Spring 5-15-2019

A Visual Political World: Determinants and Effects of Visual Content

Silvia Michelle Torres Pacheco
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Political Science

Dissertation Examination Committee:
Betsy Sinclair, Co-Chair
Jeff Gill, Co-Chair
Sanmay Das
Jacob Montgomery
Steven S. Smith

A Visual Political World:
Determinants and Effects of Visual Content
by
Silvia Michelle Torres Pacheco

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation is the end of a strenuous and fascinating path that objectively lasted 6 years, but that in reality started developing a bit before then. I have always liked describing the whole graduate school adventure as a roller coaster of emotions. The downs were scary and very frustrating, but the ups have been so wonderful that they definitely made up for the bumpy parts. The highlights were that 1) during this path I got to meet a list of fantastic people, and 2) I re-affirmed that I have the best support team composed of mentors, friends and family. To all of you who have walked this rocky road by my side: I don't have words to express my gratitude and my love to you. Thanks for taking me to this point!

As I said, this all started a bit before grad school, so I really need to thank my mentors from CIDE and Buendía y Laredo not only for training and sending me here, but also for all the opportunities that you gave me: Javier Aparicio, Allyson Benton, Jorge Buendía, Lorena Ruano, Salvador Vázquez del Mercado and Daniel Yanes. A special thanks to Javier Márquez who inspired me to be a methodologist. Javier, I hope one day I can have at least one tenth of your brilliance and kindness; thanks for letting me learn from your passion and wisdom.

I also need to thank a crucial team of my grad school experience: the staff of the political science department and Weidenbaum Center at WashU (Gloria Lucy, Christine Moseley, Colleen Skaggs, Heather Sloan-Randick, Sue Tuhro and Melinda Warren). I really appreciate your support and help with my infinite number of requests.

I would also like to say thank you to the many scholars that read my work, discussed it, or gave me feedback during all of these years in conferences, workshops, etc. Your knowledge and guidance in the academic dimension was invaluable, but your support and help in other dimensions made everything better. In no particular order, thanks to Michael Bechtel, Sarah Brierley, George Ofosu, Andrew Reeves, Dan Butler, Adam Dynes, Ariela Schachter, David

Cunningham, Jim Gibson, Adriana Crespo-Tenorio, Margit Tavits, Francisco Cantú, Leslie Schwindt-Bayer, Matthew Hayes, Diana O'Brien, Erin Hartman, Santiago Olivella, Ashley Leeds, Kosuke Imai, Brendan Nyhan, Inés Levin, Kentaro Fukumoto, Justin Fox, Vera Troeger, Lonna Atkeson, Natalie Jackson, Justin Esarey, Matt Blackwell, Guillermo Rosas, Molly Roberts, Jon Rogowski, and Bill Lowry (Bill, I am going to miss our Cardinals/Blues talks in the hallway so much!)

I was able to stay somewhat sane thanks to the infinite support and love of my friends back home, who have been there for me in good and bad times since ever. Thanks to the fantastic Lucys (Ale, Chivis, Gisy, Paola, Lucy C., Lucy M., Lile, Fa, Ros and Susy), Onawa, Liz, Ale Nava, Karlis Cruz and Dan for always making me feel that you were there with me regardless of the circumstances.

I have to say that I have always struggled with home-sickness, but I have been extremely lucky to meet amazing friends that shared this crazy stage with me. Thanks for your encouragement, for listening to 1,000 presentations of my work, for tolerating my infinite conversation loops, and for eating ice-cream (or sushi!) with me when I needed it. It is the coolest thing that you all crazy smart people who I deeply admire are also my friends. Thanks Betül, Connie, David, Jonas, Adrián, Andy, Jeff, Jae-Hee, Jeong, Viktoryia, and Patrick for the infinite lessons and good times at WashU. Also, thanks Mayya, Connor, Chris and Sören for making academia so fantastic! Despite all the problems that this world still needs to take care of, you all give me hope and make me excited about the adventures ahead. Finally, special thanks to the methods wonder women at WashU. Erin, Min Hee and Luwei: you inspire me to be a better scholar and a better person. You will never know how much impact your kindness and your support had on the scholar I am now. Thanks for giving me the fuel and motivation to keep fighting. I can't way to keep working with you to make the world a better place.

I also need to give special thanks to the people that became my family in these foreign

lands. My very beloved Júlio, Miguel, Christina, Arthur, Charlie, Josh and Kaitlin (and Lucie!). This whole adventure would have been extremely hard, if not impossible, without you and your amazing personalities. Thanks for all the shared passions, the food, the trips, for listening to my rants and dealing with those episodes of extreme negativity, for hugging me when I was crying and for making me cry with joy, for traveling to the exotic Mexico with your babies, and for all the great moments. You made St. Louis amazing, and I am deeply grateful for the fact that you opened your homes and hearts to this weird lady. Just know that you have my eternal gratitude and love for that.

A million thanks to my Mexican and German families. Dear Homolas and Co., I will forever be grateful for all your support during these years. Thanks for adopting me and making me feel so welcome all the time; your love (paired with chocolate and bike trips) have been a huge determinant of my happiness and development through grad school. Thanks, Schatzi and Papapap for always tracking my events, successes, deadlines and anxiety. I won the jackpot with you as my in-laws! And to my lovely Mexican family: you know that being far from you has been the hardest part of the experience. I will never be able to express how much strength you gave me through these years: every message, FaceTime call, smile, and detail has been incredibly powerful. I love that despite the distance, we are closer than ever. I love you with all my heart families Ferrer Pacheco, Pacheco Alonso, Pacheco Canto, Pacheco Maldonado, Pacheco Torres, and Torres Rodríguez. Special thanks to Mimi and Hugo for your prayers, support and even the million Facebook messages with videos and memes. And to my incredible godchildren and cousins: I cannot imagine better hands to put the world's future in than yours.

As I am approaching the end, I would like to switch to Spanish to thank my parents and siblings. Furiosos de mi alma, gracias por ser mi soporte más fuerte y simplemente la mejor familia del mundo. Hermano oso, gracias por las lecciones de vida y por siempre estar apoyándome, siempre seré la admiradora más grande de tu inteligencia y habilidades.

Chio, siempre serás la persona que más admire y mi mayor fuente de inspiración. Gracias por los videos, los mensajes y por llenarme de energía con tu amor y tu pureza. Papito, tu sabiduría e inteligencia son inspiradoras. Gracias por cada acción que me ha llevado hasta donde estoy y que se enfoca en garantizar mi felicidad: desde comprarme un comal cuando me quejo de mis tortillas, hasta los viajes de emergencia para volverme a la vida. Mamita, no hay palabras para agradecerte ese amor tan inmenso que me das. No hay forma de que hubiera terminado esto sin tu apoyo, llamadas y cariño. Estar lejos de ti siempre será lo mas difícil de todo este proceso. Gracias furiosos porque siempre han estado ahí para apoyarme en mis locuras, aún cuando no estén muy convencidos de ellas. Ustedes son mi razón de ser.

I left my committee members, the Dream Team, almost until the end because everything I have accomplished in the past years is entirely attributable to their efforts, support and trust in me. Sanmay, thanks for reading the work of a political scientist who truly admires your work and who tries really hard to understand a bit of the amazingness of computer science. Thanks for joining the team without hesitation. Steve, thanks a million for your endless support. You always brought me back to Earth when I was losing the substantive focus, and your insights gave my work and my research interests a whole new perspective. I will always be in awe of your knowledge and wisdom, and therefore I will deeply miss our long conversations where I learned more about American politics than in all my classes and books combined. Jeff, I came to WashU wanting to work with you but also very concerned that I was not good enough to work with someone I admired so much. Thanks for trusting me and for giving me so many opportunities. I am very proud of saying that I belong to your legion of students. Jacob, words are not enough to say thank you for what you have done for me. I could not have done this without your encouragement and guidance. Thanks for convincing me that I could be a methodologist, and for all the countless hours you spent working on my papers, ideas, presentations, and even rants. Thank you for also sharing some of the best and worst moments of my life: having you next to my hospital bed and

dancing with me in my wedding meant the world to me. Betsy, I will be forever grateful to God, destiny and the universe for crossing our paths. I have not met a person with your energy, kindness, intelligence and brilliance. Thanks for sharing your magic with me, and for letting me learn from the incredible human being and scholar that you are. Some of my best memories from gad school involve our conversations and the feeling of overwhelming passion and excitement for my work and for science after those. Thanks for always having the most magical comment, for giving spark to my work, listening to my frustrations, and wiping my tears; thanks for making me believe in myself. You inspire me in so many dimensions. You all are really the Dream Team! I honestly cannot imagine a better committee than this and I cannot thank you enough for taking the risk of coaching me.

Finally, I would like to thank the most crucial person in this process. My office mate, cohort mate, co-author, future colleague and husband, Jonathan. It is a privilege that I got to share this experience with the smartest and most amazing person I know in the planet. The truth is that without you, this would have been impossible. Thanks for being an editor, reviewer, discussant, therapist, cheerleader, voice of reason, devil's advocate and best friend. Thanks for sharing your knowledge, spark and charm with me, and for your infinite patience during this stage that was very hard for me, and that you managed to make magic and amazing. I really cannot wait for all the new adventures, successes, rejections (hopefully just a few), travels and scientific journeys ahead. I love you from the 3 to the last digit of π!

All the strengths and good qualities of my work (and of me as a scholar) are thanks to you all. Any potential weaknesses are due to my own shortcomings. I owe everything to you all and I hope that we will keep creating fantastic moments together. Hopefully to also give back to you a little bit of what you have done for me. Thanks from the bottom of my heart!

Michelle

Washington University in St. Louis

May 2019

ABSTRACT OF THE DISSERTATION

A Visual Political World:

Determinants and Effects of Visual Content

by

Silvia Michelle Torres Pacheco

Doctor of Philosophy in Political Science

Washington University in St. Louis, 2019

Political communication is a central element of several political dynamics. Its visual component is crucial in understanding the origin, characteristics and consequences of the messages sent between political figures, media and citizens. However, visual features have been largely overlooked in Political Science. Thus, in this dissertation, I introduce, describe and apply computer vision techniques for the analysis and processing of visual material, in order to not only improve data collection and visual content extraction, but also the understanding of the effects that visual components have on relevant political variables. In the first main chapter of this project, I implement computer vision and image retrieval techniques to measure and understand messages conveyed in pictures. The chapter presents and details the implementation of a Bag of Visual Words (BoVW), an intuitive and accessible technique for the extraction and quantification of visual features that allows researchers to build an Image-Visual Word matrix that emulates the Document-Term matrix in text analysis. For the purposes of this chapter, I validate the BoVW approach using a structural topic model to identify relevant political features of images of the migrant caravan. The second main chapter introduces and describes the implementation of a popular tool in Computer Vision, Convolutional Neural Networks (CNN), for the processing and extraction of political information. I apply the CNN for the extraction of handwritten numbers in electoral tallies,

and enumerate the benefits and drawbacks that this technique has for the study of political events. Finally, the third main chapter studies the factors behind the generation of visual frames, and the impact that these have on political attitudes. By focusing on the depictions and visual framing of protests, I find that conservative newspapers depict the protests in darker and nocturnal settings more often than liberal outlets. Further, the framing of the mood of the environment with conflict-related elements have an impact on the opinions and attitudes towards social movements: depictions of violence negatively affect identification and engagement with the movement, and these effects are moderated by who is the actor behind the violent events. Overall, the dissertation focuses on the importance that visuals have on the way that citizens engage with political information, and provides a framework that allows researchers to have a better understanding of several political dynamics.

Chapter 1

Introduction

"Photographs furnish evidence," reflects Susan Sontag (1977, p. 6) on the importance of visual information. "Something we hear about, but doubt, seems proven when we're shown a photograph of it." The communicative power of images comes from humans' capacity to process images faster and with less cognitive effort than text or speech (Hockley 2008; Madigan 1983). As a result, the prevalence of visual information in everyday social interactions should not be surprising: ads and posters in campaigns, photos in newspapers, documentation of contemporary events across news channels, political cartoons, and even artistic manifestations with social messages. The importance of this form of communication has recently became more evident with the ease of taking and sharing images, increasing the flow of visual images at a rate of at least 95 million new pictures per day.[1]

This goes beyond just the amount of visual material surrounding us. "The widespread use of cameras by people around the world has created more than a mass of images; it has created a new form of encounter, an encounter between people who take, watch, and show other people's photographs, with or without their consent, thus opening new possibilities of political action and forming new conditions for its visibility" (Azoulay 2008, p. 24).

---

[1] "45 Visual Content Marketing Statistics You Should Know in 2018", Jesse Mawhinney, *HubSpot*; available at https://blog.hubspot.com/marketing/visual-content-marketing-strategy, August 20, 2018)

Social scientists have recognized the importance of images and visual content in political dynamics. Several authors have focused on the role of visual material in candidates' evaluations (Bauer and Carpinella 2018; Valentino, Hutchings and White 2002), perceptions of social movements (Corrigall-Brown and Wilkes 2012; Fiske and Hancock 2016; Valenzuela 2013), depictions of war (Howe 2002; Keith, Schwalbe and Silcock 2006; Zelizer 2004), engagement with painful and impactful issues (Reinhardt 2012), and many other outcomes. However, the challenges that the multidimensionality and amount of images pose to their study and analysis have limited our knowledge about the process behind the generation and diffusion of political images, and their effect on attitudes and behavior.

However, the combination of technological advancements and collaboration with other fields like Computer Science, has provided social scientists with a variety of tools that allow them to study topics that were unthinkable in past decades. We can now gather fine-grained information on public sentiment and ideology using spatial relationships (Monogan and Gill 2016), collect and study complex data like social networks (Sinclair 2012), produce relevant demographic estimates at a low level of aggregation using tree-based models (Montgomery and Olivella 2018), analyze change in citizens' opinions across long spans of time (Tucker, Montgomery and Smith 2018), and even detect and measure how information providers like Wikipedia shape public opinion (Das, Lavoie and Magdon-Ismail 2016).

Along the same lines, important advances in the fields of computer vision and machine learning have allowed a more systematized, accurate and efficient way of analyzing large numbers of images and videos. However, in very few cases these have been applied to the study of political visual content. This is concerning, considering the prevalence of visual content and its key role in relevant debates in the areas of political communication, attitude formation, and others.

In this dissertation, I contribute to this debate by introducing and applying computer vision tools to the study of images relevant to social scientists. The dissertation has two

objectives. The first one, a methodological one, is to describe and illustrate the logic and implementation of two tools imported from computer vision: the Bag of Visual Words and Convolutional Neural Networks. As the use of these tools increases across disciplines, I intend to provide a comprehensive guide of these methods in an attempt to: 1) make them accessible to social scientists, 2) alleviate the concerns about their "black box" nature by providing a detailed description of their foundations and operation, and 3) discuss the challenges that researchers face when using these methods, as well as the limits to their scope and applicability. The second objective, a substantive one, is to illustrate the application of these tools to relevant political events such as social movements. This objective is the heart of this dissertation. The intention is to provide a framework for the analysis of images to reach new insights and acquire new knowledge about fundamental political dynamics like protests. What can we learn from the images surrounding us? What are the factors behind the generation of pictures of protests and their use as frames? What is the impact of these images in the opinion formation process of individuals?

I will briefly describe each of the three main chapters in detail.

1.1 The Bag of Visual Words: Using computer vision to understand visual frames and political communication

Political communication is a central element of several political dynamics. Its visual component is crucial in understanding the origin, characteristics and consequences of the messages sent between political figures, media and citizens. In this chapter, I introduce a tool to dissect the structure and content of visual material in order to assess its relationship with political variables: the Bag of Visual Words (BoVW), an approach that represents an image

as a collection of "patches" that emulate words in a text.

I present a survey of the literature regarding the impact of visual material on political attitudes, and the challenges that researchers face when quantifying pictures and video. I also introduce the BoVW as a tool to quantify images. In this section, I detail the steps to implement this technique in order to obtain a count of "visual words" per picture that emulates a Document-Term matrix in text analysis, and that could feed a wide variety of classification algorithms. In a nutshell, the BoVW consists on the identification of key points, the extraction of features from these points, measured as changes in pixel intensity, the construction of a visual vocabulary based on clusters of such features, and the association of all the features in the sample with the words belonging to the visual vocabulary.

The chapter includes the validation of this method using images of the caravan of Central American migrants trying to arrive and seek refugee in the U.S. For this purpose, I use a structural topic model to identify meaningful political components of the images of the caravan such as "crowds", "fences" and "camps", and conduct some descriptive analysis using those frames. The final part includes a discussion of the advantages and weaknesses of the method, especially in comparison with other tools popular in computer vision such as Convolutional Neural Networks, which I detail in Chapter 3.

1.2 Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data

Computers have increasingly taken over tedious and repetitive tasks previously assigned to humans. From counting words in a file to performing long and complex calculations, the computers' convenience is their capacity to follow a set of instructions in a repetitive way with no fatigue or cognitive bias. Their capacity to perform quickly and reliably allows us to analyze information from a large amount of data such as roll-call votes (McCarty, Poole and

Rowenthal 2006; Poole and Rosenthal 1997), congressional floor and supreme court speeches (Dietrich, Enos and Sen 2018; Dietrich, Hayes and O'Brien 2019), or social media posts (Barberá 2015).

And yet, despite their capacity to analyze textual or numeric information, computers have historically underperformed when classifying visual data. To overcome this, recent developments in computer science propose an alternative approach to classify images and retrieve information from them more efficiently and accurately. Rather than following a set of given rules, computers are now able to identify and transform image features—such as edges, textures, and motion—into a label after being exposed to multiple examples (the learning process).

This chapter presents the functioning, implementation, and challenges of Convolutional Neural Networks (CNNs)—one of the most popular tools for classifying visual information. The chapter states that the adequate use of CNNs reduces the resources necessary for the tedious task of classifying images and extracting information from them.

To illustrate this methodology, I apply a CNN to a relevant issue in election science: coding voting results. The chapter presents a way to collect the handwritten results registered in the vote tallies from the 2015 federal election in Mexico, captured in a set of comprehensive images. This example allows me to illustrate the characteristics of this method and the intuition behind its performance, providing an exhaustive guide for its implementation while still highlighting the challenges it poses, as well as offering recommendations and alternatives to overcome them.

1.3 Framing a Protest: Determinants and effects of visual frames

The information that media provides to citizens fuels their attitudes, opinions and behavior. Although scholars have extensively studied the way in which media communicates political events such as protests and social movements, existing analyses have mostly focused on the verbal component of news and have overlooked a crucial element of the communication process: the visual material.

Although most news outlets are committed to neutral and objective journalism, the variance in the way that they report political events, like protests, is high. The literature has mostly focused on issues related to coverage, provision of factual information and even style. However, the framing of stories using visual content has received less attention despite the impact that images have on the development of attitudes and behavior.

This chapter focuses on visual frames of protests and applies the Bag of Visual Wordes method (BoVW) developed in Chapter 2 to the study of this topic. First, I analyze the differences in the frames of the mood and environment that liberal and conservative outlets use when they talk about protests by identifying and capturing visual frames in the images posted in the Facebook feeds of prominent newspapers in the U.S. In order to achieve this, I extract an Image-Visual Word matrix from the corpus of images using the BoVW approach, that I subsequently feed into a Structural Topic Model. Results show that in the pictures that media outlets publish about protests of the Black Lives Matter movement, conservative newspapers tend to show a higher proportion of nocturnal and dark elements than liberal outlets.

The chapter also studies how the visual framing of violence affects attitudes towards social movements. In order to achieve this, I conduct two sets of experiments varying the level and type of visual and textual violence. The results show that violent depictions of

protests negatively affect evaluations of and engagement with social movements when the protesters are the perpetrators. This effect is larger when the violence is depicted through images instead of text.

These results allow us to have a better understanding of the effects of visual framing on political attitudes and participation, and the variables associated with the generation of these visual frames.

This dissertation provides a framework for the analysis of visual content relevant to social sciences, and offer rigorous and yet accessible tools that can help researchers with a deeper understanding of all the elements involved in political communication processes, and their impact on attitudes and behavior.

Chapter 2

The Bag of Visual Words: Using computer vision to understand visual frames and political communication

Citizens form their attitudes and act according to the information that their own experience and the sources surrounding them provide. Further, the diversity of these sources leads to different attitudes, opinions and behavior. For example, while former Rep. Beto O'Rourke condemned the U.S. government's response to the migrant caravan from Central America by stating that "[i]t should tell us something about her home country that a mother is willing to travel 2,000 miles with her 4-month old son to come here [to the U.S.]," President Donald Trump mobilized the military to stop the migrants and referred to the caravan as an "invasion" where "criminals and unknown Middle Easterners are mixed in." Such opposing perceptions of the same phenomenon warrant the question about what factors form these opinions.

Several studies use the analysis of messages exchanged between political actors to provide a better understanding of political events and behavior, as well as of attitude formation. How-

ever, with a few recent exceptions using audio-visual material (Anastasopoulos and Williams 2016; Bauer and Carpinella 2018; Casas and Williams 2018; Dietrich, Enos and Sen 2018; Knox and Lucas 2019; Lucas 2018), most of these studies focus solely on verbal communication and text analysis (Cho et al. 2003; Gamson and Modigliani 1989; Grimmer and Stewart 2013; Lecheler, Schuck and de Vreese 2013; Lecheler and de Vreese 2013). Meanwhile, visual material is an important element of human communication that has remained overlooked.

This omission is concerning given that 1) vision is a crucial sense involved in information processing via both conscious and unconscious paths, and 2) we are constantly exposed to political messages containing visual material. These facts motivate the following questions: what can we learn from the massive amount of images illustrating political events that surround us, and how can we quantify that visual material? The purpose of this article is to shed light on these questions by focusing on the caravan of Central American migrants and the pictures used to illustrate its activities, pilgrimage and arrival to the U.S.

What can we learn about the environment, size and mood of such a social movement based on the photos taken of it? What information does an image provide about the way in which a communicator frames an event? The large amount of images and the subjectivity of human coding are, among several others, two important challenges that complicate answering most of these questions. However, drawing from the computer vision field, I introduce a technique to political science that helps to summarize the content of visual material, and that serves as the basis for both supervised and unsupervised classifiers. This tool allows the dissection of a wide pool of images, and the identification of the structure and components of pictures of a given event. To conduct this analysis I use a Bag of Visual Words (BoVW) approach that represents an image as a collection of "patches" that emulate words in a text.

First, I present a survey of the literature regarding the impact of visual material on political attitudes, and the challenges that researchers face when quantifying pictures and video. Second, I introduce the Bag of Visual Words (BoVW) method as a tool to quantify images.

In this section, I detail the steps to implement this technique in order to obtain a count of "visual words" per picture that emulates a Document-Term matrix in text analysis, and that could feed a wide variety of classification algorithms. Then, I present the validation of this method in which I use a structural topic model to identify meaningful political components of the images of the caravan such as "crowds", "fences" and "camps", and conduct some descriptive analysis using those frames. Fourth, I test the relationship of these components with factors like the political leaning of news outlets, and find that right-leaning outlets tend to use pictures with large crowds more often than the rest of the outlets. Then, I discuss some of the advantages and limitations of this method, especially in comparison to other computer vision tools like Convolutional Neural Networks. Finally, I conclude with a list of steps for further research designed to improve the BoVW, its predictive power and accuracy, and potential applications of this method as well as its impact on the social sciences field.

The BoVW is a useful technique for the identification of the structure and the classification of visual content, and it provides social scientists with a comprehensive framework for uncovering messages in images. Further, this article highlights the differences in the way that media frames social phenomena and brings attention to the consequences of visual information for attitude formation.

2.1 Beyond words: images, frames and political attitudes

The impact of the amount and content of media messages on political opinions and attitudes has been widely explored (Davenport 2009; Downing 2000; Gerber, Karlan and Bergan 2009; Iyengar and Kinder 2010; Levendusky and Malhotra 2016; Newton 1999). However, most of the literature on this issue focuses on the textual messages that media send and does not consider the visual material that accompanies the text. There are several reasons to

be concerned about this omission. First, the eyes are our main source of information about the world. They send more data more quickly to the nervous system than any other sense (Barry 1997). The sensory signals that the eyes receive *first* travel to the thalamus and to the amygdala before a *second* signal is sent to the cortex (Zajonc 1984). The main implication of this circuit is that we begin to respond emotionally to visual stimuli *before* we can even process them in a conscious manner. Thus, without proper realization, emotional responses to visual sources influence attitudes, thinking and behavior (Erisen, Lodge and Taber 2014; LeDoux 1986).

Second, we are exposed to a large flow of visual stimuli. Some researchers suggest that we live in a visual age where our primary mode of communication are images (Kress, Van Leeuwen et al. 1996). Images are everywhere and constantly flowing: each year, television sends 48 million hours of original programming (Lyman and Varian 2001), Photoworld estimates that Snapchat users share 8,796 photos every second, and an average American is caught on surveillance cameras 75 times a day.[1] The reliance of organizations, parties, governments and activists on social media as a means for communicating their messages exponentiates the amount of visual material that society encounters.

Third, visuals can act as symbols that provide extra and sometimes implicit information that influences the way in which recipients understand the message these convey (Butz 2009; Mendelberg 1997, 2001; Valentino, Hutchings and White 2002). This information also helps to reinforce or highlight a message: the idea of "seeing to believing". Therefore, images are useful tools to *frame* a story for persuasion, agenda setting or other purposes (Iyengar 1994) through several pathways including the activation of emotions and predispositions (Butz, Plant and Doerr 2007; Ehrlinger et al. 2011; Valentino, Hutchings and White 2002). Thus, a visual frame is an element that an actor uses to relay information, and that reveals what she

---

[1] "How many photographs of you are out there in the world?", Rose Eveleth, *The Atlantic*, November 2, 2015; available at http://www.theatlantic.com/technology/archive/2015/11/how-many-photographs-of-you-are-out-there-in-the-world/413389/ (accessed October 16, 2016).

sees as relevant to the topic at hand (Chong 1996; Chong and Druckman 2007; Druckman 2003; Druckman and Nelson 2003; Gamson and Modigliani 1989)

To illustrate the existence, characteristics and analysis of these visual frames, this article focuses on the depictions of the caravan of Central American migrants asking for refugee in the United States. Also known as the *"Viacrucis del Migrante"*, these caravans are composed of people fleeing from gang violence, poverty and political repression in their countries of origin (mostly Guatemala, Honduras, and El Salvador), traveling from the Guatemala-Mexico border throughout Mexico with the objective of reaching the U.S. and seeking asylum. The first caravan, with an estimated number of around 700 migrants, started its journey on March 25, 2018, and reached the Mexico-U.S. border on April 29 after traveling 2,500 miles across Mexico. Other groups of about 1,000 people followed the same journey in the next months.

The caravans have triggered and intensified immigration debates, especially between the Mexican and American governments. Both countries have condemned the plans and actions of the migrants, and even used tactics to discourage mobilization (e.g. use of tear gas, deployment of troops, etc.). Further, citizens from both countries have also taken sides on the debate that either advocate for the migrants, their rights and their safety, or that evaluate them as a threatening source of crime and instability. Media coverage of the activities of this movement also reflects the variability in the perceptions of the caravans, especially regarding the visuals used to illustrate news pieces. For example, on October 5th of 2018, several news outlet covered the reaction of President Trump to the caravans and the threat he made to cut aid to countries like Honduras if the governments did not stop the formation and flow of caravans. Several media outlets used the text and facts from a news wire source but the pictures differed from outlet to outlet. While the *Columbus Dispatch* illustrated its article with the photo of a young girl walking in the caravan with her mom (top panel of Figure 2.1), other outlets showed greater densities of people. The *St. Louis Post-Dispatch* illustrated the article with a long line of people walking on the street (middle panel of Figure 2.1), and

12

the *Chicago Tribune* paired the text with an image showing a large flow of people walking on the street (bottom panel). The size and composition of the migrant groups portrayed in each image differ significantly between sources and this motivates the question: how can we quantify and explain these differences?

2.2   Quantifying images: the Bag of (Visual) Words

How do computers interpret images? An image is a set of pixels. A pixel is the finest unit defining an image and is considered the "color" or "intensity" of light that appears in a given place of the image. If we characterize an image as a grid, each square would contain a single pixel. These units are represented in two ways: grayscale and color. In grayscale, a pixel takes values between 0 and 255 representing the intensity of light. Thus, 0 is the darkest tone (black), while 255 is the brightest (white). At the same time, color pixels are generally represented in the RGB color space: they will take one value for the *Red* channel, one for the *Green*, and another for the *Blue* channel. These values also range from 0 to 255 which indicates the "amount" of each particular color in each pixel. Thus, images are represented as follows: 1) grayscale images are matrices with the number of columns and rows representing the width and height of the image in pixels respectively, while each cell entry contains the intensity of the pixel; and 2) color images are arrays comprising three matrices, each corresponding to the red, green and blue channels, and each cell denoting the intensity of the respective color. The similarities and differences between the intensity values of each pixel and its "neighbors" are the basis for the detection of shapes, edges, textures and objects in a picture. But how can we interpret these pixel intensities in meaningful ways?

Figure 2.1: One caravan, three perspectives: Pictures used in the October 5, 2018 coverage of the migrant caravan

(a) *Columbus Dispatch*

(b) *St. Louis Post-Dispatch*

(c) *Chicago Tribune*

2.2.1 Speaking the *image* language

In contrast to images, texts are composed of identifiable "tokens" like words, sentences or n-grams which make the text meaningful. Although images do not have these clearly defined tokens, the objects, edges and colors help us to identify their components and to make sense of their content. If we quantify and represent these features of an image (e.g. objects, lines, and patches) as "visual words" then we can use an analog variant of the Bag of Words, a popular technique used for text classification: the Bag of Visual Words (BoVW) (Grauman and Darrell 2005; Grauman and Leibe 2011; Grauman and Darrell 2007).

Consider this *very* simplified example in which we have four images A, B, C and D each showing a school bus, a car, a bicycle and a dog respectively. If we "break" the images into pieces to obtain a puzzle, and then we mix these pieces, we are no longer able to recognize the objects. However, we will still be able to identify certain elements. For example, we will have 10 pieces each showing a tire. A "tire" will therefore be a word in our "visual vocabulary". Then, during the classification of the images we will observe that pictures A and B (the school bus and the car) have four "tire" words each, picture C (the bicycle) has two, and picture D (the dog) has zero. If we compare the pictures based on this count of visual words then we will determine that picture A is the most similar to picture B, while picture D is the most contrasting. Visual word counts are the basis of the equivalent of the document-term matrix in text: the *Image-Visual Word matrix* (IVWM), that serves as the main input of a wide variety of classification and modeling techniques.

One of the main challenges that the analysis of visual material implies is the high dimensionality of the units of analysis. First, images contain multiple pixels that can be quantified based on several criteria and techniques. Second, the use of individual pixel values does not yield useful information for the categorization of images: content in an image cannot be captured with the intensity of individual pixels, but with the identification of connections and patterns that those pixels form. Finally, looking at different patterns of variation be-

tween images instead of raw pixel intensities helps to account for differences that might be independent from the actual content of the image, such as lighting conditions.

The BoVW involves a series of dimension reduction steps that ease the analysis and digestion of visual material. In a nutshell, the steps of the BoVW process are intuitive and straightforward: 1) identify local key points in the images under analysis and describe each of them using feature extraction, 2) cluster those features and quantify their centroids in order to form a codebook of "visual words", 3) measure similarity between the features of the images and the centroids, and identify the nearest visual word to each of an image's features, and finally, 4) summarize the image by counting the times that the visual words defined in the vocabulary appear in it (Csurka et al. 2004; Grauman and Darrell 2005; Sivic and Zisserman 2003; Sivic et al. 2005). These counts of words per image constitute the Image-Visual Word matrix (IVWM). In the following section I detail each of these steps.

