Give me the full picture: Using computer vision to understand visual frames and political communication*

Michelle Torres

Rice University†

February 2, 2020

Abstract

Political communication is a central element of political dynamics. Its visual component is crucial in understanding the messages sent between elites, media, and citizens. However, visual features have been overlooked in political science. Thus, this article introduces a tool to dissect the content of visual material in order to assess its relationship with political variables: the Bag of Visual Words. The article details and validates the implementation of this technique for the extraction of visual features that allows researchers to build an Image-Visual Word matrix that emulates the Document-Term matrix in text analysis. I illustrate its applicability by focusing on the identification of visual frames using a visual topic model. More specifically, the article studies the different depictions of the caravan of Central American migrants and finds that in news related to the caravan, right-leaning media outlets are more likely to use pictures with higher proportions of topic "crowd."

Word count: 11,926

*I would like to thank Michael Bailey, Matt Blackwell, Sanmay Das, Bryce Dietrich, Katie Einstein, Justin Esarey, Emily Gade, Jeff Gill, Soeren Henn, Jonathan Homola, Gary King, Mayya Komisarchik, Chris Lucas, Jacob Montgomery, Betsy Sinclair, Alex Theodoridis, Dustin Tingley, members of the Political Data Science Lab at Washington University in St. Louis and attendants of the Applied Statistics Workshop at Harvard, New Faces in Political Methodology X, Polmeth XXXV, and the Methods Workshops at UCLA, Stanford and Iowa State for helpful comments.

†Assistant Professor, Department of Political Science, smtorres@rice.edu

1 Introduction

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 opinions of the same phenomenon warrant the question about what factors form them.

Several studies use the analysis of political messages to provide a better understanding of political events and behavior, as well as of attitude formation (Dilliplane, Goldman, and Mutz 2013). However, with a few recent exceptions using audio-visual material (Bauer and Carpinella 2018; Casas and Williams 2018; Dietrich, Enos, and Sen 2019; Knox and Lucas 2019; Lucas 2019; Mutz 2007), most of these studies focus solely on verbal communication and text analysis (Cho et al. 2003; Chong and Druckman 2007; Druckman and Nelson 2003; Gamson and Modigliani 1989; Grimmer and Stewart 2013; Lecheler and de Vreese 2013; Lecheler, Schuck, and de Vreese 2013). Meanwhile, visual material is an important element of human communication that has remained overlooked.

This omission is concerning given that vision is a crucial sense involved in information processing via both conscious and unconscious paths, and 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 offer some answers and a tool to answer these questions.

Take, for example, 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? Can we identify topical clusters among these images based on content similarities? 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. To address these, I introduce a computer vision 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. Further, I present a novel use of this dimension reduction technique in a semi-supervised setting,

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 (CNNs).

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.

2 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,

[1] "How many photographs of you are out there in the world?", Rose Eveleth, *The Atlantic*, November 2,

governments, and activists on social media as a means for communicating their messages increases the amount of visual material that individuals encounter.

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; Mutz 1998) 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 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 seeking refugee in the United States. The 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 through Mexico with the objective of reaching the U.S. and seeking asylum.[2]

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 and either advocate for the migrants, their rights and their safety, or evaluate them as a threatening source of crime and instability. Media coverage of the movement's

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).

[2]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.

activities also reflects the variability in the perceptions of the caravans, especially regarding the visuals used to illustrate news pieces. For example, on October 5^{th} of 2018, several news outlet covered the reaction of President Trump to the caravans and the threat he made to cut aid to countries such as Honduras if the governments did not stop the formation and flow of caravans. Several media outlets used the exact same text and facts from a news wire source to inform their audiences about the caravans. However, in contrast to the identical text, 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 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 1), and 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 outlets and this motivates the question: how can we quantify and explain these differences?

3 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

Figure 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*

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?

3.1 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 patch corresponding to 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 (Deselaers, Pimenidis, and Ney 2008; Yang et al. 2007; Zhang et al. 2009).

One of the main challenges that the analysis of visual material implies is the high dimensionality of the units of analysis. First, images contain a large number of 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 between images instead of raw pixel intensities helps to account for differences that might be independent from the actual content of the image, such as light 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 subsection I detail each of these steps.

3.2 Step 1: Extracting and describing local key points

The first step of the process of building a BoVW 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.[3] For the purposes of this article, I use the FAST Hessian detector and the RootSIFT descriptor that I detail below.

