Title
Automating Classification of Nonverbal Cues from Leader Figures

Author
Seidel, Claudia

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Automating Classification of Nonverbal Cues from Leader Figures

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Computer Science

by

Claudia Seidel

2019

ABSTRACT OF THE THESIS

Automating Classification of Nonverbal Cues from Leader Figures

by

Claudia Seidel

Master of Science in Computer Science
University of California, Los Angeles, 2019
Professor Song Chun Zhu, Chair

The need for accurate measures to evaluate and study human expression has grown exponentially in recent years, especially with the proliferation of video and image content across the Internet landscape. However, the study of nonverbal communication still often relies on the creation of data by hand, with humans manually labeling video footage or images. This thesis explores automation of the process through computer vision and machine learning, allowing for better speed and precision. The developed automated classification pipeline is run on video footage of the first and third presidential debates between Donald Trump and Hillary Clinton to gauge its accuracy. Results show the automated pipeline is viable as an easily upscaled replacement for human work, able to both accurately reproduce the results of human labeling of the footage and allow for insight into the various nonverbal idiosyncrasies of the speakers.

The thesis of Claudia Seidel is approved.

Yizhou Sun

Jungseock Joo

Song Chun Zhu, Chair

University of California, Los Angeles

2019

Table of Contents

List of Figures

List of Tables

Acknowledgements

This thesis is based on work done for *Automated Coding of Televised Leader Displays: Detecting Nonverbal Political Behavior with Computer Vision*, forthcoming in the International Journal of Communication. Many thanks to my co-authors, Dr. Jungseock Joo and Dr. Erik Bucy, for setup of the communication-based coding and methodology, and for valuable comments and insight during the process of this work.

I'm also profoundly grateful to my parents for their unending love and support through this incredible time in my life. And finally, to my closest friends: you guys were my rocks when I needed it most! I will never forget it.

1. Introduction

Nonverbal cues are a highly multifaceted way of conveying information, consisting not only of dynamic factors, like facial expressions and gestures with different parts of the body, but also static traits like height. The televising on a massive scale of important events for leaders, such as speeches or debates between political candidates, allows such factors to now influence very wide audiences. It also adds extra dimensions to what can affect viewers (through video editing steps that affect the delivery of candidates' points, like camera angles or cutting of reactions) and opens up potential for expansive secondary research, such as examination of social media responses during such events. This puts a previously unprecedented spotlight on visual analysis in areas like politics and news reporting, where research is still in the process of catching up as the political environment becomes increasingly overwhelmed with visual content.

In this niche of communication studies, there are various existing approaches to analyzing the effect of the face and the emotion it conveys. They are often *ethological* or behavior-based in nature, examining how the visual presentation of political candidate/their behaviors affects public perception of and reaction to them. However, this usually also involves work to manually label the intensity of different emotions over many time intervals, as well as extra steps to ensure that work is reliable. The difficulties and time involved in completing the whole process properly can frequently hold back study in this area. This has caused many to look towards a simpler solution, found in the form of computer vision techniques, which can perform such tasks automatically and more reliably.

This work explores a computational approach to the analysis of nonverbal cues. After first determining the communication-based approach to categorizing different sentiments or key

actions shown by leader figures, a pipeline is established for objective data about subjects' facial expressions, gestures, and posture to flow through to into a learning classifier, which can then predict values in those categories. Footage of the first and third U.S. presidential debates between Donald Trump and Hillary Clinton serves as input to validate the classifier's accuracy. Results from the classifier show that this fully automated system can label video as accurately as manual labeling in significantly less time. They can also help to note certain individual traits of speakers, such as aspects of their speaking style or motions they frequently perform.

Six sections comprise this thesis. With section 1 introducing, section 2 gives background on the application of computer vision to visual analysis, especially the relevant areas of facial analysis and emotion recognition. It also outlines more reasoning behind experimenting with automation in this context. Section 3 discusses the tools used to build a solid foundation of data, and the composition of that data. Section 4 then covers the specifics of classification, as well as the steps taken to feed the data into the classifier, with section 5 explaining the results of this work in further detail. Finally, the paper concludes with remarks in section 6.

2. Background and Motivation

2.1 Analyzing the Face and Emotions

The face has always been one of the biggest points of interest in the study of nonverbal communication. Arguably the most prominent approach to evaluating it is the Facial Action Coding System, or FACS (Ekman and Friesen 1975, 2003), which strictly measures the movement of different muscles in the face, referred to as Action Units (AU). Certain AU

combinations can denote specific emotional displays or states in the subject. This system has not only been used consistently in communication research since its publishing in 1975, but has also been part of computer vision work, applied for anything from detecting drowsy drivers (Vural et al. 2008) to monitoring childrens' expressions as they solve problems (Littlewort, Bartlett, and Lee 2011). Even prior to more modern advances in video quality and processing, researchers could, for example, apply small plastic spheres to a subject's face to mark reference points for FACS AU measurement (Kaiser and Wehrle 1992). After collecting FACS data, the numbers can then be fed into a classifier that can try to label the presence of certain emotions or other expressions based on what it has learned.