2.2.2   Step 1: Extracting and describing local key points

The first step of the process of building a Bag of Visual Words consists of detecting local key points in either the full sample or a subsample of the pictures under analysis, and extracting their features. A "key point" is a salient region in the image generally representing edges, corners or significant changes in pixel intensity between the point and its surrounding neighbors. Identifying key points is the first step to simplify the data by discarding regions that will not offer useful information for classification purposes. For example, in most cases, a solid background of an image will not provide helpful information about its content. However, the edges, corners and salient points provide hints about the objects depicted in a picture and will serve as indicators of what is contained in them. In text-as-data language, we could understand this step as an initial removal of "stop words". The words or regions that will not aid in the classification and labeling process are discarded. Once the key regions are identified, we proceed to "describe" them through the extraction of the features that define

them. For the identification part we use a "locator", and for the feature extraction we use a "descriptor."

There are multiple classes of locators and descriptors that can be categorized along several dimensions such as speed, threshold criteria, sensitivity to transformations or accuracy.[2] For the purposes of this article, I use the FAST Hessian detector and the RootSIFT descriptor that I detail below.

Detecting key points

The FAST Hessian detector is used to locate edges and corners in an image (Bay, Tuytelaars and Van Gool 2006). It is suitable for the purposes of this article given its two key properties: scale invariance (i.e. key points should be both repeatable and recognizable at different scales of the image), and high computational speed. The logic behind this detector is to identify the points and regions where significant changes in pixel intensity occur, which generally corresponds to edges and corners. These elements define the objects found in a picture, and in turn are crucial for the description of its content. A more detailed description of the procedure in which the FAST Hessian identifies key points can be found in the Appendix. Figure 2.2 highlights in green the key points identified in the photo. The points appear in salient regions of the image, and match lines, contours and edges of the most prominent elements of the picture.[3]

Describing the key points

Any classification task requires features associated with labels. When using texts, features are words, sentences, or n-grams describing each document. In a normal regression setting, these would be equal to covariates. However, the identification of comparable features in

---

[2]For a detailed comparison and description of descriptors performance, please refer to Mikolajczyk and Schmid (2005) and Canclini et al. (2013).

[3]For illustrative purposes, the points in this image were detected using a FAST locator.

Figure 2.2: Location of key points

(a) Original image

(b) Image with key points identified

images poses some challenges. Although intuitively it is easy to think of a "visual word" as a patch of an image, in practice the actual quantification of such patch is problematic given the multi-dimensionality of a picture (i.e. intensities, location, color channels) and the absence of semantic meaning for "patches" or areas of a picture. Feature descriptors help in the task of measuring or representing image characteristics in mathematical forms that can subsequently be fed to classifiers or models. As in the case of detectors, there is a wide variety of alternatives that vary in computational costs, efficiency, complexity and accuracy. Researchers interested in image classification should select from these tools based on substantive knowledge of the problem under analysis, size, type and characteristics of their data, and resource constraints.[4]

In this project, I implement a RootSIFT descriptor which extracts and quantifies the region surrounding the key points identified in the previous step. This descriptor is an extension of one of the most popular descriptors in computer science: the Scale Invariant Feature Transform (SIFT, Lowe 1999) which has the advantage of being invariant to image translation, scaling, rotation, and even partially invariant to illumination changes. The RootSIFT was developed by Arandjelović and Zisserman (2012) who added two extra steps to the regular SIFT implementation to drastically improve accuracy: a L1-normalization of the SIFT vectors, and the calculation of the square root of the elements in each of those normalized vectors. However, we should go back to the beginning of the process and first understand what is the definition of an image feature, and how we can capture it. Then, I will review some of the details and operation of the SIFT descriptor.

The SIFT descriptor considers that the defining features of a key point are the direction and size of the changes in pixel intensity in different areas of its neighborhood. We can measure these changes using gradients: vectors that capture both the *direction* and *magnitude* in

---

[4]In the section "Strengths and weaknesses of the BoVW" I discuss some of the consequences of selecting certain parameters or descriptors over others.

which pixel intensities are changing. This method focuses on the calculation and summary of those elements. The following steps are not applied to the original image, I, but to a "blurred" version of it, A, using a Gaussian-smoothing filter. This processing step helps to clean the image by decreasing the sharpness of irrelevant elements like blobs or stains.

The feature extraction on the blurred image proceed as follows. First, for each of the key points identified in Section 2.2.2, the descriptor takes its 16×16 pixel surrounding area, and then divides it into 4×4 pixel cells. This leads to a grid composed of 16 cells, each with a width and a height of 4 pixels (Panel (a) of Figure 2.3). Then follows the key step of the process: the computation of the image gradients of the 16 pixels in each cell, and a subsequent reduction of these gradients into an 8-bin histogram. Note that a histogram is computed for each of the 16 cells that comprise the neighborhood of the key point. Intuitively, this step consists of exploring how the intensity of a given pixel compares to its surrounding neighbors (Panel (b) of Figure 2.3), followed by a summary of this information with gradients (Panel (c) of Figure 2.3). Formally, we estimate the gradients in both the $x-$direction (G_x) and the $y-$direction (G_y) at pixel $A(x, y)$ with the formulas:

$$G_x = A(x, y) - A(x + 1, y) \qquad\qquad G_y = A(x, y) - A(x, y + 1)$$

Then, we calculate the *magnitude* and the *orientation* as follows:

$$M_{x,y} = \sqrt{G_x^2 + G_y^2}$$
$$\theta_{x,y} = arctan2(G_y, G_x) \times \left(\frac{180}{\pi}\right)$$

Once again, if we focus on a single cell out of the 16 that we defined in the first step, this process yields 16 gradients with their respective magnitude and orientation that we summarize using a weighted count. To do this, first, we collapse all the potential gradient

Figure 2.3: Computing pixel intensity changes in the neighborhood of a key point

Panel (a)	Panel (b)	Panel (c)
16×16 neighborhood	Computation of	Representation of change
of key point	pixel intensity changes	with gradients
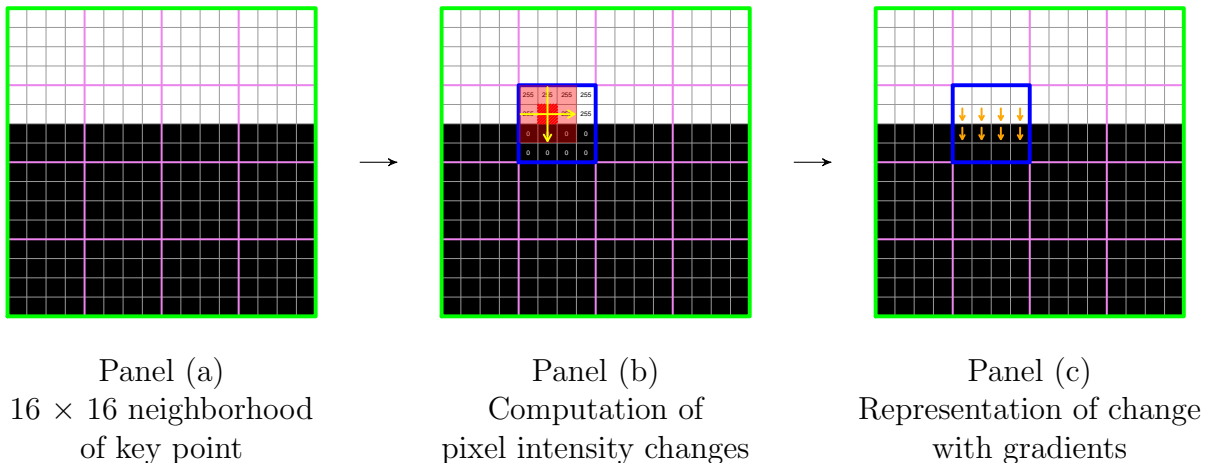
angles into 8 bins for the histograms. These angles are in the range of [0, 180] when unsigned[5] so we end up with bins that include around 20 potential angles. Then, we count the number of orientation values that fall into each of the bins, and weight them by their respective magnitude, and the distance to the key point. In other words, stronger pixel changes that are closer to the key point will be more relevant in the histogram construction.

After this process, each of the 4×4 cells is represented with an 8-element vector (Figure 2.4). The last step involves concatenating the 16 histograms, and taking the root of each of the elements of this new "flattened" long vector, in order to improve accuracy. At the end, the surrounding area of a key point is represented by a $4 \times 4 \times 8 = 128$ *feature vector* corresponding to the 8 gradient bins $\times$ the 16 cells of the neighborhood. Thus, a single image in our sample can now be represented with a number of vectors of length 128 equal to the number of key points that were detected in the first stage.

---

[5]When signed, the range of the angle values is [0, 360]. In general, it is common to use unsigned gradients, but researchers can opt for the signed range and also set a different number of bins.

Figure 2.4: Representation of the neighborhood of the key point with histograms
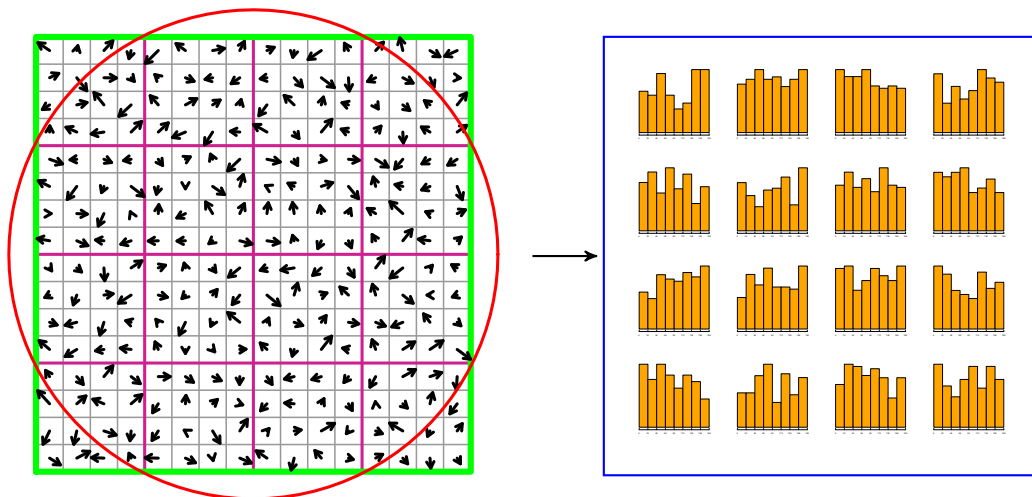
2.2.3 Step 2: Defining a vocabulary

As I discussed previously, the patches and features found in images do not have labels or semantic meaning as words. Therefore, we must define our own codebook or "visual vocabulary". To do this, we will cluster a randomly selected sample of features extracted from the key points of the images in our pool. Once we identify the v clusters, the features associated with each cluster's centroid serve as the representation of a word. Why do we lump together different patches instead of using the full set of features? Suppose that a sample of interest contains images of dogs, flowers and humans and that we are interested in classifying this pool according to the actor that each element depicts. For simplification purposes, imagine that after completing the steps above we found that one common neighborhood across human photos is (unsurprisingly) a human nose. However, although similar, it is extremely hard to find two identical noses; even two pictures of the same person would look different due to lighting, position, angles, etc. Therefore, we need the *average* of those noses to accurately represent a general concept of a nose. Thus, we can cluster the features associated with the nose and take the feature vector of the centroid as the representation

of our "visual word". Mathematically, this is going to be a vector with 128-elements, and graphically we can interpret it as a collection of the mini patches contained inside the cluster. This process is illustrated in Figure 2.5.

Figure 2.5: Creating the visual vocabulary: clustering and centroids
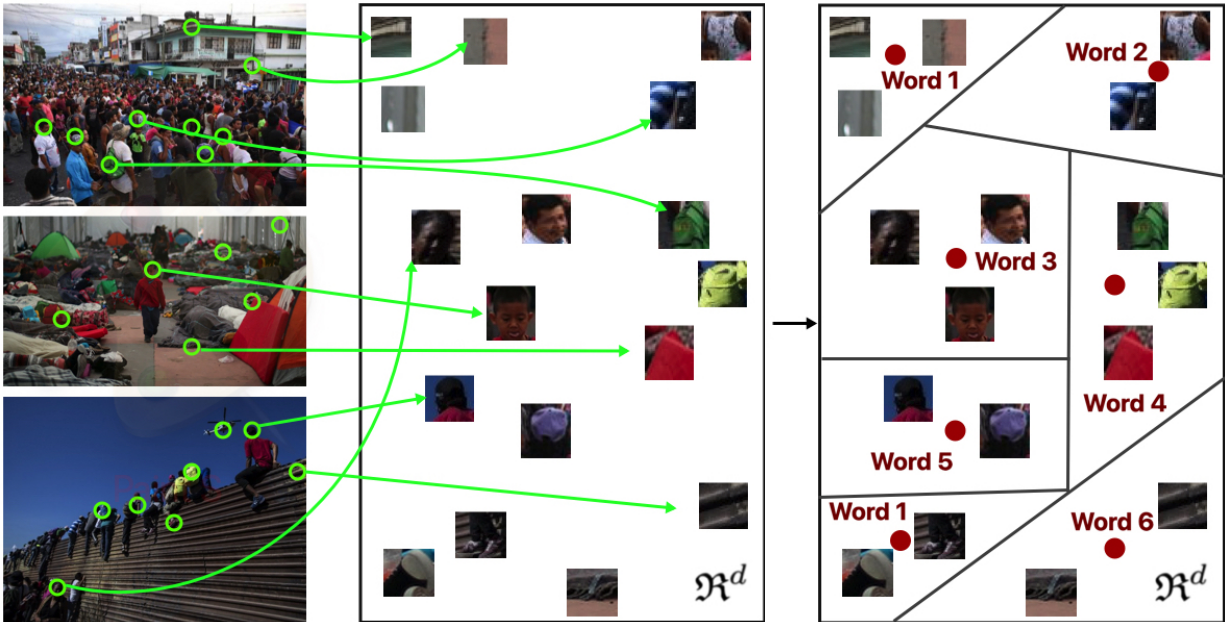
For the clustering process, I use a mini batch k-means algorithm that optimizes the distance between the feature vectors. This method requires that the user specifies the number of clusters to be generated. That is, the size of the vocabulary V will be equal to this parameter. In order to achieve higher levels of speed and efficiency we form the vocabulary based on a random sample of the feature vectors.[6] The structure and characteristics of the BoVW, and the accuracy of a given classifier are sensitive to these parameters. Therefore, it is advisable to vary these quantities and evaluate the differences in the results that each trial yields. Figure 2.6 shows a few examples of visual words that are generated by the clustering step. Numerically, each of these words will be represented by the feature vector of the cluster

---

[6]In general, taking 10-25% of the feature vectors is accepted as a common practice. However, this number will depend on computational capacity, speed necessities, and size of the data. For all the models in this article, I sampled between 30 and 35% of the features.

centroid to which they belong: the average of the feature vectors in that cluster. Notice that in some visual words, small patches seem to be repeated. This occurs because in many cases, the key points are very close to each other, and if that is the case, the neighborhoods (represented by the patches) will look almost identical.

Figure 2.6: Examples of visual words

(a) Water   (b) Large groups of people

(c) Wired fence
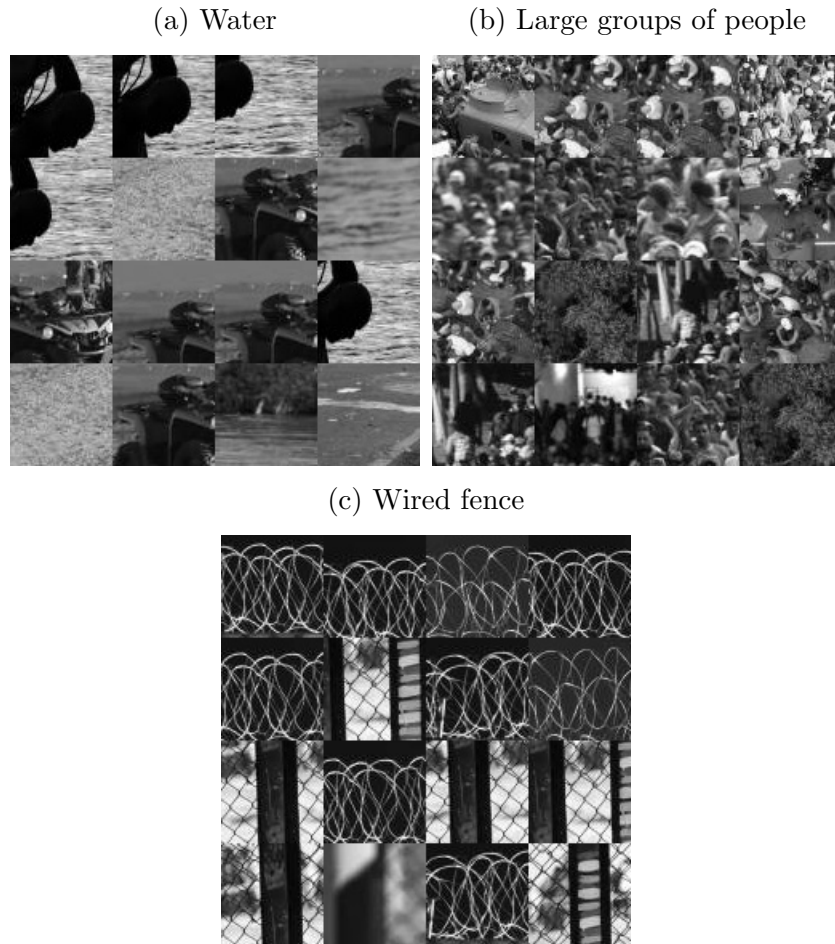
2.2.4   Step 3: Building the Image-Visual Word Matrix

Once we define a vocabulary, the last step consists of counting the number of times that each of the V "visual words" in the vocabulary appears in an image. While this closely emulates the building of the document-term matrix in text analysis, the multi-dimensional

structure of the features and visual words demands additional steps. Let I_n be one of the N images in the sample. If the image was used to build the vocabulary, the key points and their respective feature vectors have already been computed. If the image was not used to build the vocabulary, it is necessary to apply the detection and description steps detailed above. Suppose that 15 key points could be identified in I_n. Then, this image is represented by $M = 15$ feature vectors, $\mathbf{w} = [\overrightarrow{w_1}, \overrightarrow{w_2}, ..., \overrightarrow{w_{15}}]$. For each feature vector $\overrightarrow{w_m}$, we compute the Euclidean distance between it and the words in the vocabulary or, in other words, the feature vector of the centroid of the clusters we identified in the previous step. We add 1 to the count of word v in image I_n if:

$$\|\overrightarrow{w_m}, \overrightarrow{v}\| < \|\overrightarrow{w_m}, \overrightarrow{u}\| \qquad \qquad \text{for } u \neq v$$

In this way, each patch of an image is associated with a visual word in the vocabulary and we can identify the number of times a particular word appears in every photo. This constitutes our Image-Visual Word matrix.

2.3 Validating the BoVW: the migrant caravan in pictures

To illustrate the process outlined in the previous section and evaluate its performance, I build a BoVW and an Image-Visual Word matrix from images of the Central American migrant caravan. The objective is to use this BoVW to detect meaningful political components of the pictures of the caravan that could provide us with relevant information about dimensions of the movement such as size, composition, mood, environment or central actors related to it. More specifically, I focus on the analysis and identification of large groups or crowds of people given the information that this feature provides about the size and impact of the

caravan.

For the validation, I compiled a dataset with around 6,500 images of the caravan from multiple datasets including *Getty Images*, and 35 media outlets. This dataset includes photographs and most of them metadata covering the author of the picture, source, caption, dimensions, and keywords associated with each image.

2.3.1 Detecting underlying messages

With an appropriate training sample, the BoVW can also be set as the basis of supervised classifying and labeling exercises. However, its applicability covers multiple purposes. For example, we can conduct an exploratory analysis of the categories underlying a pool of images of interest (Feng and Lapata 2010; Monay and Gatica-Perez 2007). If the BoVW is correctly summarizing and quantifying the visual data, then the expectations are that, after an exploratory process, we should be able to identify latent topics in the images that 1) are cohesive and semantically sensible, 2) provide information about characteristics of the caravan, and 3) can be used to identify elements that the author or publisher of a picture uses to frame or depict the social phenomenon represented by the full pool of images.

In order to uncover these topics, I implement a Structural Topic Model (Roberts et al. 2014) based on the Image-Visual Word matrix obtained using the BoVW approach.

First, I build a visual vocabulary of 500 "visual words" based on the clustering of features of 5,952 photos from *Getty Images*. The images were collected using the tag "migrant caravan," and the search was restricted to pictures from Central America, Mexico and the U.S. between March 20, 2018 and November 18, 2018. The images that the *Getty* collection contains come from different photographers and sources, thus alleviating the concerns of potential individual biases in the coverage and content of the images published about the caravan. These images provide a rich composition of the events under analysis, and therefore are useful for building a vocabulary that contains as many frames as possible.

After building the vocabulary, I extract the Image-Visual Word matrix which I subsequently feed to a structural topic model initialized with 6 topics and three prevalence covariates: source (agency to which the photographer belongs), date and author of the photography. All of these account for the idiosyncrasies of the photographer and the particular characteristics of each event covered.

Overall, the results from the STM seem to identify coherent topics in the content of the images: border/fence, crowd, water/sky, shadows/darkness, small groups/portraits, and camps. These labels were manually assigned based on the most representative and exclusive visual words in each topic, as well as the most representative images per topic. Most of the topics, like "Border/Fence", "crowd" and "Camps" are correctly capturing the content they represent and giving information about the composition and dynamics of the caravan while accurately clustering similar patches of images. Further, other topics give some indications about the environment and mood of the movement: "Darkness" and "Water/Sky." Figure 2.7 shows 4 of the most frequent and exclusive visual words (FREX) from the six topics. The labeling of the topics was based on the ten most important words according to different measures.

Notice that the most representative visual words of most topics contain mini patches that clearly represent the group. For example, the topic "crowd" has visual words with patches showing large groups, dense conglomerations of people and granular textures. In contrast, a few topics such as "Small groups" reveal other less obvious features. In this case, it is possible to observe patches with human figures and body parts, but the most prominent patches are light and solid pieces generally found in the background of portraits. However, the topic becomes more obvious when we observe its most representative images (see Table 2.1).
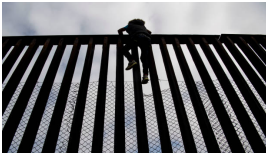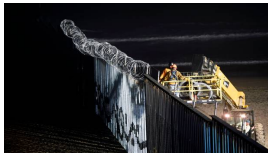
Therefore, it is important to also study and determine whether the most representative images per topic represent a cohesive and sensible theme. The most representative images of

27

Figure 2.7: FREX Visual Words per Topic (Getty model)

Topic 1:
Border/Fence

Topic 2:
Crowd

Topic 3:
Water/Sky

Topic 4:
Shadows/
Darkness

Topic 5:
Small groups/
Individuals
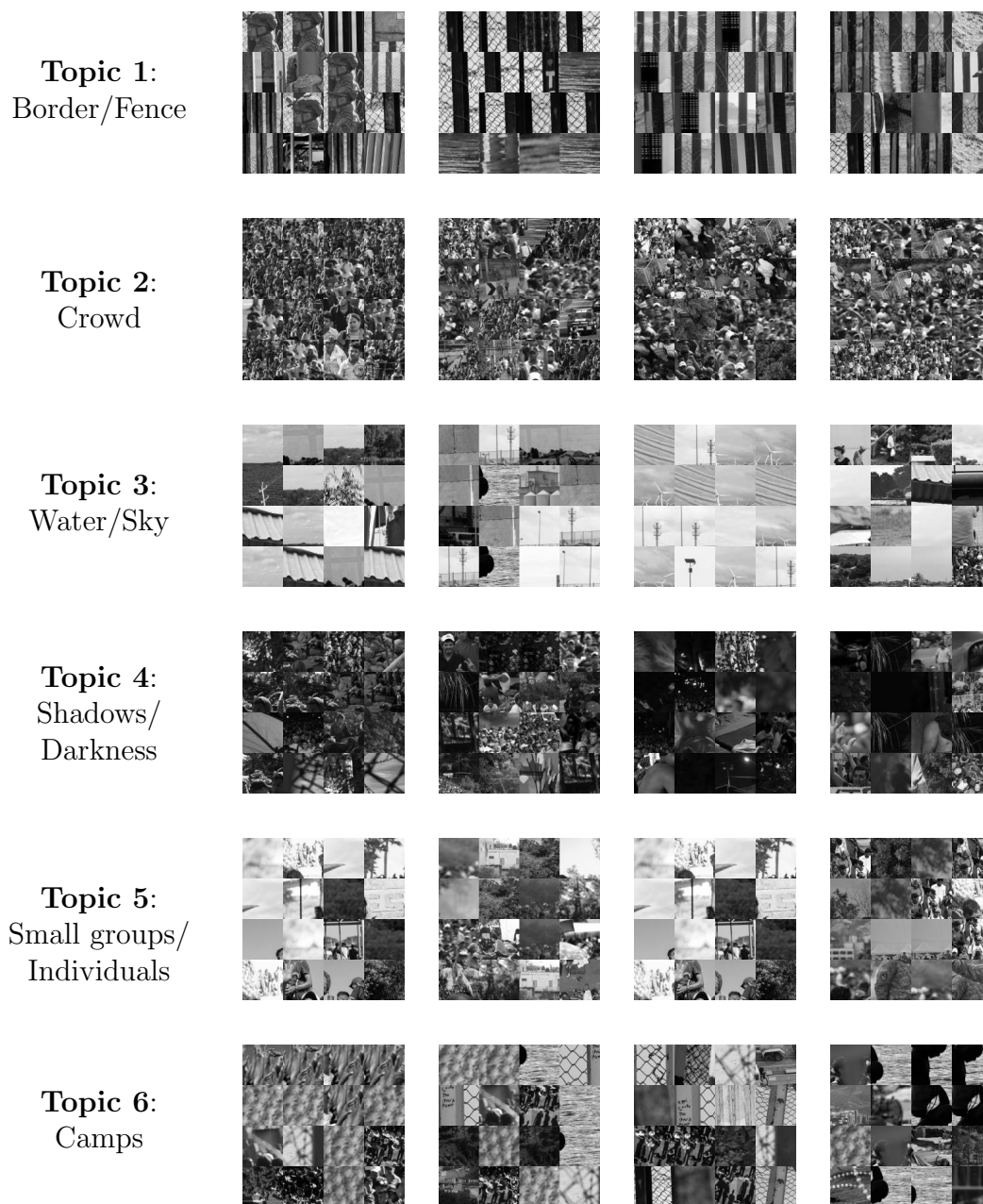
Topic 6:
Camps

a topic k are those photos with high proportions of such topic k. Table 2.1 presents examples of these. We can clearly identify coherent patterns in the data that not only validate the construction of the BoVW, but that also increase our knowledge of the data at hand, and that allow us to measure relevant dimensions for further analysis.

Table 2.1: Most representative images per Topic (Getty model)

Border/Fence

Crowd

Water/Sky

Shadows/Darkness

Small groups/Individuals

Camps

For example, we can analyze the use of the topic "crowd" through time to identify

whether there is any variation in the coverage of the magnitude dimension of the caravan. Figure 2.8 shows on the y-axis the proportion of topic "crowd" in the full corpus of images at different points in time (x-axis). The red dashed lines show relevant events that received wide coverage in the U.S. media, such as the arrival of the caravan to the U.S. border. It is interesting to notice that these events correspond to peaks in the dataset, suggesting a stronger focus on the size and magnitude of the caravan when its salience in the media market is higher.

2.3.2 Framing a political event: The Migrant Caravan

Perceptions of threat

The method and modeling approach outlined above allows us to extract and analyze structures and patterns in the visual material under analysis. However, there is still the question whether these dimensions and topics are associated with relevant information regarding the nature of political phenomena, and, more importantly, whether these are associated with political variables.

The topics that I identified using the BoVW and the STM give information about the characteristics of the migrant caravan such as time ("Darkness"), place ("Water/Sky", "Camps") and also magnitude of the group ("crowd"). In particular, this last feature is relevant given the effect that it has on the formation of opinions about issues related to immigration. More specifically, the size of an out-group impacts the perceptions of threat that this group poses, and these perceptions, in turn, shape attitudes towards the members of the group and the policies related to them.

The literature on attitudes towards immigration identifies several sources of threat that impact the attitudes of individuals towards immigration: cultural, economic, and security-related (Hainmueller and Hopkins 2014; Quillian 1995). The strength and origins of threat

30

Figure 2.8: Proportion of topic "crowd" in time

Use of Topic 'Crowd' in time
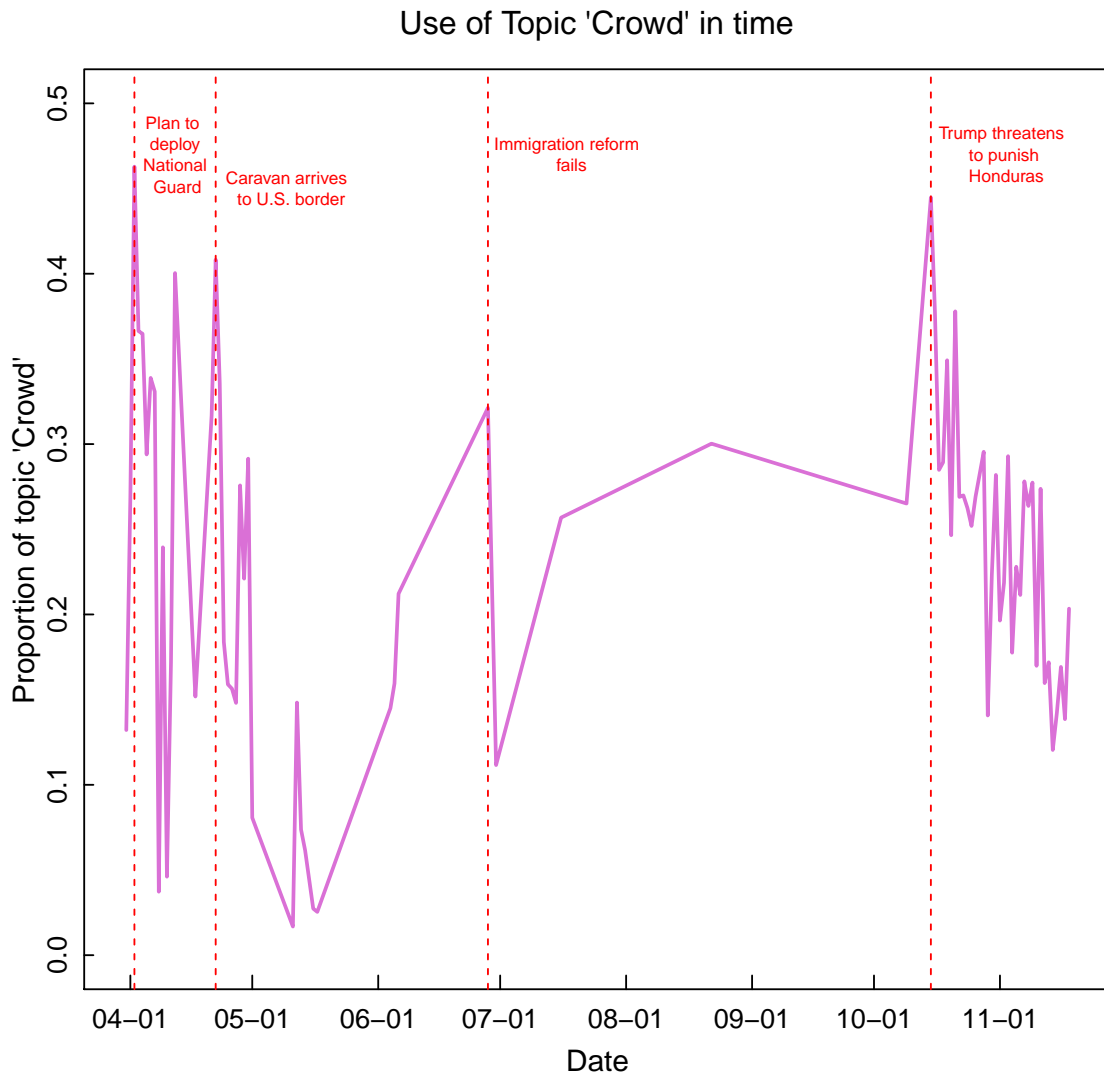
depend on multiple dimensions including situational and personal triggers like ideology (Homola and Tavits 2018; Lahav and Courtemanche 2012), predispositions (Sniderman, Hagendoorn and Prior 2004), and the ways in which threat is framed (Lahav and Courtemanche 2012). However, there are two fundamental ideas underlying the group threat theory: 1)

31

the struggle over scarce resources makes people more likely to favor their own group instead of the out-group, and 2) the potential for collective action against the majority increases disapproval of the out-group members. Thus, the relative size of a minority (or out-group) has an effect on threat (Schneider 2008): "the larger the minority group(s), the greater the threat and, correspondingly, the greater the antipathy felt towards it/them" (Hjerm 2007, p.1255).