3.2.1 Detecting key points

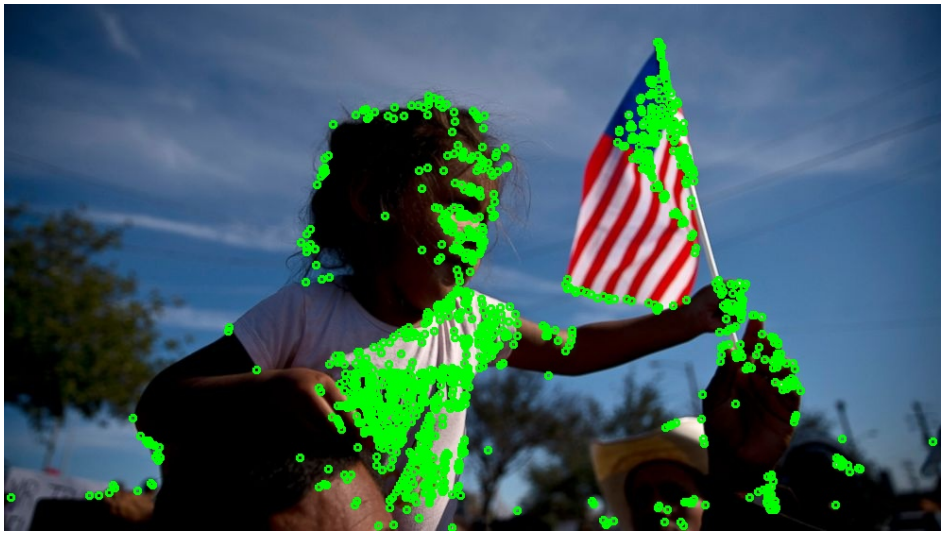
The FAST Hessian detector is used to locate edges and corners in an image (Bay, Tuytelaars, and Van Gool 2006). It is suitable for the purposes of this article given its two key properties: scale invariance (i.e. key points should be both repeatable and recognizable at different scales of the image), 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 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

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

elements of the picture.[4]

Figure 2: Location of key points

(a) Original image



(b) Image with key points identified

3.2.2 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

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

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 this 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.[5]

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 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,

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

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 proceeds as follows. First, for each of the key points identified in Section 3.2.1, 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 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 3), followed by a summary of this information with gradients (Panel (c) of Figure 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:

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

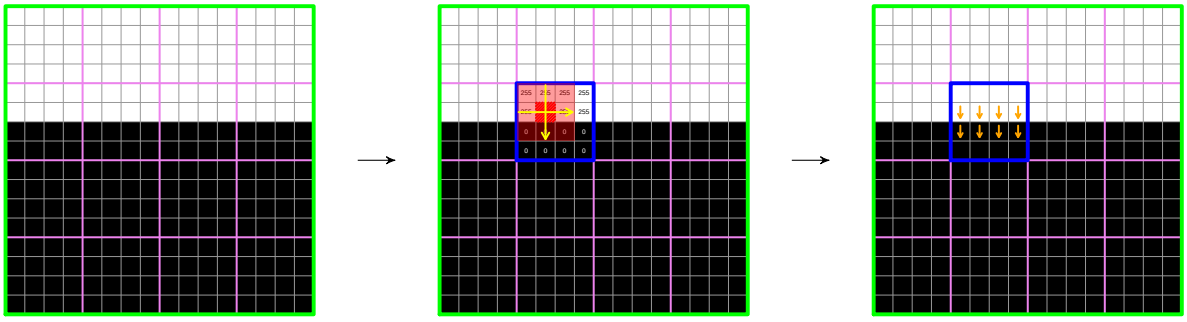
Then, we calculate the *magnitude* and the *orientation* as follows:

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

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 angles into 8 bins for the histograms. These angles are in the range of [0, 180] when unsigned[6]

[6]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 3: Computing pixel intensity changes in the neighborhood of a key point



