International Journal of Communication 13(2019), 4044-4066 1932–8036/20190005

Automated Coding of Televised Leader Displays: Detecting Nonverbal Political Behavior With Computer Vision and Deep Learning

JUNGSEOCK JOO[1]
University of California, Los Angeles, USA

ERIK P. BUCY
Texas Tech University, USA

CLAUDIA SEIDEL
University of California, Los Angeles, USA

For decades, nonverbal communication scholars have employed manual coding as the primary research methodology for systematic content analysis of nonverbal behaviors such as facial expressions and gestures. Manual coding of visual data, however, is expensive and time consuming and therefore not suitable for studies relying on large-scale data. This article introduces a novel computational methodology that can automatically analyze visual content of human communication from visual data. Based on computer vision techniques, the method allows to automatic detection and classification of diverse facial expressions and communicative gestures that have been manually coded in traditional work. To demonstrate the new method, we develop a computational pipeline to classify fine-grained facial expressions and physical gestures and apply our technique to the first 2016 U.S. presidential debates between Donald Trump and Hillary Clinton. The results confirm that computational methods can replicate human coding with a high degree of accuracy for bodily movements and facial expressions, as well as nonverbal tics and signature displays unique to individual candidates. Automated coding should soon facilitate rapid progress in quantitative visual communication research by dramatically scaling up existing manual studies.

Keywords: computational communication science, computer vision, deep learning, nonverbal behavior, facial expressions, gestures, 2016 presidential debates

Jungseock Joo: jjoo@comm.ucla.edu
Erik P. Bucy: erik.bucy@ttu.edu
Claudia Seidel: cseidel2@ucla.edu
Date submitted: 2018–10–17

[1] This research was supported by National Science Foundation Grant 1831848, a Hellman Fellowship, and a UCLA Faculty Career Development Award. We wish to thank Bingbing Zhang and Duncan Prettyman for their assistance with manual coding.

Advances in multimodal media analysis, which involve systematic coding of nonverbal, tonal, and rhetorical elements of political behavior, have enabled researchers to characterize the landscape of political performance at a level of richness and detail that has surpassed earlier efforts to content analyze news or other forms of political communication such as presidential debates (see Bucy & Stewart, 2018; Grabe & Bucy, 2009; Joo, Li, Steen, & Zhu, 2014; Shah et al., 2016; Shah, Hanna, Bucy, Wells, & Quevedo, 2015; Stewart, Bucy, & Mehu, 2015). Although communication researchers have spent much time and effort on the analysis of news frames and other content elements of media coverage (see D'Angelo, 2018), as well as on the rhetorical themes and functions of political debates (e.g., Benoit, 2013; Hart & Jarvis, 1997), the visual analysis of news and politics has been given scant research attention until recently.

Research conducted over the past decade has begun to make up for this deficit in the literature, with longitudinal and cross-national studies of visual framing and leader display behavior (see Bucy & Grabe, 2008; Esser, 2008; Grabe & Bucy, 2009), along with detailed coding of candidate nonverbal communication in presidential debates (Shah et al., 2015, 2016). Although this work is more encompassing than what has come before, and is built on a solid foundation of ethological (i.e., behavioral) analysis, the coding categories lack the descriptive precision of computational approaches and are time consuming because they are performed by hand, relatively limited in scope, and cumbersome to apply. The possibility of human error also requires close training and continuous efforts at quality control and consultation to maintain a high degree of intercoder reliability (Bucy & Gong, 2016). Although useful on a small scale, manual coding does not allow the documentation of behavior on larger scales, which typically limits analysis to single-year studies. As a consequence, researchers working in the area have pointed optimistically to a time when coding of televised political behavior may be captured and recorded algorithmically (Bucy & Stewart, 2018).

This project introduces a novel computational methodology that uses machine classification techniques for analyzing nonverbal behavior, including facial expressions, gestures, and related movements, from recordings of televised political events. Using computer vision and machine learning techniques, the new method allows researchers to automatically detect and classify diverse communicative gestures or any other nonverbal signals that have been manually coded in traditional work. To demonstrate the new computational method, we developed a fully automated computational pipeline that can classify fine-grained facial expressions and body movements. We then applied the method to a C-SPAN video recording of the first 2016 U.S. presidential debates between Donald Trump and Hillary Clinton. The results confirm that automated methods can replicate human coding with a high degree of accuracy for evocative gestures and facial expressions, as well as nonverbal tics and signature displays unique to individual candidates. Automated coding should soon facilitate rapid progress in quantitative research in visual communication by dramatically scaling up existing manual studies.

Studying Nonverbal Cues

Building a conceptual case for systematically studying nonverbal cues in presidential debates begins with an appreciation for the social information that expressive communication imparts (Grabe & Bucy, 2009). Nonverbal information consists of both static capacity cues, including stable candidate characteristics such as height and attractiveness, and dynamic behavioral signals such as facial expressions, eye blinks, head movements, evocative gestures, vocalics, and other communication behaviors. Production features that

affect candidate presentation such as camera angles and shot lengths also influence how viewers see the candidates. Even audience applause and exhortations, particularly in primary debates, have figured into presidential debate research (Stewart, 2012). Particular effects on second screen activity—namely, Twitter messages in response to real-time presidential debate dynamics—have been found for candidate facial displays, evocative gestures, and voice tone (Bucy et al., forthcoming; Shah et al., 2016; Wells et al., 2016). In these studies, nonverbal behaviors are generally found to drive Twitter response more than verbal or rhetorical indicators. Experimental research of candidate exchanges during debates using eye-tracking methodology confirms these findings and documents how display appropriateness elicits viewer attention in the form of gaze fixation and frequency, especially expressions that are categorized as inappropriate (Gong & Bucy, 2016).

The priority that the information processing system gives to nonverbal behavior, particularly facial expressions and displays of emotion, derives from the long evolutionary history of vision in relation to speech and the efficiency with which the brain recognizes and responds to visual signals (see Grabe & Bucy, 2009). Compared with institutions, abstract concepts, and other political structures that are difficult to visualize and hold in memory, images of leaders "are easily recognized and function as effective information processing cues" (Masters, Frey, & Bente, 1991, p. 374). Facial expressions are especially reliable indicators of a communicator's emotional and motivational state, transmitting important social signals to observers (Bucy & Bradley, 2004). Both independently and in conjunction with gestures and paravocal cues such as voice tone, facial expressions serve as the basis of judgments about politically relevant traits, including competence, integrity, political viability, dominance, and appropriateness (Benjamin & Shapiro, 2009; Bucy, 2011; Joo et al., 2014; Joo, Steen, & Zhu, 2015; Joo, Steen, & Turner, 2017). Judgments of competence that predict vote choice may be made from very thin slices of expressive behavior, including just 100-millisecond exposures to candidate photographs (Olivola & Todorov, 2010). The communicative efficiency of expressive displays derives from the extraordinary sensitivity humans, beginning in infancy, show to differences in the facial behavior they observe (Babchuk, Hames, & Thompson, 1985).