This directly illustrate the relevance of studying the information that media provides about the size and characteristics of immigrant groups like the caravan. On a substantive and factual level, the depiction of a crowd provides queues about the magnitude of the movement and affects the evaluations of costs and benefits of receiving or supporting immigrants: is the competition for jobs going to increase? Are they going to establish new communities that alter the social and cultural tissue of the region? Are there criminals in such big groups of people? However, they also trigger other processes that occur in a less conscious manner. For example, Brunyé, Howe and Mahoney (2014) find that observers heavily rely on crowd size and density to estimate risk levels, while others suggest that humans are remarkably good at detecting (and being more attentive) to anger and conflict elicited by facial expressions and body language of individuals in a crowd (Green and Phillips 2004; McHugh et al. 2010; Öhman, Lundqvist and Esteves 2001). Even in an early stage of the visual processing, authors find that "feature congestion" and "display clutter," likely to be found in saturated images of crowds, have a negative effect on the attention and digestion of visual information (Rosenholtz et al. 2005).

Now recall the pictures in Figure 2.1. The pictures that media outlets use to illustrate an event show some variation in the use of the "crowd" element. What factors explain this variation in visual frames? There is evidence that media outlets define the coverage, content and style of the information they provide based on their audiences demands, marketing considerations, and their own ideologies and values (Earl et al. 2004; Fiske and Hancock

2016; Iyengar and Hahn 2009; Oliver and Myers 1999). More specifically, 1) media outlets are more likely to cover issues that fit their own and their customers' agenda, and 2) the content is going to be filtered through ideological lenses. Thus, we expect more negative framing of an issue or event when its ideological meaning lies further from the ideal point of a news outlet. For example, Oliver and Myers (1999) and Kriesi (1995) find that more left-wing newspapers cover more movement-related events. This leads to the expectation that, for the case of the caravan, we expect right leaning outlets to depict it in more threatening ways through the use of photos showing larger crowds than other outlets. This is line with the idea that, especially in the current political discussion, conservatives and right-leaning actors are more likely to hold negative views about immigration (Homola and Tavits 2018; Schemer 2012).

Data

To test the expectation that right-leaning outlets are more likely to depict the caravan as threatening through the use of crowds in pictures, I analyze the images of 451 articles covering the caravan of Central American migrants. These articles come from 35 news outlets and were published between October 3rd, 2018 and November 1st, 2018. They were compiled using the `News API`.[7]

Further, I complement this data with information regarding the ideological leaning of the outlet. The data comes from *All sides*, an organization that provides ratings of "media bias" (right, center-right, center, center-left, and left). The scores are based on surveys asking respondents about their own bias and how they rate the bias of news sites. Then they use this information and the aggregation of the rankings by ideological group and news

---

[7]The `News API` is a tool that allows to search for and retrieve information of events and news from more than 30,000 sources worldwide. I limited the search to only sources in the U.S. The reports and news are extracted from websites of several prominent outlets such as ABC, Politico, The New York Times, Fox News, Huffington Post, etc. The metadata includes date, author, image, headline, the truncated text of the article, original length of the article, and its URL.

outlet to determine the average bias rating of a source. Figure 2.9 shows the distribution of number of articles for the following groups: left leaning (center-left and left), center, and right leaning (center-right and right). The numbers next to each distribution indicate the percentage of the group in the sample (i.e. 13.3% of the outlets in the sample are evaluated as right leaning).

Figure 2.9: Number of images by ideological group
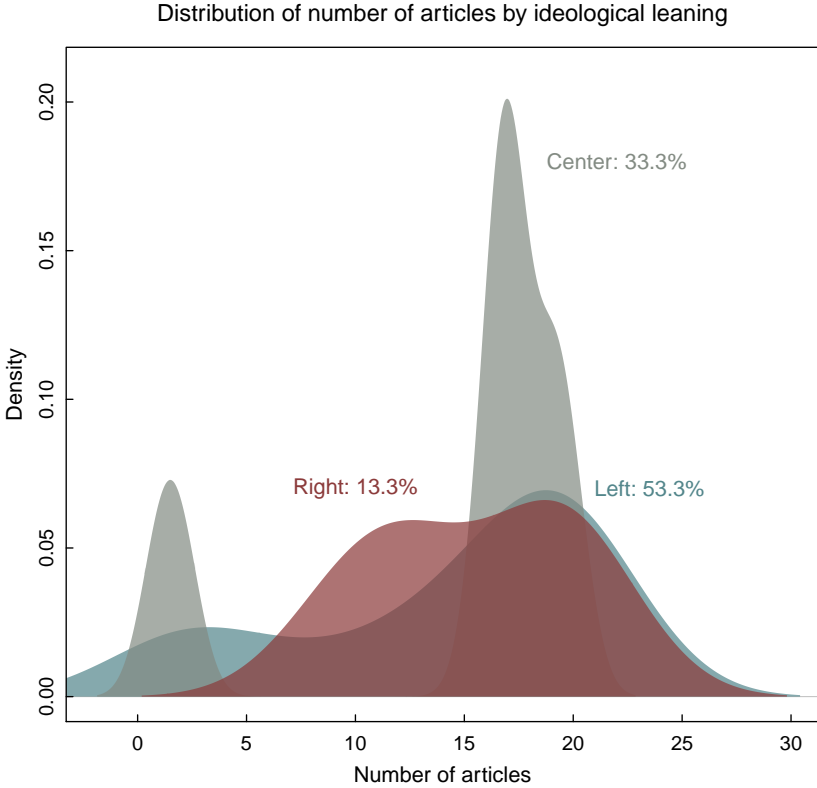
To detect the topics in this corpus of images from media outlets, I extracted an Image Visual Word-Matrix using the BoVW approach described above. I generate the codebook to build such matrix (i.e. the visual words in columns) using the images from *Getty*. This is in order to have a more neutral source of the visual patterns that can be found in the events related to the caravan. Although the coverage might still be biased, the number and

diversity of photographers generating the images ameliorate the concerns. As explained in previous sections, the codebook has 500 visual words.

I conduct a Structural Topic Model with 5 topics and two prevalence covariates: date on which the article was released, and ideology of the newspaper measured with the *All sides* media-bias score. The results are presented in the following section.

Results

Although some topics are different from the *Getty* example due to the initialization process (less topics and different prevalence covariates), the model recovers relevant topics like "crowd", "Darkness", or "Small groups/Individuals."[8] The findings indicate that there is variation in the use of crowds in the images of the caravan across the different media outlets. In Figure 2.10 we see the newspapers along the x-axis, and the mean proportions of the "crowd" topic by news outlet. The darker points highlight some prominent outlets in terms of circulation like the *New York Times* and the *Washington Post*, or in terms of ideological leaning like *MSNBC* or *Breitbart News*.

Is this variance associated with ideology? To study this question, I analyzed the effect of the ideological leaning of the newspapers, the prevalence covariate of the STM, on the generation of the topic "crowd." Figure 2.11 shows the means of this topic by ideological group. Here we can observe that the news outlets with right-leaning biases show significantly higher proportions of this topic than the rest of the groups (all of these differences are positive and reliable). On average, right leaning outlets tend to publish images with 16% more content of the topic "Crowd."

---

[8]The most representative words and images for each topic are presented in the Appendix.

Figure 2.10: Crowd topic by media outlet
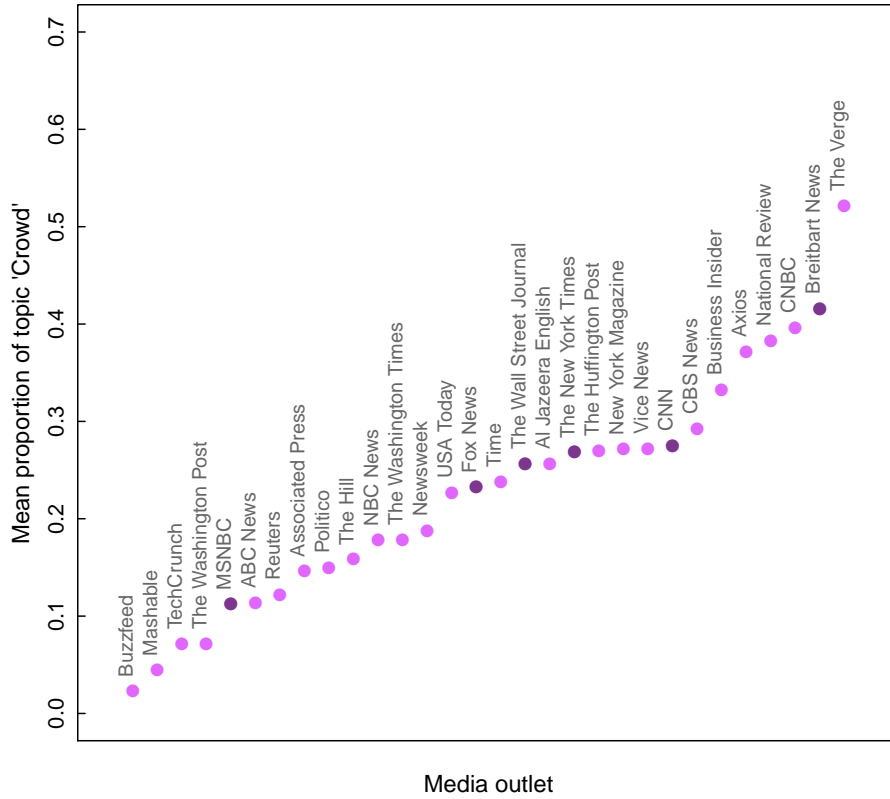
2.4 Practical considerations: strengths, challenges and diagnosis

Throughout this article, I have elaborated on the logic, implementation and applicability of the Bag of Visual Words. However, in order to fully exploit its benefits, it is important to understand its limits and scope, as well as the impact of key decisions during its implementation.

Figure 2.11: Ideological leanings and portrayal of crowds
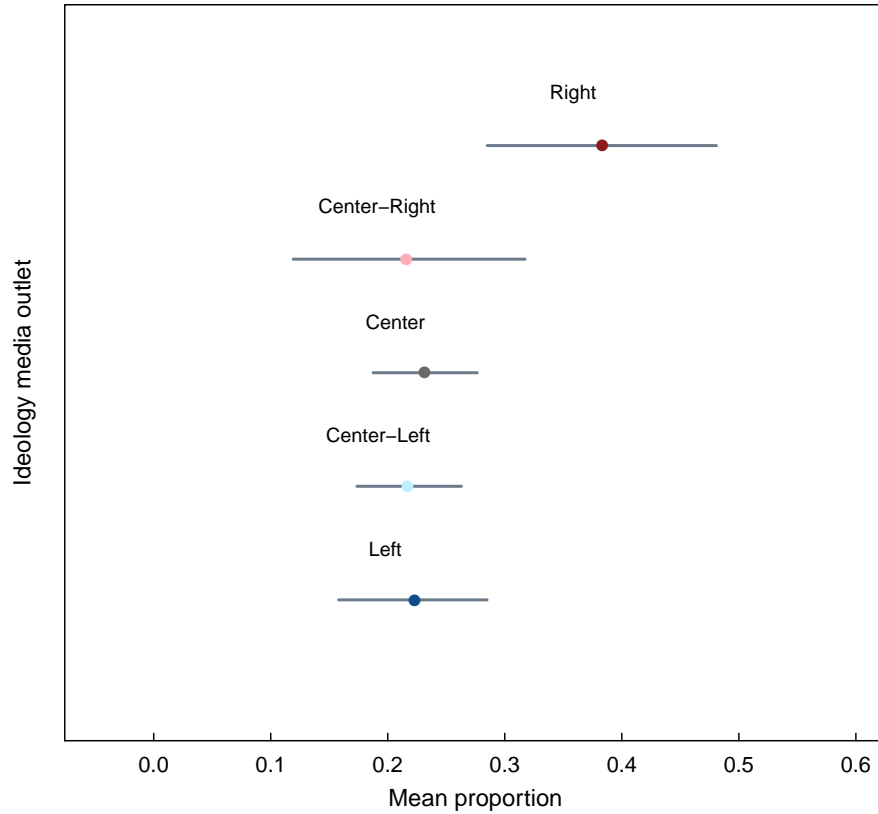
2.4.1 Strengths and weaknesses

A good way to understand the strengths and weaknesses of the BoVW is to compare this technique to other more popular and commonly used tools in the field of computer vision. In particular, there is an increasing number of applications of Convolutional Neural Networks (CNNs) to the analysis of images in political contexts (Cantú Forthcoming; Casas and Williams 2018; Dietrich 2015; Lucas 2018; Won, Steinert-Threlkeld and Joo 2017; Zhang and Pan 2019). CNNs are the state-of-the-art tools for visual recognition and detection. They are models with a directed graph structure composed of *layers* with nodes and connections. The underlying intuition of this structure is that the nodes compute simple tasks (in the context of images, these are related to feature detection and extraction), and the output

information from these tasks is transferred throughout the network and used to predict an outcome of interest. In order to reach these predictions, a process of error minimization takes place just as is in other models well-known to political scientists like linear regression (Krizhevsky, Sutskever and Hinton 2012; LeCun et al. 1998; LeCun, Bengio et al. 1995). This process, however, requires large amounts of labeled data that allows the user to train and test the algorithm.

The sophistication of CNNs allows them to have a high predictive power. However, this sophistication also comes with less parsimonious and transparent mechanisms, and high computational costs and need for training data. In contrast, the BoVW is a method with an intuitive and simple procedure, and potential for applications beyond prediction.

First, as illustrated throughout this article, each step of the BoVW is easy to track and understand. The data reduction process involves clear steps with basic mathematical foundations. Further, the computational costs of using it are low, especially compared to CNNs. Using special infrastructure like several graphics processing units (GPUs) or high performance computing clusters (HPC) is not necessary even when dealing with large pools of images. As a reference, the entire routine of building a BoVW with 15,000 images (of a maximum size of 616×612) and 2,000 words takes approximately 5 hours on a laptop with 4 processors.

Second, the BoVW and the corresponding IVWM that it produces can be used in both supervised and unsupervised methods. For example, researchers can link the rows in the IVWM to specified labels and use classifiers like support vector machine (SVM) or regression trees to conduct out-of sample predictions, or to assess the effect of "visual words" and other covariates on the outcome of interest. However, as I illustrated in this text, the BoVW also has interesting applications to other unsupervised methods like topic models or other clustering or representation learning methods. This relaxes the need of labeled training data, a step that in several applications is particularly hard to fulfill. Beyond the utility of these

methods to detect and measure interesting patterns in the data at hand, they can also be used in the initial steps of a project for exploratory purposes. Knowing and understanding your data is a fundamental step in any study, and the BoVW can facilitate it.

The parsimony and low cost of the BoVW have implications for the quality of its performance in certain applications. In previous sections, I showed that the BoVW can identify relevant topics and discriminate certain objects from others. However, this level of discrimination might not be enough for certain purposes that require finer distinctions and high predictive accuracy and precision. CNNs aim to identify a large number of specific features in each of the different layers (e.g. lines, complex figures, etc.), and this in turn makes the data reduction process more efficient and accurate. Further, in combination with appropriate and sufficient labeled data, these features are associated with more meaningful concepts that the researcher devises. This does not imply that CNNs are able to distinguish and accurately predict abstract and complex phenomena based on the features with which they work (see Chapter 4), but they will perform better in prediction tasks than the BoVW.

Finally, the BoVW is sensitive to decisions that researchers make when defining the parameters at different stages of the process: accuracy of key point detection, number of features to extract, size of the vocabulary, etc. There is a strong need for tools and tests that facilitate the diagnosis and evaluation of the impact of those decisions but for now, the section below discusses some elements to consider when using the BoVW.

2.4.2   Practical considerations

The process of building a BoVW requires certain specifications that are subject to the researcher's needs and criteria. I would like to emphasize that, to the extent possible, substantive knowledge and theoretical insights should guide the definition of some of these parameters. However, the process also involves trial and error runs to correctly tune some of the parameters. Below I present a list of a few practical things to consider.

Detecting key points

There are two important things to consider during the key point detection: 1) the type of salient regions to identify, and 2) the precision of this identification. The first one relates to the definition of what constitutes a salient region. For example, certain applications require an accurate identification of "corners" in an image (e.g. consider a case where the identification of buildings and houses, captured with rectangular shapes, is part of the objective of the study). Others, however, rely more heavily on " blobs" (e.g. emphasis on the classification of texture). This distinction might appear trivial but plays an important role especially in some of the traditional computer science applications focused on classifying basic elements like dogs, flowers, and objects. For more complex content, as in the case of most social sciences applications, the detection of all edges, blobs and corners is in general more fitting. The final definition of a salient region determines the type of detector to use. For example, the FAST and GTTT detectors are used to detect corners in images, while DoG and FAST Hessian focus on the detection of corners, edges and the combination of both.

The second feature, the precision, will have an enormous impact on the number of key points that the detector identifies in each image. The decision should be based on the answer to the question "What features are substantively relevant to the objective of this study?" For example, the FAST Hessian with a low threshold captures even small changes in pixel intensity and therefore yields a large number of key points. The literature suggests a threshold between 300 and 500, but higher thresholds might be more fitting given the specific problem at hand. Consider the picture in Figure 2.2: is it necessary to have key points around clouds in the sky? Or do we exclusively care about the salient features of the girl given that she is the most prominent figure in the image? Figure 2.12 shows the results for another example photo when the Hessian threshold varies. While for some applications the accuracy and precision of the classification depends on all the features found in an image,

others demand a broader definition.

Figure 2.12: Comparison of key point detection outputs with different thresholds

(a) Original image

(b) Threshold = 300

(c) Threshold = 1000

(d) Threshold = 3000
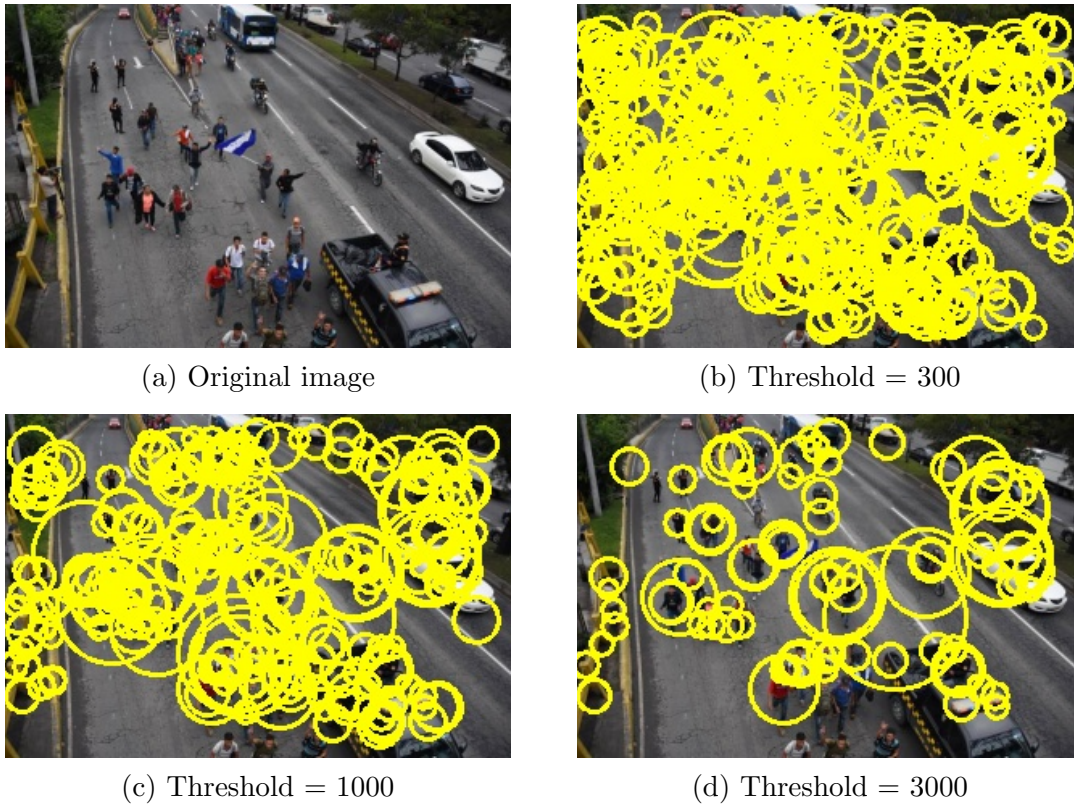
Building a vocabulary

The construction of the visual vocabulary is one of the crucial steps of the process, and one in which several decisions need to be made. First and most importantly, the researcher has to determine the images that will be considered as the basis of the vocabulary. The process outlined throughout this article considers that each of the feature vectors in an image is associated with a visual word: we consider that a word v is present in an image i if after calculating the distances between a given feature vector fv of image i, and the full set of visual words, we observe that the distance between fv and v is the shortest. Thus, it could be the case that a given feature is associated with a visual word that does not properly

represent it if there are no better candidates.[9] Therefore, it is crucial to build a vocabulary with images relevant to the target pool under study. If the vocabulary is built with pictures of flowers, the IMVW matrix that we extract from images of faces is not going to be as helpful as having a more representative corpus.

Another consideration is the number of clusters or "visual words" to extract, and the process to do so. A richer vocabulary has more power to discriminate and distinguish features, but a more parsimonious one focuses less on the details that each visual word is capturing. Once again, the decision to have a more fine-grained vocabulary depends on the substantive motivation of a project: is it relevant to have two visual words with the patch of a hand, one with a dark background and one with a light background, or do we consider it more useful to only have *one* reference for a hand? If the objective is, for example, to distinguish people in pictures the second alternative might be more sensible. In contrast, if we are interested in clustering images based on whether they happen at night or day, then the darkness of the background becomes relevant. Researchers can also rely on tests and statistics providing an "optimal" number of clusters, and use visual inspection to assess the composition and formation of visual words.

Inspect and visualize

Most of the tools designed for visual inspection lack guidance on how to proceed with diagnosis or validation procedures. This, in part, is a result of the complexity and multidimensionality of the data, and the absence of concrete tokens and concepts to consider: it is harder to find a synonym for a patch of an image than for a word. However, images provide an advantage over other types of data: they offer more and better opportunities to visualize information. This helps with the identification of "errors" and "inconsistencies" in the model, and with a more optimal tuning of relevant parameters. The code that accompanies

---

[9]Although this could be improved by using "acceptance thresholds", it is still advisable to build sensible and conceptually coherent vocabularies.

the text covers the construction of visual words as well as their visualization. The visual inspection of these clusters is fundamental to understand some of the patterns that the computer identifies. In some cases, the consistency is obvious and straightforward but in others the clustering process produces puzzling results. For example, a visual word with radically different mini-patches is a symptom of a low number of key points or a small number of clusters. Similarly, one with almost all of the mini-patches from the same image indicates that the clustering is too specific or that the precision of the detector is too high.

Some of the errors or things to change become obvious in a post-BoVW stage. For example, the picture in the left panel of Figure 2.13 tends to have a high percentage of topic "crowd" although it is just a shot of pavement. While a human can easily distinguish that it is not a crowd, the granularity and texture of the pavement resembles that of a big crowd in terms of pixel intensity changes. Similarly, the picture on the right is clustered with the "border/fence" pictures due to the flag behind Donald Trump. The changes in pixel intensity of the stripes of the flag are very similar to those found in pictures of the fence and border. Thus, the deletion of those pictures or the removal of customized "visual words" (like those with pavement) are alternatives that help to improve the extraction of the BoVW and the analysis of the images of interest. I cannot stress enough the importance of visualizing and inspecting the results, not only as a way of detecting inconsistencies, but also as a way of understanding and getting to know the complexity and depth of the data under study.

Finally, it is important to highlight that while these methods are helpful to digest, quantify and classify visual material, they cannot replace the knowledge and expertise of humans when it comes to coding or identifying more complex messages underlying it. Therefore, validation and human involvement in the classification process are crucial steps that should not be underestimated.

Figure 2.13: Visualizing mistakes

(a) High proportion topic of "crowd"

(b) High proportion of topic "border"

2.5   Conclusion and further research

The BoVW is a useful technique that provides researchers with a tool to quantify and digest information as the first step in the process of understanding visual content. The underlying logic is intuitive and the procedure to implement it accessible. Further, it is able to handle and process large pools of data with speed and efficiency. However, as this article shows, there are several aspects that can be improved in order to achieve higher levels of precision and accuracy, as well as consistent and unbiased results.

The BoVW is solely based on pixel intensities, and therefore, all images are converted to gray scale. Although intensities and change in them are capturing a lot of the information regarding the content of a picture, color is another important source of information that should not be ignored in applications like the current one (Vigo et al. 2010). Therefore, researchers should consider the inclusion of "color statistics" as part of the feature vectors of the key points for clustering and comparison purposes. This is designed to improve the predictive power of models by building a richer matrix with more detailed information regarding the color dimension.

Further, the applications of this method to visual framing should be extended to include

text and other relevant information at the news article and outlet levels. In particular, the analysis of whether visual content reinforces, complements or contradicts factual information provided in texts is fundamental for a proper understanding of the political communication process. Finally, the BoVW could also help in addressing questions regarding the differences between sources of images: do content and framing differ between media and other actors like activists or non-profit organizations? Can we learn something about the demand for certain frames from a public opinion perspective?

The BoVW can be used to address a variety of questions in multiple fields: electoral campaigns, social movements, migration flows, media coverage of political figures, etc. Images overcome one of the main challenges when studying events or issues in different countries: the language is universal and can be captured and synthesized with methods like the BoVW. Thus, the comparison of campaigns, protests, and communication strategies between countries becomes more viable. Issues like the way in which leaders in each country visually present to their nations the interactions they have with other leaders, or the different frames of protests about similar topics across countries are examples of questions that deserve attention. However, there are other questions that are relevant in local contexts and within a country such as the differences in visual depictions of actors based on characteristics like race or gender (e.g. media coverage of female and male candidates).

This article addresses a few of the multiple issues regarding image analysis and visual framing, and intends to contribute to a blooming literature focused on the extraction and analysis of information that pictures and videos provide. These are efforts oriented towards achieving a better understanding, a "full picture", of multiple political events and phenomena, and the way in which that information reaches hearts and minds.

Chapter 3

Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data

Computers have increasingly taken over tedious and repetitive tasks previously assigned to humans. From counting words in a file to performing long and complex calculations, the computers' convenience is their capacity to follow a set of instructions in a repetitive way with no fatigue or cognitive bias. Their capacity to perform quickly and reliably allows us to analyze information from a large amount of data such as roll-call votes (McCarty, Poole and Rowenthal 2006; Poole and Rosenthal 1997), congressional floor speeches (Dietrich, Enos and Sen 2018; Dietrich, Hayes and O'Brien 2019), or social media posts (Barberá 2015).

And yet, despite their capacity to analyze textual or numeric information, computers have historically underperformed when classifying visual data. In principle, we would expect computers to, for example, find all images including a specific candidate, if we provide specific rules describing the physical characteristics of the individual—e.g., the form of her nose, the distance of her nose to her eyes, and the size of her forehead. But even when

the computer would strictly follow such instructions, these directives would be insufficient to identify those pictures in which the individual is backwards or wearing sunglasses. In an even subtler dimension, the computer might struggle to identify the individual if the light conditions differ substantially from one picture to another. Of course, we could increase the number of instructions to follow, but the list would be as expansive as the ways in which an individual may appear in a picture. As a result, when classifying images, computers' ability to follow a set of rules appears to actually be its main limitation.

To overcome those challenges, recent developments in computer science propose an alternative approach to classify images and retrieve information from them more efficiently and accurately. Rather than following a set of given rules, computers are now able to identify and transform image features—such as edges, textures, and motion—into a label after being exposed to multiple examples. This process, inspired by the way in which humans learn to digest visual content, allows computers to gradually learn that a particular combination of colors, edges, contours and textures correspond to a given object. They are therefore able to identify and track that object under various conditions.

To address several of the challenges that visual processing implies, this paper introduces to political science a reliable, cost-effective way to classify pictures. In particular, we present Convolutional Neural Networks (CNNs) as an alternative tool for the tedious task of analyzing, coding and classifying large-scale image collections. Our main goal is to provide general guidelines on how CNNs work, and to discuss the challenges and problems that researchers might encounter when using this tool by focusing on a canonic example of data collection and processing. To illustrate this methodology, we apply a CNN to a relevant issue in election science: coding voting results. We present a way to collect the handwritten results registered in the vote tallies from the 2015 federal election in Mexico, captured in a set of comprehensive images. This example allows us to illustrate the characteristics of this method and the intuition behind its performance, providing an exhaustive guide for its implementation while

47

still highlighting the challenges it poses, as well as offering recommendations and alternatives to overcome them.

The use of CNNs can help social scientists to expand the research on topics like the visual tone of campaign coverage (Druckman and Parkin 2005; Lawson and McCann 2004), the portrayals of media outlets regarding immigrants and protests (Casas and Williams 2018; Farris and Mohamed Forthcoming; Torres 2018; Wilmott 2017), and the camera-recorded interactions between police officers and citizens (Makin et al. Forthcoming). Further, analyzing images of documents allows researchers to recover and collect information such as signatures, annotations or signs of alterations to answer questions related to, for example, the decision to continue fighting in a war (Huff 2018), historical political participation (Homola 2018), or electoral fraud (Cantú Forthcoming).

We first introduce what CNNs are and explain each stage in the process. Next, we illustrate the practicability of this tool by applying CNNs to capture the vote results of the the 2015 election in Mexico directly from vote tallies. Finally, we provide a list of challenges and recommendations when applying these tools, as well as a discussion of the limitations and scope of CNNs for certain measurement and classification tasks.

3.1   A Primer on Convolutional Neural Networks (CNNs)

A CNN is a special case of a neural network, which consists of a set of inter-related *nodes*. Each node, or neuron, receives an input, transforms it, and transmits the output to other neurons. The transformation of the output is defined by the neuron's *activation function*, and it represents a simple operation. For instance, the activation function of a neuron might follow a logistic form, which would transform the input $\eta = \boldsymbol{X}\boldsymbol{\beta}$ into a probability output. The resulting operation becomes a *signal* that each other neuron will use as input for its own computation. The relevance of every released signal for the final outcome is determined

48

by the *weights* between the emitting and receiving neurons. Positive weights amplify the signals and highlight their contribution to the output, while negative weights weaken the signal to which they are attached. At the end of the network, there are as many nodes as labels of interest. Each final node delivers the probability for the original input to belong to the specific label. The neural network tries to minimize the errors in predictions by gradually modifying the weights of the preceding nodes (Lucas 2018; Schrodt 2004).

The particularity of CNNs in comparison to other type of neural networks lies in their capacity to process visual inputs. CNNs are organized into layers where each layer is a set of filters, or neurons, that represents specific visual features such as edges, blobs or color combinations. In essence, this is a technique aiming to reduce and condense the high dimensionality of images, and to address the issue of having a large number of ways in which simple concepts or objects can be represented visually. Thus, instead of using pixel intensities as "features" or "covariates" of an output label that in isolation are meaningless, the CNN attempts to summarize the content of an image using meaningful features and patterns.

The more layers a CNN has, the more complex features of the image will be recognized (Qin et al. 2018). To describe the functioning of CNNs, we divide the process into three parts. First, the *preprocessing* stage transforms the image into a format that can be read by the computer. Next, the *feature extraction* stage deconstructs the image into multiple visual components, each of them representing a specific visual feature. Finally, the *classification* stage uses the image's components to classify the image into one of the categories available. Figure 3.1 illustrates these stages and provides a road map for all the figures in this text detailing and illustrating each step. We explain below the logic behind each of the stages, as well as the decisions available to the researcher at each of them.