As a natural extension of facial analysis, detecting the presence of different emotions has also been a topic of great interest in computer vision research (see Black and Yacoob 1995; Busso et al. 2004; Kahou et al. 2013 for examples over the years). After noting movements of a subject's facial features, that same data can help to gain insight on their mental state beyond what they are saying out loud. FACS does venture into this territory, as mentioned above, but the evaluation of emotions can be a subjective process, and to most effectively study them would require some perspective from the field of communication. The ethological coding scheme by Masters, Sullivan, Lanzetta, McHugo, and Englis (1986) is a prime example of this, directly labeling facial expressions in terms of the more complex emotions and sentiments they display. This approach forms the main basis of the automated classification pipeline in this work, which is explained in detail in section 4.1 later.

2.2 Problem Definition

The necessity of automation is visible just through a look at the methodology of existing techniques. Consider FACS as an example. Its manual labeling requirements (manual work is still one of the more commonly used methods when the system is involved in research) demand annotations by trained FACS coders, who must pass a test for certification. This test is estimated by the system's creators to take 50 to 100 hours to prepare for if self-taught, and they recommend preparing for it in groups to decrease the tedium of the process. The only other approved alternative to become a verified FACS coder is an intensive and monetarily costly five-day workshop with limited availability, which also culminates in taking the test (Rosenberg 2013). If researchers wish to work with this system, they must either take the test themselves, or involve third parties whose eligibility and completed training must be verified.

This exemplifies a common trade-off in this field: human labelers and their intuition can better understand nuance and ambiguity, resulting in more "accurate" labeling of presented emotions, but quality control of those human labelers is often a large time sink in the research process. Even after labelers are verified, their work can be tedious and prone to errors. For example, full televised debates can last 1-2 hours, which results in a very large number of potential ratings to settle by hand, particularly if measurements are in fine-grained intervals. Such intervals can mean at least one image every few seconds, maybe even multiple for every second of video at a very high level of precision, to be reviewed and cross-checked by multiple parties. This raises more issues with the process's reliability, given that the human eye can't always capture subtle shifts in expression even given a few seconds to work with, and makes automation of face and emotion classification an ideal alternative to look towards.

With a problem or goal defined and the appropriate images or video collected for analysis, an automated face or expression classification model will require only a solid set of "ground truth" annotations to learn from. This is particularly effective if the classifier uses a pre-annotated dataset like the CelebA dataset (Liu et al. 2015) or the ExpW dataset (Zhang et al. 2015) to learn about the human face, as training data will not need custom labels before beginning experiments. After classification is completed, there will be enough data present to move into the validation stage, where the model's accuracy is verified. A properly tested classifier can outdo manual annotation of datasets in speed and precision all while keeping the same accuracy as a human labeler, effectively replacing the manual stage of work and still leaving other key aspects of the research process unchanged.

2.3 Computational Approaches

In computer vision, automating this kind of classification is approached through *artificial neural networks*, programmed constructs that emulate the flow of information through a human brain, which allows systems to adapt to different input situations and familiarize themselves with the human face like a person might. One of the first forays into artificial neural networks was done by LeCun et al. (1989), using an early type of "learning network" to recognize valid US zip codes in images of handwritten number sequences. In the earlier days of computer vision, researchers would use simpler techniques such as multilayer perceptrons, one of the most basic neural networks, essentially consisting of many layers of "neurons" with binary 0/1 output that link up to perform more complex calculations. When detecting or analyzing the face, these would usually take basic information like head orientation and key points on the face to monitor

as input, which could then be used to identify certain facial expressions on a person (Zhang 1999). Another common method for these tasks was support vector machines (SVMs), which were commonly employed for face detection tasks because of their success with pattern recognition. This strength meant they could make the classification between "face" and "not a face" categories faster than other alternatives at the time (Guo 2000).

While such methods are still used in computer vision today, their simpler nature now often leaves them better suited for tasks like the earlier problem of text analysis. More refined methods to evaluate human beings have evolved to match the growing performance of contemporary machines. One of the first more complex systems used in this area is the convolutional neural network (CNN), a regularized version the earlier multilayer perceptron. Both CNNs and their ancestor can be categorized as *feed-forward neural networks.* In such networks, computation is based only on a single initial input, which the system simply performs different operations on in sequence to arrive at its output. For this reason, feed-forward networks are best suited for tasks such as image classification, where input can be a given image and output can consist of a set number of categories or labels, in this case things like basic emotions (e.g. "happy," "angry," "sad") or facial expressions ("smiling", for example). Not only is the input dimension always at a fixed size, but input images can also be processed independently of each other, making the CNN and its kin suited for applying such label sets to varieties of images depicting different situations.

