2021

# FACE DETECTOR: A REAL-TIME FACE RECOGNITION SYSTEM FOR LIVE BROADCASTING

Y. Kaburagi, Y. Manya and T. Aoki

NHK, Japan

## ABSTRACT

We have developed a real-time face recognition system, Face Detector, that identifies and displays the names of people appearing in an input video stream. The detector has been developed for use in such broadcasting operations as providing support for the commentators of live sports events and assistance for editors who superimpose people’s names on images. Face Detector achieves high, real-time accuracy for use in live broadcasting and is also straightforward to use. The system can recognize people from the side or above and even when the subject is wearing a medical mask or sunglasses. It can also reliably differentiate identical twins. The processing speed of three frames per second is fast enough for use during live broadcasting, and sufficient accuracy is obtained even from training data consisting of only one image per person. The system can also be used at relay sites with no network connection. We have already used Face Detector to assist commentators on a live radio sports relay.

## INTRODUCTION

We developed “A System of Creating Information of Performers Using Face Recognition” [1]in 2018. The system uses face recognition technology to extract as metadata performer information from the program automatically. By recognizing faces in accumulated program files broadcasted in the past, the system reduces manual labor of archiving operations. Through the development and testing of this system, we were convinced that there would be a greater demand for applying face recognition technology to real-time video processing in the broadcasting business. Therefore, in 2020, we extended this system and developed a new system, called “Face Detector,” that recognizes faces in real-time for broadcasting use. The correct identification and delivery of people’s names is very important in broadcasting. When adding superimpositions to programs, for example, the subject must be identified instantly. In sports commentaries, the commentators have to identify the athletes correctly even from difficult viewing angles, such as from the side or when the athlete is looking down. In editing work, too, editors must locate and identify their chosen subject quickly. Often, however, much special background knowledge and experience are required to make these judgments quickly and accurately. Face Detector supports these operations by identifying the names of people who appear in the input video stream. Face recognition technology has spread rapidly in recent years in such areas as analyzing images from security cameras in criminal investigations and identity verification. The broadcasting field does, however, present some unique challenges. One is that the processing must be achieved in real-time to convey the information quickly. The information also has to be correct. Further, the system must be usable not only at the broadcasting station but even at locations that have no network connection. In addition, the training data for facial recognition has to be generated quickly because of the rapid changes of the subject. This training process, too, must be efficient.  In regard to each of these issues, Face Detector has achieved the following results:

(i) High accuracy for broadcasting – it not only recognizes subjects wearing a mask or sunglasses but also differentiates identical twins

(ii) Real-time processing for live broadcasts

(iii) Generation of training data from a single image per person

(iv) Generation of ZIP file recognition models consisting of images and Excel files containing people's names

(v) Use on sites with no network connection

(vi) A special logic system for coping with crowded scenes

(vii) A chroma key display function (for cameraman assistance etc.)

## FUNCTIONS

The presence of people is detected on an input video stream and face recognition processing is performed on each person. The results are then displayed on the screen. Figure 1 illustrates an example of how Face Detector works. It is possible to identify the subject not only from the front but also from the side, and the subject's name is displayed in real-time at a processing speed of about three frames per second. Furthermore, the subject's metadata (affiliation, birthplace, age etc.), when registered in advance, can be displayed by clicking on the bounding box, as shown in the figure. The degree of similarity – confidence level –can also be displayed as a percentage.