Panel (a)
16 × 16 neighborhood
of key point

Panel (b)
Computation of
pixel intensity changes

Panel (c)
Representation of change
with gradients

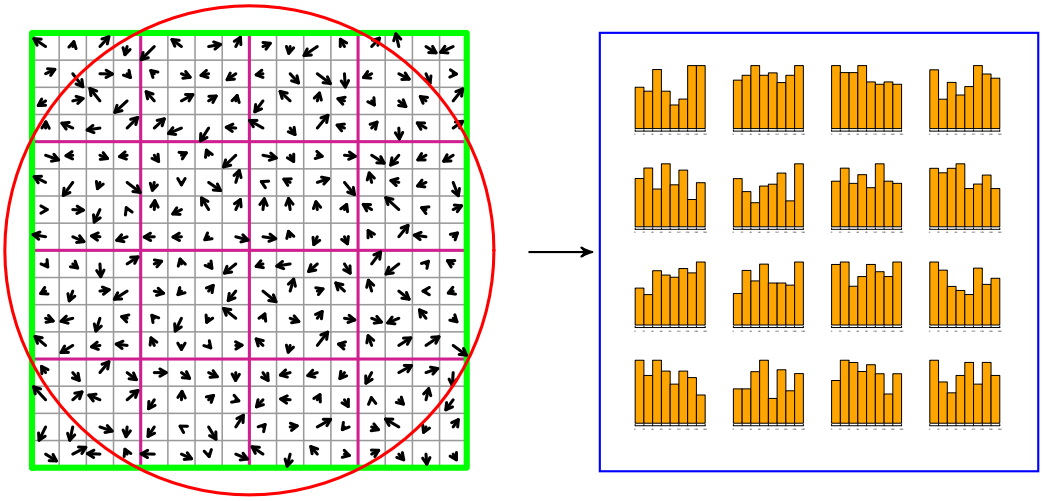
so we end up with bins that each include around 20 potential angles. Then, we count the number of orientation values that fall into each of the bins, and weight them by their respective magnitude, and the distance to the key point. In other words, stronger pixel changes that are closer to the key point 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 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.

3.3 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

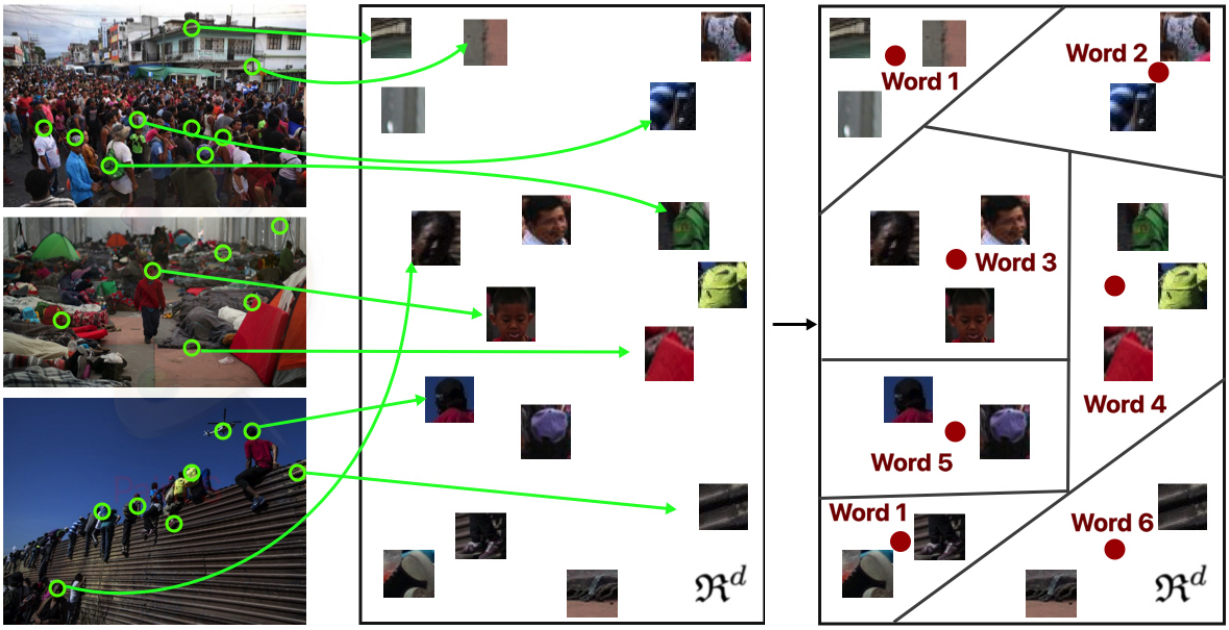
Figure 4: Representation of the neighborhood of the key point with histograms



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 5.

