Images as Data: Computer Vision for Social Science Research

Andreu Casas and Nora Webb Williams

March 2017

Prepared for MPSA 2017, April 6-9

WORK IN PROGRESS: PLEASE DO NOT CITE WITHOUT PERMISSION

Abstract

Social scientists have long argued that images play a crucial role in politics. This role is heightened by the bombardment of images that people experience today. Digitization has both increased the presence of images in daily life and made it easier for scholars to access and collect large quantities of pictures and videos. However, using images as data for social science inference is an arduous task. Political scientists have therefore often turned to other data sources and puzzles, leaving substantive theoretical questions unanswered. Fortunately, recent innovations in computer vision can reduce the costs of using images as data. The goals of this project are twofold. First, we build on existing computer vision methods to present a set of automatic techniques that will aid political scientists working with images. We highlight the potential of Convolutional Neural Nets for automatic object detection and recognition; for face detection and recognition; and for visual sentiment analysis. Second, we apply these techniques to a novel dataset of Black Lives Matter Twitter protest images, demonstrating the ability of computer vision methods to replicate gold standard manual image labels.

1 Introduction

Images play a crucial role in politics. For example, newspaper images raise awareness about political topics and contribute to issue framing (Corrigall-Brown and Wilkes 2012; Brantner, Lobinger, and Wetzstein 2011), images of political candidates affect people's perceptions and votes (Rosenberg et al. 1986), and protest images shared via social media contribute to the diffusion of social movements and protests online (Kharroub and Bas 2015; Casas and Webb Williams 2017). Compared to other forms of communication, images reduce information-processing costs and have a larger impact on people's attitudes and behavior (Graber 1996; Kraidy 2002; Barry 2002; Grabe and Bucy 2009; Iyer and Oldmeadow 2006; Gazzaniga 1998). However, although people today are bombarded with more images than ever before in human history, systematic large-n studies of images in political life are rare (see for some preliminary studies Kharroub and Bas 2015; Anastasopoulos et al. 2016; Casas and Webb Williams 2017). This is in part due to the large costs associated with manually labeling even small datasets of images for political science inference. We present a set of computer vision methods to efficiently and automatically label thousands and even millions of images. We use a sample dataset of social media images related to a Black Lives Matter protest to illustrate how computer vision techniques can be used to study *what* is in an image, *who* is in an image, and *what emotions* an image evokes.

Our paper provides an introduction to the cutting edge of computer vision research, with an emphasis on deep machine learning.[1] In particular, we provide a high-level explanation convolutional neural networks (CNNs). Our focus is on the potential applications of CNNs in political science research. Accessing these tools does require some start-up effort, but as we demonstrate, the potential payoff is quite high. Following the conventions of computer science, we test the CNN image analysis method using computational "experiments." We provide an algorithm with a sample of images for which we already have human-generated, manual, gold standard labels. Prior to labeling, pre-processing steps include standardizing image size and removing duplicate images from the sample. The generated labels and corresponding images are then partitioned into train and tests for the supervised learning algorithms. The algorithm learns from the training set which image features are associated with a particular label (or classification). The trained algorithm then provides a

[1] Computer vision is the broad branch of computer science that deals with images. Automatic image processing is a growing portion of the field.

predicted label for the held-out test set images. Comparing the predicted labels for the test set to their known, gold standard labels allows us to evaluate the accuracy of the algorithm. Once an algorithm is satisfactorily validated by predicting held-out labeled images, researchers can use the trained algorithm to automatically label any remaining, unlabeled images.[2]

To demonstrate the potential for CNNs in political science research, we first highlight a few areas of study where labeling and understanding images is of particular importance. We then describe CNNs, with a focus on three particular applications: object recognition, face recognition, and visual sentiment analysis. Finally, we test the methods on observational data from a Black Lives Matter protest, conducting three separate experiments. Our goal is to see to what extent we are able to replicate our manual image labels using automatic, CNN techniques.

1.1 Images in Social Science Research

Why might social scientists wish to work with images as data? And why might they want to work with large quantities of images, such that manually labeling images becomes infeasible? Many possible goals spring to mind. Researchers might want to use images and the data they contain as outcome variables, whether their aim is descriptive, predictive, or causal. Or researchers might consider images as independent variables, as inputs that relate to some attitudinal or behavioral outcome of interest. We cover these two broad categories in turn below.

The study of images is not a new phenomenon in social science research. Scholars in subfields of communications, geography, sociology, history, economics and political science are well aware of the power of images in shaping and reflecting the human experience. What is new, however, is the digitization of images, the mass adoption of smart phones, and the omnipresence of the Internet, which combine forces to provide vast corpora of images for research. Accordingly, our focus here is less on clean uses of single images as data (e.g. showing an image as an experimental treatment or using a historical map to demonstrate migration patterns) and more on the messy masses of observational digital images. These masses, if properly treated, open up new avenues for broad, big-picture research.

[2]The train-test strategy is a common approach and will be familiar to those used to working with text as data and other machine learning subfields, see for example Benoit et al. 2016.

1.1.1 Images as Dependent Variables

At a very basic level, the presence or absence of an image might be a political outcome of interest. Which news stories are accompanied by a picture, and which are not? When do politicians include a image with their tweets, and when do they not? More complex, perhaps more interesting, questions might consider the content of images. When do news stories about protests include an image of the police? When do politicians include a picture of their family with campaign materials?

Scholars could also use images as a source of information in building a separate dependent variable. Images can help us estimate the size of an inauguration crowd, for example: pictures from a variety of angles and times of day could be automatically compiled and analyzed to produce an aggregate person count. And scholars of economic development increasingly use nighttime satellite images to proxy for development, where brighter footprint indicates a better-off town or village (newer analyses of luminosity data use CNNs, see Jean et al. 2016.)

1.1.2 Images as Independent Variables: Attitude and Behavior

As above, a basic treatment of images as an independent variable might wonder about the descriptive, predictive, or causal effect that the simple presence of an image could have on an outcome of interest. A more complex research question would ask which features of images have particularly strong effects. Following a traditional divide in political science, we might ask how the type and content of images are associated with variation in attitudes and behaviors. Prior work has considered how images of war may have swayed public opinion, or how images of endangered species contribute to viral activist campaigns. Altered images are a staple of conspiratorial discussions, used to convince new adherents of the truth a some claim. Images might also impact concrete behaviors, such as willingness to protest, vote, or donate money to a cause.

Small-n studies have touched upon many of the research questions above. But studying images systematically to learn about their effects and determinants requires labeling a large quantity of pictures. Manually labeling images is costly, making the potential to automatically label images appealing. Recent computer science developments in deep learning now make this feasible (LeCun, Bengio, and G. Hinton 2015). In the following section, we introduce the field of computer vision and Convolutional Neural Nets.