A very recent power player in this field is the residual learning framework, or ResNet (He et al. 2016). ResNet centers around a problem with many deep neural networks, which is that learning eventually starts to degrade the further down in network layers one gets. When data gets back-propagated back to earlier layers, the constant multiplication involved has the potential to get to the point that no weights are updated, and the network may not continue to serve its purpose at

all. The ResNet authors observed that in plain, unaltered networks, shallower versions would have worse train and test error than deeper versions. The resulting work to counter this is an improvement to the CNN architecture that works by bypassing certain layers of the model, allowing for unprecedented deepness in neural network layers through "skipped connections." In a follow-up paper, they demonstrated that a 1001-layer ResNet could significantly outperform a less expansive 200-layer ResNet in both test and training error (He et al. 2016).

In contrast to the operations mentioned above is the recurrent neural network (RNN), which is a non-feedforward network with a cyclical, more closed nature that made it the learning system of choice for this paper. In an RNN, output from a previous time step's computation is used as input for the next time step, which allows the network to operate with all previous computations kept in mind. This makes them ideal to use on input sequences where elements depend on each other somehow, exactly like the debate videos that are analyzed in this work. It becomes more important to have the bigger picture that RNN provides because a video's frames are connected *temporally*, making it critical to analyze each frame in the context of all the frames before it. Input size is also not fixed in RNN, which means that videos of different lengths could easily be fed into the same classifier to perform the same type of analysis. The details of this work's specific RNN setup are explained in full in section 4.

3. Feature Extraction

3.1 OpenFace

The open-source toolkit OpenFace (Amos, Ludwiczuk, and Satyanarayanan 2016), was used to collect feature data on the faces of the candidates. It can estimate and quantify features such as

what direction the gaze is aimed and how the head is oriented, as well as facial landmarks, which are keypoints denoting the outlines of major facial features like the eyebrows, nose, and lips.

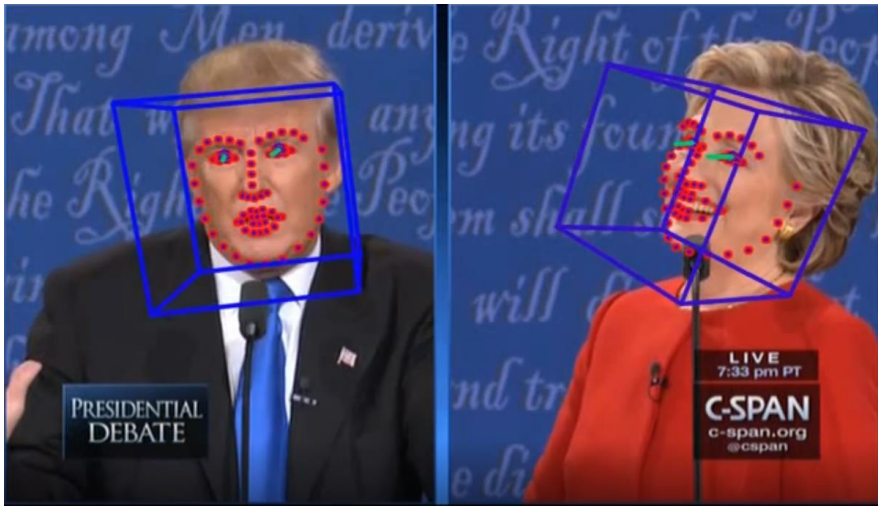

Figure 1*:* A screen capture of OpenFace feature extraction. Facial landmarks are in red points, gaze direction in green lines, and head orientation shown through a blue cube.

An extra data factor not visible above is OpenFace's extraction of facial Action Unit (AU) data, a process automated by Baltrušaitis, Mahmoud, and Robinson (2015). OpenFace collects a subset of the AUs used in FACS encoding, as defined in Section 2, and scores them in two ways: whether they are present, conveyed with a binary 0 or 1, and how intense the AU action is on a scale of 1 (lowest) to 5 (highest). Both scores are included in the classification process later, but it is worth noting that this automatic AU measurement may not lead to correct conclusions if used as a sole basis of information. When analyzing footage like these political debates, AU measurements can be thrown off by how often candidates are talking, since speaking can result in facial muscle movements which could be falsely interpreted as conveying certain emotions. As such, nothing is inferred directly from the generated AU scores. The numbers are instead

simply used for classification along with the rest of the more straightforward OpenFace data.

Figure 2 below shows examples of how the presence of certain AU appears on monitored faces.