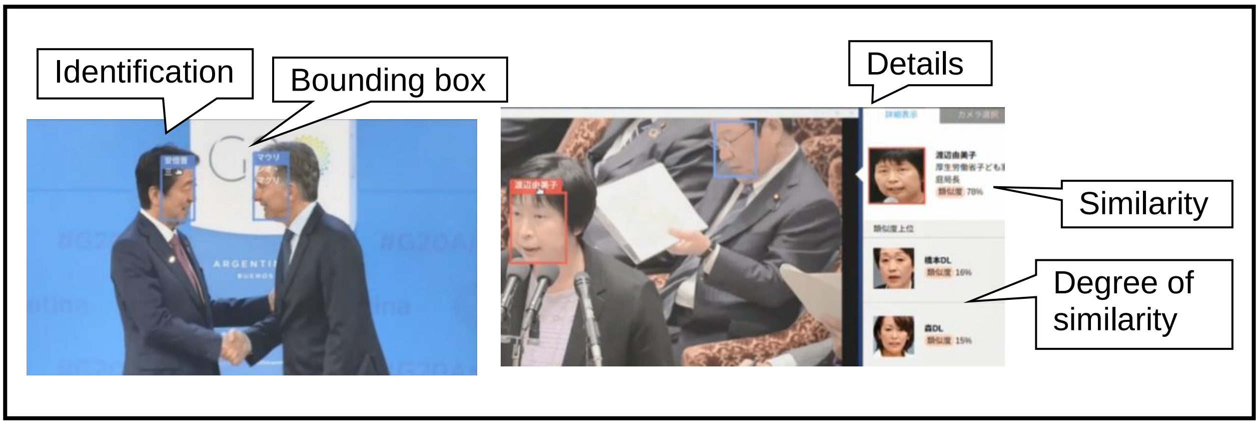


Figure 1 - Example showing how Face Detector works

## MOUNTING SUBSTRATE

Table 1 on the right shows the PC specifications. The mounting substrate is built on a local PC for ease of use not only at the broadcasting station but also at locations with no network connection. This reduces the overhead of network data transfer compared with cloud use, thereby facilitating high-speed processing. It also has the advantage of bypassing security risks. Installation of a general GPGPU secures the availability of advanced computing resources at low cost and assures real-time processing for use in live broadcasting.

| | PC Specs. |
|---|---|
| CPU | Core i7-9700K, 3.6GHz |
| Memory | 32GB |
| GPU | NVIDIA RTX2060 |
| OS | Ubuntu (Linux) |

Table 1 – PC Specifications

## APPLICATION SOFTWARE

We used open-source software to minimize the development and operating costs and to apply the latest technology. In building the face recognition process, we verified and compared multiple open-source software options and adopted “Insightface”[2], which is a high-precision engine suitable for broadcasting operations. The use of Docker, an open-source software program for container virtualization, made it easy to develop and update applications for use in diverse PC environments. The container virtualization technology combines applications and libraries as a virtualization project on the Container OS.

## SYSTEM OVERVIEW

We show the processing flow of Face Detector. To perform identification, we need to generate the Person-specific and People-set Models in advance. While other face recognition systems require multiple images for each person to generate these models, Face Detector needs only one sample image per person. This greatly reduces the model generation work. In addition, the model-generation process involves no complex operations and can be performed easily on a simple user interface. Face Detector can be used with a USB port. When inputting SDI, we convert from SDI to USB. The processing flow is as follows:

(i) The image input to the PC is cut out frame by frame.

(ii) A person is detected in the images, image by image

(iii) The person’s face is detected

(iv) Individual feature points are extracted

(v) The data so obtained is collated with the learning data of the Person-specific and People-set Models generated in advance, and the identification process performed in real-time

(vi) The person’s name identified in this manner is superimposed on the display

(vii) In cases of crowd footage, it is also possible to apply a special logic for coping with crowd scenes developed for this purpose

(This is a logic system for preventing the processing of unnecessary faces)

## MODEL GENERATION

### Person-specific Model

The model for each person is generated according to the following flow:

(i) Reading of any number of images showing the subject

(ii) Extraction of 512-dimensional feature points from each image

＊Just one image per person is sufficient

(iii) Cluster classification for each subject based on the extracted features

(iv) Attachment of a person’s name to each cluster

(The user then names the person on the GUI)

Figure 2 shows an example of the distribution of person-specific models. 512-dimensional feature points and the results of cluster classification are converted here into a two-dimensional image. The feature points are aggregated and distributed for each subject. When generating a model, Face Detector uses only representative feature points obtained from either multiple or single images.