In most political communication settings, persuasive influence rests at least as much in the social information carried by nonverbal behavior as the semantic information in candidate pronouncements (Bucy, 2017; Shah et al., 2016). The use of the continuous split-screen presentation format by U.S. television networks in recent presidential debates, in which both candidates are shown without pause regardless of whether they are speaking, has augmented this trend and highlighted nonverbal aspects of candidate behavior to an unprecedented extent. Until recently, however, it was much more common to analyze debates from the perspective of rhetorical argumentation, voter learning about candidate issue stands, or candidate character than to appreciate the wealth of social information that a biobehavioral understanding of debates imparts. Perhaps not surprising, the influence of debate viewing on candidate preference in general elections has been viewed as somewhat limited, with attitude reinforcement, evaluations of candidate traits, and voter learning identified as major outcomes (McKinney & Warner, 2013), owing in part at least to this incomplete conceptualization.

New studies, more nonverbally oriented and methodologically sophisticated than previous waves of presidential debate research, are documenting a wider range of effects—and differences among candidates—using biologically based measures and real-time outcomes (for an overview, see Bucy &

Stewart, 2018). Yet, almost all of this coding to date has been conducted manually with human coders, typically at 30-second intervals (although recent work is looking at 10-second intervals), and recorded for the presence or absence, rather than frequency and duration, of nonverbal behaviors (see Bucy, 2016). Manual coding of presidential debates and other audiovisual content is a time-consuming, painstaking process. To ensure accuracy, multiple passes of the same content are required, and for every hour of content, it takes multiple hours to perform reliable manual coding; the time required only intensifies with coding performed at the level of facial muscle movement (e.g., using the Facial Action Coding System).

In the coding of political facial display behavior, two coding schemes are particularly influential: the political ethology approach developed by Masters, Sullivan, Lanzetta, McHugo, and Englis (1986), which uses emotion/behavioral intention pairs (e.g., happiness/reassurance, anger/threat, fear/evasion), and the Facial Action Coding System (FACS) introduced by Ekman and Friesen (1976), which identifies face movements at the level of individual muscle changes. Both of these systems have been employed in recent political communication research. Given its more general frame of reference, the ethological approach has been more widely applied to news coverage of elections and presidential debates (see Bucy, 2016; Bucy et al., forthcoming; Grabe & Bucy, 2009; Shah et al., 2015, 2016), and FACS coding has been used in experimental and content analytic research of political speeches examining smaller slices of expressive behavior (see Stewart et al., 2015; Stewart & Dowe, 2013; Stewart, Waller, & Schubert, 2009).

Thus far, coding of political communication using either FACS or the ethological approach has been performed manually. Human coding is of course sensitive to nuance in expression and ambiguity, but computer analysis of candidate behavior can bring a much higher degree of precision through continuous tracking of facial display behavior compared with the intermittent coding of manual observation. Instead of one data point for every 30-second (or 10-second) interval, automated coding can produce 30 data points per second, and hence 900 data points for every 30-second interval (or 300 for every 10-second interval). Once a classifier is trained to recognize nonverbal cues based on human coding, the computational approach can offer exponentially more precision and provide a much closer numeric representation of expressed reality. At the same time, computational coding can track considerably more display variability for statistical testing, and exact durations of expressive behaviors can be calculated for a fuller descriptive portrait of the candidates' on-stage performances.

From a practical standpoint, televised presidential debates offer an advantageous testing ground for automated coding because, at least when the candidates are positioned at a podium or sitting at a table, they feature an unobstructed, well-lit view with the candidates in relatively fixed positions and shown against an invariant background. In fixed-podium debate formats, the candidates face the camera more or less continuously, while occasionally glancing down at their notes or at the opposing candidate. This consistency allows researchers to focus on movements of interest without having to track the candidates around the stage (as with a town hall debate format) and parse other figures or objects in the background. Related work by Koppensteiner and Grammer (2010) has used political speeches given in legislative chambers to distill the bodily movements of speakers into stick figure animations; findings show a correlation between the speaker's movement and personality type. In view of these advances in

political behavior analysis, we next review existing approaches to the automated coding of nonverbal data.

Automated Coding of Nonverbal Data

Evolution of Computational Tools for Visual Analysis

In the past decade, numerous computational analytic tools—for example, topic modeling from text data (Blei, 2012), opinion mining and sentiment analysis (Pang & Lee, 2008), or community detection in social media (Papadopoulos, Kompatsiaris, Vakali, & Spyridonos, 2012)—have been developed and applied to research questions in communication. These new techniques are usually innovated by computer scientists and transferred to researchers in other disciplines who then customize these tools to their own data and questions. Computational approaches enable social scientists to incorporate large-scale data sets, such as online archives of news texts or social media posts, into their studies at reduced cost. Taking full advantage of the new tools and techniques, however, typically requires a multidisciplinary effort so that the social scientific and computational elements are equally represented (see Bucy & Stewart, 2018; Joo & Steinert-Threlkeld, 2018).

Given the large volume of research in visual communication and nonverbal behavior analysis, it is also natural to apply advanced computational tools to large-scale data sets of images or videos. As mentioned, manual coding of visual data is notoriously slow and expensive. Computational tools, if they work, can facilitate research progress by substantially lowering coding costs. In computer vision and machine learning, these tools have been developed over decades, but their quality and identification rate were not reliable enough for use in actual applications until very recently, when the field made tremendous gains in accuracy by using artificial deep neural networks (LeCun, Bengio, & Hinton, 2015).

Artificial neural networks are computational models whose structure is defined by a number of internal nodes and their connections. An artificial neural network resembles a biological neural network in its structure and the way that information is passed between neurons. The model and its training algorithm were introduced in the 1980s (LeCun et al., 1989) but remained relatively underused because of computational complexity and difficulty in training. In the recent years, however, neural networks have emerged as the most popular and powerful machine learning framework after computationally efficient GPU-based algorithms were developed and large-scale training data sets became available (Krizhevsky, Sutskever, & Hinton, 2012). For more technical background about deep learning methods, we refer readers to Joo and Steinert-Threlkeld (2018).

Advanced computer vision and deep learning methods have been increasingly employed by scholars in communication and political science to ease the burden of manual coding and widen the scope of analysis. Recent examples include identifying visual framing of political leaders in mass media and its relation to the public opinion (Joo et al., 2014), predicting election outcomes from the facial appearance of candidates (Joo et al., 2015), measuring media bias from facial expressions of politicians (Boxell, 2018; Peng, 2018), predicting and characterizing political ideology from social media images (Xi et al., 2019), estimating the polarization in Congress from human motion data (Dietrich, 2019), detecting deception and assessing credibility during interviews (Pentland, Twyman, Burgoon, Nunamaker, & Diller, 2017; Yu et al., 2015), and

detecting and characterizing social movements and collective actions using social media images (Steinert-Threlkeld, Chan, & Joo, 2019; Torres, 2018; Williams, Casas, & Wilkerson, 2019; Won, Steinert-Threlkeld, & Joo, 2017; H. Zhang & Pan, forthcoming). As these examples demonstrate, automated computational methods in computer vision and machine learning can fit into social science frameworks and enable large-scale quantitative analysis of visual data.

Automated Facial Expression and Emotion Analysis by Computer Vision

Computational techniques for visual analysis have traditionally been developed and used by scholars in computer science, but recently a body of work using computer vision has been applied to communication research. This subsection reviews existing studies incorporating a fully or semiautomated computational approach to analysis, focusing on the applications in nonverbal communication.

