Political Analysis (2024), **32**, 221–239
doi:10.1017/pan.2023.33

ARTICLE

Face Detection, Tracking, and Classification from Large-Scale News Archives for Analysis of Key Political Figures

Andreu Girbau[1], Tetsuro Kobayashi[2], Benjamin Renoust[3], Yusuke Matsui[4] and Shin'ichi Satoh[1,5]

[1]Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, Japan; [2]School of Political Science and Economics, Waseda University, Tokyo, Japan; [3]Institute for Datability Science, Osaka University, Osaka, Japan; [4]Department of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan; [5]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

Corresponding authors: Andreu Girbau and Tetsuro Kobayashi; Email: agirbau@nii.ac.jp, tkobayas@waseda.jp

(Received 2 November 2022; revised 21 March 2023; accepted 25 April 2023; published online 6 November 2023)

Abstract

Analyzing the appearances of political figures in large-scale news archives is increasingly important with the growing availability of large-scale news archives and developments in computer vision. We present a deep learning-based method combining face detection, tracking, and classification, which is particularly unique because it does not require any re-training when targeting new individuals. Users can feed only a few images of target individuals to reliably detect, track, and classify them. Extensive validation of prominent political figures in two news archives spanning 10 to 20 years, one containing three U.S. cable news and the other including two major Japanese news programs, consistently shows high performance and flexibility of the proposed method. The codes are made readily available to the public.

Keywords: TV news; content analysis; computer vision; face tracking; person recognition; deep learning

Edited by: Jeff Gill

1. Introduction

Content analysis of TV news has been a primary methodology for political communication research (Riff, Lacy, and Fico 2014). While the study of media audiences has been facilitated by the rapid proliferation of online data collection, content analysis of TV news still relies heavily on human coding. Conventional human coding of TV news is unsuitable for analyzing large amounts of data and, thus, requires the timeframe of the analysis to be limited and random sampling of the data. While this method can track obvious changes between several discrete time points, it cannot capture continuous, often nonlinear changes, and sampling inevitably entails error.

However, the time is ripe to break through the limitations of such conventional content analysis of TV news. First, large-scale TV news data archives are being developed as a result of the digitization of TV broadcasting and the increasing scale and affordability of storage capacity. Furthermore, the rapid development of computer vision technology based on deep learning (DL) is making it possible to analyze TV news content without relying on human coding. To be sure, there are still many aspects of content analysis that cannot be replaced by computer vision. For example, analysis at the level of abstract meaning conveyed to the audience, rather than objective features of the image, still requires manual judgment by humans. Nevertheless, thanks to DL-based computer vision techniques, it is now possible to calculate objective indicators of whether a particular person or object appears in the news with an accuracy comparable to human coding, and in a quantity and speed that is well-beyond the

© The Author(s), 2023. Published by Cambridge University Press on behalf of the Society for Political Methodology.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

capacity of human coding. In particular, today's media environment is rapidly becoming video-centric. While television maintains its position as the primary news media, social media is also being used as news source in video format, such as YouTube, Instagram, and TikTok. Therefore, the application of such computer vision techniques to TV news analysis will make an essential contribution to the methodological development of political communication research.

Against this background, this study proposes a face detection and tracking method based on computer vision technology that is more versatile and flexible than other existing methods. The appearance of individuals in the news, especially politicians and candidates, is a highly relevant driver of media effects. As an illustration, the frequent appearance of politicians is relevant to the name recognition effect (Kam and Zechmeister 2013), as well as the mere exposure effect (Bartels 1988). Politicians' appearance in news is also crucial to the political neutrality of the media (Hopmann, Van Aelst, and Legnante 2012). Therefore, the detection and tracking of prominent political figures constitute a critical element of TV news content analysis. Face detection and tracking is an actively studied area of computer vision and are increasingly being applied to the social sciences (Joo and Steinert-Threlkeld 2022; Torres and Cantú 2022; Williams, Casas, and Wilkerson 2020). We demonstrate the performance of the proposed method by applying it to the analysis of five different TV channels, three corresponding to the U.S. cable news, and two corresponding to the Japanese news programs. The algorithm used in the analysis is made public and freely available to researchers.[1]

2. Literature Review

2.1. Content Analysis of TV News Using Human Coding

Human coding is the gold standard for TV news analysis. Clearly defined rules and reliable coding by well-trained coders can most faithfully measure contents of news. In particular, many valuable findings have been delivered by the studies that have used human coding to conduct long-term content analysis of TV news. For instance, Schulz and Zeh (2005) analyzed television news coverage of four German Bundestag elections between 1990 and 2002 and found an increasing personalization and dramatization of media coverage. Cushion, Lewis, and Kilby (2020) analyzed BBC News and other media and survey data at four time points from 2007 to 2016 and found changes in coverage of issues that were devolved to the municipalities comprising the British Commonwealth. Bucy and Grabe (2007) analyzed news from broadcast networks (ABC, CBS, and NBC) during four U.S. presidential elections from 1992 to 2004 and found that image bites had been used more frequently than sound bites. While these studies capture important changes in TV news, they are limited to discrete point-in-time comparisons, resulting in a low granularity of analysis and failing to capture continuous or nonlinear changes. The majority of TV news content analyses other than these long-term studies cover shorter time periods or focus on single events. In this regard, recent developments in computer vision technology are making it possible to analyze larger volumes of media messages over longer periods of time without relying on sampling and human coding.

2.2. Face Detection and Classification

In this article, we focus on face detection and tracking of politicians because their appearance in news is a theoretically important factor in political communication. Highly accurate automatic detection, tracking, and classification of faces in news is feasible thanks to the recent advances in machine learning (ML), more specifically in the field of DL, by using convolutional neural network (CNN) approaches.

The early methods in ML-based face detection are characterized by employing hand-crafted features, that is, manually designed filters, to extract and classify elements of interest, such as filters focusing on light contrast between areas to detect the possible location of eyes, nose, and mouth within a face. Examples of these early methods include the first real-time face detector using a cascade of

[1] https://github.com/TeleStats/KAO.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

multiple weak classifiers that result on a much stronger detector (Viola and Jones 2001), a person detector using histogram of gradients (HOG) by Dalal and Triggs (2005), and a joint face detection, pose estimation, and facial landmark localization using a tree-structured modeling to be flexible over different perspectives (Zhu and Ramanan 2012).

In contrast, DL-based methods automatically learn meaningful features from the data, instead of using manually defined instructions. State-of-the-art methods for generic object detection and classification use CNNs, which are particularly suitable for processing data with a grid-like structure, such as images. By using CNNs, new families of object detectors have been developed. Some of the early examples, which are still broadly used, are Faster RCNN (Region Based Convolutional Neural Networks) (S. Ren *et al.* 2015), SSD (Single Shot MultiBox Detector) (Liu *et al.* 2016), and YOLO (You Only Look Once) (Redmon *et al.* 2016). Given their superior performance, they were quickly adopted for face detection (Jiang and Learned-Miller 2017; Sun, Wu, and Hoi 2018), producing the current state-of-the-art methods in face detection such as MTCNN (Zhang *et al.* 2016), DFSD (Li *et al.* 2019), and YOLO-face (Qi *et al.* 2021).

In object detection, it is common to do both detection and classification (i.e., what type of object corresponds to each detection) at the same time. This also can be applied to the face detection, where the face classification (also known as face recognition) can be defined as assigning a known identifier (ID) to each detected face. While the combination of face detection and classification at the same time is a powerful approach, it is common to tackle the face recognition as a problem on its own (Schroff, Kalenichenko, and Philbin 2015; Meng *et al.* 2021), applying pre-processing techniques such as face alignment based on facial landmarks, to better focus on the face characteristics. These face recognition models are trained with thousands of faces corresponding to many different people with datasets like WiderFace (Yang *et al.* 2016) or VGG-face dataset (Cao *et al.* 2018).

2.3. Applications to Political Science

Automatic coding using computer vision technology reviewed above is steadily being incorporated into political science. For example, in the analysis of still images, it is now possible to estimate the size of political protests, the ratio of male to female participants, the percentage of people with children, and state violence, by analyzing large numbers of images collected from social media (Joo and Steinert-Threlkeld 2018, 2022; Williams *et al.* 2020). A growing number of commercial services based on CNNs are also available without the need to train algorithms themselves. Araujo, Lock, and Velde (2020) used three commercial APIs (Clarifai, Google Cloud Vision API, and Microsoft Azure Computer Vision API) to evaluate the performance of automatic content analysis of images and propose a standard protocol.

