7. The audio of the summary video is replaced with that of the lead video. Since the content of the lead video is the outline explanation of the news by the announcer, this processing can be expected to improve the explainability of the summary video.

To put the system into practical use, in addition to the function of automatically generating summary videos, it is necessary to perform confirmation and correction work to deploy it to media other than broadcasting. For example, consideration must be given to personal information and the right to use video materials outside of broadcasting. Therefore, we implemented a “video segment correction page” to easily correct the automatically generated summary video. On this screen, it is possible to replace the video segment that constitutes the summary video with another candidate and adjust the IN/OUT points of each video segment. The video summary of each news item that has undergone correction work is connected to a dedicated logo video in between to output the final video summary.

With this system, it is possible to generate summary videos of 15–30-minute news programmes by automatic processing in about 10–20 minutes Full-scale practical use of this system began in the fall of 2022, and NHK headquarters and many regional broadcasting stations are distributing news summary videos generated by this system on SNS daily.

PROGRAMME VIDEO SUMMARISATION

For general programmes other than news, there is a growing interest for Internet development, and there is a demand for a mechanism that automatically generates a summary video. In this chapter, we describe a NN for programme video summarisation that we devised (17) and an automatic programme video summarisation system that we developed using this NN.

Programme Video Summarisation NN

Figure 3 shows the structure of the Programme Video Summarisation NN (P-VSNN) for estimating the importance score of each video segment in a programme video. Many programme videos are long, from several tens of minutes to several hours. In the process of summarising long videos, it is common to evaluate the importance of each video segment by considering the video content of the section, nearby sections, and the video content of the entire video. Therefore, P-VSNN is input with feature datasets calculated on multiple timescales, such as nearby shots and the entire video, instead of the feature dataset for the video segment for which the score is to be estimated.

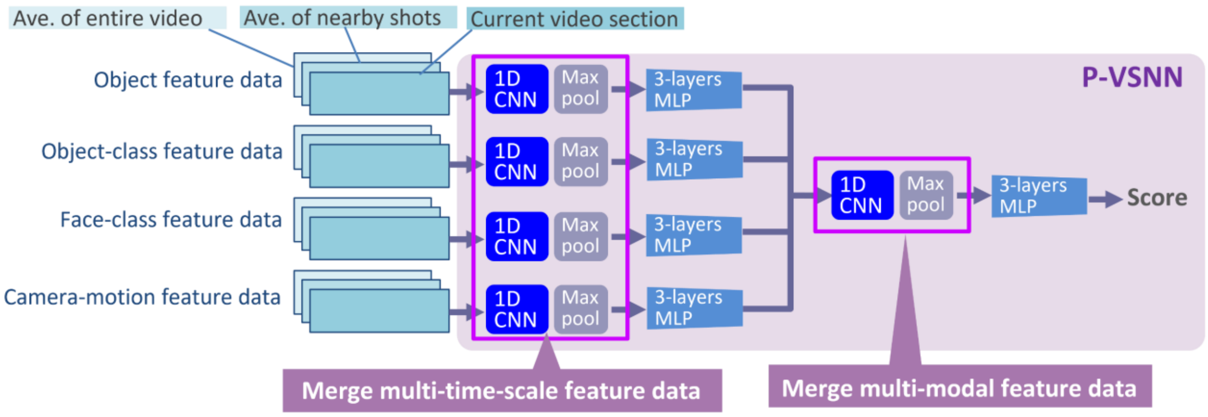

Figure 3 – Programme Video Summarisation NN

In the network of the first stage, for each modal, feature data of multiple time scales are integrated by a 1-dimensional (1D) CNN and Max pooling, and 256-dimensional intermediate data are output through a 3-layer MLP. In the latter network, the intermediate data of each modal is integrated by a 1D CNN and Max pooling, and the score of the video segment is output through a 3-layer MLP.

P-VSNN was trained using dozens of programme summary videos created by professional video editing staff as well as each programme video as learning data. By using P-VSNN, it is possible to extract video segments with picture-making (how to shoot the subject, composition, camera work, etc.) unique to the important scenes of the programme video.

Programme Video Auto-Summarisation System

Using the aforementioned P-VSNN, we have developed an automatic programme video summarisation system. Figure 4 shows the flow of programme summary video generation by this system. The user uploads a programme video to be summarised on the video input page. At that time, it is possible to specify the approximate length of the generated summary video (=T_{SUM}). After uploading the programme video, a summary video is automatically generated by the following process.