Figure 3.1: Convolutional Neural Network Structure

| PRE-PROCESSING | FEATURE EXTRACTION | CLASSIFICATION |

Input:
13 x 13

Filters:
8 @ 3 x 3

Feature map:
8 @
5 X 5

Activation layer:
Tanh

Pooling layer:
max

Fully connected neural network

Hidden units:
128

Outputs:
10

Zero padding (Figure 2)

Convolution (Figure 4)

Non-linear activation (Figure 6)

Pooling (Figure 6)
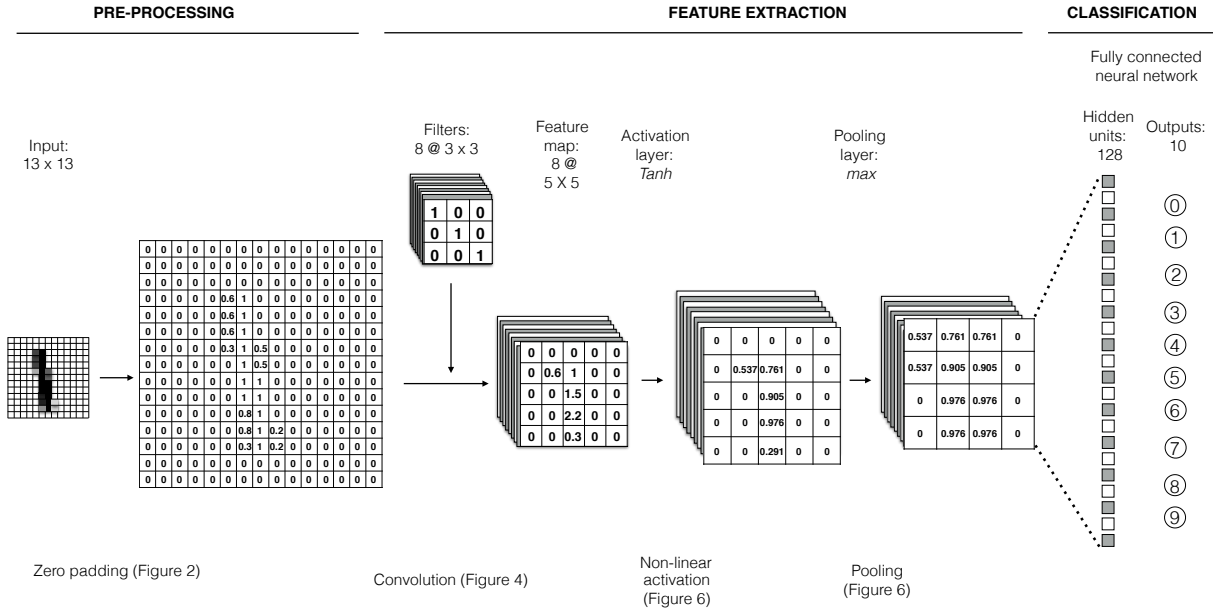
3.1.1 Image Pre-Processing

For the computer to analyze visual information, it is first necessary to represent the image as a numerical array, where each of the entries has a specific pixel value. Figure 3.2(a), for example, illustrates how a handwritten number, 1, of 13 (height) × 13 (width) pixels can be transformed into a matrix of $13 \times 13 = 169$ units, each of them specifying the light intensity of a specific pixel.[1] In the case of a color image, the transformation would produce three matrices of the same size, one for each of the three color channels (red, green, and blue).

The input matrix is the core unit of analysis, and its dimensionality decreases as it passes through the network. To make sure that we can keep even those features that are closer to the borders in the process, we apply *zero padding*. This technique appends a perimeter of

---

[1]The concept of "amount of light" might seem counterintuitive when expressed in mathematical form: In practice, a value of "0" corresponds to a black pixel, while "255" represents a white pixel. To avoid confusion and for illustrative purposes, we take higher numbers in the matrixes presented as higher concentrations of "ink". Therefore, higher numbers correspond to darker pixels.

zeros to the input matrix. The example in Figure 3.2(b) shows a zero padding of $p = 1$, increasing the numerical array to 15×15.

Figure 3.2: Image Pre-processing

(a) Image Transformation

(b) Zero Padding
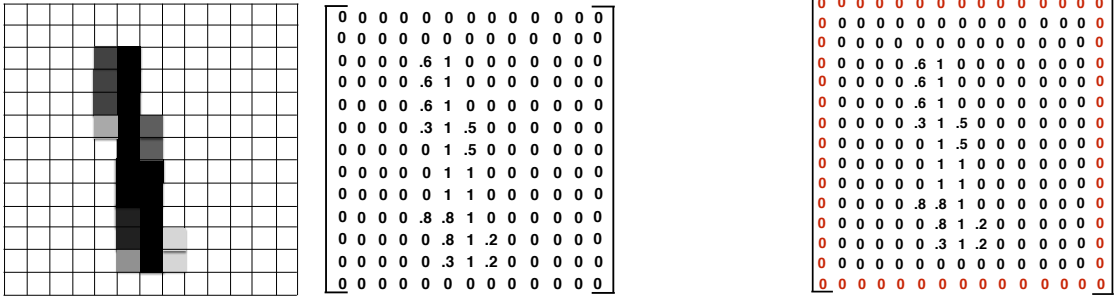
3.1.2 Feature Extraction

Once the image enters into the network, it is decomposed into single components. This decomposition consists of multiplying the input matrix with a set of smaller matrices, called *filters*. The filters slide, or convolute, across the width and height of the input matrix and obtain the dot product of every region of the image. Each filter matrix represents a particular feature: a horizontal line, a corner, a combination of both, etc. The basic intuition behind this process is to place each filter on top of different points of an image and register how similar it is to each particular area. The goal is to detect how prevalent the feature represented in a given filter is in the image. Therefore, a convolutional layer is a collection of filters extracting different information from the same input matrix. The first layer generally has filters that are randomly initialized and that tend to correspond to basic patterns like straight edges (see Figure 3.3). The subsequent layers will learn the types of

Figure 3.3: Examples of filters
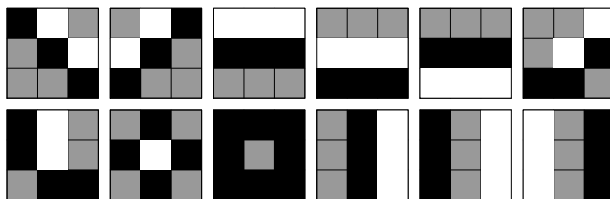
Note: These are simplified examples of filters randomly initialized in the first layer of a CNN. The size is 9, corresponding to their width and height in pixels.

features that are more prominent in the images based on the feature maps.

Before describing how convolution works, we need to specify three parameters of this operation. First, the filter *size* is the product of the width and height of the filters in the layer. This parameter sets out the type of features identified during the convolution. Small filters capture fine-grained details, but they are likely to mix up the relevant information from an image with its noise. On the other hand, large filters look for details of a larger size at the cost of a lower specificity. Second, the filter *stride* defines the number of pixels that the filter will slide through the image. This value must be a positive integer[2]. Finally, the layer *depth* defines the number of filters in the layer. This value, therefore, indicates how many features will be searched for in the layer. The optimal choice for these elements depends on our data and classification goal in every case. Figure 3.3 presents simplified examples of filters of size 9.[3]

We illustrate how a convolution process works in Figure 3.4. In this example, we use a 3×3 filter that slides through the image using a stride of 3. The convolution process then involves computing the dot products of the filter and the values of every equivalent pixel

---

[2]The magnitude of this parameter depends on the size, dimensions and characteristics of the data and CNN. The smaller the stride, the more information that you can keep from the image. For a comparison of model performances using different strides and filter size, see Simonyan and Zisserman (2014).

[3]For the CNN used in this article, the filters in the first layer are initialized randomly using the Glorot uniform method. Also known as Xavier, this initializer draws samples from a uniform distribution: $W \sim \mathcal{U}\left(\frac{-6}{u_{in}+u_{out}}, \frac{6}{u_{in}+u_{out}}\right)$, where u_{in} is the number of input units in the weight tensor, and u_{out} is the number of output units in the weight tensor. In most canned architectures, it is not necessary to define the initialization of these filters.

Figure 3.4: Illustration of the Convolution Stage
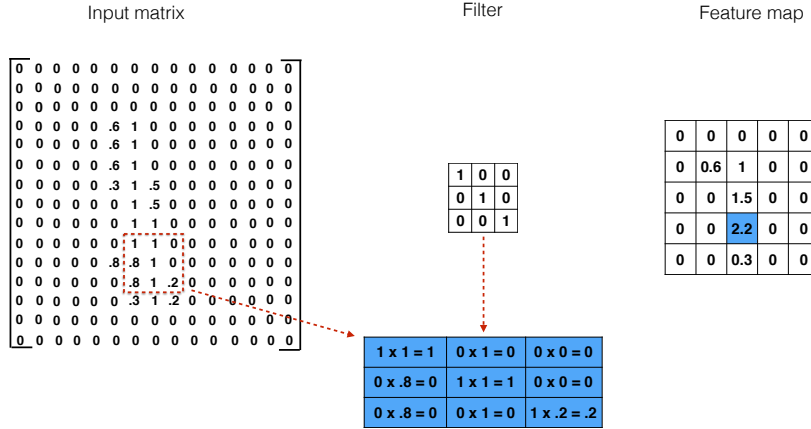
space in the image. In this example, the dot product between the entries of the filter and the input of the highlighted image area is 2.2. The filter then slides three steps to the right and computes again the dot product of its entries. The result of this operation is the *feature map* at the right of Figure 3.4, which shows the image regions with the largest dot products for this filter. The convolution process will create as many feature maps as filters specified in the layer depth, and the size of each feature map is defined by:

$$\frac{(\text{width*length} - \text{filter size} + (2 * \text{zero padding})}{\text{stride} + 1} \tag{3.1}$$

At this point, the feature maps are just a set of linear transformations of the input matrix. This implies that adding more convolutional layers only produces a single linear product, limiting the type of information that the network can extract from the images. Moreover, linear transformations are unlikely to produce smooth gradients, a necessary input for the learning phase described below. To address this problem, it is necessary to include a non-

linear *activation layer*, which performs a fixed mathematical operation upon the input values. Non-linearity allows us to stack multiple layers and extract more information from the image.

We briefly describe three of the most common activation functions, whose shape and function are presented in Figure 3.5. The first one, *Sigmoid*, is simply the inverse of the logistic function. As the figure shows, this function bounds the activation values to the $[0, 1]$ range. The second activation function is *Tanh*, and it can be understood as a linear transformation of *Sigmoid* that zero-centers the outputs and bounds them to the $[-1, 1]$ interval. While both functions are very sensitive to input values closer to 0, their output becomes flat near its boundaries, limiting the network to learn from inputs with either very low or very high activation values. Addressing these issues, the *Rectified Linear Unit* (*ReLU*) is a non-saturated function that keeps the original input value when it is positive and transforms all negative values to 0. *ReLU* usually increases the learning speed of the network, and is now the standard activation function in practice (Nair and Hinton 2010). Again, the choice of any of these or other activation functions depends on the performance of each of them to the specific data in question.

Figure 3.5: Example of Activation Functions
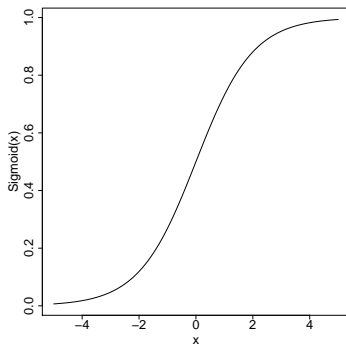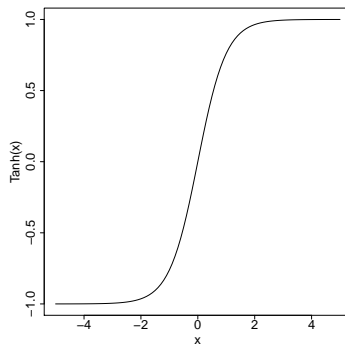
(a) $Sigmoid(x) = \frac{1}{1+e^{-x}}$ (b) $Tanh(x) = \frac{2}{1+e^{-2x}}$ (c)$ReLU(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{otherwise.} \end{cases}$
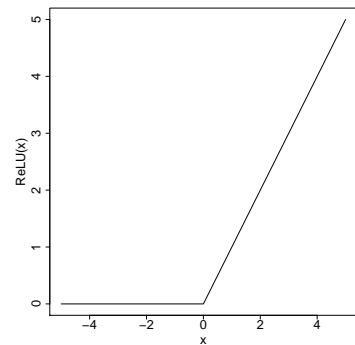
Once the activation map includes non-linear outputs, we proceed to reduce its dimen-

Figure 3.6: Illustration of the non-linear activation and pooling
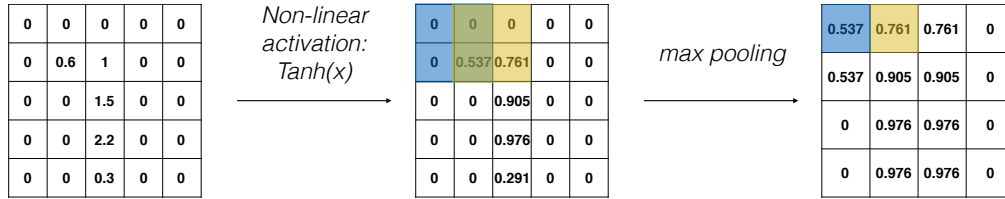
sionality using a *pooling layer*. A pooling layer shrinks the size of the matrix while keeping the most important information in the feature map. The information of a submatrix is summarized differently depending on the type of pooling layer applied. For example, it can get the largest value (*max pooling*), the smallest value (*min pooling*), or the average value (*mean pooling*) of a specific pixel area.

In Figure 3.6, we illustrate first the nonlinear transformation of our feature map using a *Tanh* function. Then, we apply *max-pooling* to keep the largest value from every 2x2 pixel area of the matrix. The reduced matrix generalizes the properties of the image, forcing the CNNs to pay more attention to whether a feature fits in the image and not too much attention to the specific location of that feature.

The process is repeated for each of the filters in the first layer, and each of the resultant feature maps becomes the input for the second convolutional layer. The process repeats all of the steps described above, but the new filters slide through each of the representations to now look for more complex features, such as a combination of lines or edges. Based on the feature maps, the new features in each layer are updated. The more layers we include in the network, the more complex features it is able to extract and learn from the images.

3.1.3 Learning

In this last stage, the network estimates the probabilities that the image belongs to each of the categories provided. It does so by gleaning the feature maps that are more likely to belong to a given label. The feature maps are "flattened" or stacked into a single vector that is then associated with a particular label. The probabilities of belonging to each of the potential outcome categories are estimated by a multinomial logistic function at the last layer of the network. This last layer is a fully connected neural network given that each of the potential outputs depends on each of the potential inputs: for the last layer, the inputs are the elements of the flattened vector. In the example we described above, if we want to identify the image from Figure 3.2(a) as a digit value, the last layer of our network has 10 neurons or outputs, one for every digit from 0 to 9. Each neuron will provide a probability that the image belongs to each digit, and the CNN will attach the label with the highest estimated probability.

The "learning process" of the model is equivalent to how infants learn to recognize object categories. As they are exposed to multiple examples, young children subconsciously assimilate the distinctive properties of an object until they can identify an object without observing all of its features. This knowledge makes us able to, for example, identify a bird from a cat after observing a beak and feathers, even if we cannot see other elements like the paws or the tail. Similarly, CNNs review multiple examples of an object to gradually determine the relevance of its features and how much each of them would help to identify the object across other images. These examples come from a *training sample.*

The technical name of the learning process is *backpropagation* (Rumelhart, Hinton and Williams 1988). Its goal is to decrease the error of the model's predictions by gradually calibrating the weights among the neurons in the network. The core of this learning process should be familiar to social scientists: it is an optimization process that aims to minimize the squared error over all of the training examples available. Minimizing the error, then,

consists of finding the points in the multi-dimensional plane where the derivative of the error function is zero.

This optimization process is not trivial when we are in a high-dimensional space composed of multiple features and a high number of potential functional forms. The CNN addresses this problem by focusing on how fast the error changes as we alter the weight of an individual connection within the layer. The estimation of this change, or *gradient descent*, allows us to figure out the steepest path that leads to the bottom of the error function. In this way the CNN tracks how the error function moves, and it optimizes the estimation of the weights and connections. We present more technical and mathematical descriptions of this process in the Appendix.

As we mentioned above, the learning process involves a training phase. To start the training stage, we need to specify a few practical features. First, it is necessary to define the number of *epochs*, or the number of times that all training examples pass through the network once. Since the gradient descent is an iterative process, we need several epochs to optimally fit the weights to the model. Too many epochs, however, are likely to produce overfitting (see Subsection 3.3.1). Unfortunately, there is not an "optimal" number of epochs with which to train a model, and its final setting depends on the specific characteristics of the training database and a close tracking by the researcher.

Second, we need to define the *batch size*, or the number of training images for the gradient descent to update the weights in the network. We can set the batch size to the total number of training images. In this case, the model updates its parameters only once per epoch. This modality requires accumulating the prediction errors across all of the images in the training set, a task that can demand a lot of computational memory. It also produces a static error surface, where the gradient descent is likely to get stuck in a saddle point or local minimum. We can also set the batch size to one, where the model updates its weights for each training example. This modality produces a dynamic error surface, decreasing the risk

57

of getting stuck on a flat region. However, updating the model with every example produces a very noisy signal, making the gradient descent jump around. The sweet spot between both extremes splits the training set into *mini-batches*, allowing the model to update its parameters several times during an epoch. Research on the topic suggests setting the batch size to 32, balancing the noisy training problems of smaller batches as well as the slower convergence of larger batches (Bengio 2012; Masters and Luschi 2018). The batch size then determines the *iterations*, or the number of times the weights of the model are updated after an epoch.

Finally, we need to set the *learning rate*, or the speed at which the gradient descent travels along the downward slope. This rate specifies the degree to which the CNN will update its weights after every iteration. A large learning rate will produce large-scale updates on the network weights, jumping around the function and overshooting its minimum. In contrast, a very small learning rate is more likely to find a local minimum, but it will take a long time to converge. It is suggested, then, to start with a large learning rate and gradually decrease it at every iteration (Buduma 2017). Finding the optimal learning rate for every parameter can be a demanding task. Fortunately, there is a variety of optimizers that adaptively tune the learning rates for all parameters in the model.[4]

3.1.4   Software

There are multiple sources of software to design and run a Convolutional Neural Network. All the analyses in this paper, including the manipulation and processing of visual material, were conducted using Python 3, and within it, `OpenCV` and `Keras` (with a `TensorFlow` backend). Keras is a neural networks API written in Python that supports models like CNNs and recurrent networks and allows a very accessible, efficient and user-friendly interaction with

---

[4]For a very helpful comparison of the most common optimizers, see `https://cs231n.github.io/neural-networks-3/`.

libraries like `TensorFlow`, `CNTK` and `Theano`. These are libraries that allow the design, training and implementation of machine learning models. However, there are other tools that facilitate the design of the architecture of a CNN, and that also allow researchers to take advantage of pre-trained models. These include, but are not limited to, *Amazon AWS Machine Learning Training*, *Google Cloud AutoML* or *Google Cloud Machine Learning Engine*.

3.2   Implementation

3.2.1   Coding Electoral Results from Vote Tallies

We illustrate the use of CNNs for coding factual content from images. Examples of this type of information involve handwritten notes in treaties and documents, signatures, annotations, or vote counts. The human transcription of such material is historically either delegated to multiple coders or ignored given the high costs that its processing implies.

In particular, we apply CNNs to code the vote results for Mexico's 2015 federal election. This example demonstrates the benefits of visual analysis not only to scholars but also to policy practitioners and election officials looking for a cost-efficient way to speed up the vote tabulation process. For the specific case of Mexico, automatic capturing of the electoral results can decrease the rate of tallies with accidental errors when adding up the votes, which occurs in almost two out of five tallies in the country (Challú, Seira and Simpser 2018). Moreover, the results of Mexico's federal elections are reported on election night after poll workers tabulate the results at every polling station, deliver the information to the district council, and election officials capture the information. In principle, the time length for publicizing the results of a given tally depends on the speed of the poll workers to count the votes and the commuting distance between the polling station to the district council. Nevertheless, the delay in announcing the results is a common complaint of political

parties, pundits, and voters. In 2006, for example, the frontrunner candidate discredited the impartiality of the report of preliminary results (PREP), claiming that the results were sorted to put his rival ahead early in the vote count (Tello 2007). Therefore, the application can help enhance increased public perceptions of electoral integrity in the country.

Unfortunately, this problem is not exclusive to the Mexican case. From Florida to Arizona, and from Haiti to Argentina, complaints about delays when announcing the results suggest the importance to register the results directly from the polling stations, and doing so in a fast and accurate way.[5] Further extensions of this project include the exportation of the methodology we propose to an app or website to code the vote results from any image of the tally taken at the polling station.

Our task starts with the extraction of the handwritten numbers of the vote results in the tally for each polling station. Then, we identify each number in the picture of the tally, and associate it with a specific digit using a CNN approach. Finally, we register that information and build a dataset that we subsequently validate with official information from the election.

Figure 3.7 shows an example of one of the tallies under analysis.[6] We compiled a dataset

---

[5] *The New York Times,* "Races in Arizona Still Hang in the Balance." November 9, 2012. (`http://www.nytimes.com/2012/11/10/us/politics/arizona-races-still-hang-in-the-balance-over-uncounted-votes.html`); *Los Angeles Times,* "Arizona ballots finally counted – and Latinos ask, Why so long?" November 21, 2012. (`http://articles.latimes.com/2012/nov/21/nation/la-na-nn-arizona-latinos-voting-20121121`); and *Tucson Sentinel* "Why is Arizona still counting votes?" November 21, 2012. (`http://www.tucsonsentinel.com/local/report/112012_az_vote_count/why-arizona-still-counting-votes/`); *The New York Times*, "Vote Count Confirms Obama Win in Florida." November 10, 2012. (`http://www.nytimes.com/2012/11/11/us/politics/florida-to-address-delays-as-it-confirms-obama-victory.html`); *National Public Radio*, "Four Days Later, Florida Declares For Obama." November 10, 2012. (`http://www.npr.org/sections/thetwo-way/2012/11/10/164859656/florida-finishes-counting-obama-wins`); *BBC*, "Haiti starts counting votes in long-delayed election." November 21, 2016. (`http://www.bbc.com/news/world-latin-america-38042585`); *Reuters*, "Haiti police clash with demonstrators ahead of election results." November 22, 2016. (`https://www.reuters.com/article/us-haiti-election/haiti-police-clash-with-demonstrators-ahead-of-election-results-idUSKBN13I05K`); *Clarín*, "Elecciones PASO 2017: Cristina Kirchner denunciará la "trampa electoral" del Gobierno y apuntará a todos los votos peronistas." August 14, 2017. (`https://www.clarin.com/politica/elecciones-paso-2017-cristina-kirchner-denunciara-trampa-electoral-gobierno-apuntara-votos-peronistas_0_SJghMNJ_Z.html`).

[6] For the purposes of our example, we only show the center panel of the full tally. The original full tally contains a horizontal panel composed of three sheets: the first one with information about the polling station,

60

with 104,979 images of tallies. All ballots have the same content, structure and format. One of the few differences is the number of parties competing in each district, so therefore the number of rows with handwritten digits range between 13 and 15.

In order to focus on the relevant part of the image that contains the numbers, first we extract the table with numbers from each tally. Because the alignment and orientation might differ from image to image, we decided to develop a function that identifies the coordinates of three focal points of the tally: the yellow banner at the top of the page, the bright pink rectangle at the bottom left of the tally, and the pink circle below the table. The coordinates of these elements, shown inside red rectangles in the first element of Figure 3.7, allow us to identify the bottom, top and left lines of the table containing the digits. The green dashed lines and yellow area in the second element of the diagram illustrates this process. Once we isolate the table, we divide it into $3 \times$ *the number of parties/candidates in the district* cells. We then cut and save each cell under the assumption that it contains a digit.[7]

The architecture of the CNN for this problem replicates the one described in Figure 3.8: two convolutional and pooling layers interleaved, followed by two fully connected layers and a terminal *softmax* (the multinomial function that allows having probabilities that sum up to 1). Because the tallies are real world data with flaws and mistakes, we have two alternatives when training this network. The first one involves training the model from scratch, using only the data extracted from the tallies. This approach, however, requires a lot of time for the network to learn the image features and a large number of observations to avoid overfitting the model. The second alternative is to use *transfer learning*, which involves using an existing trained model and tuning it to the new database. Transfer learning allows the model to adapt information previously learned to reach an acceptable accuracy rate with

the second one with the tabulation of the votes per party, and the third one with relevant signatures from party representatives and polling station authorities.

[7]This, however is not fulfilled in some cases. Although polling staff is supposed to fill all cells and use leading zeros for 1 and 2-digit numbers, or parties with no support, several ballots have empty cells.

Figure 3.7: Example of the image of a tally
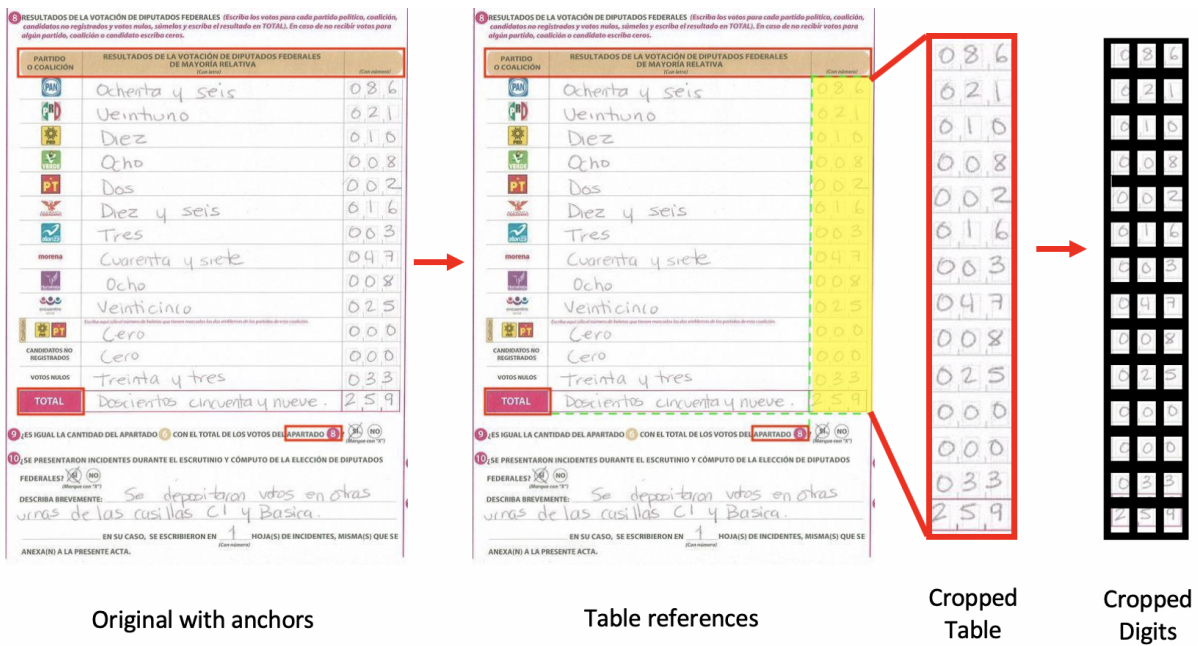
fewer data and in a shorter span of time.

When using transfer learning, the main decision involves how much information from the original model we want to import to the new CNN. This choice depends on the size and similarity of the original database with respect to the new data. It is recommended[8] to only train the linear classifier at the end of the network when the new dataset is small and very similar to the original model. In contrast, when the new dataset is large enough and different from the data in the original model, it is convenient to train the entire network again, using only the weights from the original model as initializing values for the training. In the case of our database, we pre-train our model using the Modified National Institute of Standards and Technology's (MNIST) database of handwritten digits (LeCun et al. 1989). This is one of the seminal databases on visual recognition, and it includes a training set of 70,000 examples

---

[8]Ananthram, Aditya. 2018. "Deep Learning for Beginners Using Transfer Learning in Keras." Towards Data Science. https://towardsdatascience.com/keras-transfer-learning-for-beginners-6c9b8b7143e (accessed March 21, 2019).

Figure 3.8: Network Architecture
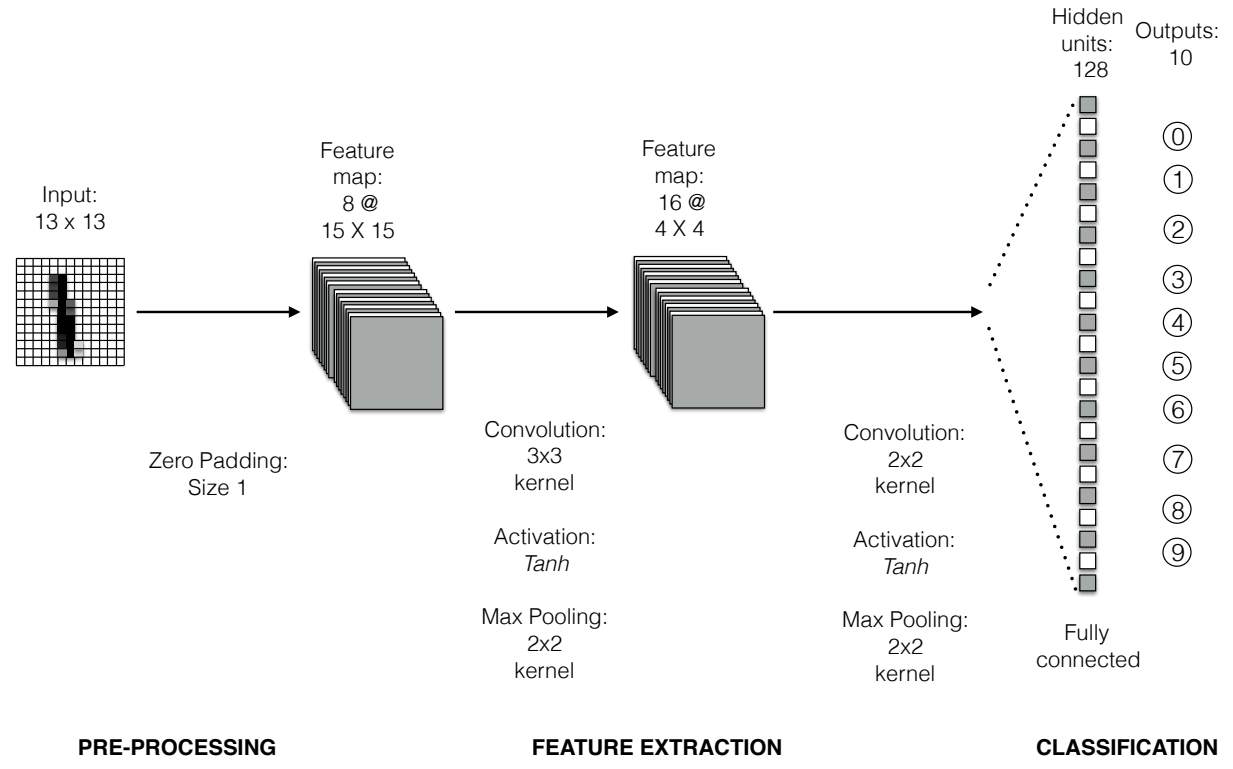
Notes: Figure 3.8 illustrates the CNN structure applied to identify digit numbers. The inputs of the images consist of numerical arrays of 28 (height) $\times$ 28 (width) pixel values. The network contains two convoluted layers of 16 and 32 filters, respectively.

of digits written by about 250 writers. Since our data is similar to the MNIST database, we just freeze the first convolutional layer and allow training for the rest of the components in the network.

In a first step, we train the CNN on 60,000 of the MNIST digits,[9] while the rest is used for testing. Recall that this step *does not* include any digits from the tallies. We simply replicate the textbook example of recognizing handwritten digits using an exceptionally clean dataset and we reach high levels of accuracy (of about 98.3%). This exercise helps as a validation that the structure of the CNN is able to identify the digits from the MNIST data accurately.