More recently, computer vision techniques have been applied to the analysis of video data. Dietrich (2021) applied computer vision techniques to video recordings of legislators' movements in the U.S. Congress and found that it is possible to predict legislator's partisan votes from automatically detected physical movements in Congress. A particularly active area of video research in political science is the analysis of debates in presidential elections. Joo, Bucy, and Seidel (2019) used recurrent neural networks to analyze videos of debates in the 2016 U.S. presidential election and successfully detected facial expressions, emotions, gestures, and related movements (see also Bucy and Gong 2016; Bucy and Stewart 2018). However, the movements of legislators and presidential debates are relatively straightforward to analyze because time is limited and the formats are somewhat standardized.

More challenging is the analysis of general news videos. Political news videos feature a wide variety of characters and backgrounds. Hong *et al.* (2021) conducted a large-scale analysis to detect the faces of relevant individuals, mainly political figures but also considering TV anchors, from the top three cable news channels in the US (CNN, FOX, and MSNBC) for 10 years from 2010 to 2019. The quantitative analysis of 244,038 hours of news videos leads to valuable findings, such as that men have more screen time than women. Tarr, Hwang, and Imai (2023) attempted CNNs-based face detection using videos of political campaign ads. They succeeded in automatically extracting summary images of the videos and detecting the faces of the people in them with high accuracy. However, they were not able to accurately

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

Figure 1. Example of publicly available downloaded images for the analysis. Left: The 53 target individuals in the U.S. TV. Right: The 41 politicians to analyze in Japanese TV.

measure the screen time of the political actors because the analysis was done on a per summary image basis. Screen time is an indispensable measure for calculating indicators of political fairness in TV programs, such as "stopwatch fairness," which requires continuous face detection and tracking. In addition, many of the existing studies applying CNNs to the analysis of political videos require person-by-person training, and are not easily used by social scientists who are not necessarily familiar with computer vision. Therefore, more versatile and flexible methods are needed for the widespread use of DL-based studies of video in political science. This study proposes a method that enables both face detection and tracking from a few template images without re-training, and demonstrates its performance.

3. Method

To help on the automatic coding of several thousand hours of videos for specific public figures appearance on TV programs, we build a system that detects, tracks, and classifies these individuals based on the face appearance on the screen, providing frame-level accurate information that can then be exploited for further analysis.

3.1. Data Preparation

First, we prepare the data corresponding to the individuals to be analyzed within the TV news broadcast. We first download a few images—ranging from 3 to 5 per person—from the internet where the individuals to detect on the TV videos appear. Each face detected in these online images constitutes part of the template of that specific individual. In Figure 1, we portray examples of downloaded images from the internet for the target individuals. On the left panel, we show the 53 target individuals in the U.S. TV. On the right panel, the 41 politicians to analyze in Japanese TV. Detailed information of the key actors can be seen in Tables 1 and 2 for U.S. TV and Japanese TV, respectively.

3.2. System Pipeline

To detect and classify the specified individuals, we follow a two-stage approach. Figure 2 represents the first stage of our system pipeline. The second stage is face tracking and classification, which is described in Section 3.4.

First, we generate target individuals' templates by detecting their faces from downloaded images and extracting a corresponding feature vector for each detected face. Then, we detect all faces in the frames of a video sampled at one frame per second, as detecting the faces in all frames is computationally unfeasible. We then extract discriminative features, that is, face embeddings, that will be used to

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

Table 1. Target individuals for the U.S. TV channels CNN, FOX, and MSNBC.

Amy Klobuchar	Hillary Clinton	Mitch McConnell
Barack Obama	Jeb Bush	Mitt Romney
Ben Carson	Jim Gilmore	Nancy Pelosi
Bernie Sanders	Jim Webb	Newt Gingrich
Beto O'Rourke	Joe Biden	Orrin Hatch
Bill Clinton	John Boehner	Paul Ryan
Bill De Blasio	John McCain	Pete Buttigieg
Bobby Jindal	Jon Huntsman Jr	Rand Paul
Carly Fiorina	Kamala Harris	Rick Perry
Chris Christie	Kellyanne Conway	Rick Santorum
Dick Durbin	Kevin McCarthy	Ron Paul
Donald Trump	Lincoln Chafee	Sarah Palin
Elizabeth Warren	Lindsey Graham	Steve Scalise
Gary Johnson	Marco Rubio	Ted Cruz
George W Bush	Martin O'Malley	Tim Kaine
George Zimmerman	Michele Bachmann	Trayvon Martin
Harry Reid	Michelle Obama	Tulsi Gabbard
Herman Cain	Mike Huckabee	

Table 2. Target individuals for the Japanese TV news programs NHK News 7 and HODO station.

Akihiro Ota	Nariaki Nakayama	Taro Aso
Banri Kaieda	Natsuo Yamaguchi	Taro Yamamoto
Fumio Kishida	Renho Renho	Toru Hashimoto
Ichiro Matsui	Sadakazu Tanigaki	Yasuo Fukuda
Ichiro Ozawa	Seiji Maehara	Yorihisa Matsuno
Junichiro Koizumi	Seiji Mataichi	Yoshihide Suga
Katsuya Okada	Shigefumi Matsuzawa	Yoshihiko Noda
Kazuo Shii	Shintaro Ishihara	Yoshimi Watanabe
Kenji Eda	Shinzo Abe	Yuichiro Tamaki
Kohei Otsuka	Tadatomo Yoshida	Yukiko Kada
Kyoko Nakayama	Takako Doi	Yukio Edano
Masashi Nakano	Takashi Tachibana	Yukio Hatoyama
Mizuho Fukushima	Takenori Kanzaki	Yuko Mori
Naoto Kan	Takeo Hiranuma	

compare the detected faces in a video against the individuals' templates face embeddings, extracted from the previously downloaded images from the internet. If two feature vectors are close, there is a high chance that they correspond to the same person. The right panel of Figure 2 demonstrates that the facial images of the same individual are positioned relatively close to one another in the feature

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

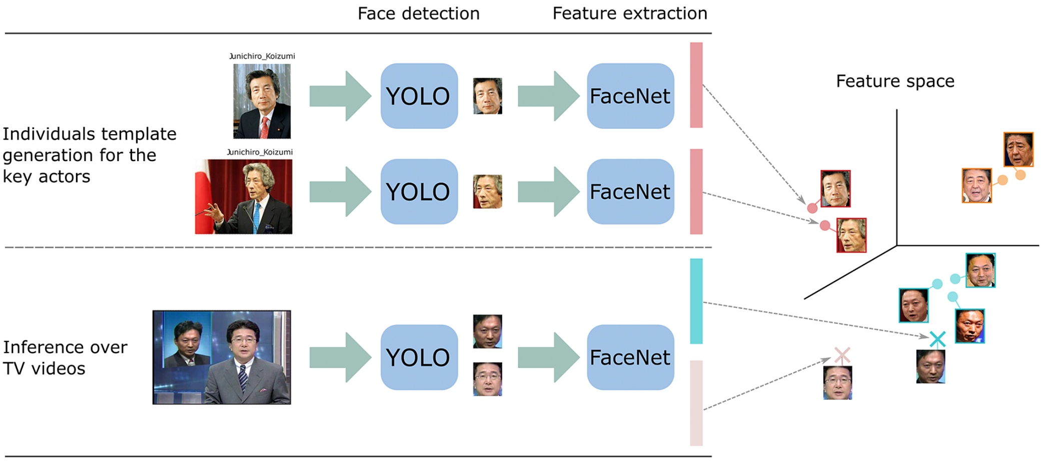


Figure 2. Pipeline for the first stage.

space. As an illustration, the three blue circles correspond to the model images that correspond to Yukio Hatoyama (a former prime minister) gathered from the Internet. The newly detected Hatoyama image from the system on the left panel of Figure 2 is marked as a blue X and is positioned near the three model images, indicating that it is highly likely Hatoyama. Conversely, the pink X that represents the news anchor's face in the same frame is far away from the three Hatoyama model images and therefore, it cannot be classified as Hatoyama, Koizumi, or Abe. In the second stage, we track and compare the face embeddings of the individuals' templates against the faces detected in the video, and classify the video face detections.