2 Computer Vision for Social Scientists

The field of computer vision has grown by leaps and bounds in the last decade. Some common goals in computer vision research include object detection/recognition, face detection/recognition, and visual sentiment analysis. While there are certainly other branches of computer vision that may be of interest to social scientists, we have found these three tasks to be particularly relevant to political science research. Recent advances in deep learning have contributed to radically improved accuracy in each of these three subfields. Today most computer vision research is based on convolutional neural networks (CNNs - ConvNets), a type of neural network perfectly designed to use images as inputs. In the rest of this section we go over the basics of deep learning and CNNs, and highlight the most relevant research on object recognition, face recognition, and visual sentiment analysis.[3]

2.1 Deep Learning and Convolutional Neutral Networks: The Basics

Deep learning algorithms use non-linear models to transform raw input (e.g. any data matrix, known as the *input layer*) into abstract representations (*hidden layers*) in order to learn from new features and better predict outcomes. Imagine a 100x2 matrix X with information about the height and gender (columns) of 100 people (rows). Imagine another 100x1 matrix Y with information about how fast these 100 people run a half-marathon. In conventional machine learning we could try to predict their finish time given their height and gender using for example a simple linear model ($Y = X\beta$) and finding a 2x1 coefficient matrix β that minimizes predictive squared error (Ordinary Least Squares regression). However, if our interest is prediction, and not to study the specific associations (βs) between height and gender on the outcome, we could also use a neutral network to improve predictive accuracy.

First we would transform the raw data X into an intermediate abstract representation X_2 by applying the dot product between X and a 2x100 (e.g.) parameter matrix W_1, and then we would apply the dot product between the resulting X_2 and a 100x1 matrix W_2 to create a vector of predictions $\hat{Y}$ of the same size (100x1) as the outcome Y. We would then find W_1 and W_2 that minimize predictive error (*loss*), and take advantage of the new 100x100 abstract intermediate feature matrix X_2 to learn more about the input and perform more accurate predictions. We would

[3]Resources to learn more about CNNs are increasingly available online. The authors particularly recommend the posted materials accompanying Stanford University's CS231n course at http://cs231n.stanford.edu/.

also add a non-linear transformation of X_2 in order to improve fit, for a final model that could look as follows: $Y = max(0, XW_1)W_2$, where $X_2 = max(0, XW_1)$ and $max(0, X_2)$ is a non-linearity often used in deep learning. The parameter matrices W_1 and W_2 (known as *weights*) are usually learned via Stochastic Gradient Descent (SGD) and chain rule is used to derive the gradient. This means that we run the model multiple times (*iterations*), updating the weight parameters (*forward propagation*) and calculating the gradient (*backward propagation*) each time, until the model loss reaches a point of convergence.

A neural net has as many layers as the number of intermediate representations plus the *output* layer. The previous example is a two-layer network because it only has one intermediate representation (X_2) plus the model predictions ($\hat{Y}$). The previous is also an example of a network with *fully-connected* layers because we apply the dot product between all units of a given layer (e.g. each data row in the *input* layer: $\{x_1, x_2, ..., x_{100}\}$) and each unit of the following layer (e.g. each parameter in the parameter matrix W_1: $\{w_1, w_2, ..., \}$). [4]

A convolutional neural network (CNN) is simply a type of neutral net with two main particularities. First, the inputs have 3 dimensions. In the previous example, the input had only two: the number of rows (100) and columns (2) of the dataset. However, computer vision researchers parse images into pixels and then represent each pixel as a triplet of red, green, and blue (RGB) intensities (depth). This means that each image x has three dimensions: pixel width, pixel height, and three color channels (e.g. 224x224x3), where $x_{1,1,1}$ contains for example information about the red intensity of the pixel in the top left corner of the image and $x_{1,1,2}$ contains information about the green intensity of the same pixel. Each pixel intensity representation $x_{i,j,z}$ is usually a standardized integer ranging from 0 to 255. Figure 1 provides an illustration of how a one-dimensional image is translated into a three-dimensional RGB input.

[4] See LeCun et al. (2015) and Schmidhuber (2015) for a more extensive overview of deep learning.

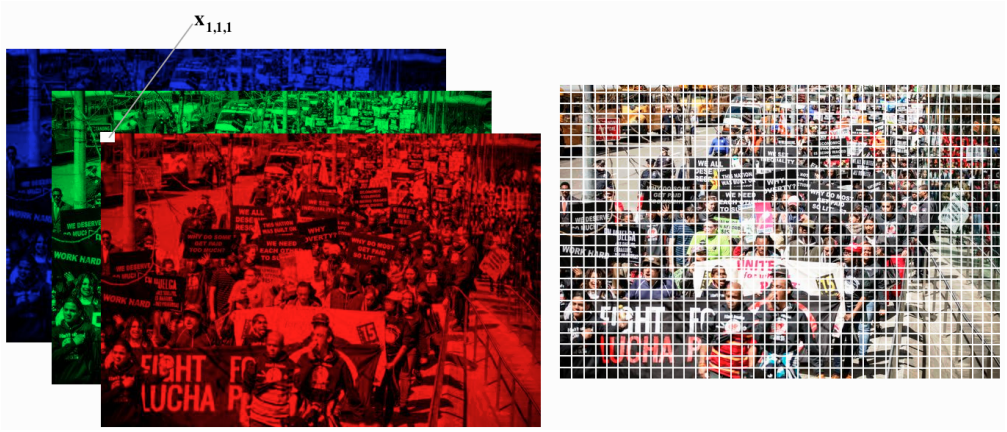


Figure 1: An example of an image represented as a 3-dimension input

A second characteristic of a CNN is that most layers of the network are not fully connected to the previous layer: these are known as *convolutional* layers. Instead, the weights in these layers (known as *filters*) are of a smaller width and height than the input (e.g. 3x3x3 instead of 6x6x3) and they are only fully connected to a particular local region of the input at a time. This reduces the number of parameters to estimate and significantly decreases computation time. Dot products are computed between the filter parameters and all possible regions of the input by "sliding" the filter along its width and height, summing up the outcomes to create an output volume that becomes a new input in the following layer. Figure 2 for example shows a 3x3 filter sliding horizontally and vertically across what could be the Green pixel-intensity representation of a 6x6x3 image.

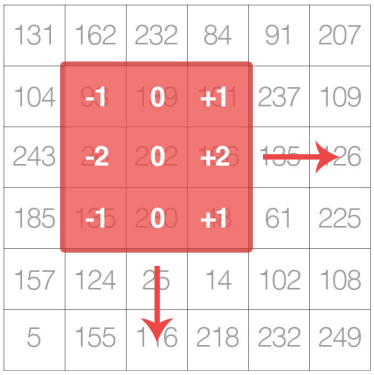


Figure 2: An example of a filter sliding in a convolutional. Credit: PyImagesSearch.

Apart from convolutional layers and some fully-connected layers at the end of the network, there are two other type of layers often used in CNNs: *pooling* and *RELU* layers. Pooling layers are used

to reduce the size of the weights (parameter matrices) in order to facilitate estimation and speed up computation. ReLu layers apply a non-linear transformation ($f(x) = max(0, x)$) to the output of convolutional layers in order to improve fit.

In sum, this complex infrastructure of weights, non-linear transformations, and intermediate representations of the raw input allow CNNs to learn about very specific pixel-level features that help predict the outcome (such as an object class, a person, and an emotion class). CNNs with high predictive accuracy are usually trained with hundreds of thousands (and even millions) of images and have a substantive number of layers. A common training set for a 1,000 class object classification task, ImageNet, has 1.3 million high-resolution images, and a CNN commonly used for object classification, VGG16, has 16 layers. The size of the training set and the depth of this CNNs mean that powerful machines with multiple graphics processing units (GPUs) and high computational power are needed for training. However, training CNNs from scratch is often unnecessary. Computer vision scholars often share their highly accurate trained CNNs (such as LeNet, AlexNet, VGG16, and VGG19). Then others can and adapt them to perform different classification tasks, such as predicting a binary instead of a 1,000 class outcome, by simply changing the last fully connected layer of the network and some minor re-training parameters. This is known as *fine tuning* and allow others to also achieve high predictive accuracy by using much smaller training sets and less training time and resources.