Facial Expression Analysis

One of the most common subjects of study in computer vision and nonverbal communication is the human face. This is not surprising because human faces carry substantial information about human subjects and provide strong signals about their emotions, communicative intents, and state of mind. Computer vision techniques have many areas of application, including security, surveillance, data collection, and other routines that involve analysis of or interaction with humans. The approach has thus attracted researchers across many different disciplines. In particular, the automated analysis of facial expressions, which refers to facial motions or changes "in response to a person's internal emotional states, intentions, or social communications" (Tian, Kanade, & Cohn, 2005, p. 247), is a well-studied topic in computer vision.

Within computer vision, a commonly accepted coding scheme for classifying facial expressions is based on FACS developed by Ekman and Friesen (1976). FACS defines a list of elementary facial movements, called action units (AUs), that can be objectively measured for systematic analysis of facial expressions. Action units are also associated with specific facial muscle movements, making the coding system anatomically grounded. Many computational approaches have therefore focused on automatically classifying facial action units from either still images or dynamic videos (Baltrušaitis, Robinson, & Morency, 2016; Tian et al., 2005).

Most automated systems for facial expression analysis involve three major steps. First, the system should detect (localize) a face from an input image. Then, the system extracts facial features, which describe the shape and appearance of the detected facial region. Finally, these data points are transferred to the classification module, which determines the presence or absence of each facial expression (e.g., an AU) in a given face.

Emotion Recognition

The problem of automatically inferring human emotions by recognizing facial expressions from an image or video recording also has been extensively discussed in the computer vision literature. As with facial expression analysis, the goal of emotion recognition is to assign an emotion category to a given facial image.

In the discrete approach to emotion research, human emotions are typically categorized into six "basic" emotions (i.e., happiness, sadness, surprise, anger, fear, and disgust; see Ekman, 1992), along with neutral affect.

Whereas facial expressions describe visible facial appearances and motions without subjective interpretation (e.g., raised eyebrows), inferred emotional states refer to the psychological orientation of the target individual. These states are often revealed by facial manifestations, such as muscle movements, head orientation, visibility of upper or lower teeth, and other indicators. Although facial expressions and emotions are different concepts, the same computational method can be used for both tasks; that is, emotional states can be estimated from the same set of facial features used for facial expression classification. In many studies, facial expressions and inferred emotional states are computed together in the same analytical pipeline (Benitez-Quiroz, Srinivasan, & Martinez, 2016; Joo et al., 2014) as they can share the same facial features.

Automated Coding of Televised Leader Displays

The goal of this study was to demonstrate the utility of new computational tools for nonverbal behavior analysis. Coding of the first U.S. presidential debate of 2016 was performed manually at 10-second intervals to generate a broad set of nonverbal, tonal, and verbal variables using the ethological approach mentioned above (see Bucy et al., forthcoming). In this analysis, we limited our focus to the candidates' nonverbal behavior and used the data from manual coding to train and validate our machine learning model. The trained model attempted to automatically replicate the manual coding for any given debate interval. Below, we describe the data collected for the study, variables used in the analysis, and methods employed to accurately detect candidate behavior from the debate video.

Data: Manual Coding Categories and Definitions

Consistent with a biobehavioral approach to nonverbal communication (see Bucy, 2017; Masters et al., 1986), the first Trump–Clinton debate was coded for nonverbal behaviors associated with leadership and contests for social dominance. Behaviors of interest were selected based on demonstrated influence in previous studies (Bucy, 2017; Bucy et al., forthcoming; Shah et al., 2015, 2016) and significance as communicative cues (Bucy & Stewart, 2018; Gong & Bucy, 2016; Stewart, Salter, & Mehu, 2009), given their capacity to convey a wealth of social information and account for a meaningful amount of variance in vote choice, particularly for undecided voters (see Olivola & Todorov, 2010). Chief among these are competitive and affiliative facial expressions and demonstrative body language or gestures.

From an ethological perspective, four categories of leader displays are of particular consequence for social organization and hierarchy maintenance: anger/threat, happiness/reassurance), fear/evasion, and sadness/appeasement (see Masters et al., 1986; Stewart et al., 2009). These dual terms describe the emotion expressed and intention signaled by the display. Anger/threat displays are characterized by lowered eyebrows, a staring gaze, the visibility of lower teeth, lowered mouth corners (frowning), facial rigidity, or lips that are tightly closed. Happiness/reassurance displays, on the other hand, feature at least one of the following: a smile with relaxed mouth position, the visibility of upper or both rows of teeth, nodding up and

down, brief eye contact to avoid staring, open or just slightly closed eyes, or "crow's feet" wrinkles around the eyes. Other expressions are characterized by equally detailed criteria and have been documented elsewhere (see Bucy & Grabe, 2008; Grabe & Bucy, 2009).

Following detailed coding criteria developed by Masters et al. (1986) and validated longitudinally over several election cycles (Grabe & Bucy, 2009), we coded gestures as body language that signaled affinity or defiance. Affinity gestures consisted of hand, body, or facial movements that suggest a friendly relationship or attempt at bonding between the candidate and the audience, opponent, or moderator. Examples include waving or giving a "thumbs-up"; winking or nodding knowingly to the camera, moderator, or other candidate; or using an open palm when referencing the audience or opponent (rather than a closed fist or pointed finger). Defiance gestures were coded as hand and arm movements that visually signaled challenge to or disregard for authority, belligerence to an adversary "out there," or threatening or dismissive actions toward the opponent (Bucy et al., forthcoming). Examples include finger pointing, wagging, or shaking; making or brandishing a fist; shaking one's head in disagreement or disapproval; prolonged stares; or other behaviors signaling aggression.

Manual coding was conducted by two trained graduate student coders. Each had advanced methods coursework, was closely supervised during the coding process, and worked from detailed variable definitions. Primary coding was performed by a Chinese graduate student fluent in English but with little knowledge of or ideological investment in American politics, minimizing the possibility of partisan bias. The secondary coder for intercoder reliability was an American graduate student, but his only task was to double-code 10% of the sample for reliability purposes, and his coding was not used in the analysis.

Primary coding involved documenting the presence or absence of each behavior of interest within each 10-second time interval throughout the debate, first for one candidate and then for the other. This process yielded 533 individual debate segments for both Trump and Clinton. The coding for this data set was conducted at a nominal level and did not produce durations of candidate behaviors; instead, manual annotations simply documented whether a given behavior was present or absent in a given time interval. Intercoder reliability for this manual coding of nonverbal behavior, reported elsewhere (see Bucy et al., forthcoming), ranged from 84% to 93% agreement, depending on the variable.

To test the robustness of our approach, we coded for the presence of additional nonverbal behaviors, including neutral expressions (the absence of an affective display), candidate gaze direction (looking at the opponent, looking directly at the camera), display appropriateness (inconsistent, exaggerated, or arbitrary expressions), nonverbal dismissals (visual disagreement, disparagement, or disbelief; also, sweeping arm movements intended to wave off or wave away the opponent), nonverbal tics (lip moistening and other tongue shows, head bobbing, podium gripping, water drinking, and nose or mouth touching), eyebrow movements (brow raises, brow furrows), and degree of candidate agency or overall activity (combined intensity of physical movement). We also coded for interruptions given that they involve intensive mouth movements and visually measurable gestural interjections. A brief summary of each measured behavior is listed in Table 1.[2]

[2] A detailed list of all coding definitions is available from the authors.