3.3. Detection and Feature Extraction

To detect the individuals and extract meaningful appearance features we use off-the-shelf, already trained, state-of-the-art DL methods. To detect faces for both internet images and video frames, we use YOLO5-face (Qi *et al.* 2021), trained on the WIDER face dataset (Yang *et al.* 2016). YOLO-face is a real-time facial detection and recognition system built on top of an object detector from the YOLO family by Redmon *et al.* (2016). The YOLO series is a progression of state-of-the-art DL algorithms that can detect objects in real time with high accuracy. It works by dividing an input image into a grid of cells, predicting the bounding boxes and class probabilities for each cell. Specifically, YOLO-face is designed for detecting and recognizing faces in real-time video streams. The output of the face detector, specified in Figure 2 as *YOLO*, per frame consists on a set of bounding boxes containing the information of the center points (c_x, c_y), and width and height (w, h) of the bounding box.

To generate discriminative face features, we crop each detected face, and project it onto an $\mathbb{R}^D$ feature space (in this work, we use $D = 512$) to generate useful representations. Specifically, for each detected face, we extract a feature vector $\mathbf{h} \in \mathbb{R}^D$ using FaceNet (Schroff *et al.* 2015), denoted as *FaceNet* in Figure 2, with Inception-ResNet-v2 introduced in Szegedy *et al.* (2017) as the backbone, trained with VGGFace2 dataset (Cao *et al.* 2018), consisting of many face annotations for 9,139 unique individuals. The set of feature vectors for a concrete video is defined as $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K\}$, where K is the amount of detected faces in a video.

3.4. Classification

Next, we classify each detected face as an individual or background. To assign an identity, it is common to train a classifier based on the extracted face features, as done by Hong *et al.* (2021). While powerful, this strategy requires re-training of the classifier every time a new person is included in the list of individuals

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

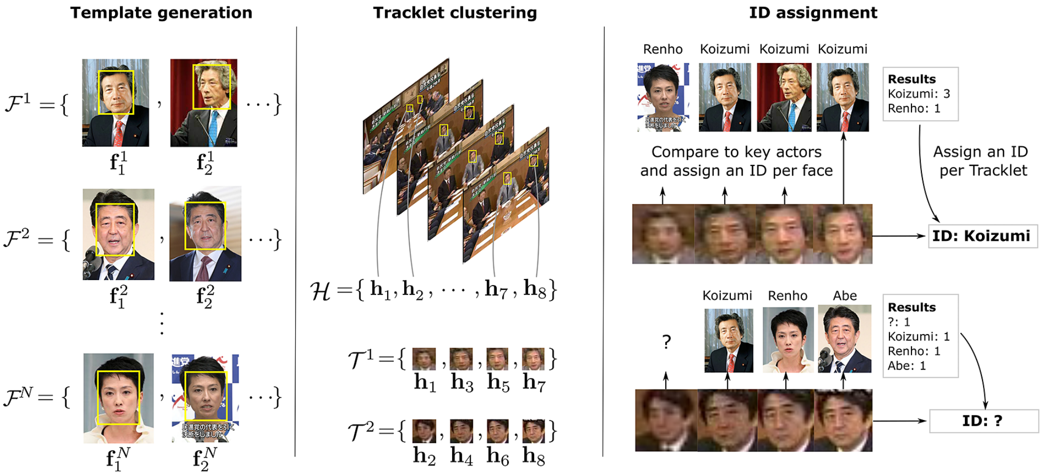

Template generation
Tracklet clustering
ID assignment
$\mathcal{F}^1 = \{$, $\dots\}$
$\mathbf{f}_1^1$ $\mathbf{f}_2^1$
$\mathcal{F}^2 = \{$, $\dots\}$
$\mathbf{f}_1^2$ $\mathbf{f}_2^2$
$\mathcal{F}^N = \{$, $\dots\}$
$\mathbf{f}_1^N$ $\mathbf{f}_2^N$
$\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_7, \mathbf{h}_8\}$
$\mathcal{T}^1 = \{$ $\mathbf{h}_1$ $\mathbf{h}_3$ $\mathbf{h}_5$ $\mathbf{h}_7$ $\}$
$\mathcal{T}^2 = \{$ $\mathbf{h}_2$ $\mathbf{h}_4$ $\mathbf{h}_6$ $\mathbf{h}_8$ $\}$
Renho Koizumi Koizumi Koizumi
Results
Koizumi: 3
Renho: 1
Compare to key actors
and assign an ID per face
Assign an ID
per Tracklet
ID: Koizumi
Koizumi Renho Abe
?
Results
?: 1
Koizumi: 1
Renho: 1
Abe: 1
ID: ?

Figure 3. Representation of template embedding extraction, tracklet clustering, and tracklet ID assignment by voting.

to analyze. To keep the system flexible, we tackle the classification problem as a direct comparison between the face features in the embedding space, assuming that these generated face embeddings are similar for a specific individual (intra-class resemblance), and are discriminative enough between different individuals (inter-class dissimilarity). We separate the classification scheme into two steps. In the first step, we track every detected face in the video, and, in the second step, we assign an identity to each tracked face by comparing the tracked faces in the video against the individuals' templates from the internet images.

Step 1: Face tracking

To pursue face tracking, we make an initial assumption, that is that the appearance of an individual is similar within neighboring frames. Following this assumption, we are able to group similar faces within a video by clustering the face embeddings extracted in the previous step.

We define these face embeddings clusters as *tracklets*, $\mathcal{T} \subseteq \mathcal{H}$, containing a subset of the face positions and embeddings within the same video. These tracklets are the representation of a detected person in a temporal window, for example, capturing the face position of a certain individual along a shot. In Figure 3, we provide a visual example where the video face embeddings in $\mathcal{H}$, are grouped into tracklets $\mathcal{T}^1$ and $\mathcal{T}^2$, which will be compared to the template face embeddings in $\mathcal{F} = \{\mathcal{F}^1, \mathcal{F}^2, \dots, \mathcal{F}^N\}$, corresponding to each target individual $\{1, 2, \dots, N\}$, to finally assign an ID per tracklet that will be extended to all the detected faces within each tracklet.

To cluster the face embeddings to generate the tracklets, we use unweighted pair group method with arithmetic mean (UPGMA) (Sneath and Sokal 1973), which iteratively fuses pairs of clusters, forming a hierarchy.

Step 2: ID assignment

Once the detected faces along a video are grouped in the form of tracklets, we proceed to assign an identity to each tracklet, corresponding to the previously selected individuals or with the "unknown" label. This assigned identity per tracklet is extended to all the face detections within the tracklet. To do so, we opt to use a majority voting scheme, where each of the face embeddings $\mathbf{h} \in \mathcal{T}$ is compared independently against all the instances of the individuals' templates $\mathbf{f} \in \mathcal{F}$ by means of the cosine distance.

Formally, let $\mathcal{Y} = \{1, 2, \dots, N, N+1\}$ be a label set representing N target individuals, and $N+1$ the "unknown" label. The face embeddings of the yth individual is denoted as follows:

$$\mathcal{F}^y = \{\mathbf{f}_1^y, \mathbf{f}_2^y, \dots\}, \qquad (1)$$

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

whereas, each $\mathbf{f}_i^y \in \mathbb{R}^D$. We use $\mathcal{F}$ to represent all faces of all individuals:

$$\mathcal{F} = \mathcal{F}^1 \cup \mathcal{F}^2 \cup \cdots \cup \mathcal{F}^N. \tag{2}$$

To compare embeddings, we define the distance function $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ as follows:

$$d(\mathbf{h}_1, \mathbf{h}_2) = 1 - \frac{\mathbf{h}_1^\top \mathbf{h}_2}{\|\mathbf{h}_1\|_2 \|\mathbf{h}_2\|_2}. \tag{3}$$

This distance function is known as the cosine distance, being 0 for the most similar embeddings, and 2 for the most dissimilar.