[9]For the purposes of this article, we use the Adam optimizer.

An interesting finding here is that the most common mis-classification is labeling a true 9 as a 4. This is not surprising given the commonalities between these two digits when the upper part of the latter is closed.

However, the digits in the MNIST dataset are perfectly centered white digits on plain black backgrounds, without stains, blobs, or inconsistencies. Therefore, this dataset is useful for teaching the CNN what the most important and core features of the digits are, but less flexible in adapting to "muddier" data like the digits in the tallies. The digits from the tallies are very different from those in MNIST: they are surrounded by stains, pencil marks, guiding lines from the tally, the quality of the alignment is lower, etc. Thus, beyond the core characteristics of the digits, we also want our CNN to learn other characteristics of the data under analysis so we can accurately classify them. For that purpose, we use the weights from this MNIST model and set them as the output of the first layer of our main CNN.

Our new training sample for the subsequent layers of the model consists of 26,271 labeled digits from our tallies, and a testing sample of 2,616 digits.[10] To address the quality issues of the digits in our tallies we implement *data augmentation*. This technique creates random variations of the existing training images by, for example, flipping, flopping, rotating, zooming them out, or combining all of these alternatives (Chatfield et al. 2014). The random transformations will force the model to pay less attention to the specific location of a feature on an image and instead grasp its relationship to other image features. For our example, every time a training image passes through the network it will be (1) rotated within a [-15,15] degrees range and (2) zoomed in or out no more than 20% of the original image size.

The accuracy of this model on the testing dataset with *actual* digits from the tallies is 96.46%. The confusion matrix is presented in Table 3.1. Notice that the number of zeros in our sample is considerably larger than the rest of the digits. This is a result of the format

---

[10]The number of images in the training and validity samples represent fewer than 0.5% of the images in our dataset.

Table 3.1: Confusion matrix for the digits in the tallies

	0	1	2	3	4	5	6	7	8	9
0	1131	7	4	0	0	1	3	2	4	0
1	8	283	4	0	5	0	3	1	0	1
2	8	2	159	3	2	3	1	3	3	0
3	2	1	3	137	0	4	0	0	0	0
4	5	1	2	1	154	0	0	1	0	6
5	3	0	1	5	0	155	0	0	0	0
6	5	0	0	0	0	0	137	0	4	0
7	2	0	1	1	1	0	0	124	0	4
8	3	0	6	2	0	2	1	0	117	2
9	2	2	0	0	2	0	0	0	3	94

of the tallies that require individuals to write the results per party using a 3-digit format. The tallies register the votes at the smallest unit, the polling station. Therefore, the number of voters they receive is relatively low with a mean of 264 and a maximum of 1,569. As a consequence, the ciphers registered in the tallies tend to be small, especially for the smaller and newer parties in each race, resulting in a high number of leading zeros.

Prediction of electoral results and trouble-shooting

Figure 3.9 shows some of the classifications that we perform on the tallies with the CNN. As we can observe, the CNN reach accurate predictions for most of the numbers. However, it also makes mistakes. The reasons behind the mistakes are various and include noisy images, inaccurate "extraction" of the digit when it is incorrectly registered, or ambiguous handwriting. In some cases, we even detected disparities between the official results and the tallies, which complicates the validation process. The post-classification process is crucial to understand the sources of errors, identify problematic cases, and conduct parameter tuning. For example, the right panel of Figure 3.9 shows that inaccurate recording of numbers (e.g. written outside the guiding box) generates some errors: a cropped 7 is classified as a 1, or a 6 as a 5. Further, the intensity of the ink also impacts the accuracy. If it is almost illegible (as

in the first cell of the "Inaccurate predictions" panel), then the model pools other elements of the image like the background square. Other mistakes, as in the case of the 4 classified as a 9, are due to handwriting styles and shared features between numbers. Some of these cases would still require human intervention and evaluation. Our goal, therefore, is not about eliminating human intervention in the process, but to reduce it as much as possible, and limit it to the most crucial or controversial decisions.

Figure 3.9: Examples of predictions
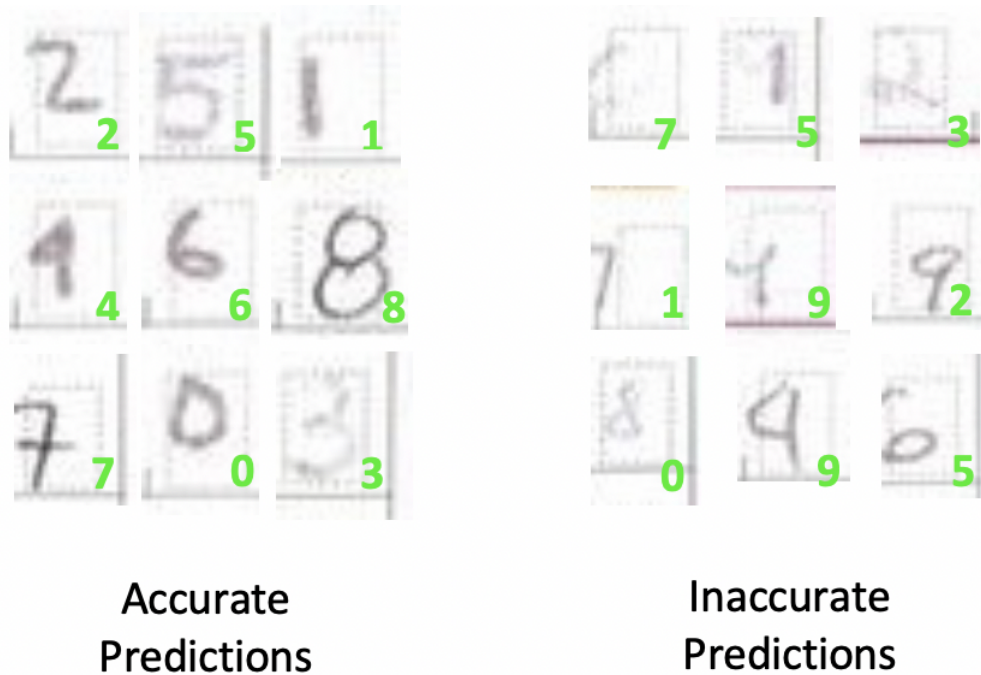
Further, it is important to remember that these errors are at the digit level while the purpose of the CNN is the extraction and registration of vote counts. In that regard, and if we assume that part of the total error is uncorrelated with the outcome, we conduct an analysis of the *bias* of vote counts and proportions per party in one electoral district of Mexico City (District #15). We chose this district because the quality of the scanning was high, and this allows us to fully focus on the performance of the model. Furthermore, focusing on

one district allows us to validate the classification of the CNN, and also to conduct a more careful analysis of the sources of errors and misclassification.

Figure 3.10 presents the comparison of detected and real vote counts in the tallies of District 15. Each point in the plot represents the vote count of each of the parties (including null votes, non-registered candidates, and coalitions) in each of the tallies. We also added to the plot information about the "quality" of the predictions of the digits. Recall that the last layer of the CNN, the *softmax* layer, outputs a list with the probabilities that each input digit has of belonging to each of the 10 possible outcomes (0-9). To classify the number, we take the category with the highest probability of the list. For most of these numbers, the maximum probabilities are pretty high (above 0.99). However, in cases where the number is ambiguous, or the model does not have enough information (e.g. the digits in the tally are not legible), then the "guesses" that the CNN makes are less likely to be accurate. Therefore, we created an indicator for each vote count registered in all tallies. We flag those digits with a probability lower than 0.7, and label each vote count (composed of three digits) as "moderate quality" when it includes at least one flagged digit.

The gray triangles in Figure 3.10 show the vote counts in the tallies identified as "moderate quality". The blue circles show the "high quality" ones. The shapes are weighted by the number of observations in each point. The dashed line is the 45 degree line close to which we expect to see most observations. We indeed observe a dense distribution of a large number of observations along the red line. This is especially true for the high quality tallies: very few deviate from the line. The "moderate quality" observations show greater deviations, but these do not follow a pattern that would suggest a systematic bias.

While these issues raise some concerns and demand further investigation and model tuning, they do not show any particular bias. However, to assess the implication of these deviations in the final vote distribution, we compared the proportions in vote counts between the official information and our predictions. This helps with a more substantive interpreta-

Figure 3.10: Number of votes registered in tallies: Official vs. Predicted

District 15 (Mexico City)

Quality of tally based on mean probabilities
● High (>0.8)   ▲ Low (<= 0.8)
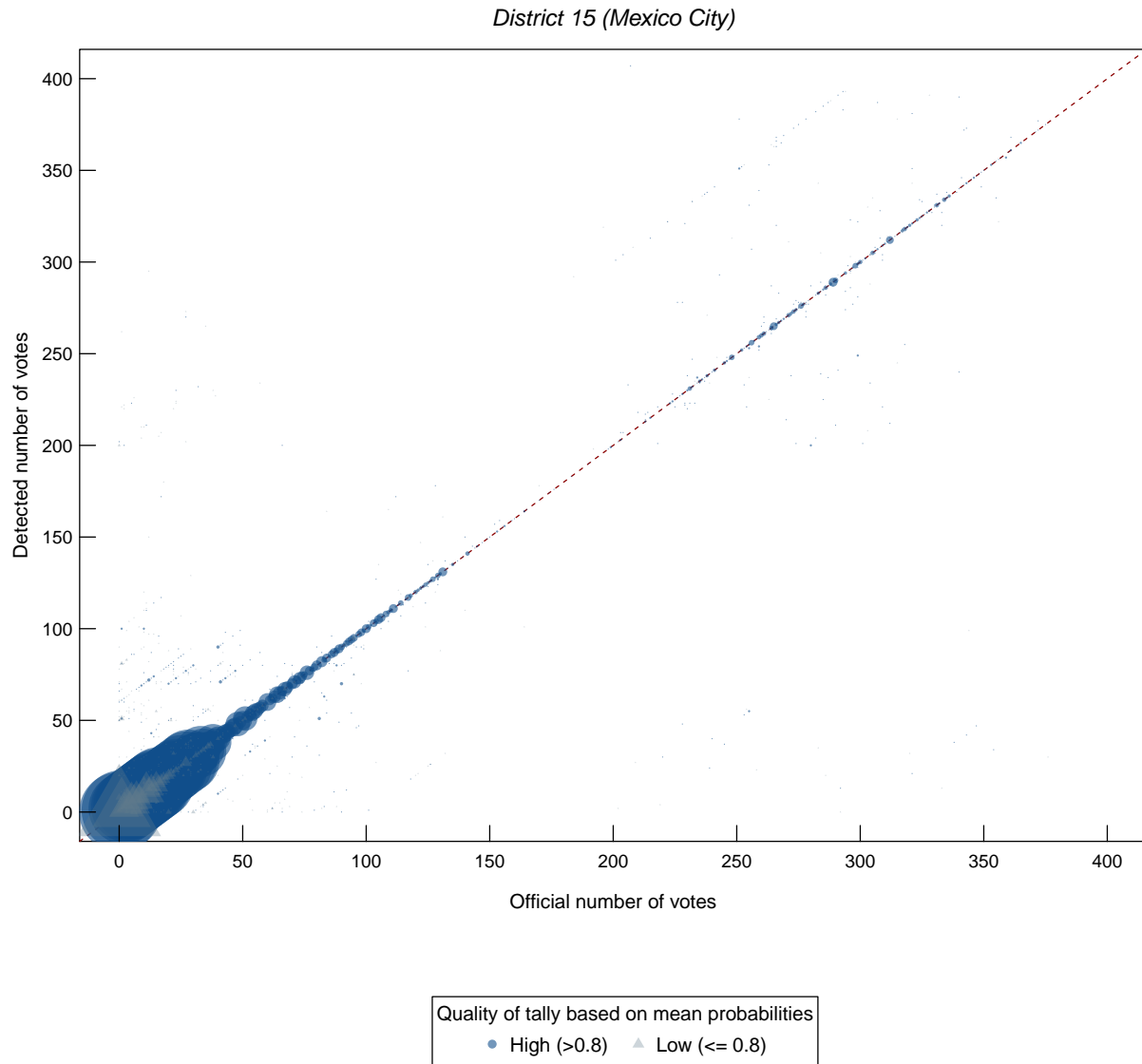
tion of how accurate the results are, and can provide evidence regarding the randomness of the error. Each of the dashed lines in Figure 3.11 represents a party competing in District 15. The symbols indicate the proportion of the votes that each party achieved in the district according to different sources: the official results (circles), and the results using our CNN on a) all tallies (crosses) and only "high quality" tallies (triangles). As the plot shows, the CNN

recovers similar proportions to those of the official results and this performance improves when using high quality tallies. Overall, the method is able to correctly identify the ranking and magnitude of vote counts. This illustrates the applicability of CNNs and their power for data collection tasks.

Figure 3.11: Vote proportions by party in District 15: Official vs. Predicted
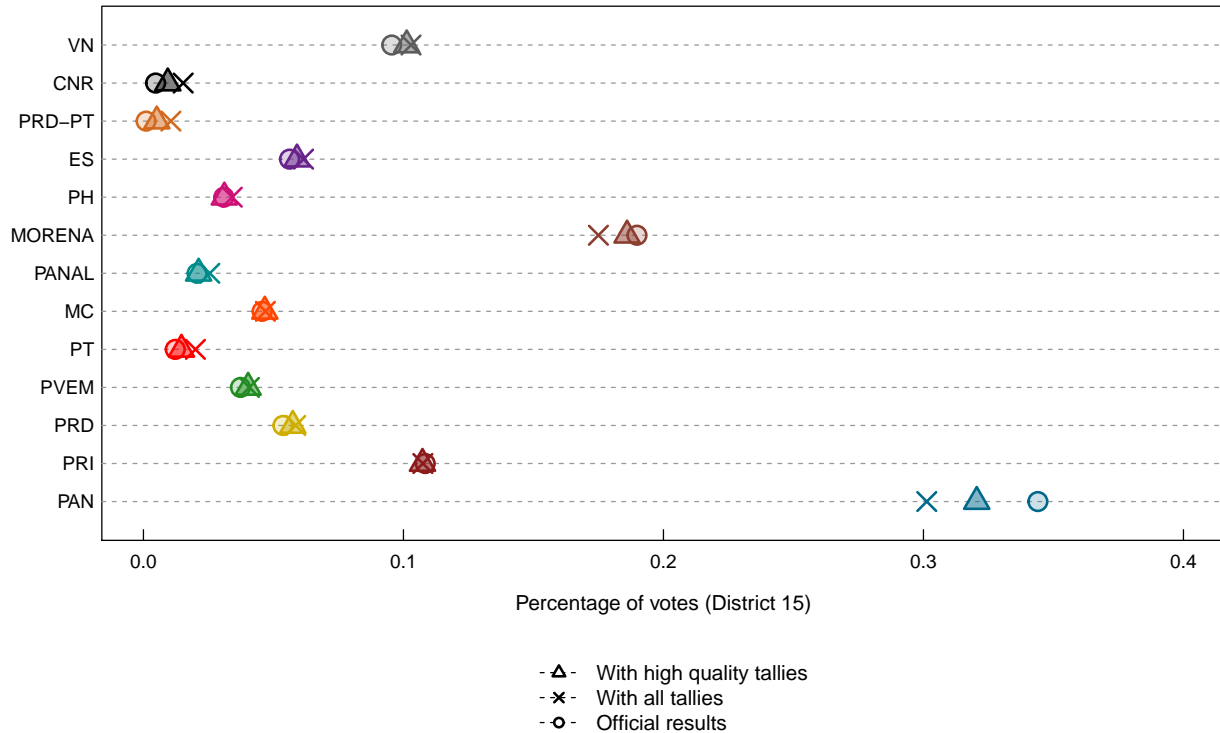
3.3 Suggestions and warnings

Just as any other method or tool designed to predict outcomes, CNNs face challenges and limitations. In some cases, these roadblocks require additional steps to ensure the quality of the results. In other cases, they demonstrate the limits of CNNs in performing complex tasks. In this section we provide a list of both practical and technical issues to consider when training and running a CNN for classification purposes. Further, we also discuss some of the

shortcomings of CNNs with respect to their scope, interpretability and validation.

3.3.1 Recommendations

Prevent overfitting

Similar to any human-performed task, prediction accuracy for a CNN model comes only with practice. Every time an image passes through the network, the model reduces its classification error by calibrating the importance it gives to each filter. The more opportunities for the model to review an image, the more accurate it becomes at classifying the examples in the training set.

Thus, training the model for too many iterations will eventually lead to overfitting. This occurs when the model "memorizes" the features of every training image in a very detailed way, decreasing its ability to generalize its predictions outside the training examples. A common approach to identify this problem involves tracking the performance of the model after every epoch (recall that an epoch is a measure of the number of times all of the training images are used *once* to train the model) and comparing the loss values in the training and validity sets. In principle, the reported loss values for both sets should decrease with the number of epochs. Overfitting takes place when the validity loss starts increasing even when the training loss continues decreasing. This situation dictates the moment to stop training the model and evaluate its overall performance on the validity set.

We suggest three ways to prevent overfitting while training the model. The first one is to increase the number of training images. Having a larger training set allows the model to focus not on the random fluctuations of a few examples, but rather on the visual features that appear repeatedly across the images. In other words, it forces the model to look for patterns that are more likely to appear outside the training set. When we cannot collect new training examples, the second suggestion involves expanding the training set with data

augmentation as argued in Section 3.2.1. A final solution involves a technique called dropout (Srivastava et al. 2014). As its name suggests, this technique "drops out" a random set of neuron activations before being transferred to the next layer. By ignoring some information units during a forward or backward pass, we increase the opportunities for the model to learn more robust features that activate multiple neurons. Dropout also extends the number of training iterations required to overfit the model.

Optimize your training set

Active learning We recommend picking the most useful instances of each class to train the model (Settles 2009). This is particularly convenient when obtaining and labeling training examples is a difficult, time-consuming, or expensive task. Selecting those examples should be based on two goals: informativeness—or how much the instances help the classifier to improve its performance—and representativeness—or how well the instances represent the overall input patterns of the entire dataset. Both are rarely achieved simultaneously, and researchers often need to choose which one to prioritize at the cost of the other (Huang, Jin and Zhou 2014).

Class balance It is also useful to make sure that all classes in the training set are represented by a similar number of examples (Buda, Maki and Mazurowski 2018). Class balance prevents skewing the predictions of the model toward the label with more training instances (Japkowicz and Stepehn 2002). This is a recurrent issue in situations where the positive cases represent a minority of all cases, such as locating oil-spills (Kubat, Holte and Matwin 1998) or identifying fraudulent bank operations (Chan and Stolf 1998).

Image cleaning Another risk when training a model is that it may learn visual features that are alien to those defining the categories of interest. To prevent this problem, we suggest

pre-processing the images to make sure they appear as similar as possible. This step may require modifying and cropping irrelevant parts of the image.

"Denoising" the images involves multiple procedures such as RGB conversion, histogram equalization, and normalization. RGB conversion converts a color image to a grayscale one and reduces its dimensionality. This alternative can be useful when, for example, the background color of an image correlates with each output label. Histogram equalization improves the contrast in images. It accomplishes this by stretching out the intensity range of the image, increasing the local contrast and enhancing the definitions of edges in each region of a picture. Normalization scales all of the images into the same pixel range.

It is also possible to homogenize the data of every training batch. In this case, batch normalization transforms the outputs of the convolutional layers to parameters with zero mean/unit variance, allowing the layer activations to be appropriately handled by any optimization method (Ioffe and Szagedy 2015). This technique keeps the network from focusing on outlying activations that decelerate its learning.

Validate and check the results

An insightful way to improve the model is to review the misclassified images in the test set. This is a practical way of identifying what types of images are most confusing for the model. In our case, this exercise allowed us to find the problems of our model in those instances where the digit-number images include some of the printed tally features. We also found out that the model is less accurate when identifying those numbers that show a soft pressure in the handwriting. These insights will help us include more training images with these characteristics in the next stage of our project.

Further, as in other fields involving prediction and measurement, validation is key in achieving better, stronger results. Validating the results will reveal potential sources of errors and provide information about the model fit. We also suggest another type of validation

72

that involves the assessment and visualization of the components of CNNs. Zeiler and Fergus (2014) provide a series of tools to identify and visualize the most relevant features driving the predictions (a process analogous to finding the most reliable and important coefficients in a regression). These tools provide information not only about the mechanisms behind the predictions, but also about the structure and composition of the data.

3.3.2   Warnings: The limits of CNNs and deep learning

Throughout this article, we present the functioning and components of CNNs, as well as an illustration of their applicability to data collection for social science purposes. The examples show that CNNs are powerful tools to automatize the extraction of information from large pools of images in an efficient, fast and reliable way. However, we have also llustrated some of the challenges and complications that arise even in relatively straightforward and textbook cases, such as handwriting detection.

While CNNs can classify image-based content in an efficient way, they can still make incorrect predictions, even in instances where the target of classification, such as a digit number, has very simple features. There are multiple sources of these errors. Researchers can ameliorate the most important technical issues using pre-processing techniques, parameter tuning, and other computational resources, as outlined in the previous section. However, other sources of errors, such as the ability to considerate context, or sensitivity to moderate changes and alterations, are related to the scope of the CNN and its limitations to complete certain tasks.

The literature has identified limitations of CNNs for solving complex tasks. First, unlike humans, CNNs do not account for the pose and orientation of objects in a picture to reach a prediction. Ideally, we would like to emulate how humans process visual content: identify an object regardless of its size or whether it is rotated. However, CNNs focus only on routing the pixel information to the neurons in charge of detecting features without adding

any information about their relative position and orientation. Similarly, the neurons ignore information about the location and surroundings of those features. For example, if a CNN is in charge of detecting faces and it identifies features associated with eyes, a nose and a mouth, then it does not matter if the eyes are below the nose and above the mouth, as in an abstract puzzle representation; the CNN would still predict a face in the image. The implication of this process is that without a comprehensive and extensive training dataset, the classification of pictures becomes inaccurate and subject to error, even in cases involving simple tasks (Sabour, Frosst and Hinton 2017).

Another criticism of CNNs lies in their lack of uncertainty measures. Unlike traditional models like regression, these tools do not yield quantities like standard errors that aid with inferences or assessments of confidence. This problem is mainly due to the predictive nature of CNNs. Further, while the last layers of the CNN provide the "probabilities" of an image belonging to a certain class, these quantities should be used with caution. For instance, Nguyen, Yosinski and Clune (2015) show how imperceptible changes to an image can cause drastic changes in the outcome probabilities. This implies that the likelihood of an image to be identified might depend not only on its basic features, but also on stochastic aspects such as its illumination or the proportion of a picture it occupies. Therefore, this change would be independent from the relevant content of the image.

It is becoming increasingly common to see applications of CNNs for the discovery and measurement of latent dimensions in data, or for the classification and scaling of abstract concepts. However, the aforementioned limitations of CNNs might make these tools unfit for such tasks. As we reviewed, CNNs can sometimes fail at even concrete, easy tasks such as number recognition due to weaknesses in the model and the challenges that real data pose. Since this limitation occurs even with factual classification, the coding of abstract or complex concepts becomes even more complicated. Recall that some crucial parts of information that give context to a visual message, such as surroundings and positions, get

lost during the classification process.

Finally, computer vision tools are able to quantify information in a way that humans cannot (e.g., get a color histogram or compute pixel intensities) and do so in a fast, reliable, and efficient way. However, CNNs cannot discover dimensions that humans cannot identify. As in other fields like text analysis, if a human cannot code or validate a trait, a CNN will not be able to do it, either.

In summary, researchers need to be aware of the limitations of CNNs, cautious about the objectives they expect CNNs to fulfill, fully knowledgeable of the data under analysis and training, and careful about the interpretation of the outputs of the CNN.

3.4   Conclusion

Using computer vision techniques for image-retrieval and classification can extend the scope of the data, theory and implications of several social phenomena. In this paper we presented a comprehensive guide for researchers interested in using Convolutional Neural Networks for visual content coding and classification. We presented the intuition behind CNNs, highlighted their potential, and described their structure and implementation.

CNNs have a wide variety of applications in multiple fields of social sciences. They can be applied to similar data collection problems like the one outlined in the text: retrieving signatures or the votes that were whipped for a given policy registered in historic documents, classifying written notes, or even the extraction and interpretation of symbols. They can also be applied to the coding of more complex political phenomena: measuring gender composition in pictures of groups, identifying the sentiment of material from electoral campaigns, recording the activities of crowds in a protest, counting the number of people waiting to vote in polling stations, etc. The extraction of information from images and visual content motivate a wider variety of questions.

The present article had three main goals. First, to illustrate the benefits of CNNs for data collection purposes and image classification by focusing on the recognition of handwritten tallies from the 2015 Mexican election. Second, to present some of the challenges and practical issues that researchers should consider when dealing with this type of the data. Finally, to discuss the strengths and limitations of CNNs.

Fortunately, the access to data and tools that allow for a richer understanding of various political phenomena is easier than ever. However, these opportunities should also be paired with a deep understanding of the characteristics, mechanisms and consequences of these models, as well as the acknowledgment of their limits and scope. The study of new data sources complements and enhances the knowledge that we already have about the political world, but also motivates innovative and interesting questions, and opens new avenues of research.

Chapter 4

Framing a Protest: Determinants and Effects of Visual Frames

Social movements are key drivers of change and tools for democratic actions (Barnes et al. 1979; Dalton 1988). However, their success depends on their ability to recruit and mobilize potential supporters, and raise sympathies among the population. The perceptions and opinions that the public forms about a social movement, and the evaluations that individuals make when deciding to join a movement are influenced by the information that media provides about the activities, composition, and objectives of a social movement's members. Media outlets offer details on the size of a protest, location and conditions of the place where the protest takes places, police presence, disturbances, and even testimonies and opinions from authorities and participants.

However, the ways in which journalists communicate stories differ depending on several considerations such as the characteristics of the audience, the ideology of the news outlet, idiosyncrasies of editors, reporters and photographers, context, resources, and others (McCarthy et al. 1999; Meyers 1996; Oliver and Maney 2000). Further, journalists and editors have a wide variety of tools to tell stories in a way that satisfies their ideologies and market

demands. Among these tools, visual material is an important and powerful element of the communication process that reveals information about a particular event, enhances the assimilation of information, and that also illustrates the vision and perception of the sender of a particular message. In contrast to text, images provide authors with a tool less subject to "fact checking" and scrutiny for objectiveness. These factors turn images into useful tools that help *framing* a story. Parry (2011) indicates that in general, media frames are patterns in which certain aspects of reality are promoted over alternatives, but visuals, and in particular press photographs, are even more selective in nature given that a single image is chosen as "emblematic" and "representative" of a particular news story.

A visual frame is an element that an actor uses to relay information, and that reveals what she sees as relevant to the topic at hand (Chong 1996; Chong and Druckman 2007; Druckman 2003; Druckman and Nelson 2003; Gamson 1989; Gamson and Modigliani 1989). Visual frames convey powerful information with the potential of triggering emotional and cognitive reactions beyond language. They are useful communicators of moods and vibes, and also impactful means of highlighting facets of a particular event. Pictures of protests can show the anger of their participants, as well as the pain of the victims they represent. They can show a large and active crowd illustrating the broad scope of a movement, or they can focus on specific individuals. Pictures can say a lot about the intensity of the actions of the protesters, and also provide information about the reaction of the authorities. Altogether, the elements and messages that images portray are crucial in shaping attitudes and behavior.

For example, visual frames can provide information about practical facts of a protest like time and space, and also other abstract characteristics like the mood of the event, by illustrating the interactions between the participants and the opposition as well as details of the type, unfolding and impact of the episodes of the event. In particular, visual material can provide information about the level of violence and conflict of the events in a protest. The use of violence in media as a marketing and communications strategy is well known. The

depictions of violence in media not only "sell" and attract attention (Howe 2002) but also have the potential of triggering emotions, alter risk calculations, shape levels of identification with the actors, and "prove" facts. Further, the illustration of violent events have the potential of drawing attention to the event itself and away from the protests issues (Smith et al. 2001). These elements will inform and shape the formation process of attitudes and opinions about social movements, and can drive or undermine mobilization and support (Muñoz and Anduiza 2019). Analyzing the political variables behind the generation of visual frames and their effect on opinions allows a better understanding of the dynamics of social movements.

In this article, I dissect some of the factors impacting the generation of visual frames and the effect of these frames on political attitudes and behavior. In the first part of this project, I implement computer vision and image retrieval techniques to measure and understand messages conveyed in pictures. I use a structural topic model to identify and measure how media outlets frame the mood and environment of a protest. Results show that in comparison to liberal outlets, conservative newspapers use a higher proportion of nocturnal and dark elements in the pictures they post in their Facebook profiles regarding protests of the Black Lives Matter movement.

In the second part of the project, I aim to understand the effect that visual frames of violence have on political attitudes towards protests and social movements, regardless of the textual content that might accompany an image. In order to do this, I conduct a set of experiments in which I vary the type and level of visual violence in photos of a protest. The results show that, on average, depictions of violent actions conducted by the protesters but not by the police decrease the perceptions of likelihood of success of the movement, weakens the identification with the movement and its members, and negatively affects the involvement with groups supporting similar causes. Further, these effects have small spillovers to other political areas such as tolerance. These effects vary between racial groups, and also groups

defined by levels of social dominance orientation (SDO).

In the following sections, I will first provide a brief overview of the factors impacting mobilization and success of social movements to highlight the aspects that can be affected by visual frames of protests. Second, I discuss the characteristics and potential impact of visual frames of mood and environment of a protest, especially with violent content, on attitudes and opinions towards social movements. Then, I introduce the methodology, data and results from the analysis of the generation of frames and its relationship to the political ideology of newspapers. Fourth, I present the design and findings of an experiment in which I vary whether the visual stimuli is violent, and also the origin of the depicted violence (protester vs. police). Finally, I present a discussion of the implications of these findings.

4.1   Shaping attitudes towards social movements

Social movements are key elements in the study of social and political change (Killian 1964). The emergence of groups typically unaligned with major political parties and with demands that respond to inequality and perceived unfairness have the potential to alter the political and social tissue. These movements and protest actions have become a permanent component of democracies (Barnes et al. 1979; Dalton 1988; Della Porta and Diani 2009) and an increasingly efficient tool for change in non-democracies. Taking part in social movements and protests is not only a strong feature of politically active citizens and an essential ingredient of healthy democracies, but also represents the development of collective political identities (Polletta and Jasper 2001). They raise awareness about issues that society faces, promote cooperation and action, and ultimately are able to change policies and the status quo (Costain and Majstorovic 1994).

However, an important determinant of a social movement's survival and success is its ability to make citizens *identify* with the movement (for mobilization purposes), promote *em-*

80

pathy with its members and causes, gather positive evaluations among the public (Kitschelt 1986) and send signals that its demands will be fulfilled as the ultimate symbol of *success*. These elements not only help to awake sympathy for the movement and increase the connection between its members and the public, but also help to build expectations of costs and utilities that form the basis of participation. The literature identifies multiple determinants of people's participation at in social movements at both the individual and group levels. However, in this article I will focus on three factors that affect mobilization and participation and that can be shaped with visual cues: 1) the perception that there is indeed a problem for which it is worth to protest for (*fairness*), 2) the belief that the members of the movement are not outsiders but citizens *like one* who are able to integrate to society (Rohrschneider 1990) and with whom they can identify (*empathy*), and 3) the perception that the movement is likely to succeed (*likelihood of success*, Campbell (2005)). As in most decisions, embedded in these elements is the notion that the utility and benefits from protesting and joining a movement will outweigh both individual and social costs. For example, the perceived likelihood that the government will be responsive to the movement's demands should be high enough to overcome the risks and costs of participating in a protest. Therefore, in order to understand the success, development, and evolution of social movements it is important to revisit the process by which society forms perceptions and evaluations of such movements, and eventually grants support.

Although the process of opinion and perception formation of social movements involves multiple factors interacting with each other at both individual and group levels, a central input of the process is information. The way in which the demands, actions, and composition of a movement are framed to the public plays a central role in the strength of the support it receives (Benford and Snow 2000) and its ability to mobilize (Noakes and Johnston 2005). The framing of the mood of a protest and the activities occurring in these events inform potential attendees about the risks of getting hurt or arrested, the reactions of the police, or

even the plausibility of bringing other participants like children.