2.2 Three Applications of CNNs for Images as Data

In this paper, we focus on three classes of algorithms that have the potential to automatically assign the type of image labels that are of interest to social scientists.

2.2.1 Object Recognition

Early computer vision object recognition methods had difficulty distinguishing between cats and dogs (e.g. Golle 2008). With the help of CNNs, they can now accurately label a wide variety of objects and even identify multiple objects within a single image (Krizhevsky, Sutskever, and G. E. Hinton 2012; Ordonez et al. 2013; H. Li et al. 2015; Russakovsky, L.-J. Li, and Fei-Fei 2015; Chen et al. 2015), see Figure 3 for an example. For social scientists, object recognition could be used to determine the presence or absence of features with theoretical importance. For example, images

could be analyzed to detect whether flags are present at a protest, perhaps as a supplement to experimental studies about support for immigration policies (Wright and Citirn 2011). The gold standard labeling process would ask annotators to label for objects that are relevant to the political mechanisms of interest (such as crowds of people carrying protest signs or flags). Researchers could also take advantages of existing corpora of labeled images, such as ImageNet, though specific political objects of interest will likely not be labeled.

Figure 3: Object Recognition using a CNN trained with ImageNet data (from Krizhevsky, Sutskever, and G. E. Hinton 2012)

2.2.2 Face Detection and Recognition

CNNs and other methods such as the Viola-Jones algorithm and Haar Cascades can be used to accurately detect and count faces (Zhu and Ramanan 2012; H. Li et al. 2015; Anastasopoulos et al. 2016). Figure 4 provides an example. These methods can be used to count the number of people present in an image.

Figure 4: Li et al's 2015 Face Detection algorithm (from H. Li et al. 2015)

Once faces have been detected in an image, researchers can use CNNs for face recognition, or noting the presence of specific persons in an image. As we demonstrate below, CNNs can be

trained to recognize specific individuals with a high degree of accuracy using the train-test method with manual labels. In our case, as a preliminary step, we are attempting to recognize individuals without face detection (essentially conducting an object recognition test but with a specific person as the object). Researchers interested in face recognition could also access existing sets of important faces, such as Guo et al.'s (Guo et al. 2016) one million images of 100,000 celebrities, to see if famous faces appear in their images. Manual face recognition is already done for Kremlinology-type analyses of images from North Korea (Fisher and Patel 2017); CNNs offer the possibility of scaling that recognition up to enormous corpora of images.

2.2.3 Visual Sentiment Analysis

Recent CNNs trained with pixel-level and image-level features (objects automatically detected in the images) now do a moderately good job of predicting the emotions humans report when observing images (60-70% accuracy) (Peng et al. 2015; You et al. 2015). This ability to connect image attributes to emotions represents a significant advance over prior sentiment methods that focused exclusively on facial expressions (such as Microsoft's Emotion API). The ability to automatically label images with evoked emotions has a great deal of potential for scholars of framing and media effects. For training, scholars might turn to existing labeled datasets such as Cornell's Emotion6. However, existing datasets and analysis are often based on very clean images, see Figure 5 for an example. When applying these methods to messy, real data, such as the Twitter images we collected from a Black Lives Matter protest, visual sentiment analysis becomes both more interesting and more difficult.

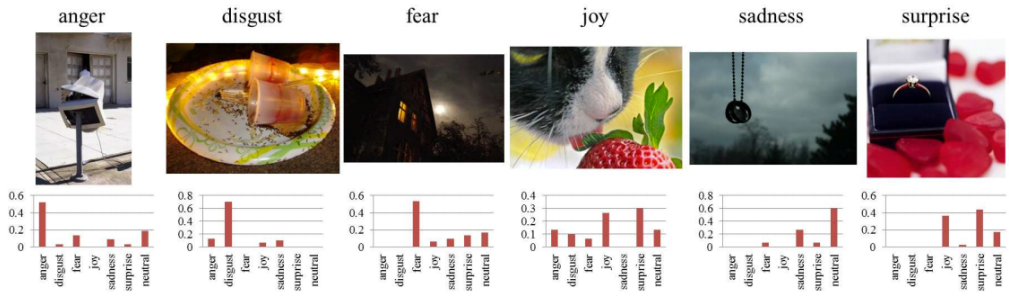


Figure 5: Visual Sentiment Analysis using CNN (from Peng et al. 2015)

3 A Political Science Application: Black Lives Matter Images

The following example, based on our own work, demonstrates the importance of studying images and their characteristics as independent variables in order to understand a political phenomena of interest. We discuss the challenges of manually labeling images and the potential of computer vision methods to replicate manually labeled results.[5]

3.1 Original study

Our aim in the project was to test the power of images in mobilizing online social movement participation. While a number of prior scholars have argued that images are instrumental for protest and social movement participation in the digital media age (Bimber, Flanagin, and Stohl 2005; Castells 2012; Bennett and Segerberg 2013; Kharroub and Bas 2015; Howard and Hussain 2013), systematic studies of political images, or images in general, in social media are rare. How might images shared via social media matter for online social movement mobilization?

To avoid selection on the dependent variable (e.g., retroactively choosing a protest or movement with particularly moving images), we decided to study a protest before knowing whether it would take off or whether it would have any images involved. The protest of interest was held on April 14, 2015 and was supported by the Black Lives Matter movement (BLM). Research in political psychology and political behavior led us to expect that images could motivate protest mobilization in three main ways, as summarized briefly below.

Emotional Trigger Mechanism: Emotions have been found to be important predictors of political behavior (Lasswell 1930; Janis and Mann 1977; Sears 1987), including support for parties, protests, and extremist groups (Melucci 1996; Victoroff 2005; Goodwin and Jasper 2006; Valentino et al. 2011). There are five politically-relevant emotions that political psychology research has been able to distinguish both theoretically and empirically: anger, fear, sadness, enthusiasm, and disgust (Ekman 1992; Marcus, Russell Neuman, and MacKuen 2000; Valentino et al. 2011; Clifford and Wendell 2016). We built on this existing literature to hypothesize that the emotionally evocative nature of images would contribute to their mobilizing effects, especially those images evoking enthusiasm, anger, and sadness.

[5]The complete paper can be accessed at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2832805.

Expectation of Success Mechanism. The prisoners' dilemma poses a puzzle for social movements: free riding should cause such movements to fail. Classic rational-choice theory predicts that individuals will invest time and resources only if they expect their efforts to significantly impact group success (Olson 1965). We therefore expected to find that images of protest crowds, involving large numbers of people, would act as a trigger for expectations of success.

Social Identity Mechanism. People are social beings and groups can provide a sense of belonging. The desire for a "positive self concept" leads people to seek out and support the in-groups with which they identify (Tajfel 1981; Tajfel 1982). Images of the Egyptian flag and the Muslim Crescent helped to increase support for dissident groups during the Arab Spring (Kharroub and Bas 2015, 7). We expected to find that group symbols, such as religious icons, raised fists, flags, and logos would increase participation among people who identify with those groups.