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

Figure 5: Creating the visual vocabulary: clustering and centroids



based on a random sample of the feature vectors.[7] 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 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 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.

3.4 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

[7] 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. Given the complexity of the images under analysis, especially in comparison to more standard canonical datasets, for all the models in this article, I sampled between 30 and 35% of the features.

Figure 6: Examples of visual words

(a) Water (b) Large groups of people

(c) Wired fence

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:

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

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.[8]

4 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 can 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 framing 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 metadata covering the author of the picture, source, caption, dimensions, and keywords associated with each image.

[8] An illustration is presented in the Appendix.

4.1 Detecting underlying messages

The BoVW can be the basis of both supervised and unsupervised exercises. 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 from this pool of images which I subsequently feed to a structural topic model initialized with six 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. The number of topics was selected based on a qualitative exploration of the data, expectations about the potential topics to find in the image pool, and a post-STM analysis to assess the composition and congruence

of the topics.

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 7 shows four 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 Figure 8).

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 a topic k are those photos with high proportions of such topic k. Table 8 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.

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

Figure 7: FREX Visual Words per Topic

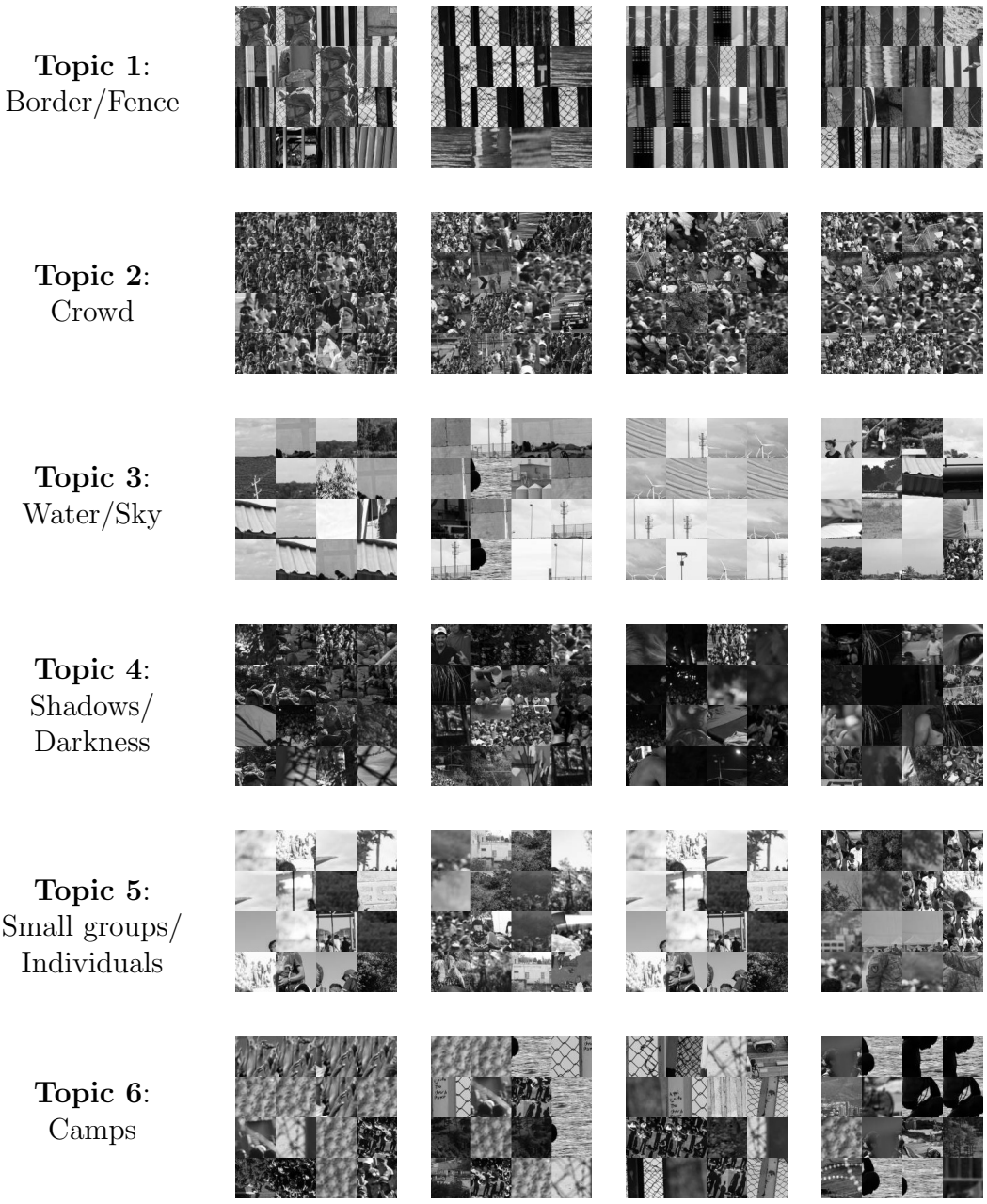
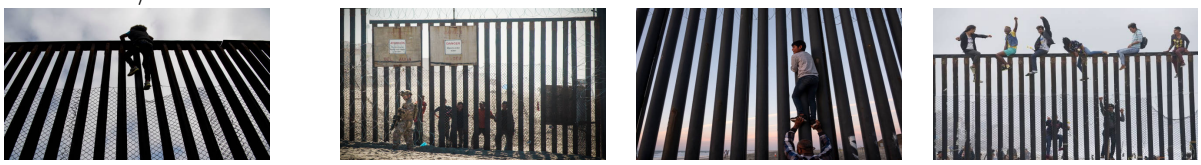