Therefore, the flow of information about protests and social movements' activities matter not only for the consolidation of the movement itself but also for shaping public attitudes about more general issues supported by the movement. Media plays a central role in providing and framing this information. Although the job of providing the public with facts of a particular event could seem straightforward, in practice journalists and communicators can use very different tools and elements to tell the same story based on their own needs, ideologies and markets (Fiske and Hancock 2016; Oliver and Myers 1999). The toolkit to frame a story not only includes basic communication elements such as text, images, videos, or interviews, but also involves communication strategies like symbolisms, rhetoric, and social media use.

The impact of some of these elements has been widely explored (Downing 2000; Gamson and Wolfsfeld 1993; Giugni 1998; Nelson, Clawson and Oxley 1997). More recently, some studies have even focused on the role of social media in the evolution of social movements (Anastasopoulos and Williams 2016; Valenzuela 2013). However, with some exceptions (Domke, Perlmutter and Spratt 2002; Griffin 2012; Linfield 2011), most of the literature on this issue focuses on the textual messages that media send and does not consider the visual material that accompanies the text. This is concerning given the important role that visual stimuli play in the process of perception and opinion formation. In this project, I analyze the effects of different visual stimuli. More specifically, I focus on the importance of the visual depiction of violence in protests, and the ways in which this affects perceptions of social movements. But, why visuals and why violence?

4.2 The role of visual frames: mood, environment and violence

The literature on social movements and political geography identifies mood, time and place as crucial elements in the analysis of the evolution and success of movements and demonstrations. Multiple authors have highlighted the importance of place and context in shaping protest actions, and highlight the need for a "spatio-temporal strategy" for mobilization purposes (Massey 1995; Pickerill and Chatterton 2006). In other words, the time and place in which the activities of a movement take place provide information to the public and impact their evaluations of costs when deciding to join a movement. Time and place also affect expectations regarding the activities and dynamics of a movement, and therefore might also impact the government's reactions towards these activities.

Within the dimension of mood and environment we can also identify the use violent to describe the events of a particular protest. The use of violence and conflict to frame an event affects the evaluations of cost , perceptions of tension between protesters and authorities, and identification with the movement. Illustrations of social movement activities and protests are likely to contain elements that can build perceptions of legitimacy, easily activate emotional responses, and motivate unconscious reactions of anxiety, opposition, empathy or denial, to just name a few (Corrigall-Brown and Wilkes 2012; Huesmann 2007; Wolfsfeld 1991). Thus, the answers regarding the magnitude and direction of the effect of visual violence on political opinions and attitudes regarding social movements are not definite.

Some authors suggest that the large loadings of violence to which people are exposed for most of their lives have led them to be more indifferent to pain, disgust, and brutality (Moeller 2018; Taylor 1998). Other scholars pose that violence generates feelings of discomfort, anxiety, and negativity that cause the audience to disengage, increase awareness of the risks of taking part in an activity perceived as violent, and can even distort the perceptions

of fairness and empathy. However, again other findings indicate that visual violence can still trigger a diversity of emotions that have the potential of altering attitudes and opinions. Visual violence might not only activate compassion through arousal and indignation, but also provide proof of atrocities and injustices that lead people to participate (Campbell 2004). Similarly, Chong (1991, p. 137) suggests that overreaction or unnecessary use of violence and repression against protesters may "motivate participation even when most people would rather abstain from action." Thus, the analysis of visual violence has multiple aspects that are worth examining such as intensity, direction, object, and even previous dispositions of the viewer.

Among these predispositions are attitudes towards equality, hierarchical structures and need for order that are strongly related to the dynamics of social movements and protests. For example, studies find that participation in and affinity to anti-globalization protests is associated with opposition to social hierarchies (Cameron and Nickerson 2009). In contrast, subjects with "hierarchy-enhancing" attitudes will tend to justify discrimination, attribute negative values to the "subordinate" or disadvantaged groups that are generally on the side of the protesters and will tend to have lower empathy, communality, and altruism: key drivers of support for and participation in social movements and collective action (Pratto et al. 1994; Sidanius et al. 1994). Thus, we expect subjects with higher levels of hierarchy-enhancing attitudes to be less susceptible to visual frames of violence in protests given the nature of these events, and to overall report negative perceptions towards protests, regardless of the visual stimuli.

4.3   Framing the mood of a protest

As explained above, the way in which the mood and environment of a protest are framed provides important information about the movement's events. There are multiple elements

that can shape the mood including the type and development of activities in the protest, the interaction between actors, and the time and place of the event.

A classic "time-space" setting is the night. Although, this is clearly a time of the day, its characteristics outline the context and environment of an event. For example, the night is considered a politically dangerous and impractical "time-space" for mobilization and political activities (Massey 1995; Shaw 2017). On the one hand, it poses logistic challenges to actual mobility: transportation is limited, service provision is unfrequent if not inexistent, and social networks are less accessible (they are already resting or taking care of their families, Shaw 2017). On the other hand, the characteristics of the night such as darkness and quietness negatively affect perceptions of safety, security and confidence; beyond the diminished sensorial capacity, night tends to be linked with crime (Edensor 2013, 2015; Morris 2011; Shaw 2015). In summary, the perception that a movement holds its events during the night has the potential of diminishing participation and engagement.

Therefore, the portrayal of protests or political events in nocturnal settings might be used as a visual frame to depict movements in more negative ways. Thus, we should expect those who are less supportive of social movements to have more incentives to depict protests and demonstrations in darker, riskier and more dangerous conditions. More specifically, we should expect conservative outlets to be more likely to subscribe to the *protest paradigm*, in which protesters are portrayed as deviant, threatening and impotent (Lee 2014) than liberal outlets. The reasoning is that in general, social movements and protest activity do not align with principles typically considered conservative: preservation of *status quo*, law and order, and patriotism (Di Cicco 2010). This is especially true for those movements, like the Black Lives Matter, which do not rally around conservative issues or values. Thus, we can pose the question of whether the use of this nocturnal frame is associated to this ideological dimension?

4.3.1 Research design

To answer these questions I analyze the pictures accompanying Facebook posts that the top 100 newspapers in the U.S. have published in their newsfeed. I collected 221 posts that mentioned the phrases "Black Lives Matter", "Black Lives", "Lives Matter" or "Ferguson" in any of their sections. I searched for those terms in eaach of their Facebook feeds and then kept the corresponding posts. Several newspapers did not publish *any* posts with these words and therefore were dropped from the analysis. At the end, the sample includes around 60 newspapers with an average of three pictures/posts per newspaper. Each post has accompanying data including date, newspaper, caption, text, etc.

In order to identify frames in the corpus of images, I conduct an analysis of the categories underlying the pool of Facebook images (Monay et al. 2009). The expectations are that some of the latent topics describing the content of the images provide information about 1) the characteristics of the protest, and 2) the elements that the author or publisher of a picture use to frame a story. More specifically, I aim to identify topics providing information about the time and place in which the protests took place (e.g. night). In order to uncover these topics and test these expectations, I implement a Structural Topic Model Roberts et al. (2014) based on an Image-Visual Word matrix that I obtained using a Bag of Visual Words approach as presented in Chapter 2.

First, I build a codebook of 2,000 "visual words" based on the clustering of features of 15,000 *Getty* images of protests from the Black Lives Matter movement. The idea behind this is to get the most comprehensive pool of images capturing as many angles and perspectives possible of a particular event (*Getty* has multiple photographers and contributors so the variety of perspectives ameliorates concerns for bias). Then, I extract an Image-Visual Word matrix that includes the number of times that each visual word in the codebook appear in the images under analysis. Subsequently, I feed this matrix to a structural topic model initialized with 12 topics. Further, I merged this data with the ideological slant scores computed by

Gentzkow and Shapiro (2010), where higher numbers indicate more conservative outlets. The STM includes ideological slant as a prevalence covariate.

4.3.2 Results

The STM was mostly consistent with some of the theoretical expectations discussed in previous sections. A topic of "Night activity" emerged from the image corpus. The most exclusive and frequent words in this topic mostly include patches of dark backgrounds with splashes of light (corresponding to night skies with lights and lamp posts glowing), and patches of dark uniforms and helmets (mainly from police and armed forces). Figure 4.1 show five examples of the most frequent and exclusive visual words. Further, Figure 4.2 shows six of the most representative images of the topic, all illustrating nocturnal settings.

Figure 4.1: Most Frequent and exclusive visual words for the topic "Night activity"

The findings indicate that there is variation in the proportion of the "Night activity" frame that the newspapers decide to include in the images accompanying their posts. In

Figure 4.2: Most representative images of the topic "Night activity"

Figure 4.3 we see the newspapers along the x-axis, and the mean proportions of the "night activity" topic by newspaper [1] on the y-axis. As we can observe, newspapers can decide to use images with 0 to almost 100% of the "night activity" topic.

---

[1] The Top 8 newspapers plus The *Boston Globe* and the *St. Louis-Post Dispatch are highlighted in the plot.*

Figure 4.3: Topics by newspaper (timeline)
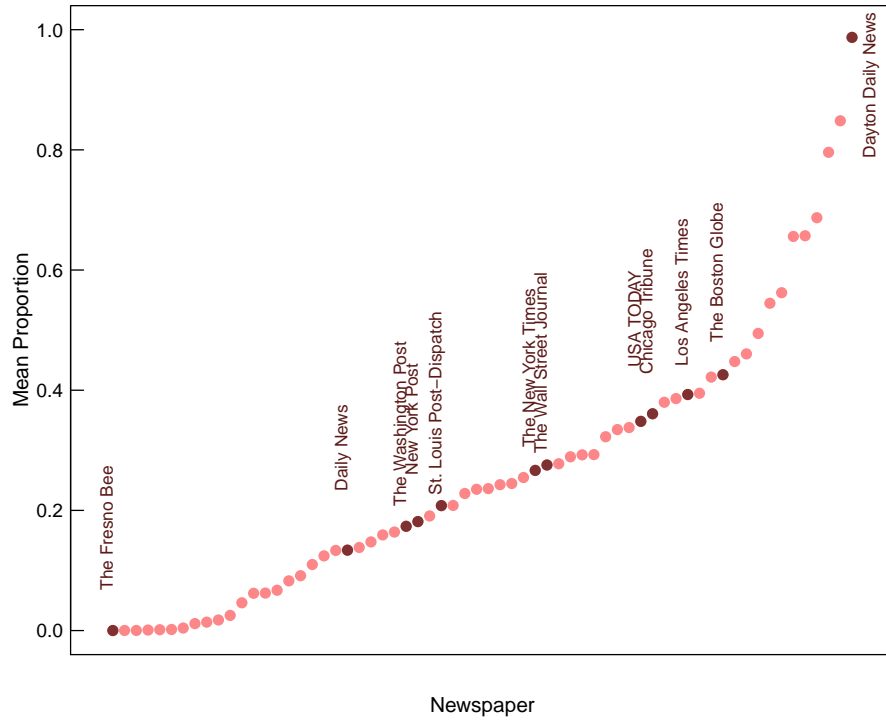
We can also observe variation in the topic that dominates each of the images that newspapers choose to publish by date. Figure 4.4 shows a timeline of the Black Lives Matter movement, with red lines indicating important events such as protests, indictment announcements or shootings. Each point represents the mean proportion of the "night activity" topic in images posted by newspaper and by date. Visual inspection suggests that there is a wide variation in the use of the "night activity" topic even for protests happening in the same day.

Finally, to test the expectation that conservative newspapers are more likely to use nocturnal frames to depict protests, I extracted the coefficient from the STM indicated the impact of "ideological slant" on the prevalence of the "Night activity" topic. The results presented in Figure 4.5 suggest that more conservative outlets are more likely to use pictures with higher proportions of the "night activity" topic when talking about the BLM protests. This is in line with the theoretical expectations.

89

Figure 4.4: Proportion of "Night activity" topics by newspaper and date
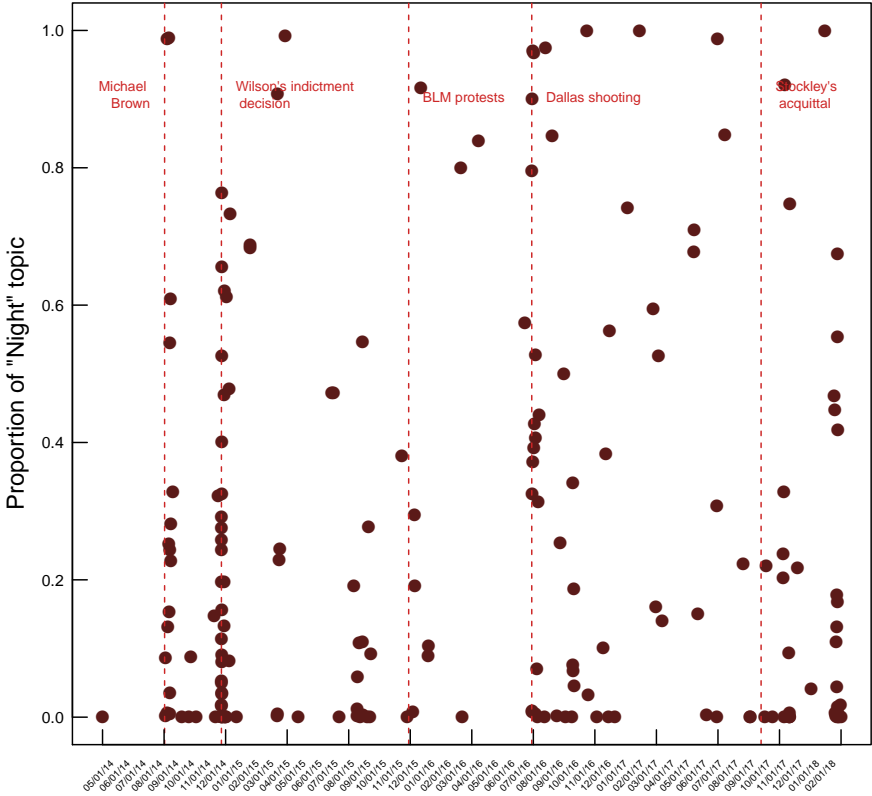
4.4 Analyzing the effect of violence as a frame

The second study includes the analysis of whether the use of violence as a frame affects opinions on social movements. More specifically, I focus on the peaceful vs. violent frame of a protest, and also on the frame of the actors behind violent acts: authority vs. protester. The expectations are that the identity of whoever is framed as the perpetrator of violence shapes the way in which individuals process visual violence and information. In other words, the frame of who originates the violence matters: a violent act of a protester is not evaluated and digested in the same way as a violent act by the police.

First, it is important to have in mind that the government has the monopoly and right

90

Figure 4.5: Ideological slant and nocturnal portrayals
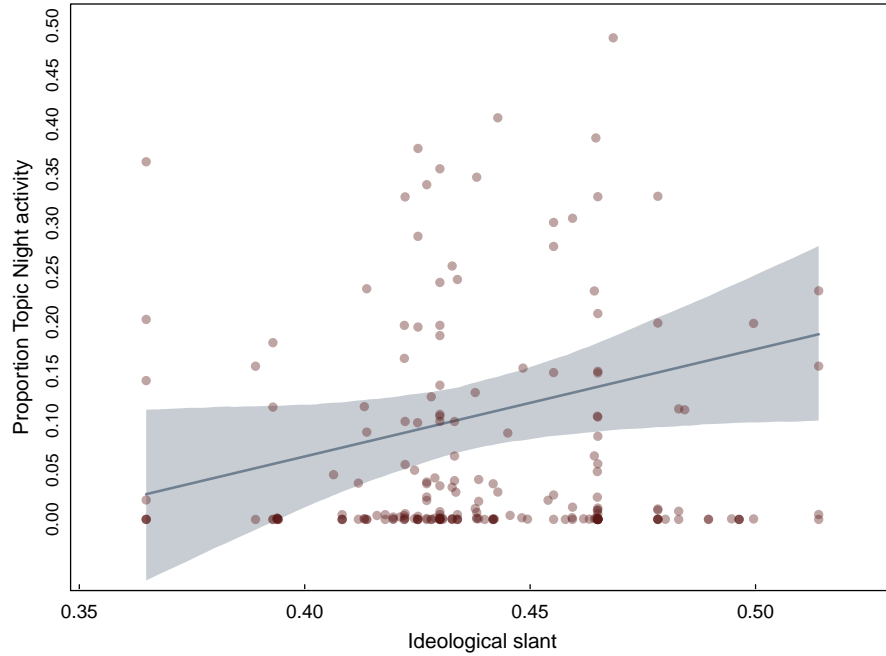
over the use of force to keep order and control. Thus, authority figures like the police enjoy a baseline level of legitimacy (especially strong among some prominent groups) that is considerably higher than protesters. Furthermore, they tend to be portrayed in the media as more credible and powerful. A study by Corrigall-Brown and Wilkes (2012) shows that news articles about protests not only tend to show authorities in dominant positions, but also tend to cite them more than protesters. Therefore, subjects are more likely to receive the authority's version of the story magnified with pictures showing officials in a position of power.

Although this could seem to be in contrast with the expectations of Chong (1991) and King Jr's (1986) in which the "aggressive non-violence" model leads to heightened public demands and legislative reform, we can still reconcile the theoretical expectations outlined above with the claim that depictions of violence by the authority have a lower effect on per-

ceptions. The model involves three stages: 1) protesters exercise their constitutional rights and assemble, 2) authorities (protesters' counterparts) incur in violent acts without provocation from the protesters, and 3) media covers the violent event and elicits compassion and sympathy towards the movement members. While intuitive, this last point is problematic. Media generally focuses on snap shots of a full event to show its evolution but without a specific order. How can we know what happened before a violent act? How do we know whether the protester provoked the violent reaction of the police? Take for example the iconic picture of Edward Crawford, a protester in Ferguson, throwing what seemed to be an object in flames. The action could be interpreted as a violent act in which the protester is actively attacking others. However, more context and information shows that he was simply returning a tear gas canister fired by the police seconds before. In other words, capturing and portraying a moment that fully fulfills point 2 (a violent act by the authorities without provocation) is a hard task that interacts with other cognitive and attitudinal filters as explained above.

The main implication of these positive pre-conceptions of the police and lack of the "full picture" is that subjects will tend to 1) be more forgiving of violent acts by the police in exchange for order and safety, and 2) perceive that a violent act by the police is likely to come as a reaction to a violent act by the protesters, and not as an instigating action. Further, the depiction of the violent act is expected to have a negative effect on the cost-benefit evaluations: more violence from the police is a clear sign of repression and therefore lower incentives to engage with the group's activities or to mobilize. Therefore, regardless of the actual evaluation of the police, the depictions of tension between activists and officials will ultimately result in greater pessimism with respect to the achievement of the movement's objectives (Earl 2003, 2006; Soule and Earl 2005).

Moreover, the portrayal of the protesters' actions, and especially the violent ones, provide information that can impact several considerations and evaluations of a subject. These

actions talk about the characteristics, motives, modes of action, and even personality of the protester that generally tends to be projected in viewers' minds as the "representative member" or "average member" of the movement. Therefore, a violent depiction of a member of a social movement can affect the perceived ideological and social distance between them and a subject, and might lead to less engagement with activities and movements similar to the ones to which they belong. Beyond the decrease in identification and empathy levels, a violent depiction of a protester increases the costs associated with participation given the means depicted in pictures (e.g. violent protesters suggest a higher probability of being arrested), and also weakens the perceptions of success by the anticipation of repression (Chenoweth and Stephan 2011; Franklin 2015).

Further, these effects are going to be reliable or stronger when the violent acts are portrayed using visuals rather than text.

To test the expectations outlined above, I conduct two sets of experiments in which I provide respondents with a news article about the Occupy movement protest.[2] The news paper articles that the subjects received did not have any identifying information of the newspaper.

4.4.1 Experiment 1: the effects of visual violence

Research design

In the first experiment, the text of the news article talks about multiple protests in different cities of the U.S. that have become violent despite the original intentions of the movement. This control condition does not include any picture. Each of the treatment conditions consist

---

[2]The reason I chose this movement is that it keeps a good balance of desirable features: it is not as salient as other movements like the Black Lives Matter movement given that it became popular several years ago, its core theme is less polarizing and controversial (income equality and better economic conditions), and the occurrence of protests in multiple cities and contexts allowed for the formulation of a neutral text, and for the identification of contrasting visual stimuli along the dimensions of violence and directionality of the violence.

of a picture accompanying the article with the exact same caption but different content: the *Peaceful* condition shows a group of people (mostly white) marching and holding signs. The *Protester Violence* treatment shows a white protester throwing an object next to a car in flames, and the *Police Violence* condition presents a white policeman pepper spraying a row of students sitting on the floor without offering any resistance. All of these images were taken from actual articles talking about the Occupy movement. Figure C.2 presents the images used for the treatments.

Figure 4.6: Pictures included in the treatment conditions (Study 2)

(a) Peaceful

(b) Violent protester

(c) Violent police

The study was conducted among 2,056 subjects from Fulcram Lucid, a platform that recruits respondents from multiple commercial sources to build a representative sample.[3] The survey included some pre-treatment questions aiming to measure tolerance to protests, Social Dominance Orientation (SDO) attitudes, and political variables like party identification, followed by the presentation of the experimental vignettes. Respondents were then asked to answer several outcome variables related to perceptions and evaluations of the protest and

---

[3] While the sampling process does not guarantee national representativity, the demographic composition of the sample resembles national parameters.

the movement, likelihood to engage with the movement, and questions related to social order and tolerance.

In order to identify the average treatment effects, I conducted several test of differences of means between the different treatment groups. The analysis also includes comparisons with different baselines. For instance, the tests of differences of means were not conducted only with respect to the "only text" condition, but also with respect to the "peaceful picture."

Results

Figure 4.7 shows the mean responses by treatment group (y-axis) for the following outcomes: perceptions of effectiveness of the protest, whether respondents share views with the protesters, their propensity to getting involved with other movements with similar objectives, and their evaluation of the police's reaction to the protesters.[4] We can observe that the means of the "No picture" and "Peaceful" groups are very similar, suggesting that the peaceful illustration of the protest does not provide information that generate a more positive or active attitude towards the protesters. However, even without a more detailed test we can observe that respondents that received the picture of the violent protester report less favorable attitudes towards the movement: they are less optimistic about the success of the movement, show less affinity to the protesters' viewpoints, a lower desire for participation, and report lower levels of agreement with the statement that the police's response to the protest (which the text stated was aggressive) went too far.

Further, the panels of Figure 4.8 show the results from multiple tests of means to determine whether the differences in each of the outcome means between the treatment groups (the average treatment effects) are statistically distinguishable from zero. Each subfigure shows the difference in means (points) and their respective 95% confidence interval (bold gray lines). Each panel in the plot indicates the baseline for the comparisons (indicated in

---

[4]All outcomes were measured on a 5-point agreement scale with each of the statements in the titles of the plot.

Figure 4.7: Means of attitudes towards the Occupy movement, by treatment group
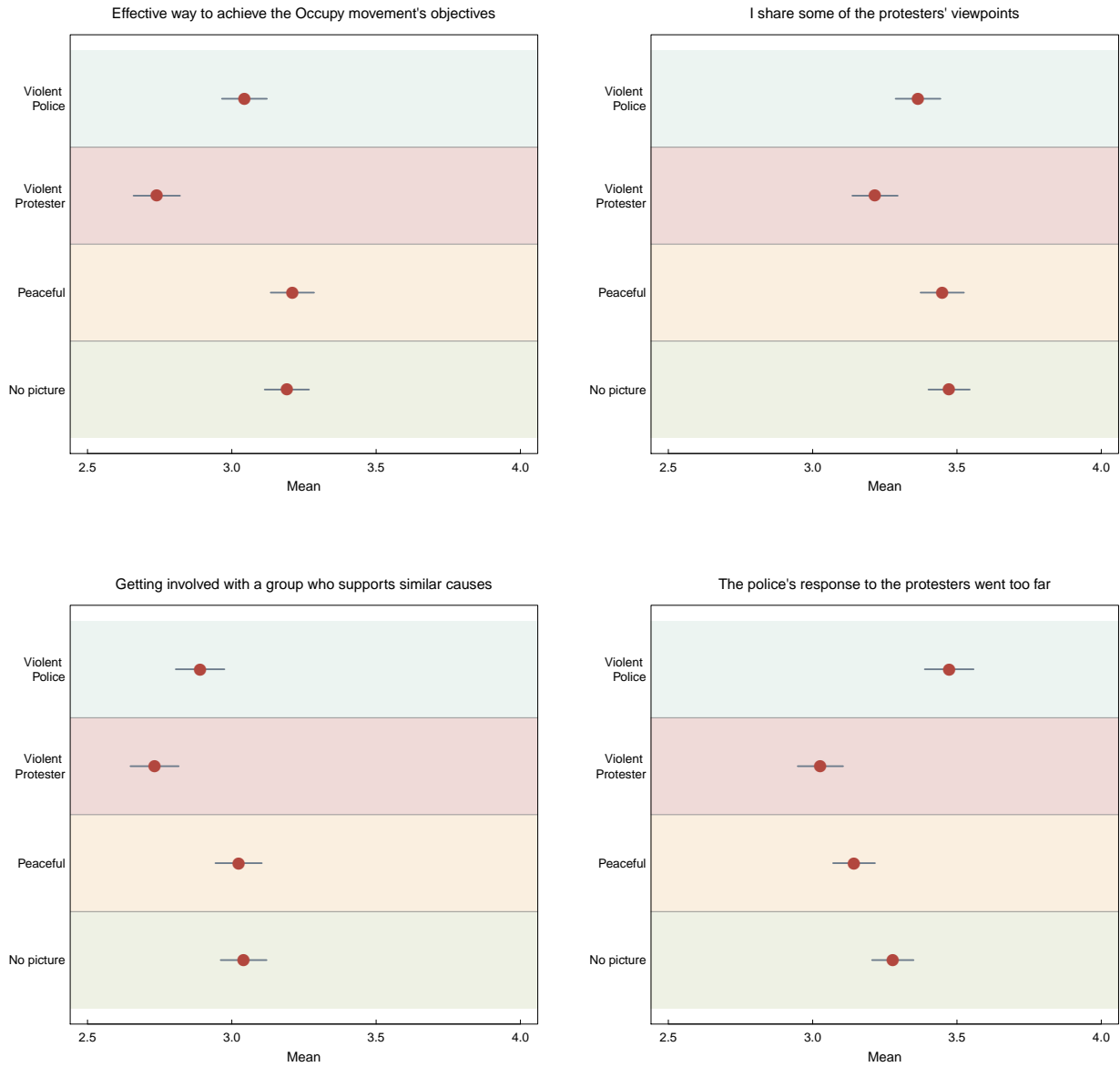
the *y*-axis). For example, the bottom panel in light blue shows the differences in means between the "Peaceful", "Violent protester" and "Violent police" conditions, and the " No picture" condition. Similarly, the top panel in green shows the differences in means of the "Violent police" and "Protester" groups. In this way we can assess whether all the possible

differences between treatment groups are distinguishable from zero (red dashed line).

Figure 4.8: The effects of directionality of visual violence on attitudes towards protests
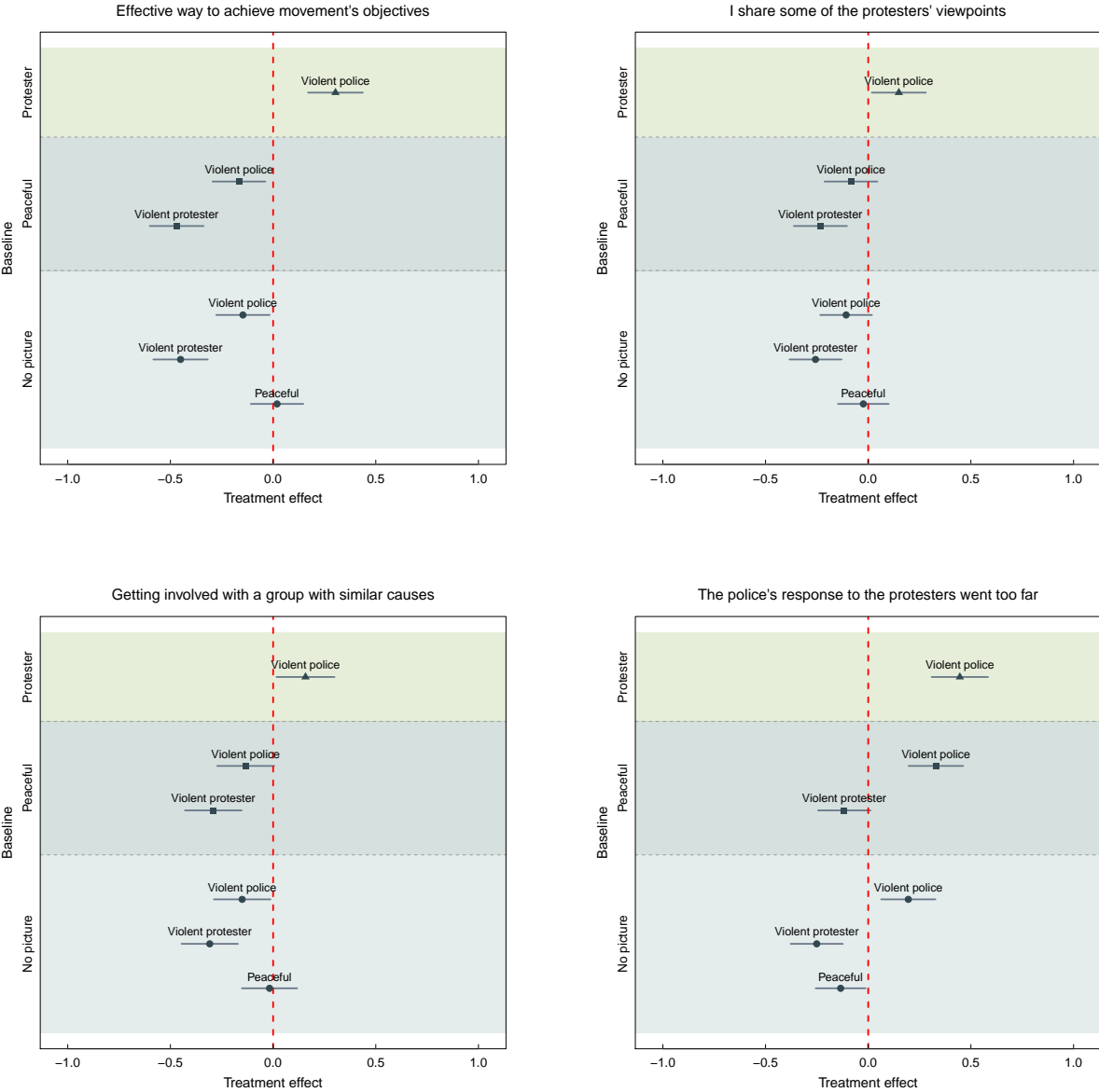
As we could infer from the previous figure, the analysis indicate that receiving the "Violent police" condition has a negative effect in all outcomes regardless of the baseline category with which we compare this group. This is in line with the expectations that violence com-

mitted by protesters will decrease empathy and identification with the movement. Further, we observe mixed results with respect to the effect of the "Violent police" treatment. Its effect is reliable on the perceptions of success, participation, and evaluation of the police's response, when compared to the "No picture" condition, and only for success and evaluation when compared to the "Peaceful" condition. Further, instead of supporting the theory regarding the "awakening" and mobilization of participants when unnecessary violence is implemented by the opposition, the results indicate mostly negative effects on opinions about the protests even in the case where the authority engages in aggressive actions. The image provides information about the consequences of getting involved with the movement and has a deterring effect possibly associated with the difficulty of attributing blame for violence *only* to the police. This is also supported by the positive and reliable effects of the "Violent police" condition with respect to the "Violent protester": violent actions committed by protesters have a bigger and negative impact on the evaluations and attitudes of citizens.

4.4.2 Experiment 2: comparing visual and textual stimuli

Research design

In the second experiment, subjects were randomized into nine potential group combinations determined by the combination of three visual treatments, and three textual treatments. Once again, the treatment was about levels of violence and actors originating it (if any): no violence (peaceful frame), violent protester, or violent police. The pictures and paragraphs for these experiments are presented in the Appendix.