These manually coded behaviors were used to train our automated classifier, which were then applied to uncoded data to generate machine coding. We elaborate the method and procedure in the following section.

Method: Recurrent Neural Network With Face and Body Features

Feature Extraction

Our automated coding system consisted of two parts: feature extraction and classification. These two steps can be combined, but we used separate modules for a clearer interpretation. For each frame in the input video, we first used OpenFace (Baltrušaitis et al., 2016), an off-the-shelf open source library, to extract features from the candidates' faces. OpenFace was developed to facilitate research in facial behavior analysis and supports many functions, including facial landmark (key point) detection, head pose estimation, eye gaze detection, and facial action unit classification. Figure 1 shows a screen shot using OpenFace on our debate video, revealing detected features from the two candidates.

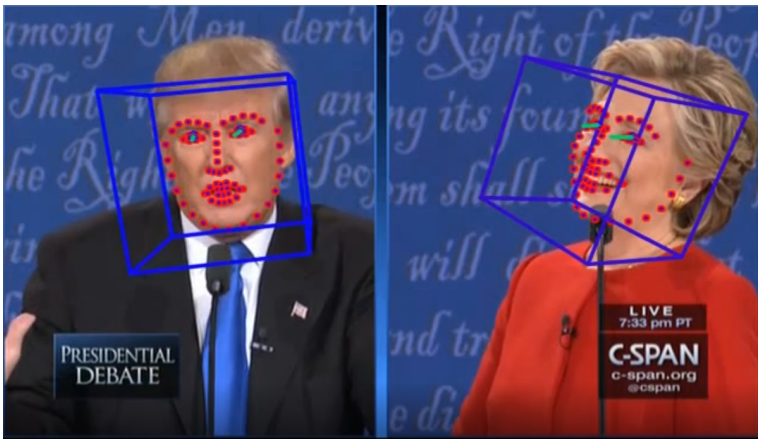

Figure 1. A screenshot from OpenFace showing facial landmarks, head orientation, and eye gaze.

We also classified facial action units (AUs) from each detected face by OpenFace. Detected results are shown in Figure 2. We retrieved two different types of AU predictions: (1) the presence of each AU (binary) and (2) its intensity (numeric). We included both terms in our final classification module. AU prediction is a challenging task and can produce inaccurate classification results. Also, because candidates in presidential debates are shown speaking, it is sometimes difficult to identify whether features such as parted lips or an open mouth should be interpreted as a signal of certain emotions or expressions (e.g., anger or happiness) or are simply because they are speaking. Therefore, we did not make any inferences directly from detected AUs, instead using them as features for the subsequent classification. The actual semantics of such facial features were automatically discovered during the training process in the classification module. Figure 2 shows cropped faces of both candidates with facial AUs detected by OpenFace.

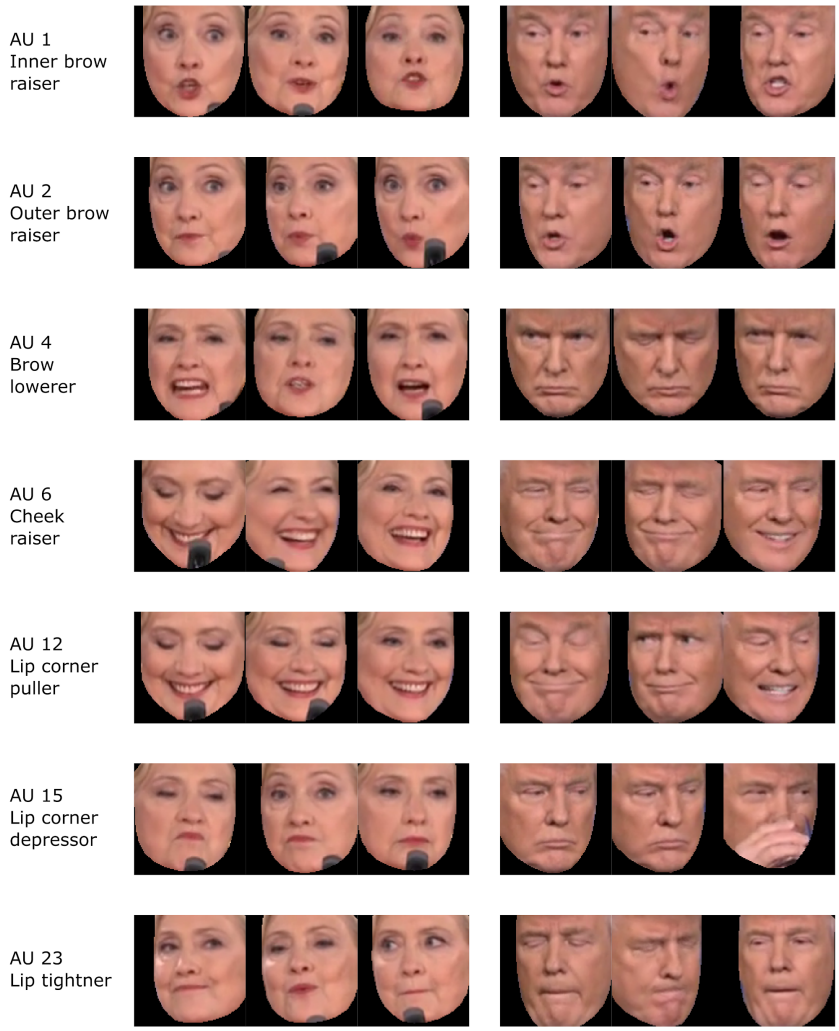

Figure 2. Cropped faces with facial action units (AUs) detected by OpenFace.

Humans also use fluent body gestures along with facial expressions in nonverbal communication (Joo et al., 2014), behavior that can be automatically coded by the computer vision technique of pose estimation. Specifically, we used OpenPose (Cao, Simon, Wei, & Sheikh, 2017; Simon, Joo, Matthews, & Sheikh, 2017) to detect a person in each frame of the video and the locations of 25 body joints including face, arm, hand, and fingers. OpenPose is another publicly available library that can be used for pose estimation from images or video recordings and is based on convolutional neural network techniques. Figure 3 shows examples of candidate hand gestures captured in both frames.

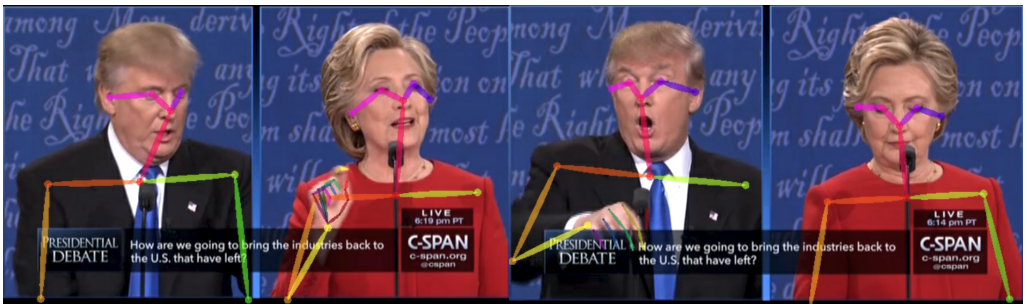

Figure 3. Screen captures illustrating body pose estimation using OpenPose.