Given a tracklet $\mathcal{T}$, let us consider the face embedding $\mathbf{h} \in \mathcal{T}$. We denote the nearest face embedding as:

$$\mathbf{f}^* = \underset{\mathbf{f} \in \mathcal{F}}{\operatorname{argmin}}\, d(\mathbf{h}, \mathbf{f}), \tag{4}$$

where $\mathbf{f}^* \in \mathcal{F}$ corresponds to the most similar face between the face embedding $\mathbf{h}$ and the face embeddings of the N target individuals contained in $\mathcal{F}$.

Here, we consider the voting function $\mathbf{v} : \mathbb{R}^D \to \{0,1\}^{N+1}$. This function receives a face embedding and estimates its label as a form of one-hot vector (the last element is for "unknown"), defined as follows:

$$\mathbf{v}(\mathbf{h}) = \begin{cases} \mathbf{i}(\mathbf{f}^*), & \text{if } d(\mathbf{h}, \mathbf{f}^*) \le \alpha, \\ [0, \ldots, 0, 1]^\top, & \text{else.} \end{cases} \tag{5}$$

Here, the indicator function $\mathbf{i}(\cdot)$ takes a face embedding as an input, and returns its label, for example, $\mathbf{i}(\mathbf{f}_1^3) = [0,0,1,0,\ldots,0]^\top$. If the distance $d(\mathbf{h}, \mathbf{f}^*)$ is greater than a threshold α, the voting function returns the "unknown" label, which is represented by a one-hot vector which final element is one.

Finally, we can accumulate the voting results of the target tracklet $\mathcal{T}$ as a single counter vector $\mathbf{c} = [c_1, c_2, \ldots c_{N+1}]^\top \in \mathbb{R}^{N+1}$ by

$$\mathbf{c} = \frac{1}{L} \sum_{\mathbf{h} \in \mathcal{H}} \mathbf{v}(\mathbf{h}), \tag{6}$$

where L is the length of the tracklet $\mathcal{T}$. The sum of the voting for the yth class is represented as c_y, therefore, the most voted label is obtained by

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, c_y. \tag{7}$$

Finally, we apply a threshold to the result to obtain the final label:

$$y_{\text{final}} = \begin{cases} y^*, & \text{if } c_{y^*} > 0.7, \\ N+1, & \text{else,} \end{cases} \tag{8}$$

meaning that the most voted identity for the tracklet $\mathcal{T}$ should have at least the 70% of the votes in order to be labeled as a specific individual, otherwise, the faces within the tracklet are labeled as "unknown," specified as the label $N+1$. The robustness check for this change in voting thresholds is reported in Section SI1 of the Supplementary Material.

In Section 4.4, we compare this tracklet voting-based classifier against a K-nearest neighbor (KNN) classifier, and a classification by directly computing the distance between the centroid of the tracklet $\mathcal{T}$, defined as the average of all the embeddings within $\mathcal{T}$, and the target individuals embeddings in $\mathcal{F}$.

4. Evaluation and Performance

In this section, we introduce the evaluation regarding our method's performance over five different TV channels from two regions, corresponding to U.S. and Japanese TV.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

4.1. Datasets

To demonstrate the performance and flexibility of our method, we tested it against two datasets, one corresponding to the U.S. cable news: CNN, FOX, and MSNBC, and the other containing Japanese TV news programs of NHK (the Japanese public broadcaster) and ANN (a commercial broadcaster). For Japanese news, we chose the news programs "NHK News 7" from NHK and "HODO (news) Station" from ANN because these two programs have the highest viewership ratings among public and commercial broadcasters, respectively. News 7 is an every-day evening hard news program, lasting half an hour per day, while HODO station, which is a weekday evening soft news consisting of news reports and talk shows, lasting an hour and a half.

The acquisition of video data is a multifaceted undertaking that entails technical as well as legal considerations. Technical complications may surface when capturing high-resolution video streams on a daily basis, whereas legal constraints such as potential copyright infringement may limit the types of content that can be recorded and made publicly available. In this study, the annotations for U.S. TV were provided by Hong *et al.* (2021), to check our system's performance on the U.S. news videos. We downloaded the specific news videos corresponding to the annotations from Archive.org, which is an archive of years of TV recordings. For Japanese data, we utilized a system developed by Katayama *et al.* (2004) to capture and save video streams from multiple Japanese television channels over the course of several years, thereby enabling us to conduct this research.

4.1.1. U.S. TV Dataset

The dataset was created by Hong *et al.* (2021), which originally covers from year 2010 to 2019, containing annotations of 53 relevant individuals (politicians, TV anchors, and other).[2] In this paper, we evaluate from year 2012 to 2019, due to our lack of access to 2010 to 2012 data.[3] The dataset contains 4,346 randomly sampled frames from programs (TV news or talk shows) of three TV channels, CNN (1,083 frames), FOX (1,674 frames), and MSNBC (1,589 frames), with a single face annotation per frame corresponding to one of the individuals of interest. The 4,346 frames in the U.S. dataset correspond to 874 randomly sampled TV programs totaling 1,140 hours (CNN: 280h, FOX NEWS 415h, MSNBC: 445h).

4.1.2. Japanese TV Dataset

The Japanese NHK dataset was originally developed by Katayama *et al.* (2004), which led to works that studied the social networks of politicians based on their co-appearance on news (Renoust *et al.* 2016), a study on the effectiveness of face detection and text-based detection for DL models (Ren *et al.* 2019), or topics associated with politicians appearance in TV news (Renoust *et al.* 2021). We significantly improved this work by adding high-quality manual annotation, extending the coverage to 2001–2021, and adding a commercial Japanese news broadcast (HODO station 2014–2021).

We randomly sampled excerpts of videos from both Japanese programs at one frame per second, and annotated 41 politicians who served as party leaders at least once from 2001 to 2021. The News 7 annotations cover from 2013 to 2021, and HODO station annotations cover from 2014 to 2021, owing to the availability of video corpus. For every sampled frame, we annotated a region of interest, known as bounding box, covering each face of the politicians of interest that are present in the frame. As a result, this dataset contains 19,823 annotated frames from NHK's News 7 (11,141 frames) and ANN's HODO (news) station (8,682 frames), resulting in 29,000 manually annotated bounding boxes (15,101 for NHK News 7 and 13,899 for HODO station). An example of annotations for both datasets can be seen in Figure 4.

[2] We appreciate Hong *et al.*'s (2021) authors for kindly sharing the dataset.

[3] Note that in Table 4, we compare our results (years 2012–2019) to Hong *et al.* (2021) (years 2010–2019), in the best effort to make the comparison as fair as possible.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

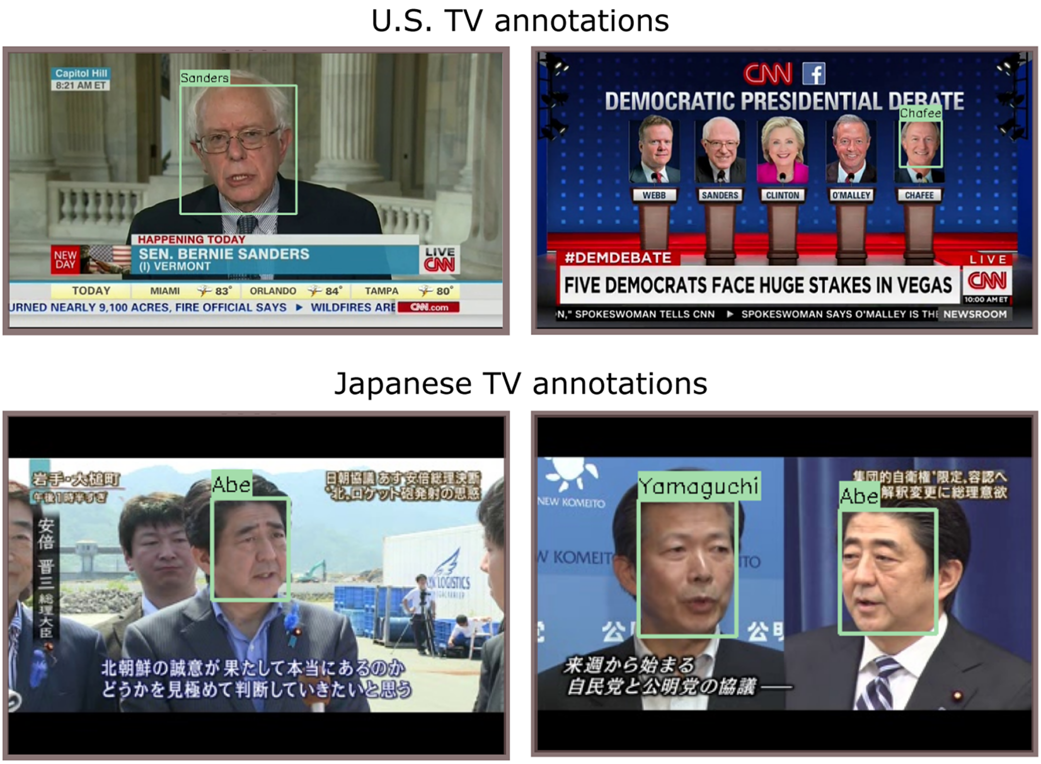


Figure 4. Annotation examples for U.S. TV and Japanese TV annotations. The U.S. TV dataset consists of a single face annotation per frame, whereas in the Japanese TV dataset, all individuals of interest are annotated in every frame. Green bounding boxes indicate the annotated ground truth data.