There is a fourth theoretical mechanism that we did not explore in the original study, the **Cue-Taking Mechanism**. Rational choice theory also predicts that rational citizens will rely on the judgments of trusted experts or opinion leaders in lieu of becoming politically informed (Downs 1957). Many subsequent studies offer evidence of the importance of heuristics and cue-taking in public opinion formation (Katz and Lazarsfeld 1955; Zaller 1992; Nisbet and Kotcher 2009), voting and political participation (Lupia 1992; Shachar and Nalebuff 1999) and protest movements (Obserschall 1973; McCarthy and Zald 1977). However, additional research finds that cue-taking may depend on prior beliefs and how different individuals process information (Kuklinski and Hurley 1994). We therefore also expect that images of admired opinion leaders could contribute to the diffusion of online mobilizations.

The goal of this paper is not to discuss the validity of these theoretical mechanisms nor the findings of the original study, but to see to what extent we are able to replicate our manual image labels using automatic techniques. We focus on using computer vision methods to automatically label images for the presence of protests (*Expectation of Success* mechanism), for the presence of a particular opinion leader, John Legend (*Cue-Taking* mechanism), and for the emotions images evoke (*Emotional Trigger* mechanism).

3.2 Data Collection and Manual Annotation

For a two week period around an April 14, 2015 Black Lives Matter (BLM) protest ("Shutdown A14"), we collected all Twitter messages related to the protest (about 150,000 tweets).[6] We then extracted and manually labeled all of the 9,500 unique images that were included in the tweets. Each of the 1000 most-tweeted images were labeled by two undergraduate research assistants and between three and eight additional crowd-sourced assistants (recruited from Amazon's Mechanical Turk). The remaining images were labeled by just one Mechanical Turk annotator. Each was labeled for the number of people guessed to be in the image, the presence of symbols and logos, and for the the degree (scale of 0-10) of fear, anger, disgust, sadness and enthusiasm it evoked in the annotator. Information about the number of people present in the image and the emotional responses for images that were labeled by more than one person were averaged across annotators to generate a single score per image. For this paper we removed image files that were of a very small size (smaller than 5kb) from the analysis since they may add confusion when training computer vision algorithms. The final image dataset contains 8,148 unique images with labels for the number of people present in the picture (numeric), whether symbols and logos are present (binary), and five emotion scores (ranging from 0 to 10 for each emotion). We use these images to perform 3 computational experiments testing the extent to which CNNs can be used to predict: (1) images of protests, (2) images of a particular opinion leader (the singer John Legend), and (3) the emotions that images evoke.

3.3 Data for Computational Experiment (1): Predicting Images of Protests

In the original study we argued that images of people protesting and actively supporting the movement on the street fostered online mobilization because they increased expectations of success about the movement. We used the manual labels provided by the annotators to detect which images showed on-street protests. We specifically considered an image to be about a protest if the annotators said the picture had numerous people plus a symbol or a sign. In this first computational experiment we

[6] We collected messages mentioning at least one of the following keywords and hashtags: #shutdownA14, murder by police, mass incarceration, shutdownA14, killer cops, police murder, #A14, stop business as usual, stolenlives, massincarceration, stolen lives, #policebrutality #stolenlives, #blacklivesmatter, black lives. We designed the list of keywords and hashtags by looking at the phrases and hashtags promoted online by the organizations responsible for the protest.

are interested in automatically detecting protest images to replicate our manual labels.

We train the algorithm with true positives and true negatives: true positive images of protests/demonstrations and true negative images that are not showing any protest. We select 100 true positives from the dataset of 8,148 images by first selecting those images annotators reported to have more than 10 people and also a sign and/or a symbol, and then by selecting from that pool 100 images that clearly show a protest/demonstration. We then select 100 true negatives from the remaining images that have less than 10 people in them and have no sign nor symbol. The top row in Figure 6 contains four examples of the true positives we use to train the protest classifier, and the bottom row contains four examples of the true negatives.

Figure 6: Examples of Images used as true positives and true negatives to train a computer vision binary classifier predicting images of protest/demonstrations

3.4 Data for Computational Experiment (2): Predicting Images of an Opinion Leader (John Legend)

When exploring the collected messages we noticed that a substantial amount of them had images of known people who explicitly supported the BLM movement prior to the protest (such as John Legend) and of known people who were actively involved in the organization of the April 14, 2015 mobilization (such as Dr. Cornel West). We want to know if we can use computer vision techniques to automatically detect the presence of an opinion leader. We focus here on predicting images of John Legend. The American singer and songwriter had given an speech against U.S. incarceration policies when accepting the Oscar for best original song at the 87th Academy Awards, about a month before the studied protest, becoming one of the front faces of the “Stop Mass Incarceration" movement during the April 2015 mobilization.

We want to train a computer vision algorithm to predict whether a given image is a picture of John Legend or not, so we construct a dataset of true positive and true negatives images to train the algorithm. The *Cue-Taking* mechanism was not part of the original study and we did not have information about whether a well-known person was present in our BLM images, but we went through the 8,148 images selected for this study one more time and searched for images of John Legend. We found 29 unique images. These were either headshots or images of him singing or standing alone. To increase the size of the training set, we also collected another 71 similar images of him from a Google image search, ending with a set of 100 true positives. We then selected similar images (headshots and pictures of a single person standing) from our BLM dataset (n = 50) and from the Internet (n = 50) to use as true negatives.

Figure 7 contains some examples of the true positives and negatives selected to train the binary computer vision classifier. The first row has pictures of John Legend from the BLM dataset, the second row has true negatives selected also from the BLM dataset, and the third row has true negatives selected from the Internet (pictures of Don Cheadle, Rihanna, Leonardo DiCaprio, and Danny Glover).

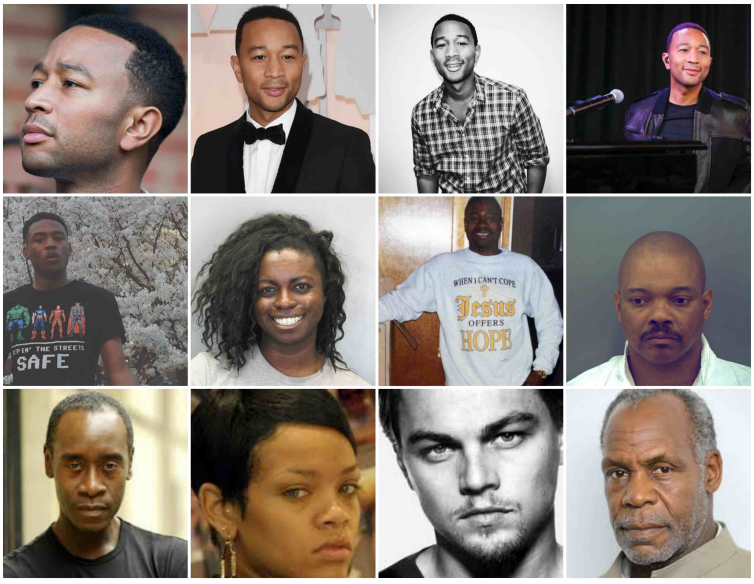
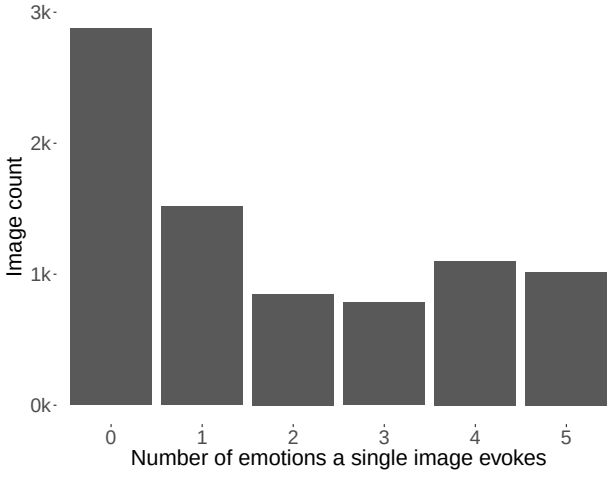