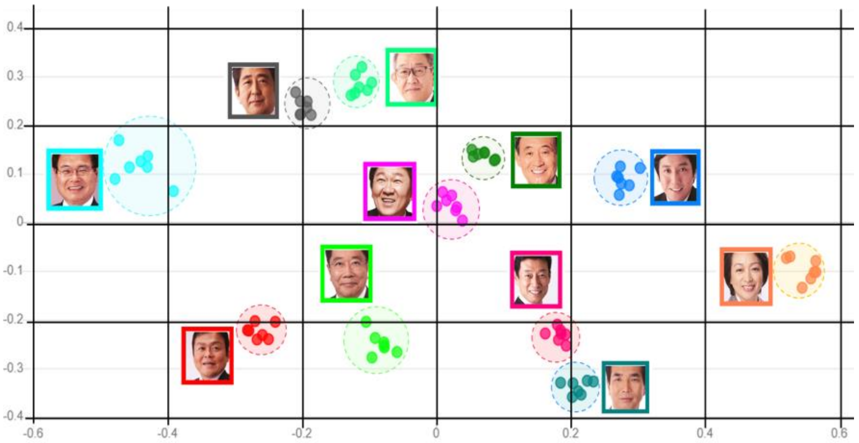


Figure 2 - Person-specific models

### Bulk Registration

Person-specific models are generated by reading the ZIP file images (one image per person) and individual name and metadata information in the EXCEL file. This use of bulk registration greatly truncates the modelling process.

### People-set Model

A People-set Model is generated by selecting multiple subjects from among the Person-specific models. In the case of a sports program, for example, we can create a set model composed of athletes participating in the event. We can also use all of the people registered in Face Detector as an identification model. The People-set Model is created by the following procedure:

(i) Calculating the center of gravity of the cluster feature points for each Person-specific model

(ii) Calculating the feature points closest to the center of gravity calculated in (i)

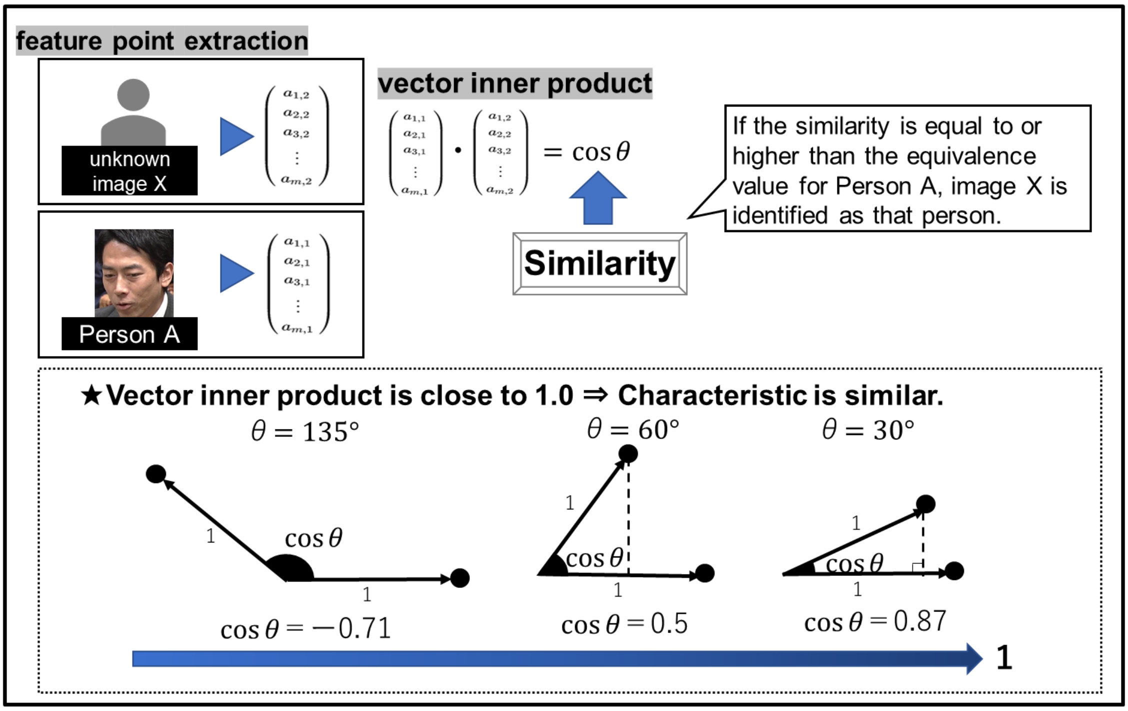
(iii) Associating those feature points to names