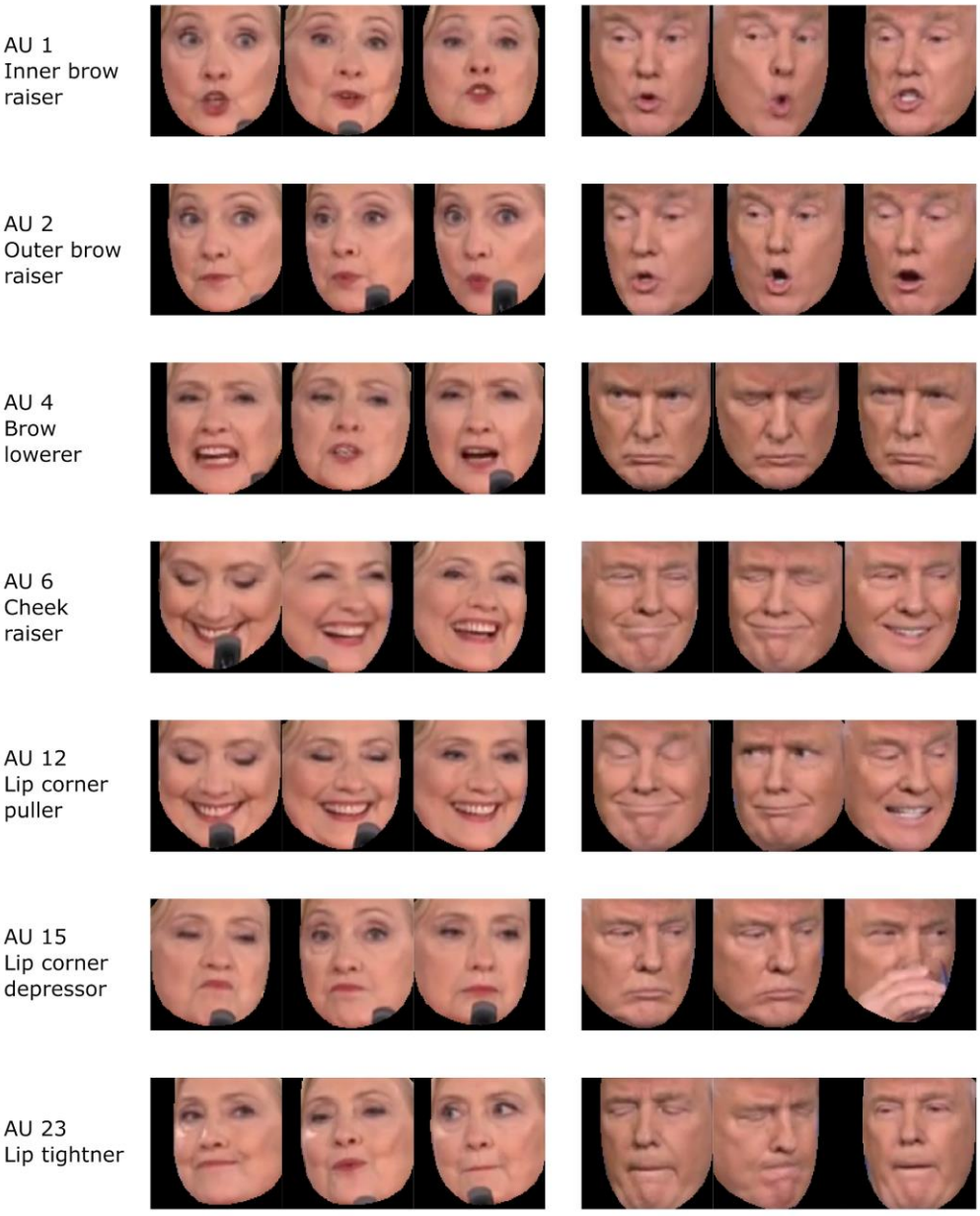

Figure 2: Candidate expressions, cropped from frames where OpenFace noted a high intensity of certain AU, with each labeled at the side.

3.2 OpenPose

Much useful information on nonverbal cues also comes from the gestures and posture that accompany facial expressions. OpenPose (Cao et al. 2017; Simon et al. 2017), another specialized tool based on convolutional neural networks, helped to acquire that information. OpenPose takes a video file as input and analyzes each frame, with output emerging in the form of XYZ coordinates of a set of body and hand keypoints, the full set of which are illustrated in Figure 3 below. All collected information is saved within a directory of JSON files, with each frame having one JSON file that contains keypoints for all detected bodies. These files end up proliferating in very large quantities, with the 90-minute debate videos used for analysis each running up about 130,000-150,000 JSON files that each represented analyzed frames. However, this file format let them later be easily converted to more efficient datasets in Python, this work's programming language of choice.

The camera work during the active argument exchange of the debate, which this study was most interested in, essentially rendered the candidates visible only from the torso and above. As such, any numbers relating to the lower body (about half of the 25 keypoints OpenPose divides the body into, as can be seen in the figure below) were not factored into later classification due to a complete absence of data. However, all available hand keypoints were used during classification to ensure total coverage of any gestures that appeared within frame. This data was concatenated with the previously acquired facial data and fed into the classifier defined in Section 4.