1. A programme moving image is divided into shot units, and each shot is divided into fixed-length moving image sections.
2. For each video segment n, feature datasets at three time scales (this video segment, average of nearby shots, average of entire video) are computed and input into P-VSNN to calculate importance S(n).
3. The video segments are cut out in descending order of S(n). At that time, to reduce the sense of the incongruity of the sound at the connection point of the moving image section, if the cut-out point is in the middle of the speech section, it is changed to the start or end point of the speech. An existing library (18) was used to detect speech segments from videos. When the total length of the clipped sections exceeds T_{SUM}, the process ends, and the video segments are connected to generate a summary video.

In this system, we implemented a “video segment correction page” in the same way as the news summarisation system. After performing correction work as necessary, the final summary video is output. The user can use this system to generate a summary video with approximately half the programme time length and only about 5–10 minutes of correction work. Currently, several NHK regional broadcasting stations are using this system on a trial

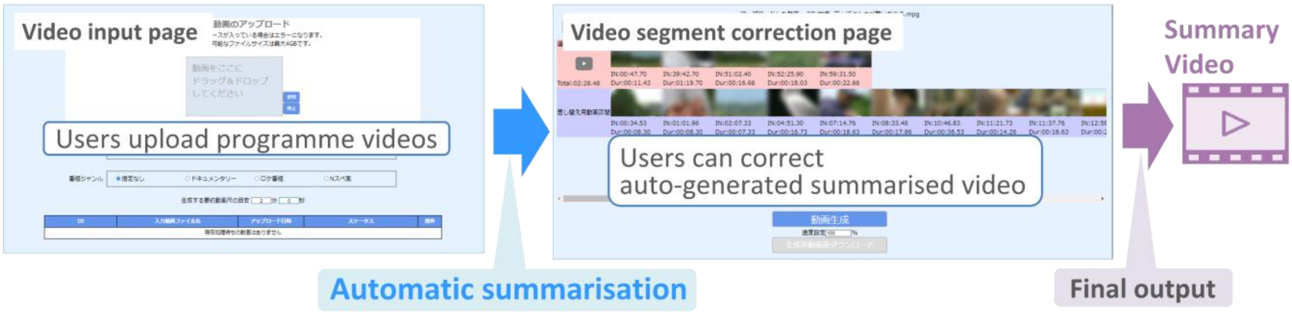
Video input page
Users upload programme videos
Automatic summarisation
Video segment correction page
Users can correct
auto-generated summarised video
Summary
Video
Final output

Figure 4 – Programme Video Summarisation System

basis, and video summaries of cameraman-led programmes and other programmes, generated on this system, are being distributed to SNS.

PROGRAMME THUMBNAIL EXTRACTION

Similar to video summaries, thumbnail images of programmes posted on programme websites and SNS are indispensable items for improving the rate of contact with broadcast content. The presentation of eye-catching thumbnails can be expected to increase the number of viewers accessing programmes. However, the selection of thumbnails is not an easy task because it is necessary to carefully consider the colour, composition, content of the subject, and so on. NHK disseminates information using many programme websites and SNS, and improving the efficiency of thumbnail selection work is necessary. This chapter describes the thumbnail extraction technology and programme website creation support system that we have developed.

Thumbnail Extraction NN

The programme genre has a great effect on the thumbnail selection criteria. For example, for a drama programme, an image showing the performers is preferred, and for a travel programme, the beauty of the scenery is given priority. Therefore, we developed a method to select thumbnails from programme videos by scoring images using a NN that has learned both images and programme genre information (19). This method makes it possible to select images considering the trend of thumbnails for each programme genre.

Figure 5 shows an overview of our developed Thumbnail Extraction NN. The input is the images sampled from the programme video and the genre information vectors representing the genre of the programme (drama, travelogue, variety, etc.), and the output is the score representing the suitability of the image as a thumbnail. A genre information vector is a binary vector representing whether a programme belongs to each genre by 1/0, and the number of dimensions is 8, which is the number of genres. If the output score is equal to or greater than a predetermined threshold value, it is selected as a thumbnail candidate.

Thumbnail Extraction NN consists of three networks: an image feature extraction network that computes image features, a genre feature extraction network, and a score computation network.