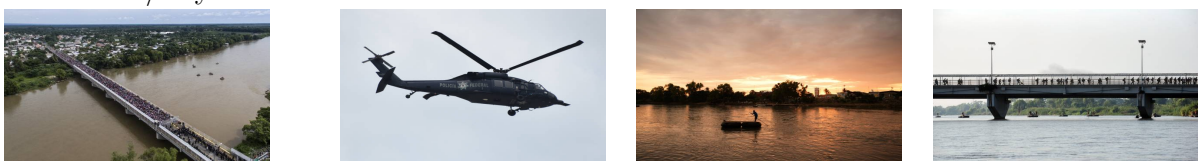
Figure 8: Most representative images per Topic

Border/Fence



Crowd

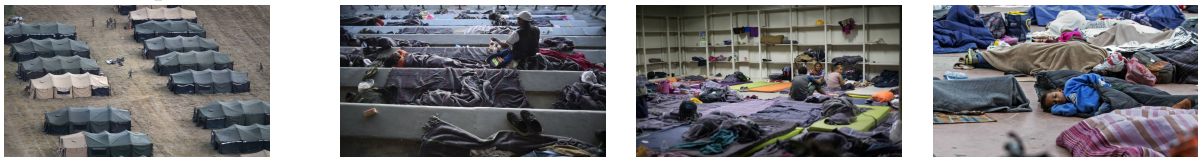
Water/Sky



Shadows/Darkness

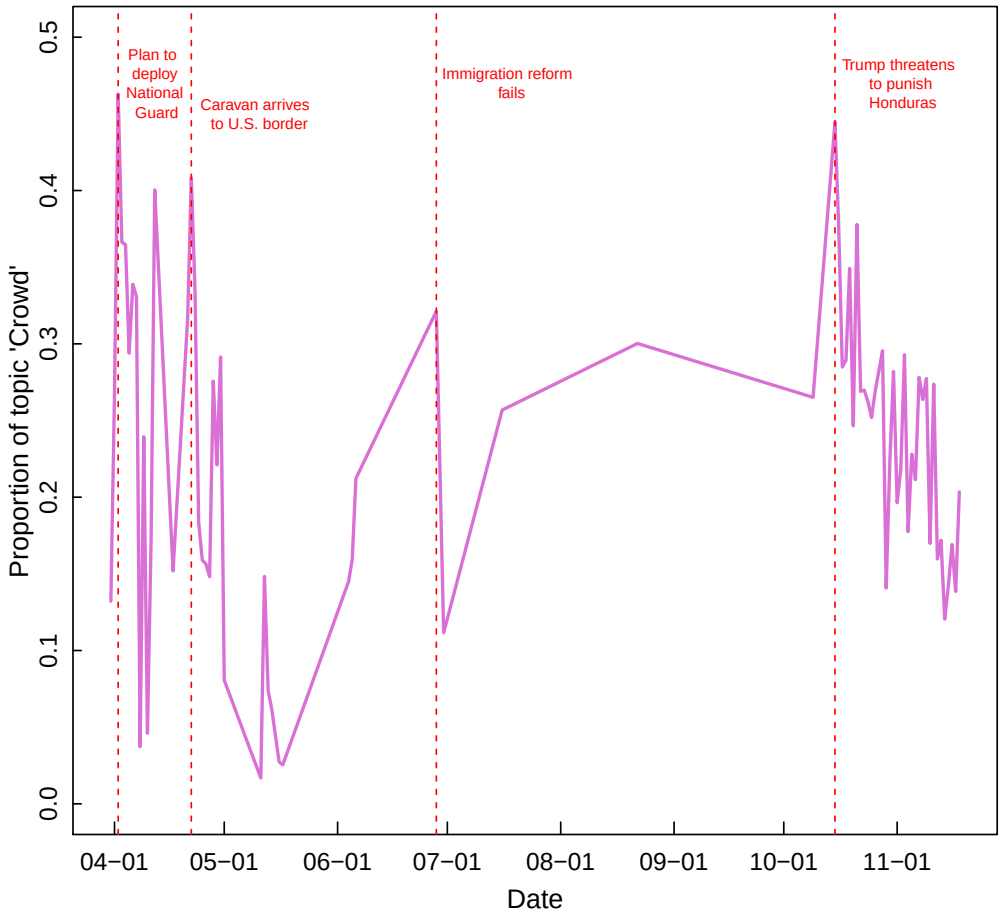
Small groups/Individuals

Camps



there is any variation in the coverage of the magnitude dimension of the caravan. Figure 9 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.

Figure 9: Use of topic "crowd" over time (2018)



Note: The purple line shows the trend of the topic "crowd" from March to December of 2018. The gaps between points indicate that there was no coverage in that period. The dashed lines indicate important dates in the time line of the migrant caravan coverage and development.

It is important to highlight that if we compare two images each showing, for example, a crowd of the same size, the proportions of the "crowd" topic in those images may vary

depending on how the crowd is captured in each of them. Figure 10 illustrates this: although both pictures show the same group of people and it is easy to identify a crowd, the proportions of the "crowd" topic are different. The one with the low proportion is focused on the display