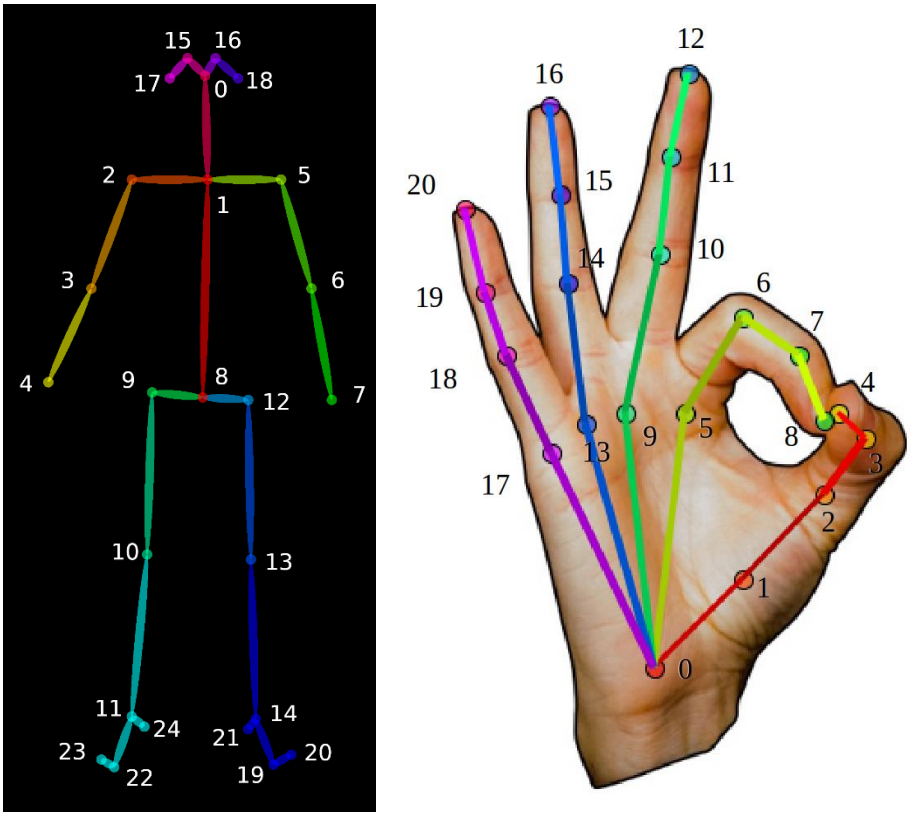

Figure 3: OpenPose's body and hand keypoints, taken from OpenPose documentation.

Figure 4: A screen capture of OpenPose extracting pose and hand data from debate footage.

4. Classification of Nonverbal Cues

4.1 Basis for Classification

Alongside the extraction of data from the debate footage, the video was divided into intervals, which were then classified with regards to various nonverbal behaviors. Deciding how to code the candidates' gestures and expressions required a perspective from the field of communication. Study of the influence of leader displays in these contexts has been active for decades (Bucy 2017), taking a "biobehavioral" approach that considers nonverbal cues just as important as participants' verbal speaking points, if not more so. The effects of these cues can be measured not only through physiological reactions, such as a viewer's heart rate or facial muscle movements (Bucy and Bradley 2004), but also through a computational lens, like analysis of the volume and sentiment of response on platforms like Twitter, a growing area of influence in the political landscape (see Shah et al. 2015, 2016 as examples).

There are four significant types of leader displays in the biobehavioral approach, denoted here with (emotion expressed/intention signaled): anger/threat (A/T), happiness/reassurance (H/R), fear/evasion (F/E), and sadness/appeasement (S/A) (see Masters et al. 1986; Stewart, Salter, and Mehu 2009). Each of these are shown through the face, specifically the eyebrows, eyelids, and mouth, as well as the head's orientation and motion. For example, S/A displays are usually indicated by raised lower eyelids and inner corners of the eyebrows, averted gaze, lowered corners of the mouth, and a general lowering of the head. The tools explained in Section 3, OpenFace and OpenPose, were able to account for all such differences in facial expression and posture in their data collection, making it easy to move forward to establishing a more detailed set of annotations than just the four basic categories.

With the essentials in mind, gesture coding was further broken down into the groupings of affinity and defiance. Affinity gestures, which encompass any movements of the hand, face, or body, signal an attempt by a candidate to act friendly/indicate a bond with other important parties in the debate. These parties could be the camera (i.e. the audience/viewers), their opposing candidate, or even the debate's moderator. Examples of this would be motions such as waving or winking. Defiance gestures, on the other hand, are arm/hand motions where a candidate is openly challenging or even dismissive. This sentiment could apply not only to their opposition, but also to any other authority within the debate's sphere. Movements such as shaking or wagging a finger (dismissive), prolonged staring at an opponent, and motions with a fist/pointed finger (aggressive or challenging) would fall under the "defiance" category.

Initial "ground truth" annotations for both the first and third presidential debates were created through manual examination of the debate footage by trained coders. Types of behaviors with documented influence in previous studies, as described above, were considered during the creation of the behavioral coding. Also taken into account were gestures or expressions noted as frequent/potentially impactful within the specific debates themselves. The names and meanings of each gesture annotation are explained in more detail, along with the results of classification, in Section 5.