Figure 7: Examples of Images used as true positives and true negatives to train a computer vision binary classifier predicting pictures of John Legend

3.5 Data for Computational Experiment (3): Predicting the Emotions Images Evoke

We face two main challenges when selecting images to train computer vision algorithms predicting emotional reactions. First, annotators reported that numerous images in our BLM dataset did not trigger any emotional reaction (that is, they received a score of 0 on every emotion). This means we can only use these images as true negatives when training emotion classifiers. Second, some images triggered more than one single emotion. These images may add confusion during training and make it more difficult for computer vision algorithms to isolate what specific image features are linked to each emotion. Figure 9a shows, for example, that annotators reported that about 35% of the BLM images (2,876 out of 8,148) did not evoke any emotion, while about 50% (3,746) of images evoked more than one emotion. They reported 1,516 images that evoked only a single emotion. Figure 9b shows how often annotators reported the an image evoked one of the given pairs of emotions. We can see for example that images triggering anger often trigger sadness and disgust as well.

(a) Number of emotions triggered by the images



(b) How often two emotions are evoked by the same image

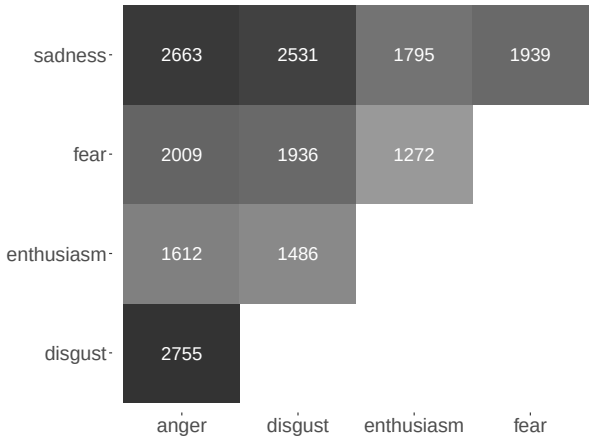
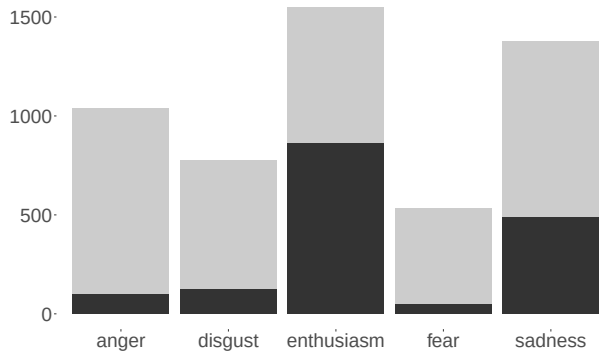


Figure 8: Images in our dataset often trigger more than one emotion

We follow previous work on visual sentiment analysis in computer science (e.g. Peng et al. 2015) and choose images that "clearly" trigger each emotion as true positive cases to train emotion classifiers. We select pictures for which the 0-10 score for the strongest emotion evoked in a picture is higher than 3 (the max class must have a value higher than 3). We then further limited the

sample to images where the difference between the highest scoring emotion and the second highest scoring emotion score is equal to or greater than 3 (these are images with a clear max class, with less overlap between emotions). We then choose “neutral" images from the set of pictures that annotators reported as not evoking any emotion as true negatives.

The advantage of this approach is that, compared to training with images that evoke multiple emotions or to use images triggering other emotions as true negatives, it should be easier for a computer vision algorithm to learn about image features that are predictive of each emotion. The drawback is that we significantly reduce the size of the training set. Only 1,630 images fulfill the criteria and they are not equally distributed across the five emotion categories. We end up with 103 true positives for anger, 127 for disgust, 862 for enthusiasm, 51 for fear, and 487 for sadness (see Figure 10).



Emotion	Max Class	Clear Max
anger	935	103
disgust	648	127
enthusiasm	686	862
fear	484	51
sadness	889	487

Figure 10: Number of images that have each emotion as a max class (gray) *versus* “clear" max class (3 point difference with the 2nd emotion score –in black–)

4 Method

We fine tune a pre-trained CNN seven times to create seven binary classifiers predicting: images of protests (Experiment 1), images of John Legend (Experiment 2), and images triggering anger, disgust, enthusiasm, fear, and sadness (Experiment 3). The pre-trained CNN was trained on 1.3 million high-resolution images from the LSVRC-2010 ImageNet dataset and was designed to predict 1,000 classes (Krizhevsky, Sutskever, and G. E. Hinton 2012). This trained CNN is commonly known in the computer vision literature as AlexNet. The network is composed of five convolutional layers,

max-pooling layers, dropout layers, and three fully connected layers at the end (see Figure 11). For each of our seven classifiers, we remove the last fully-connected layer to predict 2 instead of 1,000 classes. We allow very small updates of the weights in earlier layers and focus most of the learning in the last added layer. We run 2,000 iterations (back and forward propagation) during the training of each model. In all cases, training loss does not improve much after 2,000 iterations (see Appendix I, Figure 16).

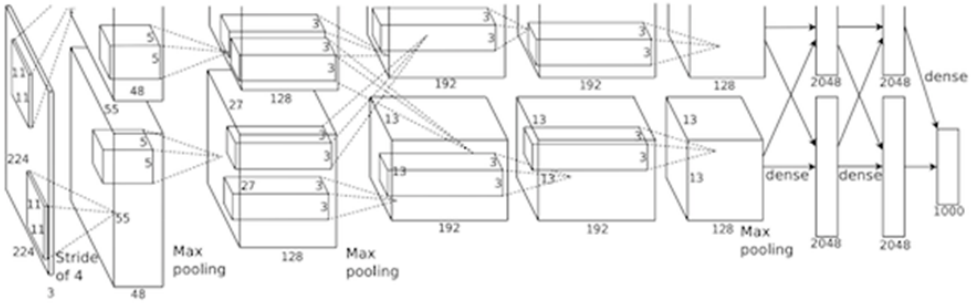


Figure 11: AlexNet architecture

Table 1 summarizes the number of training and testing observations for each CNN we train. Notice that in most cases we use only about a hundred true positives and a hundred true negatives. As mentioned, this would be insufficient to train a CNN from scratch, especially in only 2,000 iterations, but this is when the potential for fine tuning an existing nets comes in very handy. Pre-trained deep neural nets have already learned some good generic features such as edge detectors and color blob detectors. In all cases we use 80% of the true positives and true negatives for training and 20% for testing. In the following section of the paper we asses the extent to which we are able to predict the correct labels (positive/negative) in the testing sets.