(iv) Repeating Steps (i)-(iii) for each cluster

(v) Generating the People-set model through an arbitrary combination of Person-specific models

## CALCULATING SIMILARITY

The system we have developed calculates similarity not by neural network but by the method described below to simplify and accelerate the process. When examining the similarity of unidentified images of X and A, subjects registered as Person-specific models, the system calculates the vector inner products of the 512-dimensional feature points of X and A. The feature point vector is then normalized for the inner product to match a numerical value of cosθ. Accordingly, the numerical value of cosθ is taken as the degree of similarity. The closer the value of cosθ is to1, the more similar are the feature point vectors of X and A. This is illustrated in Figure 3.


feature point extraction
unknown image X
$\begin{pmatrix} a_{1,2} \\ a_{2,2} \\ a_{3,2} \\ \vdots \\ a_{m,2} \end{pmatrix}$
Person A
$\begin{pmatrix} a_{1,1} \\ a_{2,1} \\ a_{3,1} \\ \vdots \\ a_{m,1} \end{pmatrix}$
vector inner product
$\begin{pmatrix} a_{1,1} \\ a_{2,1} \\ a_{3,1} \\ \vdots \\ a_{m,1} \end{pmatrix} \cdot \begin{pmatrix} a_{1,2} \\ a_{2,2} \\ a_{3,2} \\ \vdots \\ a_{m,2} \end{pmatrix} = \cos\theta$
Similarity
$\theta = 135°$
$\theta = 60°$
$\theta = 30°$
1
$\cos\theta$
$\cos\theta = -0.71$
$\cos\theta = 0.5$
$\cos\theta = 0.87$
1

If the similarity is equal to or higher than the equivalence value for Person A, image X is identified as that person.

★Vector inner product is close to 1.0 ⇒ Characteristic is similar.

Figure 3 – Calculating similarity

## REAL-TIME SUBJECT IDENTIFICATION

The identification process is performed by the following flow:

(i) The People-set Model is selected

(ii) The input video stream is captured

(iii) The subject is detected in the video

(iv) The face range is detected

(v) The similarity of feature points is calculated with reference to the People-set Model

(vi) The most similar person above the threshold value is pinpointed

(vii) That person's name is displayed on the input video stream

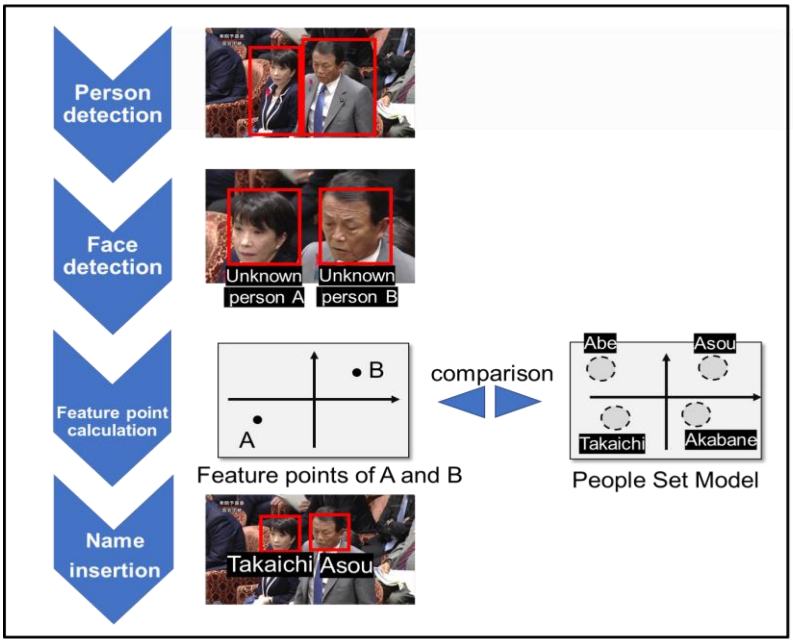


Figure 4 – Real time subject identification

## THRESHOLD FOR SIMILARITY

The degree of similarity is used to identify the subject with reference to the People-set Model. The person's name is displayed when a degree of similarity of 0.3-0.5 or higher is obtained. This threshold for similarity can be set arbitrarily. If the threshold for similarity is too high, however, the names of people who can be identified will not be displayed. Conversely, if too low, the likelihood of error becomes significant. Through use, we have established that this system works best with a threshold set in the range of 0.3-0.5.