4.2 Classifier Overview

4.2.1 Defining the Classifier

The final step in the process was to classify the data. For this study, face and pose data taken for every relevant frame of debate footage were used as input for an RNN classifier, specifically

with long short-term memory (LSTM) units (Hochreiter and Schmidhuber 1997), which then returned predicted index values for each of the nonverbal cues initially coded for. Standard RNN consists of internal state vectors, referred to as hidden states, that are updated upon processing of an input sequence. When the RNN incorporates LSTMs, another state variable called a cell state occurs in each LSTM. The cell state stores information required over the entire duration of processing of the input sequence, with an LSTM maintaining control over this cell state and updating it when necessary. To state it mathematically, RNN *f* will update hidden states at time step *t*, represented by h_t, based on input x_t at that time step, like so:

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

where the exact function pointed to by *f* is the function of the LSTM, and *T* refers to the number of elements in the input vector, e.g. how many frames are in an analyzed video. h_t is initially set to 0 in all cases, although its dimension can vary; in this work, the dimension was set to 64. Predicted values are generated at the last time step (with h_{T-1}) via a linear combination.

In this work, the input at any given timestep is a feature vector of statistics from OpenFace and OpenPose, corresponding to a frame of the analyzed video. The model begins with the first frame of the video and updates cell states and hidden states as it goes through timesteps, storing and passing in those values to the next LSTM for further computation. The values in these states are stored and passed to the computation for the next time step until every element in the input sequence has been processed. The final outputs (predicted values) are computed based on the last values in the hidden state, h_{T-1}:

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

where $\sigma(x)$ is a sigmoid function (output ranging from 0 to 1), w is a weight vector, and b is a bias term, which are learnable parameters of the model. The figure below visualizes this process.

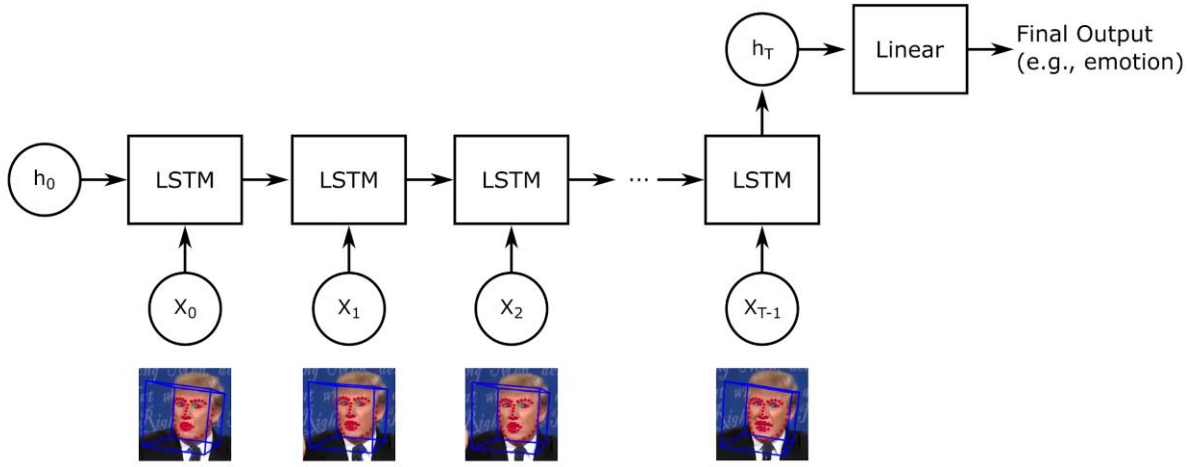

Figure 5: A simplified diagram of the workings of the classifier. Internal state value h_0 is updated through a chain of LSTMs at each time step up until step T, from input frames x_0 to x_{T-1}.

4.2.2 Classifier Training

The data gathered by OpenFace and OpenPose required some additional preprocessing before they could be used to teach the classifier. Both initially start with their own methods of denoting if measurements belong to different individuals. OpenFace assigns “face IDs” to everyone it detects in a frame, while OpenPose simply returns a set of pose keypoint coordinates for each person it captures. However, these methods were not ironclad and had caveats of slight unreliability from their developers. As such, the classifier was modified to double check the proximity of both face and pose coordinate sets to their predecessors, taking only the coordinates within appropriately minimal distances of each other as belonging to the same person: whichever candidate the data collection run happened to be focused on.

For the duration of the examined debates, there were a total of 533 time intervals for debate 1, and 547 time intervals for debate 3, with each interval lasting ten seconds. It is worth noting that while ten seconds is a somewhat coarse interval in computer vision, the manual coding of the debate (the "ground truth" markers) was established with that interval to balance human labeling with the limited time available, so the classifier had to match that for its predictions. Data at a smaller granularity was aggregated to meet this requirement, with the new data then randomly divided into train and test sets. 80% became training data, while the remaining 20%, the test set, was excluded from the training process and used later. After the classifier predicted values for the test set based on what it learned from the training data, those predictions could be compared to the test set's hand-labeled values to measure classification accuracy.