CNN-classifier	Experiment	N	True Pos (train|test)	True Neg (train|test)
Protest	Exp. 1	200	100 (85|15)	100 (85|15)
Opinion Leader	Exp. 2	200	100 (80|20)	100 (80|20)
Anger	Exp. 3	206	103 (82|21)	103 (82|21)
Disgust	Exp. 3	254	127 (102|25)	127 (102|25)
Enthusiasm	Exp. 3	1,720	860 (690|172)	858 (687|171)
Fear	Exp. 3	102	51 (41|10)	51 (41|10)
Sadness	Exp. 3	972	487 (390|97)	485 (388|97)

Table 1: Description of the Computational Experiments in the paper

5 Results

After training the seven CNNs using the true positive and true negative observations in the training sets, we use the models to predict the probability that images in the held-out test set are images of protests/demonstrations (Exp. 1), John Legend (Exp. 2), and images that evoke anger, disgust, enthusiasm, fear, and sadness (Exp. 3). The last layer of all the CNNs is a SoftMax classifier that returns the probability that a given image belongs to each of the two classes (positive/negative), with the sum of the probability adding up to 1. We transform the probability for the positive class into a dichotomous variable. Usually we use a 0.5 probability cut-off to build this dummy variable: if for example the "protest model" predicts that there is a 0.51 probability that a given image contains a protest/demonstration, then we would classify that image as being about a protest. However, using a higher probability threshold may sometimes increase model accuracy.

In the rest of this section we report the results of the seven models based on the predicted labels for the testing sets. We judge the results using three common validation measures in machine learning: recall, precision, and accuracy. *Recall* is the percentage of true positives in the test set that the model predicted as positives. *Precision* is the percentage of cases that the model predicted as positives that are actually true positives, and *Accuracy* is the percentage of both predicted positives and negatives that are actual true positives and negatives.

5.1 Computational Experiment (1): Predicting Images of Protests

In this computational experiment we trained the algorithm using 85% of the true positive and negative cases (n = 170) and we then evaluate the model's predictive accuracy by classifying the 15% of the observations that we reserved for testing (n = 30). If we use a > 0.5 probability threshold to transform the SoftMax probability outcome into a protest dichotomous variable, we observe from the confusion matrix in Table 2 that the model overestimates the presence of protests in images. Most of the true positive cases are being predicted as positive cases of protest images (93% recall), but a substantive number of true negatives (images with no protest) are being incorrectly predicted as positives, resulting in a low model precision (58%). However, by changing the probability threshold to > 0.6, we can retain the same recall and improve the precision to 74% (see Table 3). In a binary setting like this one, slightly increasing the probability threshold makes sense because it means that we want to ignore close positive predictions. In other words, we want to discard positive predictions based on 49/51% calls. However, researchers should keep in mind that a larger probability threshold increase may improve predictive accuracy for a particular test-set but not others, leading to overfitting. Figure 12 shows how changes in accuracy, precisions, and recall shift depending on which probability threshold is chosen.

Gold	pFALSE	pTRUE	G.Total	Recall
FALSE	5	10	15	
TRUE	1	14	15	(93%)
pTotal	6	24	30	
Precision		(58%)		

Table 2: Confusion matrix of predicted versus actual protest images, with > 0.5 threshold

Gold	pFALSE	pTRUE	G.Total	Recall
FALSE	10	5	15	
TRUE	1	14	15	(93%)
pTotal	11	19	30	
Precision		(74%)		

Table 3: Confusion matrix of predicted versus actual protest images, with > 0.6 threshold

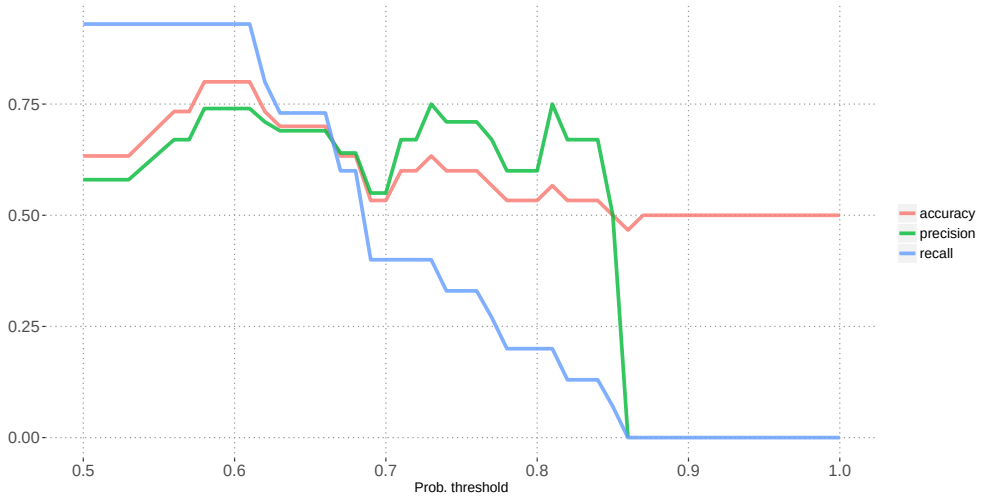


Figure 12: Precision, Recall and Accuracy of a binary classifier predicting protest images as we move the probability threshold use to create a dichotomous variable

5.2 Computational Experiment (2): Predicting Images of John Legend

The CNN predicting images of John Legend also produces very accurate results. Out of the 40 images that we do not use for training and leave out for testing (20 true positives and 20 true negatives), the model correctly predicts 36 (18 true negatives and 18 true positives) for a precision, recall, and overall accuracy of 90% (see Table4).

Gold	pFALSE	pTRUE	G.Total	Recall
FALSE	18	2	20	
TRUE	2	18	20	(90%)
pTotal	20	20	40	
Precision		(90%)		

Table 4: Confusion matrix comparing predicted versus actual images of John Legend

Table 4 shows the images that were incorrectly predicted. The two pictures on the left are false negatives: images that the model should have predicted as having John Legend in them but did not. The two pictures on the right are false positives: images that should have not been predicted as having John Legend. The two false negatives on the left were probably caused by other extra features present in the picture, such as John Legend's baby, sunglasses, and hat, and the letters in the second picture. We are less certain about what causes the two false positives. We suspect this may be related to the people in the image having a very similar position to the position John Legend has in the training images. In future iterations of the paper we will add an extra pre-processing step before training the CNN: we will first use face detection methods to extract and only compare faces present in images instead of the whole image (using for example a Viola-Jones algorithm). This is a common step when training CNNs for face detection. We thought this step was not necessary in this case because the John Legend images present in our BLM dataset were very neat. However, introducing a face detection step may help get the accuracy closer to 100% and will likely be necessary to detect notable individuals who are pictured in a crowd (such as Dr. Cornel West).

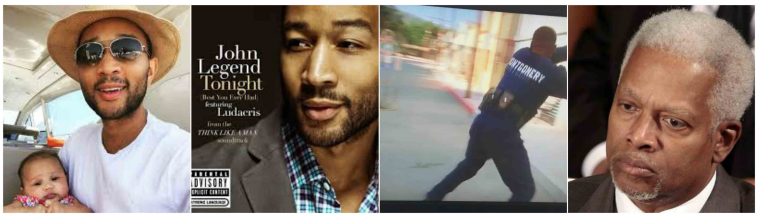


Figure 13: False negatives (left) and false positives (right) predicted by the CNN trained to predict images of John Legend

5.3 Computational Experiment (3): Predicting Emotions

Predicting the emotions images evoke is the most challenging task we undertake in this paper. Predictive accuracy is not as high as in the first and second computational experiments. Table 5 presents the results. Recall for all of the emotions except for fear is greater than 60%, with the disgust recall reaching 80%. Precision is notably lower for all of the emotions. This means that the binary classifiers have a tendency to overestimate the positive probabilities for each emotion. The classifiers predict as positives images that are actually negative (neutral) cases. As a result, predictive accuracy is around 60% for disgust, enthusiasm and sadness, and 50% (a random coin flip) for anger. The fear precision is 33% and recall is only 20%, which means that the model is doing worse than a coin-flip because it is learning some misleading features from the training set.