The image feature extraction network uses the network structure of an existing method (11) that evaluates the visual artistry of images. A structure using GoogLeNet (20) and Batch Normalization (21) generates a 1024-dimensional feature vector. The genre feature extraction network is a one-layer fully-connected NN, and generates feature vectors (1024 dimensions) with the genre information vector described later as input. The outputs of the

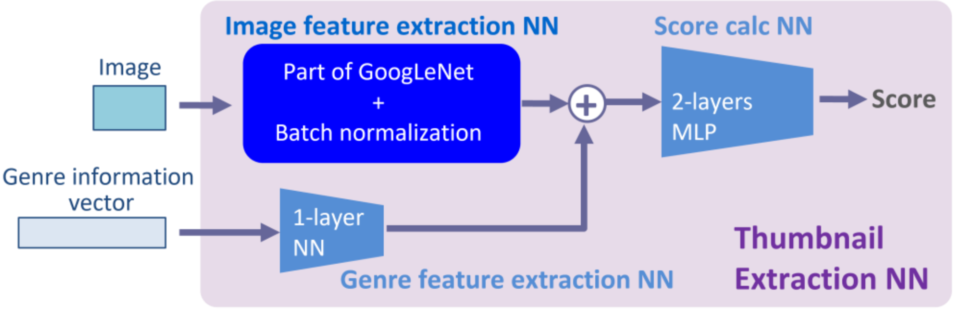
Image
Genre information vector
Image feature extraction NN
Part of GoogLeNet + Batch normalization
Score calc NN
2-layers MLP
Score
1-layer NN
Genre feature extraction NN
Thumbnail Extraction NN

Figure 5 – Thumbnail Extraction NN

image and genre feature extraction networks are added and input to the score calculation network. The score calculation network is a 2-layer MLP, and outputs a score between 0.0 and 1.0 using the sigmoid function.

As training data, we used about 6,500 images sampled from various programme videos and 3-level labels (great/good/bad) assigned to each image by programme production directors and editorial assistants regarding suitability as programme representative images. We set the correct scores for great, good, and bad images to be 1.0, 0.5, and 0.0, respectively, and trained the Thumbnail Extraction NN using the image and genre information as input. By using this NN, it becomes possible to extract thumbnails that reflect the image selection skills of professional programme production staff.

Support System for Programme Website Creation

As previously mentioned, NHK publishes information on many programme websites, and there is a demand for a mechanism to streamline the selection of thumbnail images to be posted on the website along with the creation of the website. Therefore, we have developed a system that supports the creation of programme websites using the Thumbnail Extraction NN previously described.

Figure 6 shows the process flow of this system. First, the user uploads a programme video from which a representative image is to be extracted. At that time, the genre of the programme is selected. Next, shot division of the uploaded video and image sampling processing are automatically performed. Image scores are calculated by the Thumbnail Extraction NN for the sampled images. Images with high scores are presented as candidate images, and the user selects the image they want to use. At that time, if necessary, it is possible to input text such as a programme outline to be posted on the webpage. Finally, a

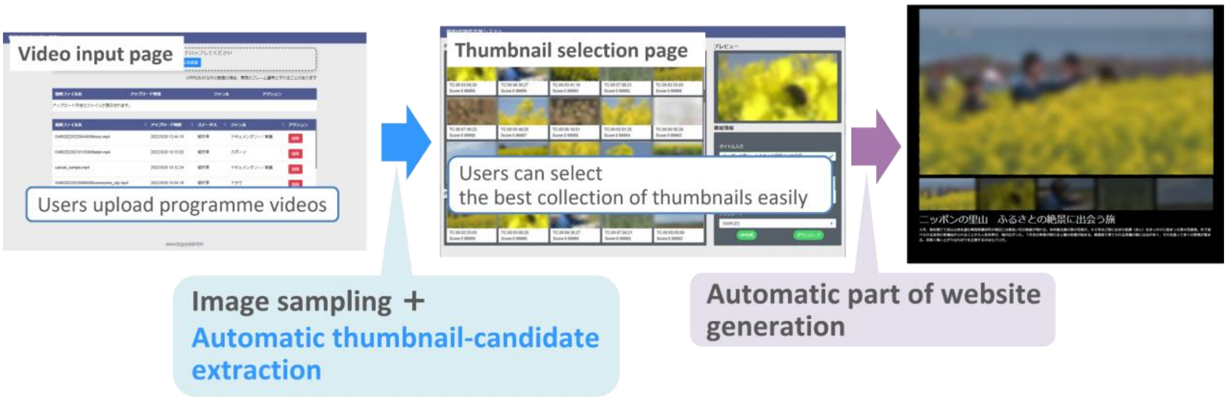
Video input page
Users upload programme videos
Thumbnail selection page
Users can select the best collection of thumbnails easily
ニッポンの里山　ふるさとの絶景に出会う旅
Image sampling +
Automatic thumbnail-candidate extraction
Automatic part of website generation

Figure 6 – Support system for programme website creation

portion of the website is automatically generated using the selected thumbnail image. Users can preview the generated website and download the HTML files and thumbnail images that comprise it.