of the flags and contains other topics like "water/sky". It therefore frames this group of people differently from the second picture which puts more emphasis on the individuals belonging to it.

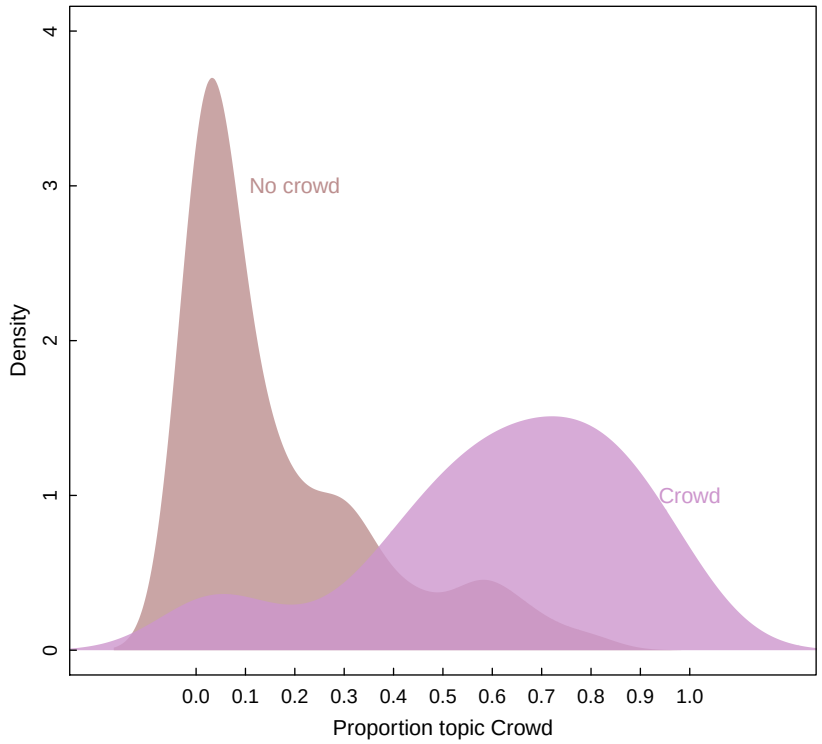
Figure 10: Comparison of different proportions of topic "crowd"

(a) Low proportion

(b) High proportion

Thus, the topic proportions provide a richer and distinct measure of framing that differs from the mere identification or binary classification of a particular object/concept, such as a crowd. To illustrate this point, I hand-coded a random sample of the images to identify the presence of a large crowd. Figure 11 shows the distribution of the "crowd" topic among those images containing a large crowd, and those without one. The first thing to note is that the modes of both distributions align with the expected "crowd" proportions: images with crowds have a high proportion of this topic, and images without a crowd show a low proportion of it. Second, the topic proportions provide more variation and flexibility regarding the depiction of the concept of interest that go beyond a binary classification and identification of an object.