## LOGIC SYSTEM FOR USE ON CROWDS

We devised our own logic system for crowd scenes to maintain the speed and accuracy of identification. When the face recognition system is used for a sports program, for example, many spectators and staff may be included in the scenes, and the system could waste time trying to identify unneeded faces (MOB), thereby also consuming a lot of computer resources and reducing overall accuracy. The logic system we devised to maintain real-time processing performance and accuracy is described below. It is also easy to turn this system on and off on the GUI according to need. In devising the logic, account was taken of the probable intentions of the cameraman and program director. In many situations, the camera shows the person of greatest interest in the center of the screen. The edges of the screen are, therefore, excluded from the identification range, which is limited to the center, thereby reducing the volume of calculations that need to be made. Also, to account for the high possibility that MOB will appear in loose shots, judgment is made on whether a shot is loose or not, and the process is adjusted accordingly. The method used here starts by detecting people within the cut's identification range. A shot is judged loose if ten or more people are detected, in which case face recognition processing is not performed.  It is only performed for nine people or less.

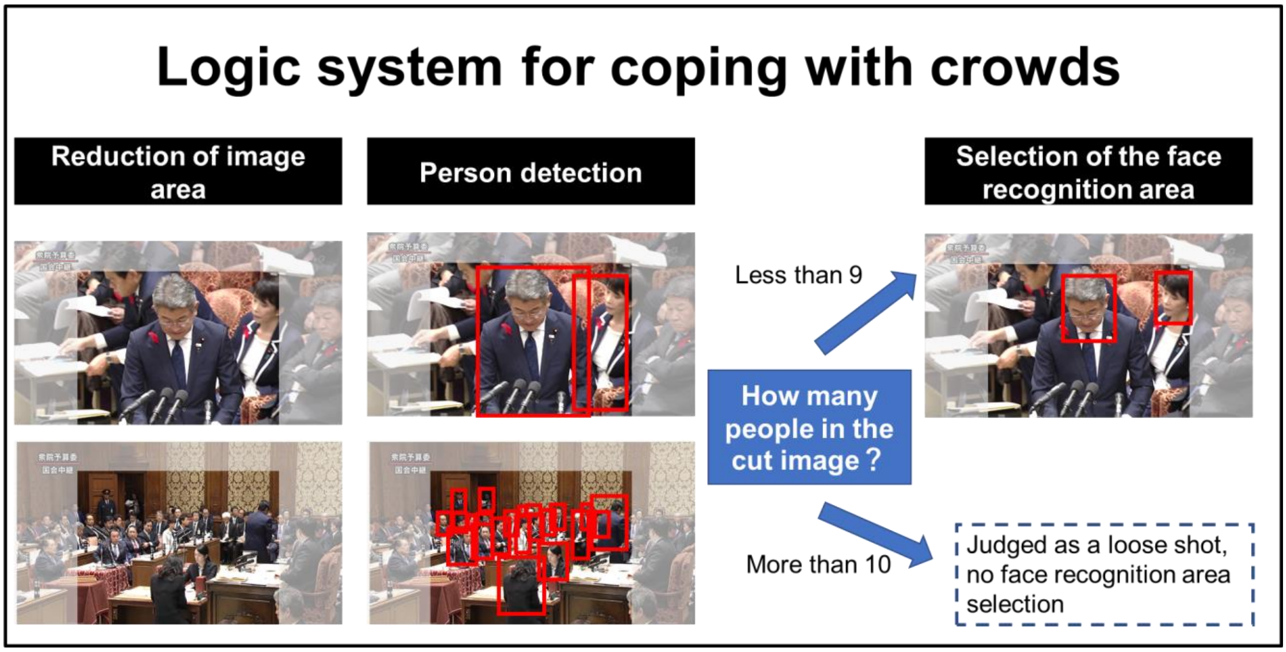

Logic system for coping with crowds
Reduction of image area
Person detection
Selection of the face recognition area
Less than 9
How many people in the cut image ?
More than 10
Judged as a loose shot, no face recognition area selection

Figure 5 – Logic system for coping with crowds

## CSV WITH IDENTIFICATION AND TIME CODE

The identity and time code are converted to CSV to record which person was detected in which frame. This data can be used in various situations where the director wishes to extract a frame in which a specific person was shown.