The classifier's specific training algorithm was Adaptive Moment Estimation, also known as Adam (Kingma and Ba 2014), an extended version of standard optimization with stochastic gradient descent, using 3 epochs (iterations through the entire dataset). Adam was implemented and carried out through PyTorch, an open-source machine learning library for Python, making it computationally lightweight and easy to configure.

5. Results

The classifier's accuracy was analyzed through 10-fold cross validation, i.e., performing the operations detailed in section 4.2.2 10 times with differently randomized train and test sets. This resulted in graphs of a ROC (receiving operator characteristic) curve whose area under curve (AUC) was taken. An AUC's range is from 0, or complete accuracy, to 1, or complete accuracy. A value of 0.5 would indicate that the classifier is about as informative as random chance,

meaning relatively inconsequential. Table 1 shows the mean (M) and standard deviation (SD) of the AUCs obtained for established features for each candidate during the first presidential debate, as well as an explanation of each annotation.

Table 1: Accuracy of automated classification of candidate nonverbal behavior in the first presidential debate. [1]

	Clinton		**Trump**		
Coding	***M***	***SD***	***M***	***SD***	**Behavior Definition**
look-at	.960	.018	.909	.025	The candidate in the reaction shot looks at the speaking candidate.
brush-off	.774	.037	.801	.053	The candidate in the reaction shot visually brushes off the opponent.
disagreement	.892	.045	.907	.017	The candidate in the reaction shot displays nonverbal disagreement.
look-into	.937	.011	.830	.040	The candidate looks directly into the camera, sometimes referred to as "breaking the fourth wall."
eyebrow	.911	.026	.754	.052	The candidate displays any noticeable eyebrow movements.
angry-face	.847	.051	.912	.035	The candidate shows an angry/threatening facial expression.
happy-face	.880	.014	.841	.041	The candidate shows a happy/reassuring facial expression.

[1] Blank entries mean a candidate was not seen to exhibit the given behavior during coding.

sad-face	.776	.034	—	—	The candidate shows a sad/appeasing facial expression.
neutral-face	.836	.025	.881	.043	The candidate shows a neutral facial expression.
affinity-gest	.821	.047	—	—	The candidate uses any affinity gestures.
defiance-gest	.745	.039	.833	.033	The candidate uses any defiance gestures.
agentic-gest	.991	.010	.939	.006	The candidate engages in an “agentic” style of behavior.
wave-off	.906	.039	—	—	The candidate “waves off” their opponent with a dismissive hand and arm swipe.
tic-lip	.736	.042	.771	.055	The candidate moistens their lip.
tic-bob	.855	.047	.713	097	The candidate bobs their head.
tic-grip	.863	.071	.821	.048	The candidate grips their podium.
tic-drink	—	—	.952	.040	The candidate drinks water.
tic-touch	—	—	.691	.049	The candidate touches their nose or mouth.
interrupt	.866	.185	.647	.059	The candidate in the reaction shot attempts to interrupt the speaking candidate.

This examination yielded clear overall statistics about the classifier’s accuracy, which averaged .825 across all statistics for Trump (with a range of .646 to .952) and .858 for Clinton (range of .774 to .991), and had the highest accuracy for both candidates when examining when

they looked at their opponents (variable look-at) as well as their overall level of activity (variable agentic). The classifier accuracy numbers also offered insights about both candidates' speaking styles. Trump's nonverbal cues and behavior stand out as significantly more demonstrative than Clinton. When factoring in what was observed during manual annotation, Trump did not show any cues that would convey a more placating impression, such as lowered mouth corners, head turned down towards body, or averted eye orientation, hence the lack of data for the sad-face and affinity-gest annotations above. In contrast, the classifier showed especially high accuracy when analyzing his *defiant* gestures or expressions, such as defiance-gest and disagreement. This demonstrates a high frequency of them that corresponds to his bombastic speaking style, which has been described as "transgressive" and "populist political performance" by other scholars in similar study. (Bucy et al. 2018)

While these results are informative, the initial dataset of the experiment was very small, consisting only of the two individuals in the debate footage. The model needed some more extensive analysis to evaluate its performance against potential alternatives, and gauge if it could generalize into application to other debates. A basis of comparison was acquired by applying a separate model trained on a larger dataset, the Expression in-the-Wild (ExpW) Dataset (Zhang et al., 2015), a public set of 91,793 faces manually labeled with each of seven fundamental expression categories: angry, disgust, fear, happy, sad, surprise, or neutral. This was done with a convolutional neural network (CNN) that takes image rather than video input; as such, the frames of the input video for the RNN were used as the CNN's classification input.