Visual Code Classification

After the feature extraction step, our system fed the features obtained at each frame of the video to our classification module, which produced the final coded behavior outputs. In this study, we used a recurrent neural network (RNN) with long short-term memory (LSTM) units (Hochreiter & Schmidhuber, 1997), which has been a popular method for video recognition in the computer vision literature to classify the final outputs.

RNNs are a special form of neural network with a cyclic connection. They are commonly used for *sequence* data, when there exists internal dependency between elements in the sequence. For instance, a text sentence is a sequence of words in which arrangement of the words is critical to understand the meaning of the entire sentence. An RNN maintains its internal state while processing the whole sequence, allowing it to leverage the context between elements. RNNs also allow inputs of variable lengths, such as sentences with different numbers of words, without changing the model structure. Of course, these are both important properties when dealing with video input because video recordings consist of a sequence of frames, temporally connected.

Specifically, an RNN is an iterative function that takes an input sequence and updates its internal states. Let x_t be the input vector at time t and h_t be the internal (hidden) state vector at time t. In our experiment, the dimension of the hidden state was set to 64. At each time step t, an RNN, f, updates states as follows:

$$h_t = f(x_{t-1}, h_{t-1}), t \in \{0, 1, 2, \ldots . T-1\},$$

where T denotes the number of elements in the input (e.g., the number of frames in a video); h_t is the output at each time step, and h_{T-1} is the last output computed. The initial state of the model, h_0, is set to zero at the beginning. The exact function f in our model is an LSTM. LSTMs have an additional state variable, called a *cell state*. The cell state maintains specific information that needs to be kept while processing the whole sequence and the LSTM controls when the information needs to be updated. This feature is useful when modeling data with a long-term dependency. At the last time step, the hidden states (h_{T-1}) are used to generate the predicted coding values through a linear combination.

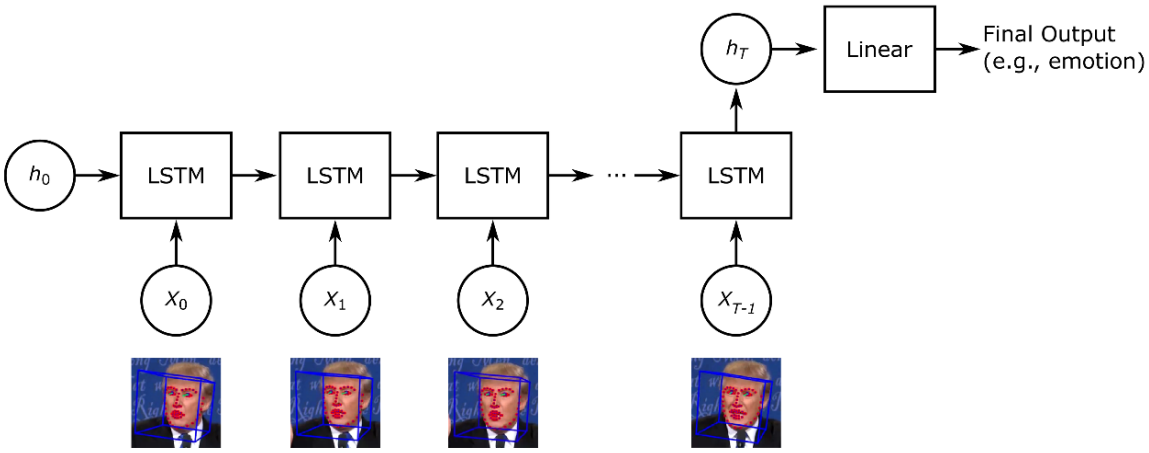

Figure 4. Our model, a recurrent neural network with long short-term memory (LSTM), processes a sequential input (facial and body features) computed from video frames.

As illustrated in Figure 4, our model started by taking the first frame of the input video clip. The actual input was the feature vector obtained by OpenFace and OpenPose. As the model proceeded through time steps, it updated internal hidden and cell states. These states' values were stored and passed to the computation for the next time step. The states were then combined with the feature vector from the next frame and the model continues updating the states. This repeated until the model finished computations for all elements in the entire sequence, and the final outputs were computed based on the last values in the hidden state:

$$y = \sigma(h_{T-1} \cdot w + b),$$

where $\sigma(x)$ is a sigmoid function whose output value ranges from 0 to 1, w is a weight vector, and b is a bias term, which are learnable parameters of the model.

Model Training

The annotations used in training were made for each 10-second time window, which resulted in 533 time intervals for the entire debate duration. Such resolution, considered coarse from a computer vision standpoint, was necessary to synchronize the machine coding with the original manual coding of the debate. We next randomly divided the entire data set into a training set (80%) and validation set (20%). The training set was used during training, and the validation set was excluded as this should be used to measure the classification accuracy of the model. The training was conducted by a popular optimization method by Kingma and Ba (2014) called Adam for three epochs[3] for each code separately.

[3] One epoch means using every data instance in the training set once.

Results

For reliable measurement of accuracy, we performed 10-fold cross validation and report the accuracy by the area under the receiver operating characteristic curve (AUC). The AUC is a standard measure of binary classification accuracy commonly used in machine learning, ranging in value from 0.0 (worst classifier) to 1.0 (perfect classifier). A value of 0.5 indicates the classifier is uninformative for the given task. Table 1 shows the means and standard deviations of AUCs for each variable obtained from the cross validation.

Table 1. Classification Accuracy for Automated Coding of Candidate Nonverbal Behavior.

	Clinton			Trump			
Coding	*M*	*SD*	Frequency	*M*	*SD*	Frequency	Behavior definition
Look at	.960	.018	.511	.909	.025	.640	The candidate in the reaction shot looks at the speaking candidate.
Brush off	.774	.037	.126	.801	.053	.083	The candidate in the reaction shot visually brushes off the opponent.
Disagreement	.892	.045	.096	.907	.017	.706	The candidate in the reaction shot displays nonverbal disagreement.
Look into	0.937	0.011	0.677	0.830	0.040	0.800	The candidate looks directly into the camera (i.e., breaks the "fourth wall").
Eyebrow	0.911	0.026	0.326	0.754	0.052	0.913	The candidate displays noticeable eyebrow movement.
Angry face	0.847	0.051	0.130	0.912	0.035	0.585	The candidate shows an angry/threatening facial expression.
Happy face	0.880	0.014	0.506	0.841	0.041	0.089	The candidate shows a happy/reassuring facial expression.
Sad face	0.776	0.034	0.017	–	–	0.000	The candidate shows a sad/appeasing facial expression.