## CHROMA KEY DISPLAY

Face Detector has a function for displaying only the subject's name together with the other identification data in the bounding box on a green background. This is for use in chroma key synthesis. This synthesis of the background image and Face Detector result provides for smooth display of the background image.

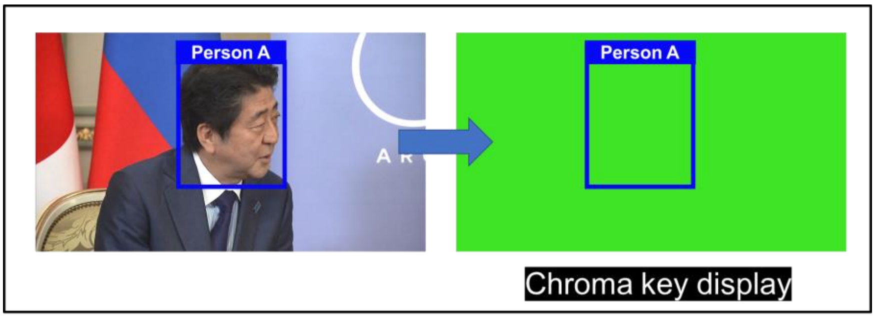


Figure 6 – Real time subject identification

## EXAMPLE USES OF FACE DETECTOR

Face Detector has already been used in the field.

## HAKONE EKIDEN

### What is the Hakone Ekiden?

The Hakone Ekiden is a 200km relay road race leading from Tokyo to the mountains of Hakone. In this event, held in January each year, 21 university teams compete. Each team consists of ten runners, and each runner runs a 20km interval and with a baton (actually a sash) and hands it to the next runner. Its origins are said to be related to the post stations of the old Tokaido Road. NHK broadcasts the Hakone Ekiden live on the radio every year.

### How the Hakone Ekiden is broadcast

The commentators take it in turns to sit in a radio booth at the broadcasting station and broadcast live while monitoring the live TV stream of the race. Seeing only the video images on the screen, each commentator must make quick judgments about who is in a video image and convey what he sees in the image by voice alone. This is no easy task with about 500 people involved in the race, including both runners and coaches. Often, too, many people overlap in the images, especially at the start of the race.

### Introduction of Face Detector

Face Detector was installed in the booth in January 2021, and the names of runners shown in the video were presented in real-time for a total of 13 hours, with almost no mistakes, to help each commentator. The system configuration is shown in Figure 7. 500 Hakone Ekiden models were generated for both runners and staff. Face Detector succeeded in providing the names of runners even in congested scenes and when they were wearing sunglasses, thereby enriching the commentary. Face Detector enabled the commentators to ascertain each person's name without further checking at the relay points and convey it in real-time. The latest AI technology enriched this traditional radio broadcast.