Figure 11: Identification of crowds and distribution of "crowd" proportions



Note: The "No crowd" and "Crowd" labels are hand-coded for a randomly chosen sample of images. The density curves show the distribution of the topic "crowd" in each group.

4.2 Framing a political event: The Migrant Caravan

4.2.1 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 and portrayal of an out-group impact 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 depend on multiple dimensions including situational and personal triggers like ideology (Homola and Tavits 2018), 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) 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 illustrates 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 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 audience's 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 (Abrajano and Hajnal 2017; Homola and Tavits 2018; Schemer 2012).

4.2.2 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`.[9]

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, this information and the aggregation of the rankings by ideological group and news outlet is used to determine the average bias rating of a source. Robertson et al. (2018) show that *All sides* scores have a strong correlation with other validated measures of media bias. The data for this article include 451 articles published by 30 sites with different ideological groups: left leaning (center-left and left, n=16), center (n=10), and right leaning (center-right and right, n=4).[10]

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 with the objective of having 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 five 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. It is important to note that this STM only used the count of visual words from the pictures collected through the `NewsAPI` and *does not* include the pictures

[9]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 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.

[10]Examples of outlets in the "Right" category include *Breitbart*, *Fox News* and the *Washington Times* whereas the "Left" includes outlets such as the *Huffington Post*, *Politico* and *MSNBC*. The "Center" covers outlets like *Bloomberg*, *CNBC* and *USA Today*. For more information regarding the distribution of number of articles per ideological group, see the Appendix.

from *Getty* used to build the vocabulary.[11]

4.2.3 Results

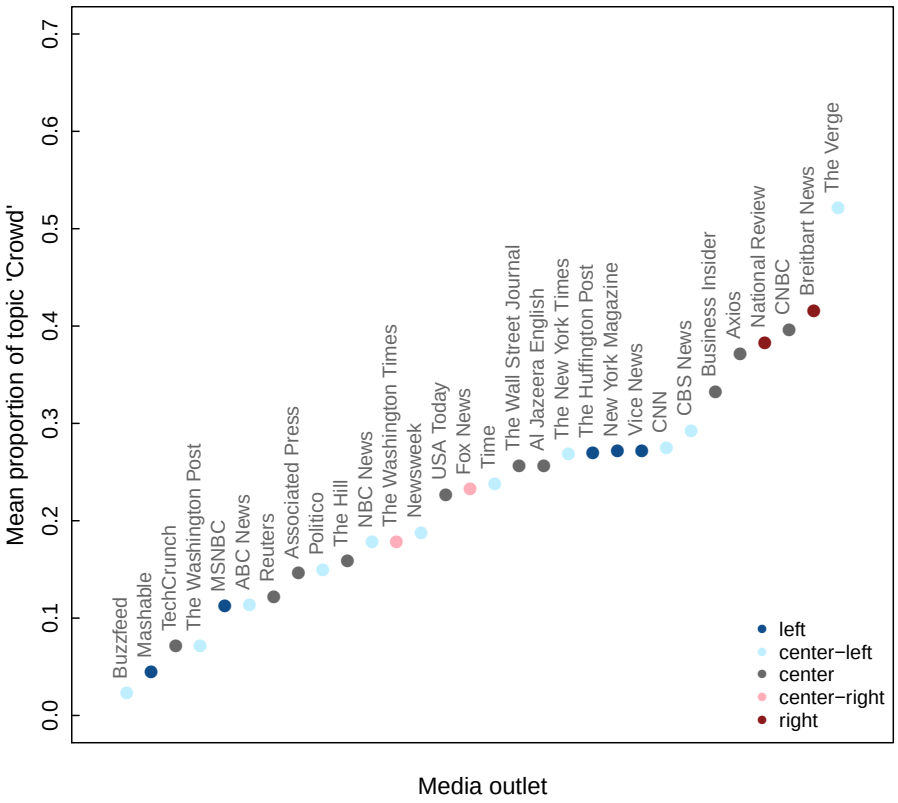
Although some topics are different from the *Getty* example due to the initialization process (less topics and different prevalence covariates), the model recovers relevant topics like "crowd", "darkness", or "small groups/individuals."[12] The findings indicate that there is variation in the use of crowds in the images of the caravan across the different media outlets. In Figure 12 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 13 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." This suggests that right-leaning outlets tend to focus on the magnitude and size of the caravan when publishing news about it, and is in line with the perception of immigrants as a major threat that a large group of actors with such ideology hold.