With these two datasets, we are able to test the ability of our method to detect and classify the same individuals within different channels, as well as its performance given different target individuals and different video settings *without the need of any kind of re-training*.

4.2. Metrics

To evaluate the performance of our method, we use the metrics for object detection employed by the widely used COCO dataset (Lin *et al.* 2014). We compute the Precision, Recall, and F1 score metrics. In addition, we study the performance of our method for different detection sizes, ranging from very small to very large, and computed the F1 score and the mean Average Precision (mAP) for each size.

Precision (P), Recall (R) and F1 scores, are defined as:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\,,\ \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\,,\ \mathrm{F1} = 2\cdot\frac{\mathrm{P}\cdot\mathrm{R}}{\mathrm{P}+\mathrm{R}}\,, \tag{9}$$

where TP corresponds to *True Positives* (correct matching between annotated ground truth (GT) and detections), FP to *False Positives* (faces wrongly classified), and FN to *False Negatives* (faces present in the GT that are not classified). Note that a face of an individual that is wrongly classified will count both as FP and FN.

The mAP score is defined as a weighted mean of precision values at different Intersection over Union (IoU) sensitivities, in the form of thresholds, of the detector.

$$\mathrm{mAP} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{AP}_k\,, \tag{10}$$

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

Table 3. Missed detections percentage per TV channel over the three different face detectors DFSD, MTCNN, and YOLO-face.

Detector	CNN	FOX	MSNBC	NHK	HODO	OVERALL
DFSD	4.62%	3.88%	5.16%	3.13%	6.28%	4.61%
MTCNN	5.36%	4.60%	6.36%	2.84%	8.55%	5.54%
YOLO	1.11%	0.96%	1.01%	0.65%	0.33%	0.81%

where AP_k is the average precision of the class k for a specific IoU threshold, and N the number of classes in the dataset.

Given the sparse annotations present in the U.S. dataset (a single annotation per frame), we follow the evaluation criteria of Hong *et al.* (2021), and only consider the detections with an associated GT. For the Japanese TV dataset, where all faces per frame are annotated, we consider all the detections and classifications from our system, that is, also considering misclassifications of random people in the background as one of the individuals of interest. Note that this difference in evaluation criterion will affect the amount of FPs, conditioning the Precision score to be higher in Table 4 and lower in Table 5.

4.3. Procedure

For the U.S. TV dataset, we download the videos containing the validation frames used in the paper presented by Hong *et al.* (2021) stored in Archive.org. The videos have a resolution of 640x360 for CNN and 640x480 for FOX and MSNBC, and are recorded at 30 or 60 frames per second, depending on the program. For the Japanese TV dataset, we gathered the data for both TV programs, NHK News 7 and HODO station, whose videos have a resolution of 352x240, and recorded at 30 fps. In our method, for each dataset, we model the set of corresponding individuals using 3 to 5 images publicly available in the internet.

To address potential dependency on different detectors and classification strategies, we tested three popular face detectors, YOLO-face (Qi *et al.* 2021), DFSD (Li *et al.* 2019), and MTCNN (Zhang *et al.* 2016), trained on WIDER face for YOLO-face and DFSD, and VGGFace2 for MTCNN, and tried three different classification strategies. The main classification strategy of our method, introduced in Section 3.4, is presented as *vote* in Tables 4 and 5. The *centroid* classification strategy averages the feature vectors within a tracklet and compares the resulting embedding to each target individual template embeddings, and *KNN* classifies each detected face in the video using the KNN algorithm (in this analysis, we set $K = 3$) introduced by Fix and Hodges (1989).

Note that throughout the analysis of our method, the evaluation criteria for US and Japanese data differ, as described in Section 4.2. Specifically, for the U.S. dataset, only detections with an associated ground truth are considered for evaluation. This is because the dataset has sparse annotations, with only a single face annotated per frame. This approach reduces the number of false positives, resulting in a more lenient evaluation and increasing the Precision score. However, for the Japanese dataset, all detections are considered, regardless of whether they are associated with ground truth. This provides a more realistic evaluation of the model's performance, as it considers all possible misclassifications within a frame.

4.4. Analysis

First, we check the detection capabilities of the different face detectors we tested in Table 3, as missed detections highly impact the performance of the overall system. From this preliminary analysis, we observe that YOLO-face outperforms DFSD and MTCNN by a considerable margin in all tested channels, providing the best possible performance for the whole detection and classification system shown in Tables 4 and 5.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

Table 4. Evaluation of our method over the three U.S. TV cable news channels.

		Overall			CNN			FOX News			MSNBC		
Detector	Classifier	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DFSD	KNN	1.0	0.63	0.773	1.0	0.63	0.773	1.0	0.66	0.796	1.0	0.60	0.750
MTCNN		1.0	0.62	0.764	1.0	0.64	0.779	1.0	0.64	0.782	1.0	0.58	0.732
YOLO		1.0	0.68	0.807	1.0	0.69	0.814	1.0	0.70	0.822	1.0	0.65	0.786
DFSD	Centroid	1.0	0.74	0.849	1.0	0.74	0.850	1.0	0.76	0.860	1.0	0.72	0.837
MTCNN		1.0	0.71	0.830	1.0	0.73	0.844	1.0	0.74	0.852	1.0	0.66	0.793
YOLO		1.0	0.77	0.873	1.0	0.80	0.891	1.0	0.79	0.883	1.0	0.73	0.846
DFSD	Vote	1.0	0.76	0.862	1.0	0.78	0.875	1.0	0.77	0.867	1.0	0.73	0.844
MTCNN		1.0	0.74	0.847	1.0	0.77	0.869	1.0	0.75	0.858	1.0	0.69	0.814
YOLO		1.0	0.79	**0.885**	1.0	0.82	**0.902**	1.0	0.80	**0.889**	1.0	0.76	**0.865**
Hong *et al.* (2021)		0.96	0.64	0.768	–	–	–	–	–	–	–	–	–

Note: For reference, we compare to Hong *et al.* (2021), which needs to be re-trained for newly added individuals. Note that Hong *et al.* (2021) use MTCNN as face detector. The best *F*1 scores overall and for each TV channel are shown in bold font.

Table 5. Comparison of our method between two Japanese TV programs, NHK News 7 and Hodo Station, for different detectors and classifiers.

		Overall			NHK News 7 (2013–2021)			HODO station (2014–2021)		
Detector	Classifier	P	R	F1	P	R	F1	P	R	F1
DFSD	KNN	0.81	0.69	0.738	0.81	0.75	0.776	0.80	0.63	0.701
MTCNN		0.78	0.69	0.731	0.79	0.74	0.766	0.77	0.63	0.696
YOLO		0.80	0.72	0.756	0.81	0.78	0.793	0.79	0.66	0.720
DFSD	Centroid	0.78	0.74	0.760	0.77	0.79	0.780	0.78	0.70	0.741
MTCNN		0.79	0.77	0.778	0.79	0.82	0.806	0.78	0.72	0.749
YOLO		0.77	0.79	0.785	0.79	0.84	0.816	0.75	0.75	0.754
DFSD	Vote	0.83	0.74	0.784	0.82	0.78	0.799	0.85	0.70	0.768
MTCNN		0.86	0.73	0.793	0.86	0.78	0.819	0.87	0.69	0.768
YOLO		0.85	0.78	**0.815**	0.84	0.81	**0.825**	0.87	0.75	**0.804**

Note: The best *F*-scores overall and for each TV channel are shown in bold font.