Neutral face	0.836	0.025	0.345	0.881	0.043	0.292	The candidate shows a neutral facial expression.
Affinity gesture	0.821	0.047	0.066	–	–	0.000	The candidate uses any affinity gesture.
Defiance gesture	0.745	0.039	0.098	0.833	0.033	0.334	The candidate uses any defiance gesture.
Agentic gesture	0.991	0.010	0.489	0.939	0.006	0.613	The candidate engages in an "agentic" style of behavior.
Wave off	0.906	0.039	0.370	–	–	0.000	The candidate "waves off" the opponent with a dismissive hand and arm swipe.
Tic–lip	0.736	0.042	0.149	0.771	0.055	0.136	The candidate moistens his/her lip.
Tic–bob	0.855	0.047	0.149	0.713	0.097	0.117	The candidate bobs his/her head.
Tic–grip	0.863	0.071	0.004	0.821	0.048	0.547	The candidate grips the podium.
Tic–drink	–	–	0.000	0.952	0.040	0.026	The candidate drinks water.
Tic–touch	–	–	0.000	0.691	0.049	0.040	The candidate touches his/her nose or mouth.
Interrupt	0.866	0.185	0.023	0.647	0.059	0.094	The candidate in the reaction shot attempts to interrupt the speaking candidate.

Note. Blank fields indicate that the candidate did not exhibit the behavior according to the manual coding.

Overall, the accuracies average 0.825 for Trump (range = 0.646 to 0.952) and 0.858 for Clinton (range = 0.774 to 0.991), showing better than chance classification (0.500). The best performing variables for both candidates were looking at the opponent (look at) and overall agency, or level of activity (agentic). Specific facial expressions were far easier for the classifier to recognize for Trump than for Clinton, who was much less expressive throughout the first debate (see Bucy et al., forthcoming). Other notable discrepancies in classification occurred for instances of nonverbal disagreement (labeled disagreement) and evocative gestures (labeled affinity gesture, defiance gesture) when, again, Trump was much more demonstrative than Clinton in his nonverbal behavior. Interestingly, the classifier was better at recognizing Clinton's affinity

gestures and Trump's defiance gestures, suggesting that a candidate's preferred or more natural mode of communicating—agreeableness for Clinton, hostility for Trump—was more reliably recognized.

The analysis also revealed certain behaviors performed by one candidate but not the other. Ever defiant and menacing, Trump did not show any sadness/appeasement (e.g., lowered mouth corners, head turned down toward body, averted eye orientation) that was coded by annotators. And, despite his wealth of gesturing, Trump did not engage in any waving off of his opponent with a dismissive hand or arm swipe through the air, a signature expression of Clinton's extending back to her 2016 presidential primary debates against Bernie Sanders. Clinton also did not reach for a glass to drink water at any time during the first debate (tic–drink), whereas Trump did on several occasions.

Figure 5. The left panel shows frames with Trump's facial expressions misclassified as "happy" (false positives). The right panel shows frames of Trump smiling while listening to Clinton, more accurately classified as "happy" (true positives).

External Validity and Generalizability

Another important criterion in model validation is to assess how well it can generalize to new, unseen data. Given that our data set consisted of only two individuals, it is difficult to demonstrate that the model can generalize to other politicians because their debate styles and nonverbal gestures may vary significantly. Therefore, we tested the generalizability of our approach using a publicly available data set of facial images with labels on emotions.

Specifically, we trained a facial expression classifier using a public face data set, called Expression-in-the-Wild (Z. Zhang, Luo, Loy, & Tang, 2015), which contains 91,793 faces. This classifier takes a facial image as input and determines the present expression in relation to seven discrete emotional displays: angry, happy, sad, fearful, disgust, surprised, and neutral. Here, we used a convolutional neural network (CNN), which takes an individual image as input rather than a video. This classifier was not trained on any part of our debate videos or the annotations made on them, so we could measure the generalization performance by applying it to our debate data for the variables related to facial gestures. Because the model processes each image (i.e., frame) independently, we applied the model to every frame and computed the average score within a 10-second time interval for each emotion category to make it comparable to our manual coding data. We used the four categories that overlapped in our data set: angry, happy, sad, and neutral.

Table 2 presents the classification accuracy. This analysis reveals several important findings. First, the classifier correctly classified the facial expressions of Trump and Clinton, despite not being trained on

either person. Second, compared with our original method based on RNN and LSTM, the face-based classifier achieved higher classification accuracy for Clinton but weaker results for Trump. Several possibilities could explain this. For example, Clinton's face might display expressions and emotions that are more commonly found among the people in the Expression-in-the-Wild data set, or Trump's face might show more unique characteristics specific to Trump.

Table 2. Classification Accuracy of a Face-Based Classifier Trained From an External Database (CNN) and Our Video-Based Classifier Trained on Our Data (RNN).

	Clinton		Trump	
Variable	RNN (video)	CNN (image)	RNN (video)	CNN (image)
Angry face	.847	**.865**	**.912**	.697
Happy face	.880	**.947**	**.841**	.670
Sad face	.776	**.846**	–	–
Neutral face	**.836**	.688	**.881**	.600

Note. CNN = convolutional neural network; RNN = recurrent neural network. Bold font indicates better classification accuracy.

Classification accuracy can be also influenced by the temporal coherence and continuity of facial expressions. For instance, we found that this image-based "happy" expression classifier was sometimes triggered by false positives for Trump when he was speaking (see Figure 5) because his face exhibits features consistent with smiling, such as a wide-open mouth. Because most facial image databases consist mainly of posed still photographs, the classifier may not be able to recognize that the person is speaking and thus his mouth is open. In this case, a video-based approach can help distinguish between a real smile and momentary activation of smile-like features. In addition, we found that Clinton smiled much more frequently than Trump during the debate. Trump smiled very infrequently, using his smile only in attempts of "laughing off" allegations made by Clinton.[4]

Interpretability

Automated approaches, especially when using a model with many internal parameters (i.e., deep learning), are often criticized for the difficulty of understanding their internal mechanisms, although the huge number of parameters can also help them achieve a better accuracy and expressiveness. Model intelligibility is indeed an important desired property to communication scholars who wish to identify fine-

[4] Peak moments from the happiness/reassurance facial expression classification for Trump coincided with these four comments made by Clinton during the debate: (1) "I call it Trumped-up, trickle-down." (2) "He owes about $650 million to Wall Street and foreign banks." (3) "Donald publicly invited Putin to hack into Americans' [computer accounts]." (4) "He said, 'You know, if they taunted our sailors, I'd blow them out of the water and start another war.'"

grained elements of nonverbal behaviors or understand which feature set is more important for the model to classify the behavior type rather than simply predict outcome variables. To mitigate the problem and help researchers better understand how the model performs internal computations, we visualize the feature importance of our anger/threat classifier for each candidate in Figure 6.