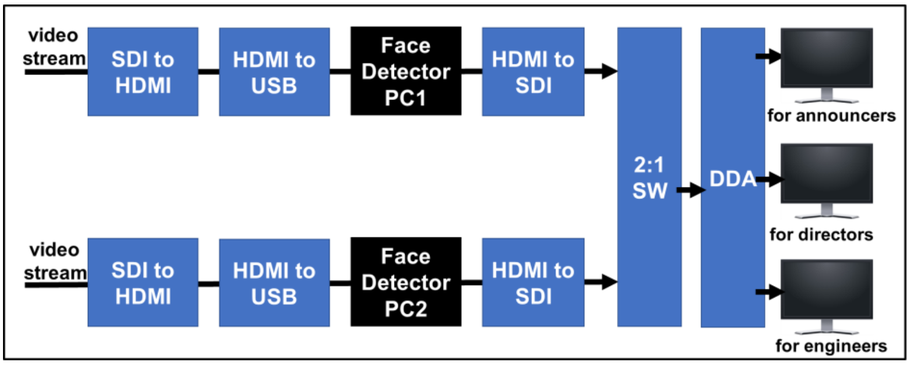


Figure 7 – System used for Hakone Ekiden

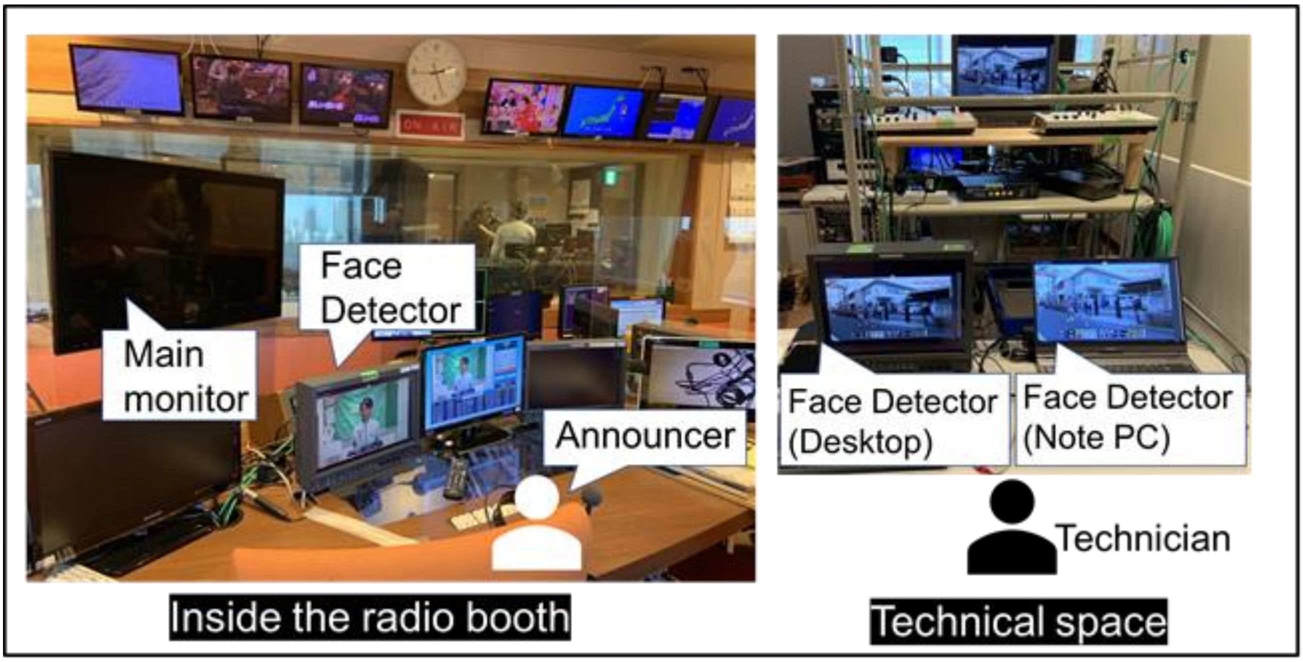


Figure 8 – Use of the system for Hakone Ekiden

## PARLIAMENTARY BROADCASTS

The system was also tested on a TV broadcast from the Japanese parliament. Since NHK's broadcasts cover a question-and-answer session, only the names of the people who are speaking need to be identified, even if many other faces (MOB) are also included in the scene of the chamber.  By using crowd logic, we were able to maintain the speed and accuracy of face recognition, even when the lawmakers were wearing medical masks. People were also identified correctly when viewed from the side. This function could be used for the automatic superimposition of names in the future.

## ASSISTING CAMERAMEN

It can be hard for the cameraman to locate a target person in a sports broadcast. We have developed a cameraman assistance function using Face Detector that displays people’s names in the camera viewer. The outline of this system is shown in Figure 9. The system was tested successfully during the Ekiden race.