Emotion	Test Set (N)	Pos (n)	Neg (n)	pPos (n)	pNeg (n)	Rec.	Prec.	Acc.
anger	42	21	21	32	10	0.76	0.50	0.50
disgust	50	25	25	32	18	0.80	0.62	0.66
enthusiasm	343	172	171	226	117	0.76	0.58	0.60
fear	20	10	10	6	14	0.20	0.33	0.40
sadness	194	97	97	103	91	0.61	0.57	0.58

Table 5: Results from emotions CNN

There are a couple points that may account for the somewhat weak visual sentiment analysis results, especially for fear. First, note that the sample sizes are quite low, meaning the algorithms do not have much to learn from. Second, it is important to notice how different the images in our dataset are from images used for visual sentiment analysis in computer science. Figure 14 shows images labeled as sad images in the *Emotion6* dataset while Figure 15 shows images from our BLM data that annotators labeled as evoking sadness. Computer vision scholars have been able to build relatively accurate (60-70% accuracy) binary emotion classifiers (Peng et al. 2015, Wang et al. 2013), but the images they have used so far have a very simple composition. They usually do not have more than one or two elements in the picture. On the contrary, politically relevant images that people share online are a lot more complex. For example, in our BLM dataset there are images of cartoons, images with several elements, and images that combine visuals with text. This means that larger

training sets of politically-relevant images and more complex models need to be build in order to improve accuracy in predicting emotions that political images evoke: such as adding text, detected objects, and detected facial expressions as features.

Figure 14: Sad images from the *Emotion6* dataset

Figure 15: Sad images from our *BLM* dataset

6 Conclusion

The main takeaway from our paper is that computer vision methods using CNNs can produce very accurate results, even when applied to messy, real world images, and even with a relatively small number of manually labeled images to work from. While there are adjustments that we could undertake in order to improve the results even further, such as including a face detection step in the second computational experiment, and include some other image-level features to improve the accuracy of the visual sentiment classifier, these preliminary results are encouraging. With samples of between 100 and 2,000 tweeted images from a Black Lives Matter protest, we were able to (1) identify whether the picture was of a protest, (2) identify whether the picture contained the singer John Legend, and (3) predict whether the picture would evoke one of five emotions.

We hope these results will encourage social science researchers who are asking important questions about the images, whether they want to derive a separate measure of interest from the data in images (e.g., the number of protesters) or whether they want to study how images affect attitudes and behaviors (e.g how images affect willingness to tweet about a social movement). Researchers

with thousands or even millions of images should not be intimidated by the prospect of having to label those images in order to get to theoretical questions of interest. An investment in manually labeling a random subsample of those images, combined with an investment in learning how to use CNNs, will unlock huge datasets of images. In our project, we have gold standard labels for all of our 9,500 tweeted images. But what if we had collected 200,000 total unique images? Instead of painstakingly labeling all of them by hand with multiple annotators, we could use the algorithms trained on a sample of images, asking the algorithm to automatically label the remaining images. The final data would allow us to build analyses from a massive corpus of images.

7 Appendix I

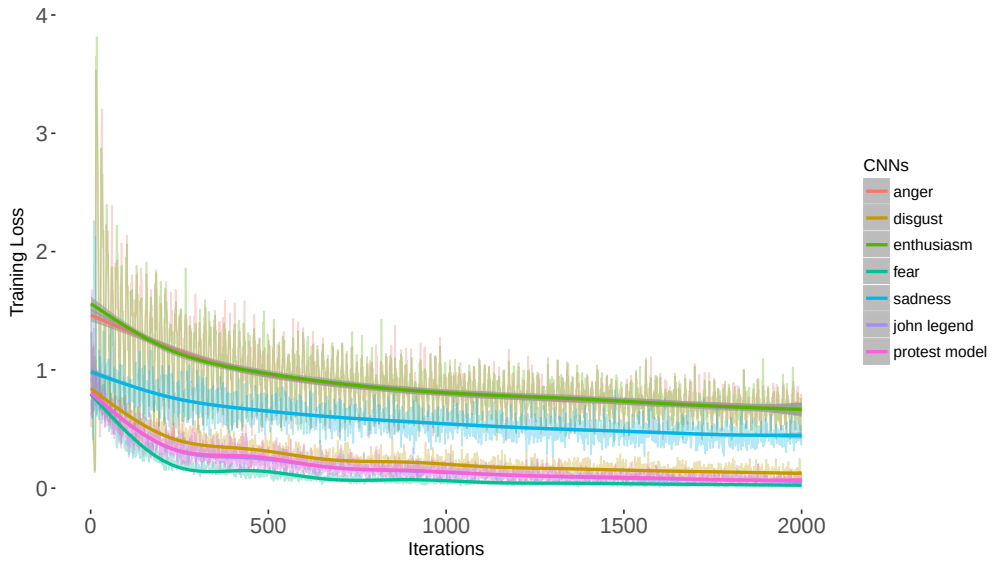


Figure 16: Training Loss of seven Convolutional Neural Nets trained for 2,000 iterations

References

Anastasopoulos, L. Jason et al. (2016). "Photographic Home Styles in Congress: A Computer Vision Approach". In: pp. 1–52. arXiv: 1611.09942. URL: http://arxiv.org/abs/1611.09942.

Barry, Ann M. (2002). "Perception and Visual Communication Theory". In: *Journal of Visual Literacy* 22.1, pp. 91–106.

Bennett, W Lance and Alexandra Segerberg (2013). *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*. New York: Cambridge University Press.

Benoit, Kenneth et al. (2016). "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data". In: *American Political Science Review* 110.2, pp. 278–295.

Bimber, Bruce, Andrew J Flanagin, and Cynthia Stohl (2005). "Reconceptualizing Collective Action in the Contemporary Media Environment". In: *Communication Theory* 15.4, pp. 365–388. ISSN: 1050-3293, 1468-2885. DOI: 10.1111/j.1468-2885.2005.tb00340.x. URL: http://doi.wiley.com/10.1111/j.1468-2885.2005.tb00340.x.

Brantner, Cornelia, Katharina Lobinger, and Irmgard Wetzstein (2011). "Effects of Visual Framing on Emotional Responses and Evaluations of News Stories about the Gaza Conflict 2009". In: *Journalism & Mass Communication Quarterly* 88.3, pp. 523–540.

Casas, Andreu and Nora Webb Williams (2017). "Images that Matter: Online Protests and the Mobilizing Role of Pictures". URL: https://ssrn.com/abstract=2832805.

Castells, Manuel (2012). *Networks of outrage and hope: social movements in the Internet Age*. Cambridge, UK ; Malden, MA: Polity Press. ISBN: 978-0-7456-6284-8.

Chen, Jianfu et al. (2015). "Déjà Image-captions: A Corpus of Expressive Descriptions in Repetition". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 504–514.

Clifford, Scott and Dane G. Wendell (2016). "How Disgust Influences Health Purity Attitudes". In: *Political Behavior* 38.1, pp. 155–178.