Table 2: Classification accuracy of a CNN classifier trained on the ExpW dataset, and the original video-based classifier trained on the data in this work (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

Using the CNN classifier also gave a chance to single out flaws or inconsistencies in the automated process. For example, despite Trump's wider range of emoting, he did not actually smile often, with the expression usually occurring in attempts to "laugh off" speaking points by Clinton. However, in his already low-accuracy classification in the CNN's "happy" category, there were also evident *false positives*: incorrect classifications of frames as "happy" when they were actually depicting other emotions. Because facial datasets like ExpW often have examples of "happy"-categorized expressions that show cues such like visible teeth and a widely opened mouth, those cues can be wrongly interpreted in different situations by classifiers that overlook context.

Figure 6: Examples of classified frames. Two are false positives, wrongly classified as “happy” (left), while two are true positives, correctly classified as “happy” (right). In the true positive frames, Trump is actually smiling while hearing Clinton speak.

6. Conclusion and Future Work

Many of the 20 different nonverbal cues established for study in this work warranted awareness of their context (a US presidential debate) and had potential for ambiguity or misinterpretation. Looking at the analysis of the classification pipeline in section 5, it is clear the system can effectively and correctly categorize those behaviors (even those which would usually warrant human labeling) with accuracy significantly better than chance, demonstrating promising results. The labels in this work encompass a wide range of emotions and actions, but will need revising and re-checking as more input data outside of the two examined debate videos are prepared for classification. The 10-second intervals used for observation can also be improved. While still better than the even coarser intervals that manual labeling sometimes resorts to, a wide variety of expressions and gestures can take place in ten seconds, and many nonverbal cues can and will occur over the space of a second or less, shifting or disappearing immediately after. As such, an important step in future work will be establishing more detailed annotations, ideally reaching multiple per second. This will require balancing reliable human labeling work for “ground truth” annotations with the granularity needed for a more accurate, precise system.

Although still not entirely foolproof, this automated technique presents a highly viable alternative to manual classification of data on nonverbal cues or behaviors. It not only can achieve a similar level of accuracy, but also promises much higher speed and precision than human work (often subjective and difficult to maintain reliably without much time and effort) could achieve. With the refinement of automated classification and how it analyzes emotions, there will also come the important shift towards working with much bigger datasets of image, video, or even audio data, allowing for analysis on unprecedentedly large scale. Such study could help to create systematic guidelines to interpreting different subjects' behaviors, as well as a much more detailed picture of what affects the large audiences their rhetoric is televised to.

References

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). OpenFace: A General Purpose Face Recognition Library with Mobile Applications. *CMU School of Computer Science*.

Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-Dataset Learning and Person-Specific Normalisation for Automatic Action Unit Detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on* (Vol. 6, pp. 1–6). IEEE.

Baltrušaitis, T., Robinson, P., & Morency, L. (2016). OpenFace: An Open Source Facial Behavior Analysis Toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–10).

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Fully Automatic Facial Action Recognition in Spontaneous Behavior. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)* (pp. 223–230).

Black, M. J., & Yacoob, Y. (1995, June). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of IEEE international conference on computer vision* (pp. 374-381). IEEE.

Bucy, E. P. (2017). 17. Media biopolitics: the emergence of a subfield. *Handbook of biology and politics*, 284.

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.

Bucy, E. P., Foley, J. M., Lukito, J., Doroshenko, L., Shah, D. V., Pevehouse, J., & Wells, C. (2018). Performing Populism: Trump's Transgressive Debate Style and the Dynamics of Twitter Response. *New Media & Society*.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., ... & Narayanan, S. (2004, October). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205-211). ACM.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields (pp. 7291–7299). Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Ekman, P., & Friesen, W. V. (1976). Measuring Facial Movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56–75.

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

Guo, G., Li, S. Z., & Chan, K. (2000). Face recognition by support vector machines. In *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. no. PR00580)* (pp. 196-201). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630-645). Springer, Cham.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

Joo, J., Bucy, E., & Seidel, C. Automated Coding of Televised Leader Displays: Detecting Nonverbal Political Behavior with Computer Vision. *International Journal of Communication. Forthcoming/to be published.*

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., ... & Mirza, M. (2013, December). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 543-550). ACM.

Kaiser, S., & Wehrle, T. (1992). Automated coding of facial behavior in human-computer interactions with FACS. *Journal of Nonverbal Behavior*, *16*(2), 67-84.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980 [Cs]*.

Littlewort, G. C., Bartlett, M. S., Salamanca, L. P., & Reilly, J. (2011, March). Automated measurement of children's facial expressions during problem solving tasks. In *Face and Gesture 2011* (pp. 30-35). IEEE.

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.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730-3738).
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.

Rosenberg, E. (2013, June 2). Is FACS Training right for you? Retrieved from http://erikarosenberg.com/facs/is-facs-training-right-for-you/

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.

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.

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).

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.

Vural, E., Çetin, M., Erçil, A., Littlewort, G., Bartlett, M., & Movellan, J. (2008). Automated drowsiness detection for improved driving safety.

Zhang, Z. (1999). Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron. *International Journal of Pattern Recognition and Artificial Intelligence*, *13*(06), 893-911.

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).