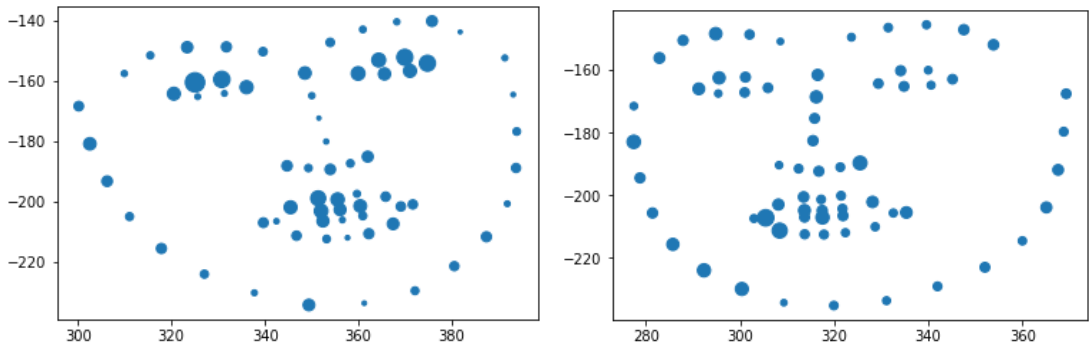

Figure 6. Feature importance for angry/threatening classification of Trump (left) and Clinton (right) computed and visualized by the gradient of the model output with respect to the features. The size of each point indicates the magnitude of the gradient.

The importance of each facial landmark was computed by the magnitude of the model output gradient with respect to each landmark's position. In this example, we see that Trump used more focused movements of individual facial components such as his eyes and mouth to signal anger, whereas in Clinton's faces, the weights are more distributed across many points, suggesting that her anger displays were more restrained or expressed by nonfacial features such as body movements. These types of diagnostic techniques can shed new light on the complex, continuous nature of nonverbal expressions and offer insight into the fine-grained characteristics of political display behavior beyond what manual coding can identify.

Conclusion

As this analysis has shown, our automated classifier is effective at categorizing a range of candidate nonverbal behavior at a rate that far exceeds chance. All 20 behaviors tested from the first presidential debate of 2016 were classified with accuracies in the 65% to 99% range (M = 84.2, SD = 8.3). Importantly, most of these nonverbal behaviors were either ambiguous in nature or required some rhetorical context to accurately classify. Nevertheless, the vast majority of cases were correctly classified, and at a high rate of accuracy, demonstrating the utility of our approach. To enhance the precision of classification in subsequent studies, manual coding at an even finer grained level (e.g., 1-second intervals) may be necessary to achieve a more sensitive set of training data.

Indeed, a limitation of the current analysis is the relatively coarse 10-second unit of observation. Manually documenting candidate behavior at 10-second intervals represents an improvement over 30-second segments used in previous manual coding research but is still an imprecise index of human behavior. In any given 10-second time increment, a candidate may exhibit a variety of expressive displays, whether

emotions of anger and happiness or gestures of affinity and defiance, within this span of time. Moreover, most facial gestures or expressions do not persist for very long. Emotions or gestures that punctuate statements, for instance, may present for only a half second, or even less, and then disappear.

A more granular annotation set with expressive displays manually coded at a rate of once per second would greatly improve the trained models' accuracy. However, this added precision would significantly increase the amount of time and resources required for annotation. As with any complex and novel research endeavor, there are trade-offs inherent to the process, in this case, having some reliable human coding of an entire presidential debate versus more precise ground truth of a much shorter duration. As an initial demonstration of the technique, the existing annotation set (even at 10-second intervals) serves the purpose of the experiment and shows the utility of the new method.

Ultimately, the ability to automate coding of televised nonverbal behavior should present new research possibilities, allowing more rigorous description and testing of nonverbal behavior and visual elements of news. The precision offered by automated coding should bring insights that were not evident before. As the small data of manual coding give way to the big data of automated analysis, scholars should also be able to develop norms for political communication performances across different settings and contexts. Once established, a more precise form of discrepancy analysis may be performed on whether candidates are communicating, say, within or beyond the average range of typical political display behavior, and how these fluctuations and norm violations affect audience response. Would a violation, say, 30% outside the norm for certain settings elicit negative viewer reactions, or would it take a more egregious violation for the audience to react punitively? Are there differences for male and female candidates, older versus younger contenders, or establishment versus "populist" campaigners? Automated coding could provide systematic guidance on political performances that, thus far, have been assessed either impressionistically or through the small *N* of hand coding that cannot keep pace with today's fast-moving—and increasingly visual—media environment.

References

Babchuk, W. A., Hames, R. B., & Thompson, R. A. (1985). Sex differences in the recognition of infant facial expressions of emotion: The primary caretaker hypothesis. *Ethology and Sociobiology*, *6*(2), 89–101. doi:10.1016/0162-3095(85)90002-0

Baltrušaitis, T., Robinson, P., & Morency, L. (2016). OpenFace: An open source facial behavior analysis toolkit. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–10). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/WACV.2016.7477553

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5562–

5570). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/CVPR.2016.600

Benjamin, D. J., & Shapiro, J. M. (2009). Thin-slice forecasts of gubernatorial elections. *The Review of Economics and Statistics*, *91*(3), 523–536. doi:10.1162/rest.91.3.523

Benoit, W. L. (2013). *Political election debates: Informing voters about policy and character*. Lanham, MD: Lexington Books.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. doi:10.1145/2133806.2133826

Boxell, L. (2018). *Slanted images: Measuring nonverbal media bias* (MPRA Working Paper 89047). Retrieved from https://mpra.ub.uni-muenchen.de/89047/

Bucy, E. P. (2011). Nonverbal communication, emotion, and political evaluation. In K. Döveling, C. von Scheve, & E. A. Konijn (Eds.), *Handbook of emotions and mass media* (pp. 195–220). Abingdon, UK: Routledge. doi:10.4324/9780203885390.ch12

Bucy, E. P. (2016). The look of losing, then and now: Nixon, Obama, and nonverbal indicators of opportunity lost. *American Behavioral Scientist*, *60*(14), 1772–1798. doi:10.1177/0002764216678279

Bucy, E. P. (2017). Media biopolitics: The emergence of a subfield. In S. A. Peterson & A. Somit (Ed.), *Handbook of biology and politics* (pp. 284–304). Cheltenham, UK: Edward Elgar.

Bucy, E. P., & Bradley, S. D. (2004). Presidential expressions and viewer emotion: Counterempathic responses to televised leader displays. *Social Science Information*, *43*(1), 59–94. doi:10.1177/05390184040689

Bucy, E. P., Foley, J. M., Lukito, J., Doroshenko, L., Shah, D. V., Pevehouse, J. C. W., & Wells, C. (forthcoming). Performing populism: Trump's transgressive debate style and the dynamics of Twitter response. *New Media & Society*.

Bucy, E. P., & Gong, Z. H. (2016). Image bite analysis of presidential debates. In R. X. Browning (Ed.), *Exploring the C-SPAN archives: Advancing the research agenda* (pp. 45–75). West Lafayette, IN: Purdue University Press.

Bucy, E. P., & Grabe, M. E. (2008). "Happy warriors" revisited: Hedonic and agonic display repertoires of presidential candidates on the evening news. *Politics and the Life Sciences*, *27*(1), 78–98. doi:10.2990/27_1_78

Bucy, E. P., & Stewart, P. A. (2018). The personalization of campaigns: Nonverbal cues in presidential debates. In W. R. Thompson (Gen. Ed.), *Oxford research encyclopedia of politics* (n.p.). Oxford Research Encyclopedias Series. New York: Oxford University Press. doi:10.1093/acrefore/9780190228637.013.52

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7291–7299). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/CVPR.2017.143

D'Angelo, P. (Ed.). (2018). *Doing news framing analysis II: Empirical and theoretical perspectives*. New York, NY: Routledge.