This experiment involves 2,011 respondents from Lucid. The characteristics of the sample are similar to those of experiment 1. Further, the survey was fielded almost with an identical structure and content to that of the first experiment.

In this section I conduct different tests of means as in the previous example in order to

register the different treatment effects of visual violence within the text and visual stimuli groups. However, I also execute a series of linear regressions and tests of differences in coefficients. This last analysis with the intention of discerning whether the effects of visuals are indeed different from those of text even if they fulfill the expectations with respect to their orientation. Results are presented below.

Results

First, Figure 4.9 presents the means for different outcome variables. After repeating the exercise of comparing means between treatment categories for each of the two types of stimuli (e.g. text vs. images), I find that almost all of the results from the first experiment hold: the "Violent police" treatment increases the perceptions that the police went too far, the "Violent protester" condition decreases the identification with the group and the likelihood of engaging with similar groups, in a significantly higher degree than the "Violent police condition", and both violent frames decreases the perception that the movement will achieve its objectives.[5] However, these reliable differences are mostly observable when the stimulus is visual and not when is textual, although the direction of the treatment is very similar between the two types of violent frames. However, notice that this is not true for two outcome variables: sharing the protesters views, and achieving the movement's objectives. For both outcomes, we can observe that the effect of textual violence has a negative direction regardless of the actor originating it. This contrasts with the visual frames where only the "Violent protester" condition evidences a negative effect, but not the "Violent police".

Further, I conduct a series of regressions that allow me to test whether the differences between textual and visual coefficients are distinct from zero. Table 4.1 show the results. Each column shows an outcome variable. The baseline category is "Peaceful text" and

---

[5]The main differences between experiment 1 and 2 are the difference in treatment effects between the police and protester conditions. This could be attributed to the intensity of the violence in the vignettes used in the second experiment. Purposefully, I increased the intensity of the violence perpetrated by the police, and decreased the violence by the protester. Substantively, this does not differ from the results above.

Figure 4.9: Means of attitudes towards the Occupy movement, by textual and visual groups
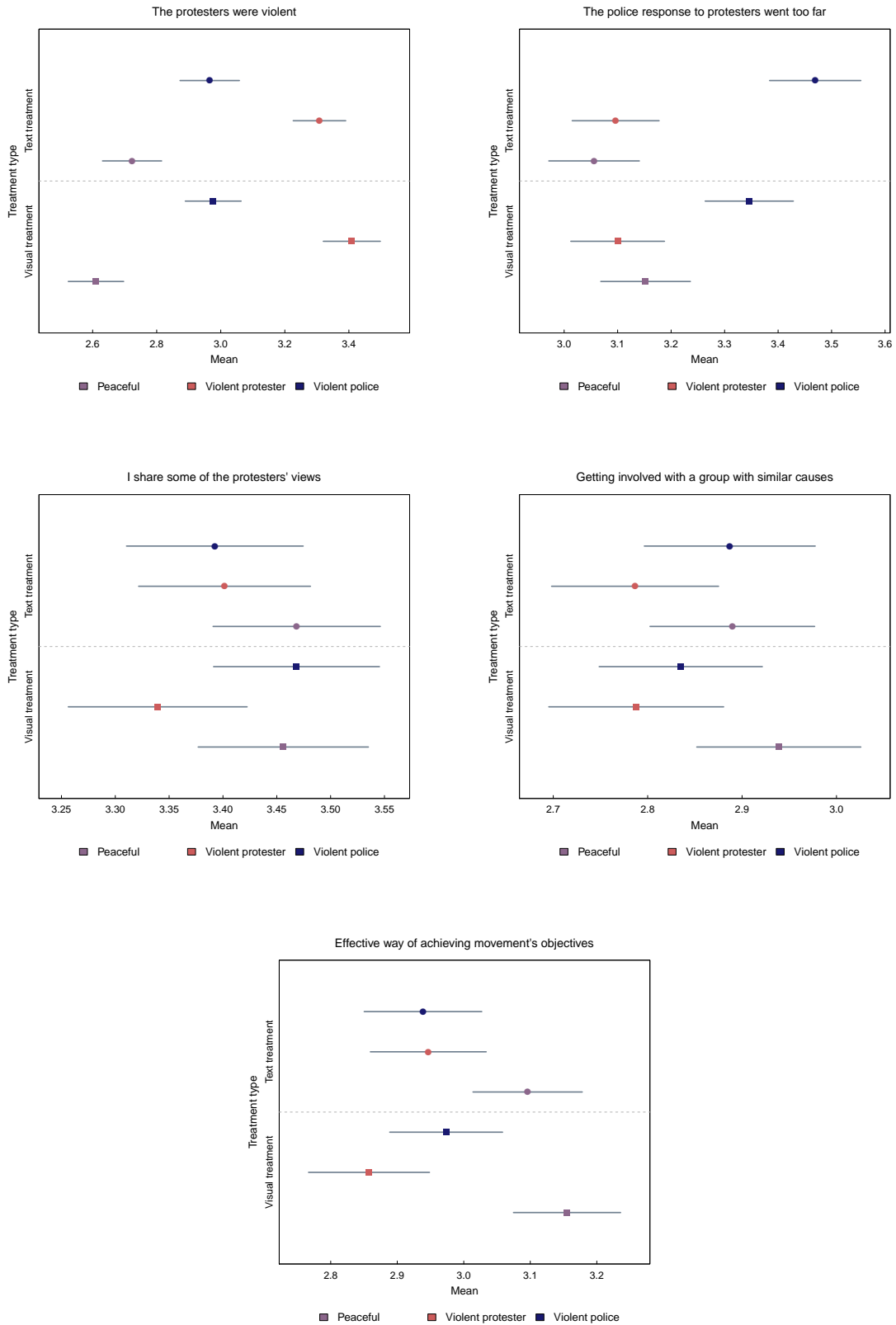
"Peaceful picture". The coefficients confirm the discussion in the previous paragraph. First, we can observe that in line with theoretical expectations, the visual frame of the "Violent protester" is reliable for most of the models presented. Further, it has a negative effect suggesting less engagement with the movement, and more negative evaluations of success and perceptions of violence. This does not hold for the textual stimuli in two cases: sharing the views and willing to participate with other similar group. For the perceptions of violence, evaluations of success and whether they should be allowed to hold a rally, we see similar effects of textual violence to those of visual content. However, the test of coefficients indicate that for columns (1) and (5), the coefficients of visual stimuli are significantly higher than those of text. Further, it is interesting to observe that there is one instance, the evaluations of whether the police went too far, where we see that the textual frame of the "Violent police" has a significantly stronger effect than the visual one. Respondents seem to trust text more when it comes to explaining police violence rather than images. This, again, could be related to the thoughts that a violent police might be reacting to protesters' provocation given the chaotic environment that the image portrays. Further research should address these issues, as well as the mechanisms in which respondents are using different sources and type of information to build perceptions and opinions about a political event.

4.5   Conclusion

Social movements are a crucial part of the political and social dynamics in which we live. While the existing literature has focused on carefully explaining the motivations and factors behind people's attitudes, opinions, and decisions to get involved with a social movement, less has been said about the information that predates those factors and that feeds the attitude formation process itself. Media outlets have means of communication that go beyond language and text. Images are part of the toolkit that allows them to frame stories, and

Table 4.1: Effects of textual and visual violence

	Violence	Police too far	Share views	Participate	Achieve objectives	Hold a rally
	(1)	(2)	(3)	(4)	(5)	(6)
Violent protester (Text)	**0.612**	0.038	−0.071	−0.108	**−0.159**	**−0.139**
	(0.060)	(0.058)	(0.056)	(0.062)	(0.060)	(0.058)
Violent police (Text)	**0.245**	**0.414**	−0.076	−0.003	**−0.158**	−0.072
	(0.061)	(0.059)	(0.057)	(0.063)	(0.060)	(0.059)
Violent protester (Image)	**0.820**	−0.058	**−0.118**	**−0.155**	**−0.301**	**−0.151**
	(0.061)	(0.059)	(0.057)	(0.063)	(0.061)	(0.059)
Violent protester (Image)	**0.375**	**0.193**	0.011	−0.106	**−0.184**	−0.037
	(0.060)	(0.059)	(0.056)	(0.062)	(0.060)	(0.058)
Constant	**2.317**	**3.010**	**3.505**	**2.978**	**3.260**	**3.727**
	(0.055)	(0.054)	(0.051)	(0.057)	(0.055)	(0.053)
N	2,009	2,009	2,010	2,010	2,009	2,009
R^2	0.123	0.038	0.004	0.005	0.017	0.006
Adjusted R^2	0.121	0.036	0.002	0.003	0.015	0.004

Bolded coefficients indicate $p \leq 0.05$

change the evaluations and perceptions of the people consuming them with regards to the event they depict.

In this project, I analyzed the origins and consequences of two visual frames of protests: 1) the use of nocturnal elements to frame a protest's mood, and 2) the framing of violence in a protest in terms of intensity and directionality.

The results suggest that political ideology is associated with the generation of darker and more negative depictions of protests. Results show that conservative newspapers use a higher proportion of these elements in images of protests than liberal newspapers. Also, I find that visual violence shapes perceptions about protests and social movements. More specifically, depictions of violent protesters have a negative and significant effect on areas like evaluations of success, identification with the members of the movement, evaluations of police response, and feelings of empathy. In contrast, while the depictions of violent authorities also affect some of these outcomes, in general we observe that the public is more likely to react to

protester violence than to police violence. Perceptions of legitimacy and expectations of the use of violence are factors that could explain this pattern. Further, these effects exist or are stronger when we the violence is visual rather than textual.

While the findings of this paper are suggestive of important effects triggered by visuals, I acknowledge that "violence" is a treatment that is difficult to isolate, especially in the context of images. While the photos might be introducing extra information about other dimensions of the protest or the participants, there are a number of ways in which I have tried to minimize such potentially confounding effects (Dafoe, Zhang and Caughey 2018). First, the experimental set-up was carefully designed to control for as many relevant cleavages and dimensions as possible. For example, the pictures do not differ along racial or gender dimensions. Second, the research design was devised to include real pictures that were published in articles talking about these protests. The original captions accompanying these pictures described and highlighted the presence of violent actions as the main object of the picture under analysis. Third, while this setting could still raise questions about the validity of the effects of violence on attitudes, the findings about the impact of visual framing stand: differences in visual depictions of protests have an effect on attitudes even when holding text constant.

Further research should focus on the way in which respondents assimilate both visual and textual stimuli, as well as on the differences in cognitive and political filters that respondents use when processing information from media outlets. The findings and implications of this research are crucial to understanding the different information flows that people consume and the way in which they eventually translate it into attitudes. A closer analysis of these dynamics may aid in developing strategies to provide objective information, close information gaps, motivate participation, and eventually decrease polarization.

Chapter 5

Concluding remarks and further research

Throughout this dissertation I presented the description, implementation and illustration of tools that allow an easier analysis of images, especially related to political phenomena. Social scientists should take advantage of the interdisciplinary bridges between fields like computer science and psychology to complement and strengthen our research frameworks and methodological tool kits.

In Chapter 2, I introduce and illustrate the use of a Bag of Visual Words (BoVW) for the construction of a Visual Word-Matrix that can serve as input of both supervised and unsupervised methods. Its characteristics and implementation make it an accessible and powerful tool for the exploration of visual content in images, and the relationship of this content with relevant political variables.

Chapter 3 offers a detailed description of the characteristics and process of Convolutional Neural Networks (CNNs): one of the most popular tools in the field of computer vision for the analysis and classification of visual material. Although political scientists are increasingly using them to address substantive questions, this chapter clarifies the intuition and

mechanisms underlying their functioning in an accessible way. The objective is to decrease the perceptions of their opaqueness and "black box" nature, but also to discuss their limits and scope to optimize their use.

Finally, Chapter 4 illustrates the application of the BoVW for the analysis of visual frames. It studies the political factors, like ideology, behind the generation of visual frames of protests and shows that more conservative news outlets tend to portray protests of the Black Lives Matter movement in darker and more nocturnal settings than liberal outlets. It also presents the results from a set of experiments testing the effect of different verbal and visual frames of violence on attitudes and opinions. The findings suggest that the information that media provides to citizens shape their evaluations of and identification with social movements. Visual frames have a higher impact on this molding of opinions and therefore should be included in any study aiming to understand political communication and information processing.

This dissertation intends to motivate and inspire not only more analyses regarding the conclusions it reaches and the mechanisms suggested, but also new questions and applications across fields.

Techniques like CNNs provide researchers with tools beyond classification purposes: they are powerful tools for data collection from sources like archival and historical documents, or hard to code settings. This type of material can be crucial for studies focused on long-term effects or with historical content, but also for contemporary cases with real-world implications like the waiting time in voting lines during election day.

Further, the use of images overcomes some of the barriers that other sources like text face, such as language differences, and therefore provides a unique and powerful feature to be exploited in fields like comparative politics and international relations. What do the visual characteristics of protests and events related to ethnic conflict tell us about the similarities and differences of these events across countries and institutional settings? How do local

105

media portray authorities in authoritarian regimes and how do these portrayals compare with those found in democracies? What are the different visual strategies of candidates running for office according to characteristics like gender or race?

There is a long list of applications and fields in which visual content plays a significant role. Needless to say, the analysis and understanding of visual content is crucial in fields like political communication where answering questions regarding the ways in which information flows between actors and the impact that it has on their behavior and attitudes can have substantive and important real world implications.

This dissertation motivates several extensions and further research in the area of political communication. These extensions cover questions about the generation, supply and assimilation of visual information. For the generation stage, there is the question regarding the differences between the material that media, activists and social media create about a movement. One angle of this question relates to the ideology and idiosyncrasies of the people in the field generating the pictures: what are the filters and criteria that photographers and photojournalists use to decide the focus and style of their material? A measure of ideology based on social media interactions with political figures could provide some insights about these questions. Further, the comparison of the material generated by different actors can improve our understanding of their agendas and objectives, as well as the interpretation of each other's material.

The analysis of the supply also raises some questions about the motivations of media outlets and their role in public opinion dynamics as either followers or shapers. What are the strategic factors, conscious and unconscious, that play into the decision of the outputs they offer to their audiences? The analysis of visual material, paired with strong causal inference methods, time-series, and surveys of prominent media figures like editors, journalists or reporters might shed some light on these dynamics.

Finally, one of the most interesting aspects of these information flows is the final stage

106

where subjects receive and process the information. For this part, two aspects are particularly interesting: 1) the mechanisms in which elements of news material, ads, posters and brochures affect viewers and the information that they retain from that material, and 2) the individual factors and characteristics that moderate different assimilation processes. The use of laboratory experiments with new technologies like eye-tracking allows researchers to address these questions in an attempt to have a better understanding of the cognitive side of the political communication process.

One of the main goals of this dissertation is to provide a framework of methods that facilitate the collection, content retrieval and study of visual data. However, the ultimate objective is to highlight the importance of including other sources of information, like visuals, in the study of political issues. The social sciences are the *hard sciences* (Gill and Torres 2019) given the high complexity of the phenomena that we analyze. Thus, it is important to make use of the increasing feasibility to access methods and data that allow us to have a rich understanding of the world.

Bibliography

Anastasopoulos, Lefteris and Jake Williams. 2016. "Identifying violent protest activity with scalable machine learning." Working paper.

Arandjelović, Relja and Andrew Zisserman. 2012. Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE pp. 2911–2918.

Azoulay, Ariella. 2008. *The Civil Contract of Photography.* New York: Zone Books.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76–91.

Barnes, Samuel H, Klaus R Allerbeck, Barbara G Farah, Felix J Heunks, Ronald F Inglehart, M Kent Jennings, Hans Dieter Klingemann, Alan Marsh and Leopold Rosenmayr. 1979. *Political action: Mass participation in five western democracies.* Beverly Hills, CA: Sage.

Barry, Ann Marie. 1997. *Visual intelligence: Perception, image, and manipulation in visual communication.* SUNY Press.

Bauer, Nichole M and Colleen Carpinella. 2018. "Visual Information and Candidate Evaluations: The Influence of Feminine and Masculine Images on Support for Female Candidates." *Political Research Quarterly* 71(2):395–407.

Bay, Herbert, Tinne Tuytelaars and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision.* Springer pp. 404–417.

Benford, Robert D and David A Snow. 2000. "Framing processes and social movements: An overview and assessment." *Annual Review of Sociology* 26(1):611–639.

Bengio, Yoshua. 2012. "Practical recommendations for gradient-based training of deep architectures." *CoRR* abs/1206.5533.
URL: *https://dblp.org/rec/bib/journals/corr/abs-1206-5533*

Brunyé, Tad T, Jessica L Howe and Caroline R Mahoney. 2014. "Seeing the crowd for the bomber: Spontaneous threat perception from static and randomly moving crowd simulations." *Journal of experimental psychology: applied* 20(4):303.

Buda, Mateusz, Atsuto Maki and Maciej A Mazurowski. 2018. "A systematic study of the class imbalance problemin convolutional neural networks." *Neural Networks* (106):249–259.

Buduma, Nikhil. 2017. *Fundamentals of Deep Learning.* O'Reilly Media.

Butz, David A. 2009. "National symbols as agents of psychological and social change." *Political Psychology* 30(5):779–804.

Butz, David A, E Ashby Plant and Celeste E Doerr. 2007. "Liberty and justice for all? Implications of exposure to the US flag for intergroup relations." *Personality and Social Psychology Bulletin* 33(3):396–408.

Cameron, James E and Shannon L Nickerson. 2009. "Predictors of Protest Among Anti-Globalization Demonstrators 1." *Journal of Applied Social Psychology* 39(3):734–761.

Campbell, David. 2004. "Horrific blindness: Images of death in contemporary media." *Journal for Cultural Research* 8(1):55–74.

Campbell, John L. 2005. "Where do we stand." *Social movements and organization theory* pp. 41–68.

Canclini, Antonio, Matteo Cesana, Alessandro Redondi, Marco Tagliasacchi, João Ascenso and R Cilla. 2013. Evaluation of low-complexity visual feature detectors and descriptors. In *2013 18th International Conference on Digital Signal Processing (DSP)*. IEEE pp. 1–7.

Cantú, Francisco. Forthcoming. "The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election." *American Political Science Review* .

Casas, Andreu and Nora Webb Williams. 2018. "Images that matter: Online protests and the mobilizing role of pictures." *Political Research Quarterly* .

Challú, Cristian, Enrique Seira and Alberto Simpser. 2018. "The Quality of Vote Tallies: Causes and Consequences." Working Paper.

Chan, Philip K and Salvatore J Stolf. 1998. Toward scalable learning with non-uniform classand cost distributions: A case study in credit card fraud detection. In *KDD 1998*.

Chatfield, Ken, Karen Simonyan, Andrea Vedaldi and Andrew Zisserman. 2014. "Return of the Devil in the Details: Delving Deep into Convolutional Nets." Working paper.

Chenoweth, Erica and Maria J Stephan. 2011. *Why civil resistance works: The strategic logic of nonviolent conflict.* New York: Columbia University Press.

Cho, Jaeho, Michael P Boyle, Heejo Keum, Mark D Shevy, Douglas M McLeod, Dhavan V Shah and Zhongdang Pan. 2003. "Media, terrorism, and emotionality: Emotional differences in media content and public reactions to the September 11th terrorist attacks." *Journal of Broadcasting & Electronic Media* 47(3):309–327.

Chong, Dennis. 1991. *Collective action and the civil rights movement.* Chicago, IL: University of Chicago Press.

Chong, Dennis. 1996. Creating common frames of reference on political issues. In *Political persuasion and attitude change*, ed. Diana Carole Mutz, Paul M Sniderman and Richard A Brody. Ann Arbor, MI: University of Michigan Press chapter 8, pp. 1995–224.

Chong, Dennis and James N Druckman. 2007. "A theory of framing and opinion formation in competitive elite environments." *Journal of Communication* 57(1):99–118.

Corrigall-Brown, Catherine and Rima Wilkes. 2012. "Picturing protest: The visual framing of collective action by First Nations in Canada." *American Behavioral Scientist* 56(2):223–243.

Costain, Anne N and Steven Majstorovic. 1994. "Congress, social movements and public opinion: multiple origins of women's rights legislation." *Political Research Quarterly* 47(1):111–135.

Csurka, Gabriella, Christopher Dance, Lixin Fan, Jutta Willamowski and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV.* Vol. 1 Prague pp. 1–2.

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. "Information equivalence in survey experiments." *Political Analysis* 26(4):399–416.

Dalton, Russell J. 1988. *Citizen Politics in Western Democracies: Public Opinion and Political Parties in the United States, Great Britain, WestGermany, and France.* Chatham House Publishers.

Das, Sanmay, Allen Lavoie and Malik Magdon-Ismail. 2016. "Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion." *ACM Transactions on the Web (TWEB)* 10(4):24.

Davenport, Christian. 2009. *Media bias, perspective, and state repression: The Black Panther Party.* New York: Cambridge University Press.

Della Porta, Donatella and Mario Diani. 2009. *Social movements: An introduction.* Malden, MA: John Wiley & Sons.

Di Cicco, Damon T. 2010. "The public nuisance paradigm: Changes in mass media coverage of political protest since the 1960s." *Journalism & Mass Communication Quarterly* 87(1):135–153.

Dietrich, Bryce. 2015. If a picture is worth a thousand words, what is a video worth? In *Exploring the C-SPAN Archives: Advancing the Research Agenda*, ed. Robert X Browning. Purdue University Press.

Dietrich, Bryce J, Ryan D Enos and Maya Sen. 2018. "Emotional arousal predicts voting on the US supreme court." *Political Analysis* pp. 1–7.

Dietrich, Bryce, Matthew Hayes and Diana O'Brien. 2019. "Pitch Perfect: Vocal pitch and the emotional intensity of congressional speech on women." Working Paper.

Domke, David, David Perlmutter and Meg Spratt. 2002. "The primes of our times? An examination of the 'power' of visual images." *Journalism* 3(2):131–159.

Downing, John DH. 2000. *Radical media: Rebellious communication and social movements.* Sage.

Druckman, James N. 2003. "The power of television images: The first Kennedy-Nixon debate revisited." *The Journal of Politics* 65(2):559–571.

Druckman, James N and Kjersten R Nelson. 2003. "Framing and deliberation: How citizens' conversations limit elite influence." *American Journal of Political Science* 47(4):729–745.

Druckman, James N. and Michael Parkin. 2005. "The Impact of Media Bias: How Editorial Slant Affects Voters." *The Journal of Politics* 67(4):1030–1049.

Earl, Jennifer. 2003. "Tanks, tear gas, and taxes: Toward a theory of movement repression." *Sociological Theory* 21(1):44–68.

Earl, Jennifer. 2006. "Introduction: Repression and the social control of protest." *Mobilization* 11(2):129–143.

Earl, Jennifer, Andrew Martin, John D McCarthy and Sarah A Soule. 2004. "The use of newspaper data in the study of collective action." *Annual Review of Sociology* 30:65–80.

Edensor, Tim. 2013. "Reconnecting with darkness: gloomy landscapes, lightless places." *Social & Cultural Geography* 14(4):446–465.

Edensor, Tim. 2015. "The gloomy city: Rethinking the relationship between light and dark." *Urban Studies* 52(3):422–438.

Ehrlinger, Joyce, E Ashby Plant, Richard P Eibach, Corey J Columb, Joanna L Goplen, Jonathan W Kunstman and David A Butz. 2011. "How exposure to the confederate flag affects willingness to vote for Barack Obama." *Political Psychology* 32(1):131–146.

Erisen, Cengiz, Milton Lodge and Charles S Taber. 2014. "Affective contagion in effortful political thinking." *Political Psychology* 35(2):187–206.

Farris, Emily M. and Heather Silber Mohamed. Forthcoming. "Picturing immigration: how the media criminalizes immigrants." *Politics, Groups, and Identities* .

Feng, Yansong and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics pp. 831–839.

Fiske, John and Black Hawk Hancock. 2016. *Media matters: Race & gender in US politics*. London: Routledge.

Franklin, James. 2015. "Persistent Challengers: Repression, Concessions, Challenger Strength, and Commitment in Latin America." *Mobilization* 20(1):61–80.

Gamson, William A. 1989. "News as framing: Comments on Graber." *American Behavioral Scientist* 33(2):157–161.

Gamson, William A and Andre Modigliani. 1989. "Media discourse and public opinion on nuclear power: A constructionist approach." *American Journal of Sociology* 95(1):1–37.

Gamson, William A and Gadi Wolfsfeld. 1993. "Movements and media as interacting systems." *The Annals of the American Academy of Political and Social Science* 528(1):114–125.

Gentzkow, Matthew and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78(1):35–71.

Gerber, Alan S, Dean Karlan and Daniel Bergan. 2009. "Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions." *American Economic Journal: Applied Economics* 1(2):35–52.

Gill, Jeff and Michelle Torres. 2019. *Generalized linear models: a unified approach*. Vol. 134 2 ed. Sage Publications.

Giugni, Marco G. 1998. "Was it worth the effort? The outcomes and consequences of social movements." *Annual Review of Sociology* 24(1):371–393.

Grauman, K and T Darrell. 2005. "The pyramid match kernel: Discriminative classification with sets of image features. ICCV (pp. 1458–1465)." *IEEE Computer Society* .

Grauman, Kristen and Bastian Leibe. 2011. Visual object recognition. In *Synthesis lectures on artificial intelligence and machine learning*. Vol. 5 Morgan & Claypool Publishers pp. 1–181.

Grauman, Kristen and Trevor Darrell. 2007. "The pyramid match kernel: Efficient learning with sets of features." *Journal of Machine Learning Research* 8(Apr):725–760.

Green, Melissa J and Mary L Phillips. 2004. "Social threat perception and the evolution of paranoia." *Neuroscience & Biobehavioral Reviews* 28(3):333–342.

Griffin, Michael. 2012. "Images from nowhere: Visuality and news in 21st century media." *Visual Cultures-Transatlantic Perspectives* pp. 197–228.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* pp. 267–297.

Hainmueller, Jens and Daniel J Hopkins. 2014. "Public attitudes toward immigration." *Annual Review of Political Science* 17:225–249.

Hjerm, Mikael. 2007. "Do numbers really count? Group threat theory revisited." *Journal of Ethnic and Migration Studies* 33(8):1253–1275.

Hockley, William E. 2008. "The picture superiority effect in associative recognition." *Memory and Cognition* 36(7):1351–1359.

Homola, Jonathan. 2018. "The Political Consequences of Group-Based Identities." Working Paper.

Homola, Jonathan and Margit Tavits. 2018. "Contact reduces immigration-related fears for leftist but not for rightist voters." *Comparative Political Studies* 51(13):1789–1820.

Howe, Peter. 2002. *Shooting under fire: The world of the war photographer.* New York: Artisan.

Huang, Sheng-Jun, Rong Jin and Zhi-Hua Zhou. 2014. "Active Learning by Querying Informative and Representative Examples." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(10):1936–1949.

Huesmann, L Rowell. 2007. "The impact of electronic media violence: Scientific theory and research." *Journal of Adolescent health* 41(6):S6–S13.

Huff, Connor D. 2018. "Why Rebels Reject Peace." Working Paper.

Ioffe, Sergey and Christian Szagedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Technical report arXiv:1502.03167.

Iyengar, Shanto. 1994. *Is anyone responsible?: How television frames political issues.* Chicago, IL: University of Chicago Press.

Iyengar, Shanto and Donald R Kinder. 2010. *News that matters: Television and American opinion.* University of Chicago Press.

Iyengar, Shanto and Kyu S Hahn. 2009. "Red media, blue media: Evidence of ideological selectivity in media use." *Journal of Communication* 59(1):19–39.

Japkowicz, Nathalie and Shaju Stepehn. 2002. "The Class Imbalance Problem: A Systematic Study." *Intelligent Data Analysis* 6(5):429–449.

Keith, Susan, Carol B Schwalbe and B William Silcock. 2006. "Images in ethics codes in an era of violence and tragedy." *Journal of mass media ethics* 21(4):245–264.

Killian, Lewis M. 1964. "Social movements." *Handbook of modern sociology* pp. 426–55.

King Jr, Martin Luther. 1986. *Behind the Selma march.* New York: Harper and Row.

Kitschelt, Herbert P. 1986. "Political opportunity structures and political protest: Antinuclear movements in four democracies." *British journal of political science* 16(1):57–85.

Knox, Dean and Christopher Lucas. 2019. "A Dynamic Model of Speech for the Social Sciences." Working paper.

Kress, Gunther R, Theo Van Leeuwen et al. 1996. *Reading images: The grammar of visual design.* Psychology Press.

Kriesi, Hanspeter. 1995. *New social movements in Western Europe: A comparative analysis.* Vol. 5 Minneapolis, MN: University of Minnesota Press.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.* pp. 1097–1105.

Kubat, Miroslav, Robert C. Holte and Stan Matwin. 1998. "Machine learning for the detection ofoil spills in satellite radar images." *Machine Learning* 30(2-3):195–215.

Lahav, Gallya and Marie Courtemanche. 2012. "The ideological effects of framing threat on immigration and civil liberties." *Political Behavior* 34(3):477–505.

Lawson, Chappell and James A. McCann. 2004. "Television News, Mexico's 2000 Elections and Media Effects in Emerging Democracies." *British Journal of Political Science* 35:1–30.

Lecheler, Sophie, Andreas R T Schuck and Claes H de Vreese. 2013. "Dealing with feelings: Positive and negative discrete emotions as mediators of news framing effects." *Communications* 38(2):189–209.

Lecheler, Sophie and Claes H de Vreese. 2013. "What a difference a day makes? The effects of repetitive and competitive news framing over time." *Communication Research* 40(2):147–175.

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard and Lawrence D Jackel. 1989. "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1(4):541–551.

LeCun, Yann, Léon Bottou, Yoshua Bengio, Patrick Haffner et al. 1998. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86(11):2278–2324.

LeCun, Yann, Yoshua Bengio et al. 1995. "Convolutional networks for images, speech, and time series." *The handbook of brain theory and neural networks* 3361(10):1995.

LeDoux, Joseph E. 1986. "Sensory systems and emotion: A model of affective processing." *Integrative psychiatry* .

Lee, Francis LF. 2014. "Triggering the protest paradigm: Examining factors affecting news coverage of protests." *International Journal of Communication* 8:2725–2746.

Levendusky, Matthew and Neil Malhotra. 2016. "Does media coverage of partisan polarization affect political attitudes?" *Political Communication* 33(2):283–301.

Linfield, Susie. 2011. *The cruel radiance: photography and political violence.* Chicago, IL: University of Chicago Press.

Lowe, David G. 1999. Object recognition from local scale-invariant features. In *Proccedings of the International Conference on Computer Vision.* IEEE Corfu: p. 1150.

Lucas, Christopher. 2018. "Neural Networks for the Social Sciences." Working Paper.

Lyman, Peter and Hal R. Varian. 2001. "The democratization of data." *Harvard Business Review* 79(1):137–139.

Madigan, Stephen. 1983. Picture Memory. In *Imagery, Memory and Cognition: Essays in Honor of Allan Paivio*, ed. John C. Yuille. Lawrence Erlbaum Associates pp. 65–90.

Makin, David A., Dale W. Willits, Wendy Koslicki, Rachael Brooks, Bryce J. Dietrich and Rachel L. Bailey. Forthcoming. "Contextual Determinants of Observed Negative Emotional States in Police-Community Interactions." *Criminal Justice and Behavior* .

Massey, Doreen. 1995. "Thinking radical democracy spatially." *Environment and Planning D: Society and Space* 13(3):283–288.

Masters, Dominic and Carlo Luschi. 2018. "Revisiting Small Batch Training for Deep Neural Networks." *CoRR* abs/1804.07612.
URL: *https://dblp.org/rec/bib/journals/corr/abs-1804-07612*

McCarthy, John D, Clark McPhail, Jackie Smith and Louis Crishock. 1999. "Electronic and print media representations of Washington DC demonstrations, 1982 and 1991.".

McCarty, Nolan, Keith T. Poole and Howard Rowenthal. 2006. *Polarized America.* The MIT Press.