By using this system, it will be possible to create a programme website with little effort, and further enhancement of the programme website can be expected. In the future, we aim for practical use after trials at the programme production site.

CONCLUSION

As AI technologies in the era where short-time viewing is preferred, we presented automatic video summarisation technologies and a thumbnail extraction technology to quickly distribute summary videos to SNS. Various NHK broadcasting stations utilise these systems as tools to activate Internet deployments of broadcast contents. In addition, the programme website creation support system developed using the thumbnail extraction technology is expected to be put into practical use as a tool to reduce work costs to create websites, which are essential to easily grasp the programme. In the future, we will work to improve the performance of each technology and refine the systems on the basis of the needs of programme production sites.

REFERENCES

1. How AI picks the highlights from Wimbledon fairly and fast, 2019. https://www.ibm.com/blogs/journey-to-ai/2019/07/how-ai-picks-the-highlights-from-wimbledon-fairly -and-fast/.
2. AI-generated Highlights Tell the Story of the Masters, 2019. https://blog.video.ibm.com/ai-video-technology/ai-generated-highlights-tell-story-of-the-masters-2019/.
3. Serie A agrees AI highlights deal with WSC Sports, 2022. https://www.sportbusiness.com/news/serie-a-agrees-ai-highlights-deal-with-wsc-sports/.
4. Providing automatic highlight creation service using NTT DoCoMo Co., Ltd. and AI - Easily extract only the highlight scenes you want and reduce the burden of video editing, 2021. https://3x3exe.com/premier/news20220601-3/.
5. Hakuhodo DY Media Partners Develops an Automatic Drama Digest Video Generation System Together with the Tokyo University of Science: Trial Operation with a New Drama Program Broadcast on Tuesdays by TBS, 2019. https://www.inter-bee.com/2019/en/magazine/production/detail/?id=871.
6. Y. Gao, T. Ahang, and H. Xiao, 2009. Thematic video thumbnail selection. Proc. of ICIP, pp. 4333 to 4336.
7. H.-C. Lian, X.-Q. Li, and B. Song, 2011. Automatic video thumbnail selection. Proc. of International Conference on Multimedia Technology, pp. 242 to 245.
8. Y. Song, M. Redi, J. Vallmitjana, and A. Jaimes, 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. Proc. of ACM international on conference on information and knowledge management, pp. 659 to 668.
9. N. Arthurs, S. Birnbaum, and N. Gruver, 2017. Selecting youtube video thumbnails via convolutional neural networks. Technical Report, Stanford.

10. N. Murray, L. Marchesotti, and F. Perronnin, 2012. AVA: A large-scale database for aesthetic visual analysis. Proc. of CVPR, pp. 2408 to 2415.

11. X. Jin, et al., 2019. ILGNet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation. IET Computer Vision, 13.2: 206 to 212.

12. T. Mochizuki, Y. Kawai, N. Fujimori, M. Maezawa, R. Endo, Y. Asami, 2023. Prototype of support system for news summary video production (in Japanese). Journal of the Institute of Image Information and Television Engineers, vol. 77, no. 2, pp. 262 to 271.

13. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, 2015. Learning spatiotemporal features with 3D convolutional networks. Proc. of ICCV, pp. 4489 to 4497.

14. Y. Kawai, R. Endo, N. Fujimori, T. Mochizuki, 2019. Study of face recognition using deep neural network (in Japanese). Proc. of Forum of Information Technology (FIT), no. 3, H-003, pp.103 to 104.

15. Y. Kawai, R. Endo, N. Fujimori, T. Mochizuki, 2018. Face detection using cascaded convolutional network (in Japanese). Proc. of ITE annual convention, no. 3, 22B-1.

16. NHK NEWS WEB, https://www3.nhk.or.jp/news/movie.html.

17. T. Mochizuki, Y. Kawai, 2022. Programme video summarisation by fusion of multi-modal/timescale features using 1D-CNN (in Japanese). Proc. of IEICE General Conference, D-12-26.

18. https://realpython.com/python-speech-recognition/.

19. M. Maezawa, R. Endo, N. Fujimori, T. Mochizuki, 2021. A study on the selection of thumbnail images considering a genre of TV programs (in Japanese). IEICE Technical Report, vol. 121, no. 155, PRMU2021-15, pp. 48 to 51.

20. C. Szegedy, et al., 2015. Going deeper with convolutions. Proc. of CVPR.

21. S. Ioffe and C. Szegedy, 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proc. of ICML.