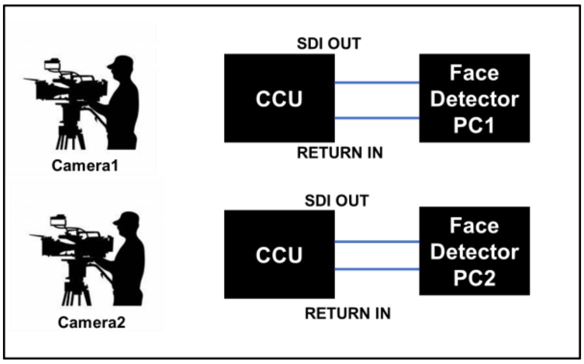


Figure 9 – Assisting Cameraman

## VALIDATION RESULTS

### Accuracy

We assessed Face Detector’s accuracy quantitatively in the parliamentary broadcast according to the two parameters of the identification rate (regarding the number of identified as against unidentified items) and precision rate (regarding the number of true as against false identifications). The identification rate averaged 88%, meaning 12% of images were not identified, and the precision rate averaged 100%, meaning no incorrect names were displayed.

| | Identification rate (%) | Precision rate (%) |
|---|---|---|
| Minister A | 94 | 100 |
| Minister B | 95 | 100 |
| Minister C | 74 | 100 |
| Average | 88 | 100 |

Table 2 - Accuracy

### Speed

This system has a processing speed of 3 frames per second. Application of respondent logic improved the processing speed to 3.1 frames per sec. We believe this system has demonstrated sufficient accuracy for broadcasting use and a high enough processing speed to handle live broadcasts.

### Masks, turbans, sunglasses, and twins

We also confirmed the identification process works when subjects are wearing masks, turbans, or sunglasses. As shown in the figure, identification can be performed successfully even when the face is partially hidden. Furthermore, the system can differentiate identical twins.

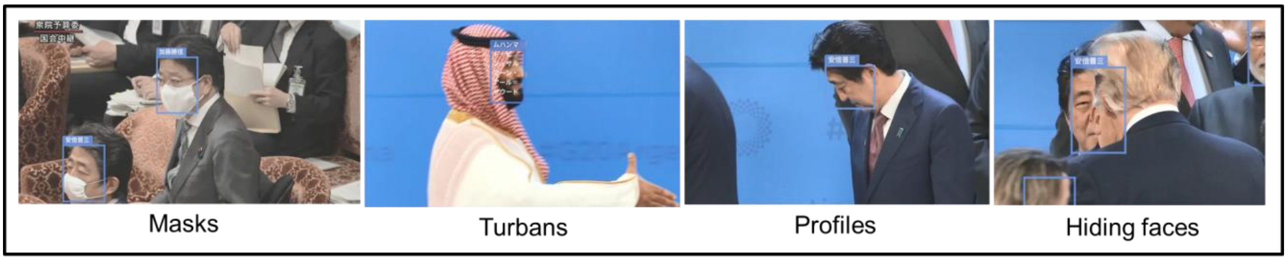


Figure 10 -Recognition results

## CONCLUSIONS

Face Detector provides for the highly accurate, high-speed display of people’s names through the application of real-time face recognition technology. Face recognition systems have been used in various ways in recent years, but this system was developed specifically for broadcast use. It works in real-time and is easy to use. Both the identification and the precision rates are high, and the processing speed of 3 frames per second is suitable for use on live broadcasts where high accuracy is a must. The system also identifies people correctly from the side and when their face is partially obscured by a mask, sunglasses, or turban, and can differentiate identical twins. We have also improved the processing speed and accuracy by developing a special logic system for use in crowded scenes, with due allowance made for the particular characteristics of program production. Regarding model generation, sufficient accuracy was obtained from just one image per person. Operation is simple on the GPU. The fact that the system has been built for PCs makes it easy to use at relay sites as well. The system has been tested successfully at the Hakone Ekiden relay race, including in the field of cameraman support. In the future, we would like to proceed with the development of video editing and highlight scene production systems featuring individual competitors with the help of Face Detector.

## ACKNOWLEDGMENTS

We would like to thank NTT Data for their enormous cooperation in the development of Face Detector. Thanks are due also to NHK for supporting this project.

## REFERENCES

1.Y. Manya. and T. Aoki. 2019, “Development of metadata automatic extraction software for sumo programs using cloud AI” *Broadcast Technology,* vol72*(2019)*: pp64

2. Jia, G., Jiankang, D., Xiang, A. and Jack, Y., “InsightFace,” https://github.com/deepinsight/insightface/, 2020