For the overall face detection, tracking, and classification of the specified individuals, we present the results over the U.S. TV dataset in Table 4. For reference, we extrapolate the results per individual from Hong *et al.* (2021).

Without the need of any kind of re-training across datasets and individuals, the proposed method presents an impressive performance. For both datasets, Tables 4 and 5, the F1 score exceeds 0.8 out of a maximum of 1 points, making the system suitable for further analysis of TV data. The best results are achieved with the YOLO-face detector, using the voting scheme presented in Section 3.4 for classification. Examples of misclassification are presented in Section SI2 of the Supplementary Material. We also probed why our system outperformed Hong *et al.* (2021). Specifically, we controlled for factors other than differences in the use of classifiers and clustering by employing the same detector as Hong *et al.* (2021) and not using tracking (see Section SI3 of the Supplementary Material for details). The results show that while clustering makes the system flexible and renders classification and retraining

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

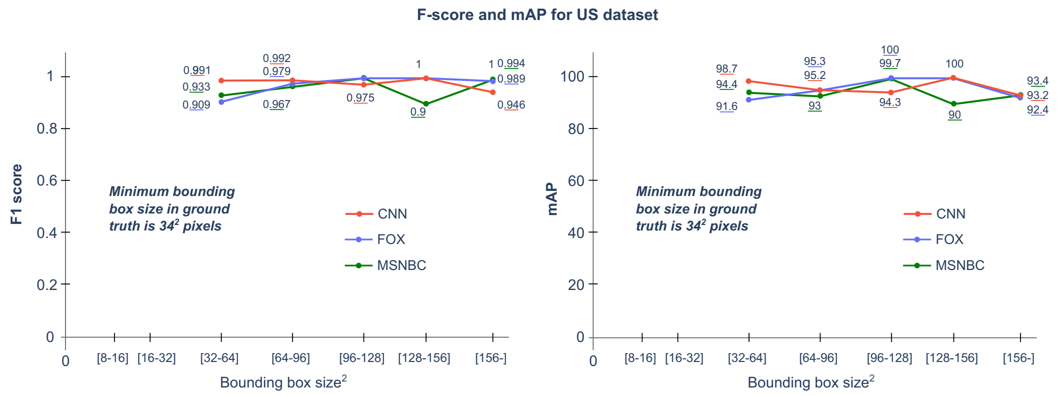


Figure 5. *F*-score and mAP evaluation for different face sizes on the three channels of the U.S. dataset. As ground truth annotations in the U.S. dataset are not tight to the faces, they produce a IoU misalignment with predictions. To compute the mAP, we modify the IoU sweep threshold to $[0.4, \ldots 0.6]$ with an increase of 0.05 per step.

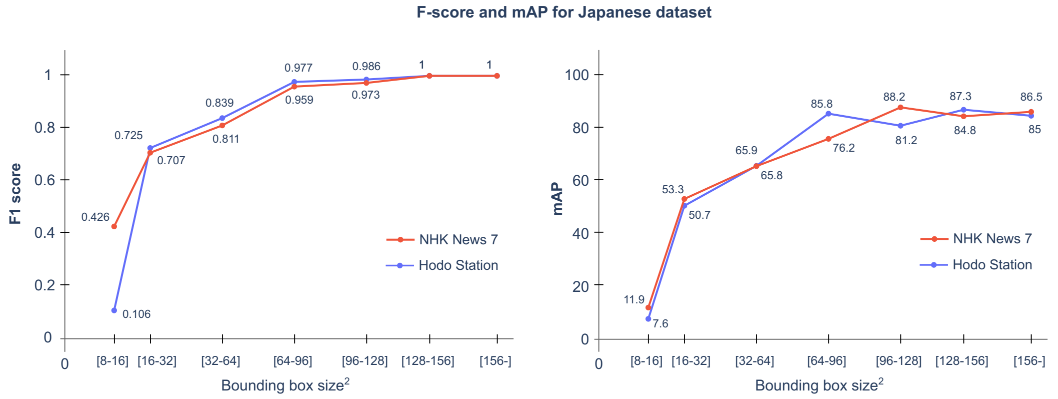


Figure 6. *F*-score and mAP evaluation for different face sizes on the two channels of the Japanese dataset. Here, we follow the COCO standard procedure of computing mAP, with the IoU sweep threshold as $[0.5, \ldots 0.95]$ with an increase of 0.05 per step.

unnecessary, which is the unique advantage of our system, it also undermines overall performance compared with classifiers. However, this performance loss is compensated for by using a more advanced backbone for FaceNet and tracking using tracklets. As a result, our flexible system pipeline outperforms Hong *et al.*'s (2021) system overall.

In Figures 5 and 6, we include a study over the $F1$ score and the mAP at different detection sizes. This analyses help to understand the limits of our system when dealing with very small faces in a video. We filter bounding boxes considering the area in pixels per bounding box for the following squared area thresholds: $[8, 16, 32, 64, 96, 128, 156]$. From these results, we observe the consistent behavior of our method across datasets and all channels, having a very high performance for medium and large faces, struggling when analyzing very small detections.

While the face detector is crucial for a good performance, face tracking, presented in Section 3.4, is also a key component of our system. In Table 6, we studied the performance using or not face tracking as part of the voting classification method. For every tested detector and across every TV channel, generating face tracklets for further classification is key for a better overall performance, both reducing false positives and false negatives, improving the Precision and Recall scores.

With the presented results, we can rely on our method to find specific key actors generically in large-scale video archives for further analysis, without needing any kind of re-training.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

Table 6. Performance with or without face tracking for different detectors.

		CNN			FOX			MSNBC			NHK			HODO		
Detector	Tracking	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DFSD	No	1.0	0.65	0.787	1.0	0.67	0.804	1.0	0.61	0.761	0.79	0.75	0.769	0.78	0.63	0.700
DFSD	Yes	1.0	0.78	**0.875**	1.0	0.77	**0.867**	1.0	0.73	**0.844**	0.82	0.78	**0.799**	0.85	0.70	**0.768**
MTCNN	No	1.0	0.65	0.787	1.0	0.66	0.793	1.0	0.59	0.744	0.78	0.75	0.765	0.76	0.64	0.695
MTCNN	Yes	1.0	0.77	**0.869**	1.0	0.75	**0.858**	1.0	0.69	**0.814**	0.86	0.78	**0.819**	0.87	0.69	**0.768**
YOLO	No	1.0	0.70	0.827	1.0	0.71	0.830	1.0	0.66	0.797	0.80	0.78	0.791	0.78	0.67	0.722
YOLO	Yes	1.0	0.82	**0.902**	1.0	0.80	**0.889**	1.0	0.76	**0.865**	0.84	0.81	**0.825**	0.87	0.75	**0.804**

Note: In bold, we highlight the best *F*-score performance per detector and TV channel.

Table 7. Screen time of major U.S. candidates over the 847 randomly sampled videos for the 2016 general elections.

		CNN	FOX	MSNBC	Total
Republican primary (2016-02-01 to 2016-06-07)	Donald Trump	6,470	2,670	4,548	13,688
	Ted Cruz	3,454	2,330	3,522	9,306
	Marco Rubio	1,838	1,025	1,435	4,298
	John Kasich	1,206	1,246	1,596	4,048
Democratic primary (2015-04-12 to 2016-06-02)	Hillary Clinton	9,953	3,744	9,916	23,613
	Bernie Sanders	8,941	1,828	7,362	18,131
General election (2016-06-02 to 2016-11-08)	Donald Trump	5,064	2,204	6,425	13,693
	Hillary Clinton	3,562	2,476	6,497	12,535

4.5. Application 1: The 2016 U.S. Presidential Election

As an example of the applications of our system pipeline, we first compare the screen time of major candidates during the 2016 primaries and the general election. Because the U.S. data we utilized are a random sample from all cable news data from CNN, FOX, and MSNBC, visualizing longitudinal trends is not suitable as it involves periods without data. Therefore, we examine the sum of each candidate's screen time, which can be found in Table 7. It should be noted that these screen time totals do not match the screen time totals from the entire period.