Corrigall-Brown, Catherine and Rima Wilkes (2012). "Picturing Protest: The Visual Framing of Collective Action by First Nations in Canada". In: *American Behavioral Scientist* 56.2, pp. 223–243.

Downs, Anthony (1957). *An Economic Theory of Democracy*. New York: Harper & Row.

Ekman, Paul (1992). "An Argument for Basic Emotions". In: *Cognition and Emotion* 6.3/4, pp. 169–200.

Fisher, Max and Jugal K. Patel (2017). *What One Photo Tells Us About North Korea's Nuclear Program*. URL: https://www.nytimes.com/interactive/2017/02/24/world/asia/north-korea-propaganda-photo.html.

Gazzaniga, Michael S. (1998). *The Mind's Past*. Berkley and Los Angeles, CA: University of California Press.

Golle, Philippe (2008). "Machine learning attacks against the Asirra CAPTCHA". In: *CCS '08 Proceedings of the 15th ACM conference on computer and communications security*, pp. 535–542.

Goodwin, Jeff and James M Jasper (2006). "Emotions and Social Movements". In: *Handbook of the Sociology of Emotions*. Ed. by Jan E Stets and Jonathan H. Turner. Boston: Springer, pp. 611–635.

Grabe, Maria Elizabeth and Erik Page Bucy (2009). *Image bite politics: news and the visual framing of elections*. Oxford; New York: Oxford University Press.

Graber, Doris A (1996). "Say it with Pictures". In: *The Annals of the American Academy of Political and Social Science* 546.The Media and Politics, pp. 85–96.

Guo, Yandong et al. (2016). "MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World". In: *Electronic Imaging*, pp. 1–6.

Howard, Philip N. and Muzammil M. Hussain (2013). *Democracy's Fourth Wave?: Digital Media and the Arab Spring*. New York, NY: Oxford University Press.

Iyer, Aarti and Julian Oldmeadow (2006). "Picture this: emotional and political responses to photographs of the Kenneth Bigley kidnapping". In: *European Journal of Social Psychology* 36.5, pp. 635–647.

Janis, Irving L. and Leon Mann (1977). *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*. New York: The Free Press.

Jean, Neal et al. (2016). "Combining satellite imagery and machine learning to predict poverty." In: *Science* 353.6301, pp. 790–4. ISSN: 1095-9203. DOI: 10.1126/science.aaf7894. URL: http://www.ncbi.nlm.nih.gov/pubmed/27540167.

Katz, Elihu and Paul F. Lazarsfeld (1955). *Personal Influence, the Part Played by People in the Flow of Mass Communications*. New Brunswick and London: Transaction Publishers.

Kharroub, Tamara and Ozen Bas (2015). "Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution". In: *New Media & Society*. ISSN: 1461-4448, 1461-7315. DOI: 10.1177/1461444815571914. URL: http://nms.sagepub.com/cgi/doi/10.1177/1461444815571914.

Kraidy, Ute (2002). "Digital Media and Education: Cognitive Impact of Information Visualization". In: *Journal of Educational Media* 27.3, pp. 95–106.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*, pp. 1106–1114.

Kuklinski, James H and Norman L. Hurley (1994). "On Hearing and Interpreting Political Messages: A Cautionary Tale of Citizen Cue-Taking". In: *Journal of Politics* 56.3, pp. 729–751.

Lasswell, Harold D. (1930). *Psychopathology and Politics*. Chicago: University of Chicago Press.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "No Title". In: *Nature* 521.7553, pp. 436–444.

Li, Haoxiang et al. (2015). "A Convolutional Neural Network Cascade for Face Detection". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition*.

Lupia, Arthur (1992). "Busy Voters, Agenda Control, and the Power of Information". In: *The American Political Science Review* 86.2, pp. 390–403.

Marcus, George E., W. Russell Neuman, and Michael MacKuen (2000). *Affective Intelligence and Political Judgement*. Chicago and London: The University of Chicago Press.

McCarthy, John D and Mayer N Zald (1977). "Resource Mobilization and Social Movements: A Partial Theory". In: *American Journal of Sociology* 82.6, pp. 1212–1241.

Melucci, Alberto (1996). *Challenging Codes: Collective Action in the Information Age*. Cambridge, UK: Cambridge University Press.

Nisbet, Matthew C. and John E. Kotcher (2009). "A Two-Step Flow of Influence?: Opinion-Leader Campaigns on Climate Change". In: *Science Communication* 30.3, pp. 328–354.

Obserschall, Anthony (1973). *Social Conflict and Social Movements*. Englewood Cliffs, NJ: Prentice-Hall.

Olson, Mancur (1965). *The Logic of Collective Action: Public Goods and the Theory of Goups*. Cambridge, MA: Harvard University Press.

Ordonez, Vicente et al. (2013). "From Large Scale Image Categorization to Entry-Level Categories". In: *2013 IEEE International Conference on Computer Vision (ICCV)*, pp. 2768–2775. ISBN: 978-1-4799-2840-8. DOI: 10.1109/ICCV.2013.344.

Peng, Kuan-chuan et al. (2015). "A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9. ISBN: 9781467369640. URL: papers3://publication/uuid/F6648898-46AD-484A-A4FC-47E1B64DD916.

Rosenberg, Shawn W. et al. (1986). "The Image and the Vote: The Effect of Candidate Presentation on Voter Preference". In: *American Journal of Political Science* 30.1, pp. 108–127.

Russakovsky, Olga, Li-Jia Li, and Li Fei-Fei (2015). "Best of both worlds: Human-machine collaboration for object annotation". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2121–2131. ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298824.

Sears, David O. (1987). "Political Psychology". In: *Annual Review of Psychology* 38.1, pp. 229–255.

Shachar, Ron and Barry Nalebuff (1999). "Follow the Leader: Theory and Evidence on Political Participation". In: *American Economic Review* 89.3, pp. 525–547.

Tajfel, Henri (1981). *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge; New York: Cambridge University Press.

— (1982). "Social Psychology of Intergroup Relations". In: *Annual Review of Psychology* 33.1, pp. 1–39.

Valentino, Nicholas A et al. (2011). "Election Night's Alright for Fighting: The Role of Emotions in Political Participation". In: *Journal of Politics* 73.1, pp. 156–170.

Victoroff, Jeff (2005). "The Mind of the Terrorist: A Review and Critique of Psychological Approaches". In: *Journal of Conflict Resolution* 49.1, pp. 3–42.

Wright, Matthew and Jack Citirn (2011). "Saved by the Stars and Stripes? Images of Protest, Salience of Threat, and Immigration Attitudes". In: *American Politics Research* 39.2, pp. 323–343.

You, Quanzeng et al. (2015). "Robust Image Sentiment Analysis using Progressively Trained and Domain Transferred Deep Networks". In: *The Twenty-Ninth AAAI Conference*, pp. 381–388.

ISBN: 9781577356998. DOI: 10.1145/2733373.2806284. arXiv: arXiv:1509.06041v1. URL: http://www.cs.rochester.edu/u/qyou/papers/sentiment%7B%5C_%7Danalysis%7B%5C_%7Dfinal.pdf.

Zaller, John (1992). *The Nature and Origins of Mass Opinion*. Cambridge; New York: Cambridge University Press.

Zhu, Xiangxin and Deva Ramanan (2012). "Face detection, pose estimation, and landmark estimation in the wild." In: *International Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2886. ISBN: 9781467312288. DOI: 10.1109/CVPR.2012.6248014.