McHugh, Joanna Edel, Rachel McDonnell, Carol O'Sullivan and Fiona N Newell. 2010. "Perceiving emotion in crowds: the role of dynamic body postures on the perception of emotion in crowded scenes." *Experimental brain research* 204(3):361–372.

Mendelberg, Tali. 1997. "Executing Hortons: Racial crime in the 1988 presidential campaign." *The Public Opinion Quarterly* 61(1):134–157.

Mendelberg, Tali. 2001. *The race card: Campaign strategy, implicit messages, and the norm of equality.* Princeton University Press.

Meyers, Marian. 1996. *News coverage of violence against women: Engendering blame.* Sage Publications.

Mikolajczyk, Krystian and Cordelia Schmid. 2005. "A performance evaluation of local descriptors." *IEEE transactions on pattern analysis and machine intelligence* 27(10):1615–1630.

Moeller, Susan D. 2018. Compassion fatigue. In *Visual Global Politics*, ed. Ronald Bleiker. New York: Routledge pp. 87–92.

Monay, Florent and Daniel Gatica-Perez. 2007. "Modeling semantic aspects for cross-media image indexing." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10):1802–1817.

Monay, Florent, Pedro Quelhas, Jean-Marc Odobez and Daniel Gatica-Perez. 2009. "Contextual classification of image patches with latent aspect models." *Journal on Image and Video Processing* 2009:3.

Monogan, James E and Jeff Gill. 2016. "Measuring State and District Ideology with Spatial Realignment." *Political Science Research and Methods* 4(1):97–121.

Montgomery, Jacob M and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Morris, Nina J. 2011. "Night walking: darkness and sensory perception in a night-time landscape installation." *cultural Geographies* 18(3):315–342.

Muñoz, Jordi and Eva Anduiza. 2019. "'If a fight starts, watch the crowd': The effect of violence on popular support for social movements." *Journal of Peace Research* pp. 1–14.

Nair, Vinod and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML'10 Proceedings of the 27th International Conference on International Conference on Machine Learning*, ed. Johannes Fürnkranz and Thorsten Joachims. pp. 807–814.

Nelson, Thomas E, Rosalee A Clawson and Zoe M Oxley. 1997. "Media framing of a civil liberties conflict and its effect on tolerance." *American Political Science Review* 91(3):567–583.

Newton, Kenneth. 1999. "Mass media effects: mobilization or media malaise?" *British Journal of Political Science* 29(4):577–599.

116

Nguyen, Anh, Jason Yosinski and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp. 427–436.

Noakes, John A and Hank Johnston. 2005. *Frames of protest: A road map to a perspective.* Lanham, MD: Rowman & Littlefield.

Öhman, Arne, Daniel Lundqvist and Francisco Esteves. 2001. "The face in the crowd revisited: a threat advantage with schematic stimuli." *Journal of personality and social psychology* 80(3):381.

Oliver, Pamela E and Daniel J Myers. 1999. "How events enter the public sphere: Conflict, location, and sponsorship in local newspaper coverage of public events." *American Journal of Sociology* 105(1):38–87.

Oliver, Pamela E and Gregory M Maney. 2000. "Political processes and local newspaper coverage of protest events: From selection bias to triadic interactions." *American Journal of Sociology* 106(2):463–505.

Parry, Katy. 2011. "Images of liberation? Visual framing, humanitarianism and British press photography during the 2003 Iraq invasion." *Media, Culture & Society* 33(8):1185–1201.

Pickerill, Jenny and Paul Chatterton. 2006. "Notes towards autonomous geographies: creation, resistance and self-management as survival tactics." *Progress in human geography* 30(6):730–746.

Polletta, Francesca and James M Jasper. 2001. "Collective identity and social movements." *Annual review of Sociology* 27(1):283–305.

Poole, Keith T. and Howard Rosenthal. 1997. *Ideology and Congress.* Transcaction Publishers.

Pratto, Felicia, Jim Sidanius, Lisa M Stallworth and Bertram F Malle. 1994. "Social dominance orientation: A personality variable predicting social and political attitudes." *Journal of Personality and Social Psychology* 67(4):741.

Qin, Zhuwei, Fuxun Yu, Chenchen Liu and Xiang Chen. 2018. "How Convolutional Neural Networks See the World — A Survey of Convolutional Neural Network Visualization Methods." *Mathematical Foundations of Computing* 1(2):149–180.

Quillian, Lincoln. 1995. "Prejudice as a response to perceived group threat: Population composition and anti-immigrant and racial prejudice in Europe." *American sociological review* 60(4):586–611.

Reinhardt, Mark. 2012. Painful photographs: From the ethics of spectatorship to visual politics. In *Ethics and images of pain.* Routledge pp. 57–80.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.

Rohrschneider, Robert. 1990. "The roots of public opinion toward new social movements: An empirical test of competing explanations." *American Journal of Political Science* pp. 1–30.

Rosenholtz, Ruth, Yuanzhen Li, Jonathan Mansfield and Zhenlan Jin. 2005. Feature congestion: a measure of display clutter. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM pp. 761–770.

Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams. 1988. "Learning representations by back-propagating errors." *Cognitive Modeling* 5(3).

Sabour, Sara, Nicholas Frosst and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems.* pp. 3856–3866.

Schemer, Christian. 2012. "The influence of news media on stereotypic attitudes toward immigrants in a political campaign." *Journal of Communication* 62(5):739–757.

Schneider, Silke L. 2008. "Anti-immigrant attitudes in Europe: Outgroup size and perceived ethnic threat." *European Sociological Review* 24(1):53–67.

Schrodt, Philip A. 2004. "Patterns, rules and learning: Computational models of international behavior." Working Paper.

Settles, Burr. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648 University of Wisconsin–Madison.

Shaw, Robert. 2015. "Controlling darkness: self, dark and the domestic night." *Cultural Geographies* 22(4):585–600.

Shaw, Robert. 2017. "Pushed to the margins of the city: The urban night as a timespace of protest at Nuit Debout, Paris." *Political Geography* 59:117–125.

Sidanius, Jim, James H Liu, John S Shaw and Felicia Pratto. 1994. "Social dominance orientation, hierarchy attenuators and hierarchy enhancers: Social dominance theory and the criminal justice system." *Journal of Applied Social Psychology* 24(4):338–366.

Simonyan, Karen and Andrew Zisserman. 2014. "Very deep convolutional networks for large-scale image recognition." *arXiv* .

Sinclair, Betsy. 2012. *The social citizen: Peer networks and political behavior.* Chicago, IL: University of Chicago Press.

Sivic, Josef and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *null.* IEEE p. 1470.

Sivic, Josef, Bryan C Russell, Alexei A Efros, Andrew Zisserman and William T Freeman. 2005. Discovering objects and their location in images. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 1 IEEE pp. 370–377.

Smith, Jackie, John D McCarthy, Clark McPhail and Boguslaw Augustyn. 2001. "From protest to agenda building: Description bias in media coverage of protest events in Washington, DC." *Social Forces* 79(4):1397–1423.

Sniderman, Paul M, Louk Hagendoorn and Markus Prior. 2004. "Predisposing factors and situational triggers: Exclusionary reactions to immigrant minorities." *American political science review* 98(1):35–49.

Sontag, Susan. 1977. *On Photography.* New York: Farrar, Straus and Giroux.

Soule, Sarah and Jennifer Earl. 2005. "A movement society evaluated: Collective protest in the United States, 1960-1986." *Mobilization* 10(3):345–364.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15:1929–1958.

Taylor, John. 1998. *Body horror: Photojournalism, catastrophe, and war.* New York: NYU Press.

Tello, Carlos. 2007. *2 de Julio: La Crónica Minuto a Minuto Del Día Más Importante de Nuestra Historia Contemporánea.* Mexico City: Editorial Planeta Mexicana.

Torres, Michelle. 2018. "Framing a Protest: Determinants and Effects of Visual Frames." Working Paper.

Tucker, Patrick D, Jacob M Montgomery and Steven S Smith. 2018. "Party Identification in the Age of Obama: Evidence on the Sources of Stability and Systematic Change in Party Identification from a Long-Term Panel Survey." *Political Research Quarterly* p. 1065912918784215.

Valentino, Nicholas A, Vincent L Hutchings and Ismail K White. 2002. "Cues that matter: How political ads prime racial attitudes during campaigns." *American Political Science Review* 96(1):75–90.

Valenzuela, Sebastián. 2013. "Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism." *American Behavioral Scientist* 57(7):920–942.

Vigo, David Augusto Rojas, Fahad Shahbaz Khan, Joost Van De Weijer and Theo Gevers. 2010. The impact of color on bag-of-words based object recognition. In *2010 20th international conference on pattern recognition*. IEEE pp. 1549–1553.

Wilmott, Annabelle Cathryn. 2017. "The Politics of Photography: Visual Depictions of Syrian Refugees in U.K. Online Media." *Visual Communication Quarterly* 24(2):67–82.

Wolfsfeld, Gadi. 1991. "Media, Protest, and Political Violence: A Transactional Analysis." *Journalism and Communication Monographs* 127.

Won, Donghyeon, Zachary C Steinert-Threlkeld and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM pp. 786–794.

Zajonc, Robert B. 1984. "On the primacy of affect." *American Psychologist* 39(2):117–123.

Zeiler, Matthew D and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer pp. 818–833.

Zelizer, Barbie. 2004. When war is reduced to a photograph. In *Reporting War*. Routledge pp. 125–145.

Zhang, Han and Jennifer Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media.".

Appendix A

The Bag of Visual Words: Using computer vision to understand visual frames and political communication

A.1 Key-point detection

In order to identify key points while preserving the scale invariance property, the FAST Hessian relies on the approximation of the Hessian matrix of a scale-space function, where space is measured by $\mathbf{x} = (x, y)$, and scale by σ. Let $I(x, y)$ be the intensity of the pixel located at coordinates (x, y). Ideally, the process starts by calculating the second order partial derivatives of the image, by convoluting it with a second order scale normalized Gaussian kernel. Thus, the "ideal" Hessian matrix has the form:

$$\mathcal{H}(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}; \sigma) & L_{xy}(\mathbf{x}; \sigma) \\ L_{xy}(\mathbf{x}; \sigma) & L_{yy}(\mathbf{x}; \sigma) \end{bmatrix},$$

where, for example, $L_{xy}(\mathbf{x}; \sigma)$ is the convolution of the Gaussian second order derivative, $\frac{\partial^2 g(\sigma)}{\partial x^2}$, with the image I in point $\mathbf{x}$.[1] The determinant of the Hessian of each pixel will then be used to determine salient points. However, the estimation of this Hessian is computationally expensive, especially as the size of the kernel grows. Thus, Bay et al. (2006), proposed an approximation of the second derivative kernels by using "box filter" representations of those matrices. Figure A.1 illustrates the original and approximated filters.

Figure A.1: Original second order derivative Gaussian filters and approximations
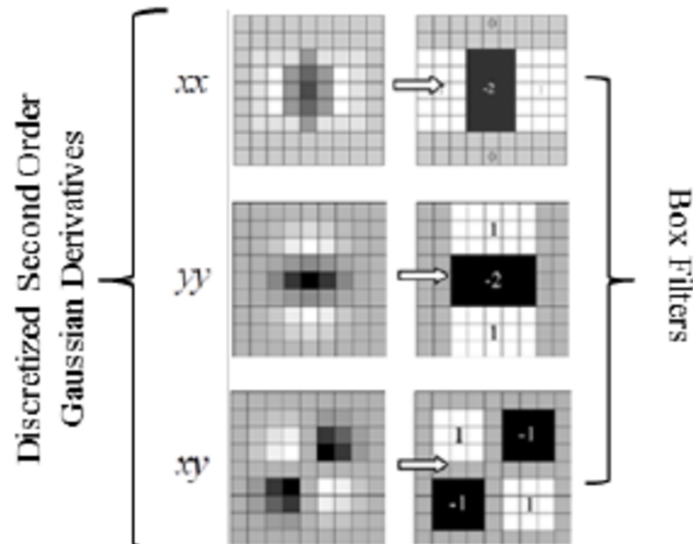
These box filter approximations of L_{xx}, L_{xy} and L_{yy}, denoted as D_{xx}, D_{xy} and D_{yy} increase efficiency and speed considerably, and allows us to estimate the determinant of the approximated Hessian as follows:

$$det(\mathcal{H}_{approx}) = D_{xx}D_{yy} - (0.9D_{xy})^2$$

In order to detect key points, we will build layers of the image by using increasing sizes of kernels as a way of varying the scale of the original picture (for example, the smallest kernels possible of size 9×9, will correspond to a real valued Gaussian with $\sigma = 1.2$). Once we build

---

[1]Where $g(\sigma)$ is the pdf of a normal distribution with $\mu = 0$ and standard deviation σ.

this scale-space 3D structure, a maximal suppression is performed to find the salient points. In other words, a pixel is considered a key point if its intensity is higher than the one of its 26 neighbors, comprised in the $3\times3\times3$ cube that surrounds it: 8 along the x and y axis plane, and 9 across scale layers. The final step involves interpolation of the data surrounding the key points in order to reach sub-pixel accuracy. Figure 2.2 shows an example of the key points that are found in one of the images in my sample. The green circles represent the coordinates of the key points. The figure illustrates how most of the key points are representing edges, corners or regions where color changes significantly. Once the key points are identified, as in the case of this image, we proceed to extract its features.

A.2 STM results: Media outlets dataset

The STM was initialized with 5 topics and 2 prevalence covariates: the ideological leaning of the news outlet, as captured by *All sides*, and the date in which the news article was published. The most representative images per topic, and most frequent and exclusive words are presented below.

Figure A.2: FREX Visual Words per Topic (Media model)

Topic 1:
Outdoor/Portraits

Topic 2:
Crowd

Topic 3:
Small Groups

Topic 4:
Shadows/
Darkness

Topic 5:
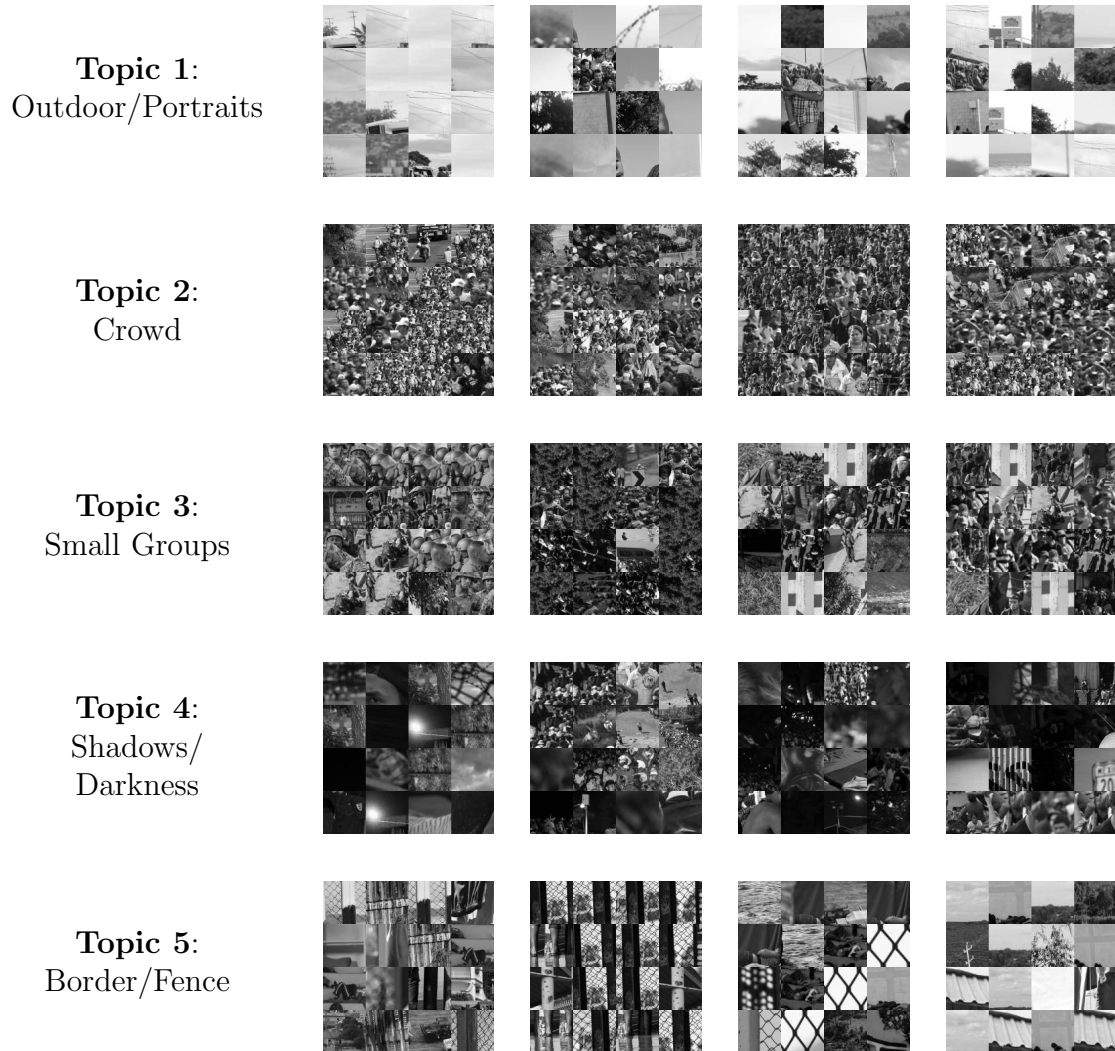Border/Fence

Table A.1: Most representative images per Topic (Media model)

Outdoor/Portraits

Crowd

Small groups

Shadows/Darkness

Border/Fence

Appendix B

Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data

B.1 Back-propagation

Suppose that a neuron j in the last layer provides a classification outcome y_j.[1] To estimate the prediction error, the model compares such an outcome with the target label, t_j. In our digit recognition example, the prediction error of the neuron for the outcome "1" is the difference between the true outcome and the model's estimated probability for the image to belong to that category. After adding up the prediction error of all the neurons in the layer, $E = \frac{1}{2} \sum_{j \in 10} (t_j - y_j)^2$, we can estimate the error function derivative of the last layer:

$$\frac{\partial E}{\partial y_j} = -(t_j - y_j) \tag{B.1}$$

Similarly, we can express the error derivatives in terms of the logit of the neuron, z_j:

---

[1]The explanation and notation of this example come from Buduma (2017).

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial z_j} = y_j(1 - y_j)\frac{\partial E}{\partial y_j} \tag{B.2}$$

To minimize this error term, the network goes back to its prior layers and identifies those weights contributing the most to this error. In other words, it estimates how the neuron outcomes in layer i affect the outputs of layer j given the weighted connection between both layers, w_{ij} :

$$\frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial y_i} = \sum_j w_{ij}\frac{\partial E}{\partial z_j} = \sum_j w_{ij}y_j(1 - y_j)\frac{\partial E}{\partial y_j} \tag{B.3}$$

These partial derivatives allow us to estimate the contribution of a specific weight to the error term:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial z_j}{\partial w_{ij}}\frac{\partial E}{\partial z_j} = y_j(1 - y_j)\frac{\partial E}{\partial y_i} \tag{B.4}$$

The partial derivative in Equation B.4 allows the model to gradually modify its weights after reviewing a set k of examples from the database K:

$$-\Delta w_{ij} = -\sum_{k \in K} y_i^{(k)}y_j^{(k)}(1 - y_j^{(k)})\frac{\partial E^{(k)}}{\partial y_i^{(k)}} \tag{B.5}$$

Appendix C

Framing a Protest: Determinants and Effects of Visual Frames

C.1 Descriptive statistics

Table C.1: Descriptive statistics (full sample)

Statistic	N	Mean/%	St. Dev.
Female	1,084	0.534	0.499
Age	2,031	45.287	16.613
White	1,382	0.681	0.466
African American	205	0.101	0.301
Hispanic	218	0.107	0.310
College or more	892	0.440	0.496
Democrat	732	0.360	0.480
Republican	603	0.297	0.457
Independent/Other	696	0.343	0.475
Ideology	2,028	3.755	1.838
SDO	2,018	34.469	12.217

C.2 Full vignette (Peaceful picture and text example)

THOUSANDS PROTEST POOR ECONOMIC CONDITIONS IN U.S. MARCHES

By THE EDITORS

In the past week, several demonstrations were held in dozens of cities including Washington, Boston, Chicago, Los Angeles, Miami and Toronto as part of the Occupy movement.

The protests are part of a series of international demonstrations that started in Asia and Europe and rippled around to the United States and Canada.

The Occupy movement has attracted thousands of people from multiple backgrounds denouncing inequality, and demanding better social and economic conditions.

The movement began with a commitment to peaceful demonstrations, and protesters behaved calm and orderly. The tone of the protests has been peaceful and subdued. Only a small number of minor disruptive events have been reported.

"The great thing about Occupy Wall Street is that they have brought the focus of the entire country on the middle-class majority," said George Aldro, 62. "We're in it together, and we're in it for the long haul."

The protesters have varied causes, but have spoken largely about unemployment and economic inequality, reserving most of their criticism for Wall Street. "We are the 99 percent," they chanted, contrasting themselves with the wealthiest 1 percent of Americans.

Source: AP Images

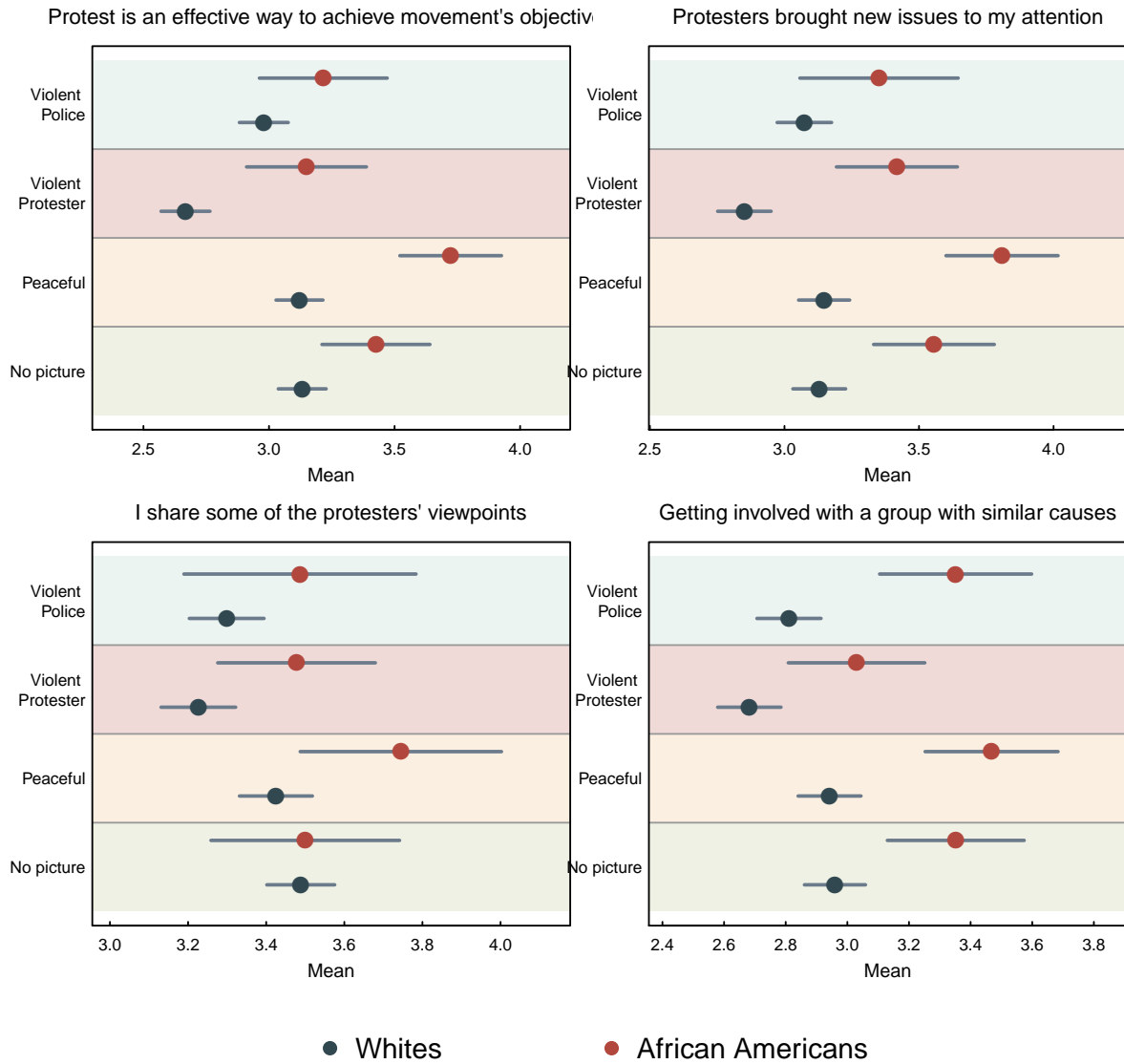**Protests demanding economic equality erupted throughout the country this past week.**

C.3 Wording

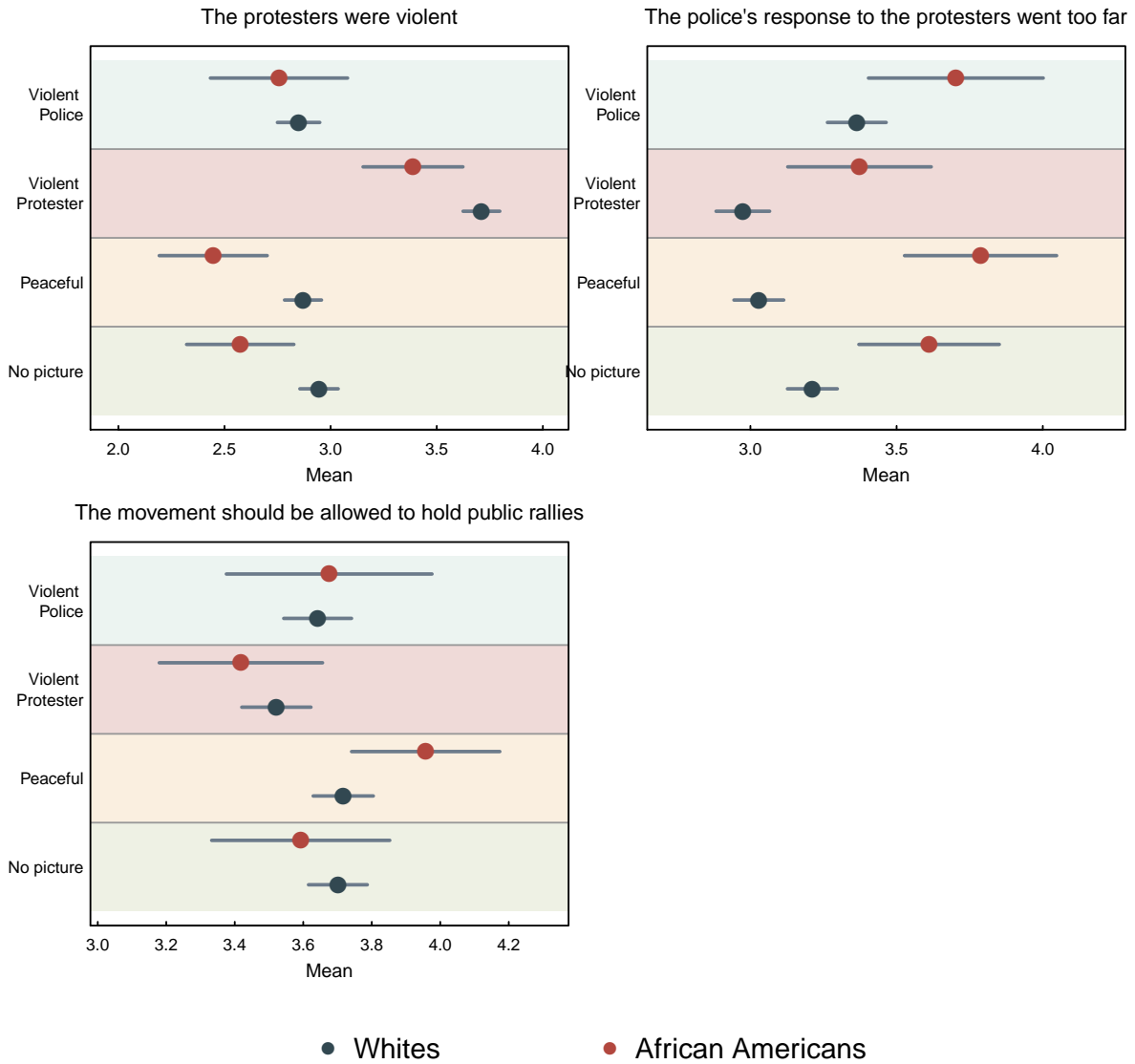| Name	Options	Values	Wording
TREAT		0= No picture 1= Peaceful protest 2= Violent protester 3= Violent police	
tol_*	blm = Black lives Matter whitesup = White supremacists occupy = Occupy movement bluelm = Blue Lives Matter	1= Strongly disagree 2= Disagree 3= Neither 4= Agree 5= Strongly agree	For each of the following groups, please indicate to which degree you agree or disagree with the statement: "This group should be allowed to hold public rallies and demonstrations"
therm_*	police = Police hisp = Hispanic muslims = Muslims whitesup = White supremacists	0-100	Rate each group or individual using the scale shown below, where 0 is the worst grade possible and 100 the best grade possible.
achieveobjs		1= Strongly disagree 2= Disagree 3= Neither 4= Agree 5= Strongly agree	To what degree do you agree or disagree with the following statements: "This protest is an effective way to achieve the Occupy movement's objectives."
broughtattn		1= Strongly disagree 2= Disagree 3= Neither 4= Agree 5= Strongly agree	To what degree do you agree or disagree with the following statements: "These protesters brought new issues to my attention."
shareprotviews		1= Strongly disagree 2= Disagree 3= Neither 4= Agree 5= Strongly agree	To what degree do you agree or disagree with the following statements: "I share some of the protesters' viewpoints."
involvegroup		1= Strongly disagree 2= Disagree 3= Neither 4= Agree 5= Strongly agree	To what degree do you agree or disagree with the following statements: "I would consider getting involved with a group who supported causes similar to those of the protesters."

protestersviolent	1= Strongly disagree 2= Disagree 3= Neither 4= Agree 5= Strongly agree	To what degree do you agree or disagree with the following statements: "The protesters were violent."
policetoofar	1= Strongly disagree 2= Disagree 3= Neither 4= Agree 5= Strongly agree	To what degree do you agree or disagree with the following statements: "The police's response to the protesters went too far."
occupyrally		To what degree do you agree or disagree with the following statements: "The Occupy Movement should be allowed to hold public rallies and demonstrations."
idcard		To what degree do you agree or disagree with the following statements: "There should be a policy requiring everyone to carry a national identity card at all times to show to a police officer on request."
investigatenonviolent		To what degree do you agree or disagree with the following statements: "There should be a policy allowing law enforcement officials to investigate people who participate in non-violent protests against the U.S. government."

C.4   Means of outcome variables, by race group

Figure C.1: The effects of directionality of visual violence on attitudes towards protests: White and African American respondents

The protesters were violent

The police's response to the protesters went too far

The movement should be allowed to hold public rallies

● Whites   ● African Americans

C.5 Vignettes for Experiment 2

Figure C.2: Visual treatment conditions (Experiment 2)

(a) Peaceful

(b) Violent protester

(c) Violent police

Table C.3: Textual treatment conditions (Experiment 2)

Peaceful	The movement began with a commitment to peaceful demonstrations, and protesters behaved calm and orderly. The tone of the protests has been peaceful and subdued. Only a small number of minor disruptive events have been reported.
Violent protester	The movement began with a commitment to peaceful demonstrations. However, some of the events turned violent when police clashed with attendants. Several protesters were arrested for violating city ordinances, unlawful occupation, attacks to the police, public disturbances and substance abuse.
Violent police	The movement began with a commitment to peaceful demonstrations. However, some of the events turned violent when police clashed with attendants. Police officers in riot gear met the mostly peaceful crowds with escalating force. There were several reports that the police officers attacked protesters with tear gas, rubber bullets and physical force.