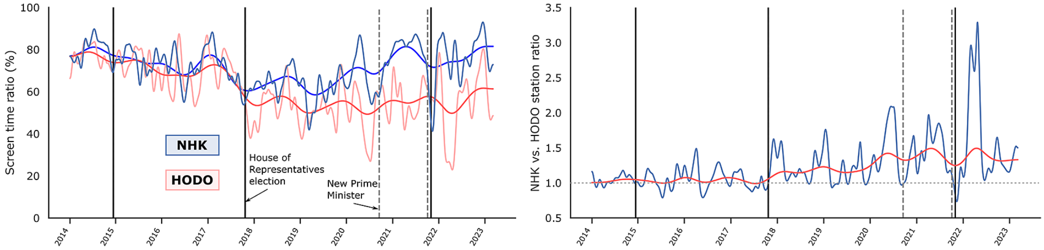
In the Republican primary, Donald Trump received the most screen time overall, which is consistent with the findings of Hong *et al.* (2021). However, while Hong *et al.* (2021) estimated that Trump received more than 2.6 times as much screen time as Cruz (342 vs. 130 hours), our estimate is approximately 1.5 times. This suggests that the disproportionate media attention to Trump may not have been as substantial as reported in prior studies. Furthermore, it is interesting to observe that the attention given to Trump was more prominent on liberal CNN and MSNBC than on conservative FOX. It is possible that during the Republican primaries, the liberal media were warier of Trump and aired him more often as a target for criticism.

On the other hand, in the Democratic primaries, Hillary Clinton had 1.3 times more screen time than Bernie Sanders. On CNN and MSNBC, the two candidates were more balanced in terms of screen time, but on FOX, Clinton had more than twice as much screen time as Sanders, a symmetrical result with the Republican primaries. In the general election, Trump and Clinton's screen time was nearly identical. Thus, utilizing the system pipeline proposed in this study, we find that the inter-candidate imbalance in screen time during the primaries was relatively large for coverage of partisan opponents (Republican primary for CNN and MSNBC and Democratic primary for FOX) and that in the general election, such inter-candidate imbalances are small even on partisan media outlets.

4.6. Application 2: Public versus Commercial Broadcasters in Japan

The next application considers the potential pro-incumbent bias of public compared with commercial broadcasters. As the public broadcaster in Japan, NHK's budget is required by the Broadcasting Law to be approved by the Diet each fiscal year. Furthermore, the prime minister appoints the governors of NHK, NHK's highest decision-making body, with the consent of both houses of the Diet. Moreover, the chairman, who is responsible for NHK's business operations, is appointed by the governors. Therefore, it has been pointed out that NHK is influenced by politics in both budget and personnel matters, which may cause its pro-incumbent bias (Krauss 2000).

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press



(a) Share of screen time of incumbent party leaders among all leaders

(b) NHK/HODO ratio of the share of incumbent party leaders

Figure 7. Share of screen time of incumbent party leaders among all leaders and its NHK/HODO ratio. The three vertical solid lines represent the timing of the House of Representatives elections, whereas the dashed lines indicate the timing of the change of prime ministers.

Figure 7 plots the shares of screen time of the incumbent party leaders in the total screen time of all party leaders since January 5, 2014, for which both NHK and HODO data are available. The incumbent party leaders are those of the LDP and the Kōmeitō party (the coalition partner of the LDP). The trends represent 15-day moving averages, and smoothed lines represent 90-day moving averages.

Figure 7a shows the trends of share of screen time of incumbent party leaders among all leaders. From 2014 to the end of 2019, NHK and HODO show similar trends, with the percentage of incumbent party leaders ranging from 60% to 70%. However, since the end of 2019, the percentage of incumbent party leaders on NHK has been on an upward trend, reaching over 80% by the end of 2022. This means that, of the screen time of all party leaders, the incumbent party leaders (mainly the prime minister) account for 80%. This upward trend is not seen on HODO. Part of the increase in the percentage of incumbent party leaders in NHK arguably reflects the increase in the number of prime ministerial press conferences related to the pandemic. However, the incumbent party holds about 60% of the seats in the House of Representatives and about 57% in the House of Councilors. The fact that 80% of the party leaders' screen time is devoted to the incumbent party leaders indicates that pro-incumbent bias is intensifying, at least in terms of "stopwatch fairness."

Figure 7b more explicitly shows the NHK/HODO ratio of the share of screen time of incumbent party leaders among all leaders. The trend shows the ratio between NHK and HODO was around 1 until 2017, indicating that the screen time share of incumbent party leaders was more or less equal between NHK and HODO. However, it began to rise in 2017 and reached around 1.3 in 2022. Figure 7a shows that it was late 2019 when the screen time share of incumbent party leaders in NHK began to rise, but the NHK/HODO ratio began to rise earlier in 2017 because the screen time share of incumbent party leaders in HODO slightly dropped after the House of Representative election in 2017.

5. Discussion

With the aim to apply recent developments in computer vision techniques to TV news content analysis, this study proposed a versatile and flexible face detection, tracking and classification method that does not need re-training. With the proposed method, feeding only a few template images allows us to analyze the appearances of previously specified individuals of interest in large-scale video settings. As television maintains its position as a major media outlet and communication on social media is increasingly becoming image and video oriented, media messages in political communication are also shifting to video-centric from text-based. Traditional manual coding has been, and will continue to be, a necessary tool that provides valuable insights by most faithfully measuring the contents of media messages. However, analyzing large amounts of "video as data" without relying on sampling and manual coding is essential for detecting long-term, continuous, nonlinear changes in political communication.

We empirically demonstrated the flexibility of the proposed method by testing it over two datasets focusing on U.S. and Japanese TV news, each dataset containing a different set of videos and target

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

individuals. The consistent high performance across the two datasets strongly indicates the robustness and the broad utility of the proposed method across different political contexts.

Before discussing the potential range of applications of the proposed method, some limitations of this study should be noted. This study focuses only on face detection and tracking and does not address other rich information in video data. In addition to face images, videos contain a variety of other information such as voice, body motion, closed caption text information, and background music, which are extremely useful for analyzing speaker emotions and message framing (Dietrich 2021; Tarr *et al.* 2023; Rheault and Borwein 2019). Although this study contributed to extending existing research by focusing on face detection and tracking, the analysis combined with other types of information is promising for future research.

The potential applications of the face detection and tracking proposed in this study are wide. For example, in many Western democracies, there is a growing trend toward presidentialization, whereby power and media attention are increasingly focused on individual presidents and prime ministers rather than on political parties and other institutions (Garzia 2011; Poguntke and Webb 2007). To demonstrate the increasing media attention to political leaders, one important aspect of presidentialization, the proposed face detection and tracking can be used to examine whether the screen time of presidents and prime ministers is increasing over time. Another future direction is to apply this system to analyze the imbalance of screen time by gender and race in the news. Furthermore, detecting the co-presence of politicians in the news will allow us to extract networks among them (Renoust, Le, and Satoh 2016) or to identify specific news formats such as talk shows.

Analyzing "video as data" is the most recent introduction of machine learning in political science. Given the growing amount of available video data and rapid development of computer vision technology, we can expect that many creative political science studies using computer vision will emerge in the future. To contribute to such future studies, we make the code and models openly available in https://github.com/TeleStats/KAO.

Acknowledgements. The authors would like to thank James Hong for sharing their annotations and providing insightful discussion on the evaluation and Roger Macdonald for providing access and support for downloading U.S. TV data from Internet archive.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.33.

Data Availability Statement. Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code, and can be viewed interactively at the reproducible capsule (Girbau *et al.* 2023). A preservation copy of the same code and data can also be accessed via Dataverse here (Girbau *et al.* 2023).

Funding Statement. This work was supported by JSPS KAKENHI (Grant No. 20K20513).

Competing Interest. The authors declare no competing interest exists.

References

Araujo, T., I. Lock, and B. van de Velde. 2020. "Automated Visual Content Analysis (AVCA) in Communication Research: A Protocol for Large Scale Image Classification with Pre-Trained Computer Vision Models." *Communication Methods and Measures* 14 (4): 239–265.

Bartels, L. M. 1988. *Presidential Primaries and the Dynamics of Public Choice*. Princeton: Princeton University Press.