Dietrich, B. J. (2019). *Using motion detection to measure social polarization in the U.S. House of Representatives* (Working Paper 45). Department of Political Science, University of Iowa, Ames, Iowa.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200. doi:10.1080/02699939208411068

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56–75. doi:10.1007/BF01115465

Esser, F. (2008). Dimensions of political news cultures: Sound bite and image bite news in France, Germany, Great Britain, and the United States. *International Journal of Press/Politics, 13*(4), 401–428. doi:10.1177/1940161208323691

Gong, Z. H., & Bucy, E. P. (2016). When style obscures substance: Visual attention to display appropriateness in the 2012 presidential debates. *Communication Monographs*, *83*(3), 349–372. doi:10.1080/03637751.2015.1119868

Grabe, M. E., & Bucy, E. P. (2009). *Image bite politics: News and the visual framing of elections*. New York, NY: Oxford University Press.

Hart, R. P., & Jarvis, S. E. (1997). Political debate: Forms, styles, and media. *American Behavioral Scientist*, *40*(8), 1095–1122. doi:10.1177/0002764297040008010

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Joo, J., Li, W., Steen, F. F., & Zhu, S. (2014). Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp.

216–223). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/CVPR.2014.35

Joo, J., Steen, F. F., & Turner, M. (2017). Red Hen Lab: Dataset and tools for multimodal human communication research. *KI-Künstliche Intelligenz*, *31*(4), 357–361.

Joo, J., Steen, F. F., & Zhu, S. (2015). Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 3712–3720). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/ICCV.2015.423

Joo, J., & Steinert-Threlkeld, Z. C. (2018). *Image as data: Automated visual content analysis for political science*. Retrieved from https://arxiv.org/abs/1810.01544

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. Retrieved from https://arxiv.org/abs/1412.6980

Koppensteiner, M., & Grammer, K. (2010). Motion patterns in political speech and their influence on personality ratings. *Journal of Research in Personality*, *44*(3), 374–379. doi:10.1016/j.jrp.2010.04.002

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Playa Del Rey, CA: Curran Associates.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. doi:10.1162/neco.1989.1.4.541

Masters, R. D., Frey, S., & Bente, G. (1991). Dominance and attention: Images of leaders in German, French, and American TV news. *Polity*, *23*(3), 373–394.

Masters, R. D., Sullivan, D. G., Lanzetta, J. T., McHugo, G. J., & Englis, B. G. (1986). The facial displays of leaders: Toward an ethology of human politics. *Journal of Social & Biological Structures*, *9*(4), 319–343. doi:10.1016/S0140-1750(86)90190-9

McKinney, M. S., & Warner, B. R. (2013). Do presidential debates matter? Examining a decade of campaign debate effects. *Argumentation and Advocacy*, *49*(4), 238–258. doi:10.1080/00028533.2013.11821800

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83–110. doi:10.1007/s10919-009-0082-1

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. doi:10.1561/1500000011

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, *24*(3), 515–554. doi:10.1007/s10618-011-0224-z

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, *68*(5), 920–941. doi:10.1093/joc/jqy041

Pentland, S. J., Twyman, N. W., Burgoon, J. K., Nunamaker, J. F., Jr., & Diller, C. B. R. (2017). A video-based screening system for automated risk assessment using nuanced facial features. *Journal of Management Information Systems*, *34*(4), 970–993. doi:10.1080/07421222.2017.1393304

Shah, D. V., Hanna, A., Bucy, E. P., Lassen, D. S., Van Thomme, J., Bialik, K., . . . Pevehouse, J. C. W. (2016). Dual screening during presidential debates: Political nonverbals and the volume and valence of online expression. *American Behavioral Scientist*, *60*(14), 1816–1843. doi:10.1177/0002764216676245

Shah, D. V., Hanna, A., Bucy, E. P., Wells, C., & Quevedo, V. (2015). The power of television images in a social media age: Linking biobehavioral and computational approaches via the second screen. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 225–245. doi:10.1177/0002716215569220

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4645–4653). doi:10.1109/CVPR.2017.494

Steinert-Threlkeld, Z. C., Chan, A., & Joo, J. (2019). *Repression, cleavages, and protest dynamics* (Working paper). University of California, Los Angeles.

Stewart, P. A. (2012). *Debatable humor: Laughing matters on the 2008 presidential primary campaign*. Lanham, MD: Lexington Books.

Stewart, P. A., Bucy, E. P., & Mehu, M. (2015). Strengthening bonds and connecting with followers. *Politics and the Life Sciences*, *34*(1), 73–92. doi:10.1017/pls.2015.5

Stewart, P. A., & Dowe, P. K. F. (2013). Interpreting president Barack Obama's facial displays of emotion: Revisiting the Dartmouth group. *Political Psychology*, *34*(3), 369–385. doi:10.1111/pops.12004

Stewart, P. A., Salter, F. K., & Mehu, M. (2009). Taking leaders at face value: Ethology and the analysis of televised leader displays. *Politics and the Life Sciences*, 28(1), 48–74.

Stewart, P. A., Waller, B. M., & Schubert, J. N. (2009). Presidential speechmaking style: Emotional response to micro-expressions of facial affect. *Motivation and Emotion*, *33*(2), 125. doi:10.1007/s11031-009-9129-1

Tian, Y.-L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In S. Z. Li & A. K. Jain (Eds.), *Handbook of face recognition* (pp. 247–275). New York, NY: Springer. doi:10.1007/0-387-27257-7_12

Torres, M. (2018). *Give me the full picture: Using computer vision to understand visual frames and political communication*. Retrieved from https://pdfs.semanticscholar.org/153b/49f10be8379e9b9ae52367f0063cb42e6b66.pdf

Wells, C., Shah, D. V., Pevehouse, J. C., Yang, J., Pelled, A., Boehm, F., ... Schmidt, J. L. (2016). How Trump drove coverage to the nomination: Hybrid media campaigning. *Political Communication*, *33*(4), 669–676.

Williams, N., Casas, A., & Wilkerson, J. (2019). *An introduction to images as data for social science research: Convolutional neural nets for image classification* (Working paper). University of Georgia.

Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017). Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM International Conference on Multimedia* (pp. 786–794). New York, NY: Association for Computing Machinery. doi:10.1145/3123266.3123282

Xi, N., Ma, D., Liou, M., Steinert-Threlkeld, Z. C., Anastasopoulos, L. J., & Joo, J. (2019). Understanding the political ideology of legislators from social media images (Working Paper). University of California, Los Angeles.

Yu, X., Zhang, S., Yan, Z., Yang, F., Huang, J., Dunbar, N. E., . . . Metaxas, D. N. (2015). Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual dues. *IEEE Transactions on Cybernetics*, *45*(3), 492–506. doi:10.1109/TCYB.2014.2329673

Zhang, H., & Pan, J. (forthcoming). CASM: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*.

Zhang, Z., Luo, P., Loy, C.-C., & Tang, X. (2015). Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3631–3639). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/ICCV.2015.414