Bucy, E. P., and Z. H. Gong. 2016. "Image Bite Analysis of Presidential Debates". In *Exploring the C-SPAN Archives: Advancing the Research Agenda*, edited by R. X. Browning, 45–75. West Lafayette, IN: Purdue University Press.

Bucy, E. P., and M. E. Grabe. 2007. "Taking Television Seriously: A Sound and Image Bite Analysis of Presidential Campaign Coverage, 1992–2004." *Journal of Communication* 57 (4): 652–675.

Bucy, E. P., and P. Stewart. 2018. "The Personalization of Campaigns: Nonverbal Cues in Presidential Debates." In *Oxford Research Encyclopedia of Politics*, edited by William R. Thompson. New York: Oxford University Press. doi:10.1093/acrefore/9780190228637.013.52.

Cao, Q., L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. "Vggface2: A Dataset for Recognising Faces across Pose and Age." In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 67–74. Xi'an, China: IEEE.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

Cushion, S., J. Lewis, and A. Kilby. 2020. “Why Context, Relevance and Repetition Matter in News Reporting: Interpreting the United Kingdom’s Political Information Environment.” *Journalism* 21 (1): 34–53.

Dalal, N., and B. Triggs. 2005. “Histograms of Oriented Gradients for Human Detection.” In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, 886–893. IEEE. https://doi.org/10.1109/CVPR.2005.177.

Dietrich, B. J. 2021. “Using Motion Detection to Measure Social Polarization in the US House of Representatives.” *Political Analysis* 29 (2): 250–259.

Fix, E., and J. L. Hodges. 1989. “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties.” *International Statistical Review/Revue Internationale de Statistique* 57 (3): 238–247.

Garzia, D. 2011. “The Personalization of Politics in Western Democracies: Causes and Consequences on Leader–Follower Relationships.” *Leadership Quarterly* 22 (4): 697–709.

Girbau, A., T. Kobayashi, B. Renoust, Y. Matsui, and S. Satoh. 2023. “Replication Data for: Face Detection, Tracking, and Classification from Large-Scale News Archives for Analysis of Key Political Figures.” Version 1. https://doi.org/10.7910/DVN/TBWMCG.

Hong, J., et al. 2021. “Analysis of Faces in a Decade of US Cable TV News.” In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery.

Hopmann, D. N., P. Van Aelst, and G. Legnante. 2012. “Political Balance in the News: A Review of Concepts, Operationalizations and Key Findings.” *Journalism* 13 (2): 240–257.

Jiang, H., and E. Learned-Miller. 2017. “Face Detection with the Faster R-CNN.” In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 650–657. Washington, DC: IEEE.

Joo, J., E. P. Bucy, and C. Seidel. 2019. “Automated Coding of Televised Leader Displays: Detecting Nonverbal Political Behavior with Computer Vision and Deep Learning.” *International Journal of Communication* 13: 4044–4066.

Joo, J., and Z. C. Steinert-Threlkeld. 2018. “Image as Data: Automated Visual Content Analysis for Political Science.” *arXiv preprint arXiv:1810.01544*.

Joo, J., and Z. C. Steinert-Threlkeld. 2022. “Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events.” *Computational Communication Research* 4 (1): 11–67.

Kam, C. D., and E. J. Zechmeister. 2013. “Name Recognition and Candidate Support.” *American Journal of Political Science* 57 (4): 971–986.

Katayama, N., H. Mo, I. Ide, and S. Satoh. 2004. “Mining Large-Scale Broadcast Video Archives towards Inter-Video Structuring.” In *Pacific-Rim Conference on Multimedia*, 489–496. Berlin, Heidelberg: Springer.

Krauss, E. S. 2000. *Broadcasting Politics in Japan: NHK and Television News*. Ithaca: Cornell University Press.

Li, J., Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang. 2019. “DSFD: Dual Shot Face Detector.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5055–5064. https://doi.org/10.1109/CVPR.2019.00520.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. “Microsoft Coco: Common Objects in Context.” In *European Conference on Computer Vision*, 740–755. Springer.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. 2016. “Ssd: Single Shot Multibox Detector.” In *European Conference on Computer Vision*, 21–37. Amsterdam, Netherlands: Springer.

Meng, Q., S. Zhao, Z. Huang, and F. Zhou. 2021. “Magface: A Universal Representation for Face Recognition and Quality Assessment.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14225–14234.

Poguntke, T., and P. Webb. 2007. *The Presidentialization of Politics: A Comparative Study of Modern Democracies*. Oxford: Oxford University Press.

Qi, D., W. Tan, Q. Yao, and J. Liu. 2021. “YOLO5Face: Why Reinventing a Face Detector.” arXiv:2105.12931.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. “You Only Look Once: Unified, Real-Time Object Detection.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, H., F. Yang, B. Renoust, Y. Matsui, T. Kobayashi, and S. Satoh. 2019. “Evaluating Face Tracking for Political Analysis in Japanese News over a Long Period of Time.” In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume*, 51–58.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” In *Advances in Neural Information Processing Systems*, 91–99.

Renoust, B., T. Kobayashi, T. D. Ngo, D.-D. Le, and S. Satoh. 2016. “When Face-Tracking Meets Social Networks: A Story of Politics in News Videos.” *Applied Network Science* 1 (1): 1–25.

Renoust, B., D.-D. Le, and S. Satoh. 2016. “Visual Analytics of Political Networks from Face-Tracking of News Video.” *IEEE Transactions on Multimedia* 18 (11): 2184–2195.

Renoust, B., H. Ren, G. Melançon, M.-L. Viaud, and S. Satoh. 2021. “A Multimedia Document Browser Based on Multilayer Networks.” *Multimedia Tools and Applications* 80 (15): 22551–22588.

Rheault, L., and S. Borwein. 2019. “Multimodal Techniques for the Study of Affect in Political Videos.” Paper presented at the 2019 PolMeth Conference, MIT, Cambridge, MA, July 18–20, 2019.

Riff, D., S. Lacy, and F. Fico. 2014. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. New York: Routledge.

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press

Schroff, F., D. Kalenichenko, and J. Philbin. 2015. “Facenet: A Unified Embedding for Face Recognition and Clustering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Schulz, W., and R. Zeh. 2005. “The Changing Election Coverage of German Television. A Content Analysis: 1990–2002.” *Communication* 30(4): 385–407.

Sneath, P. H., and R. R. Sokal. 1973. “Unweighted Pair Group Method with Arithmetic Mean.” In *Numerical Taxonomy*, 230–234. San Francisco: Freeman.

Sun, X., P. Wu, and S. C. Hoi. 2018. “Face Detection Using Deep Learning: An Improved faster RCNN Approach.” *Neurocomputing* 299: 42–50.

Szegedy, C., S. Ioffe, V. Vanhoucke, and A. A. Alemi. 2017. “Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning.” In *In Proceedings of the AAAI Conference on Artificial Intelligence* 31(1). https://doi.org/10.1609/aaai.v31i1.11231.

Tarr, A., J. Hwang, and K. Imai. 2023. “Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study.“ *Political Analysis* 31(4): 554–574.

Torres, M., and F. Cantú. 2022. “Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data.” *Political Analysis* 30 (1): 113–131.

Viola, P., and M. Jones. 2001. “Rapid Object Detection Using a Boosted Cascade of Simple Features.” In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1–9. Kauai, HI: IEEE.

Williams, N. W., A. Casas, and J. D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification*. Cambridge: Cambridge University Press.

Yang, S., P. Luo, C.-C. Loy, and X. Tang. 2016. “Wider Face: A Face Detection Benchmark.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5525–5533.

Zhang, K., Z. Zhang, Z. Li, and Y. Qiao. 2016. “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks.” *IEEE Signal Processing Letters* 23 (10): 1499–1503.

Zhu, X., and D. Ramanan. 2012. “Face Detection, Pose Estimation, and Landmark Localization in the Wild.” In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2879–2886. Providence, RI: IEEE.

Cite this article: Girbau, A., Kobayashi, T., Renoust, B., Matsui, Y. and Satoh, S., (2024). Face Detection, Tracking, and Classification from Large-Scale News Archives for Analysis of Key Political Figures. *Political Analysis*, 221–239. https://doi.org/10.1017/pan.2023.33

https://doi.org/10.1017/pan.2023.33 Published online by Cambridge University Press