[11] To ameliorate the concerns regarding the impact of the prevalence covariates on the distribution of the topics in a STM setting, I also conduct a Latent Dirichlet Allocation (LDA) process. The characteristics of the topics, and the results presented below do not differ substantially between methods, as shown in the Appendix.

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

Figure 12: Crowd topic by media outlet

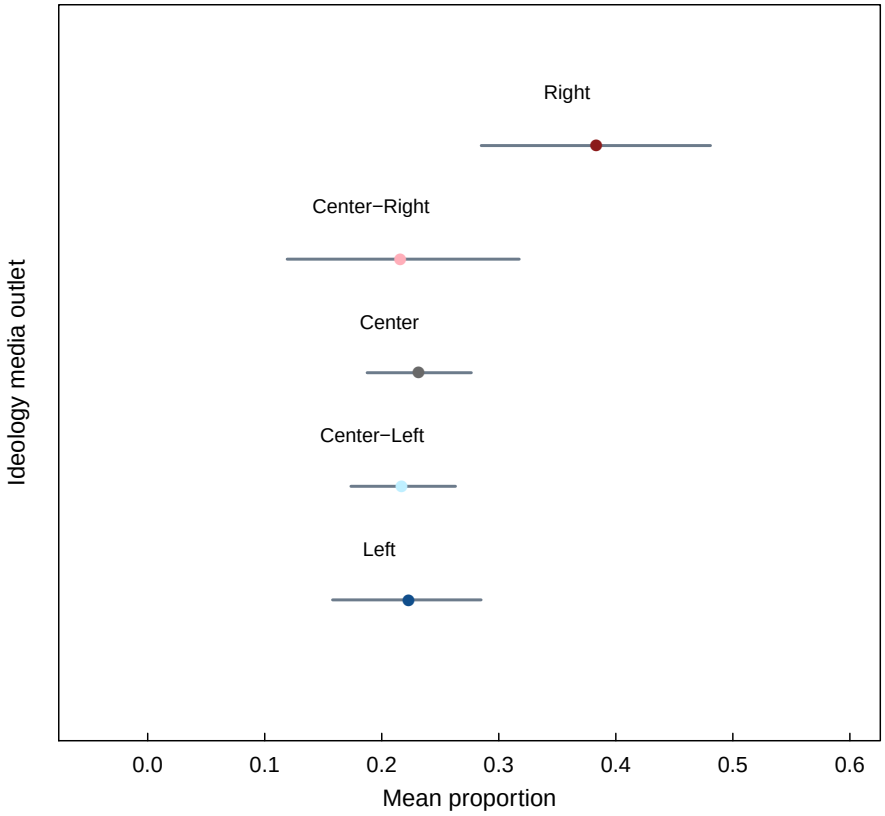


Note: Each point represents the mean "crowd" topic proportion among the images of each of the outlets in the sample. The points are ordered from lowest to highest proportion of topic "crowd". Colors indicate the ideological slant of the outlet.

5 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 13: Ideological leanings and portrayal of crowds



Note: Each point represents the mean "crowd" topic proportion among the images published by media outlets in each of the ideological slant categories. All differences with respect to the "Right" baseline are significantly different from 0 ($p<0.05$).

5.1 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 2019; 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, higher precision and accuracy also come with less parsimonious and transparent mechanisms, and high computational costs and need for training data. In contrast, the BoVW is a method with a more intuitive and relatively simpler procedure, and potential for applications beyond prediction.

First, as illustrated throughout this article, it is possible to track and understand each step of the BoVW. 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 Image-Visual word matrix 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 article, 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 some 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 sufficiently 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, 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. For now, the section below discusses some elements to consider when using the BoVW.

5.2 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.

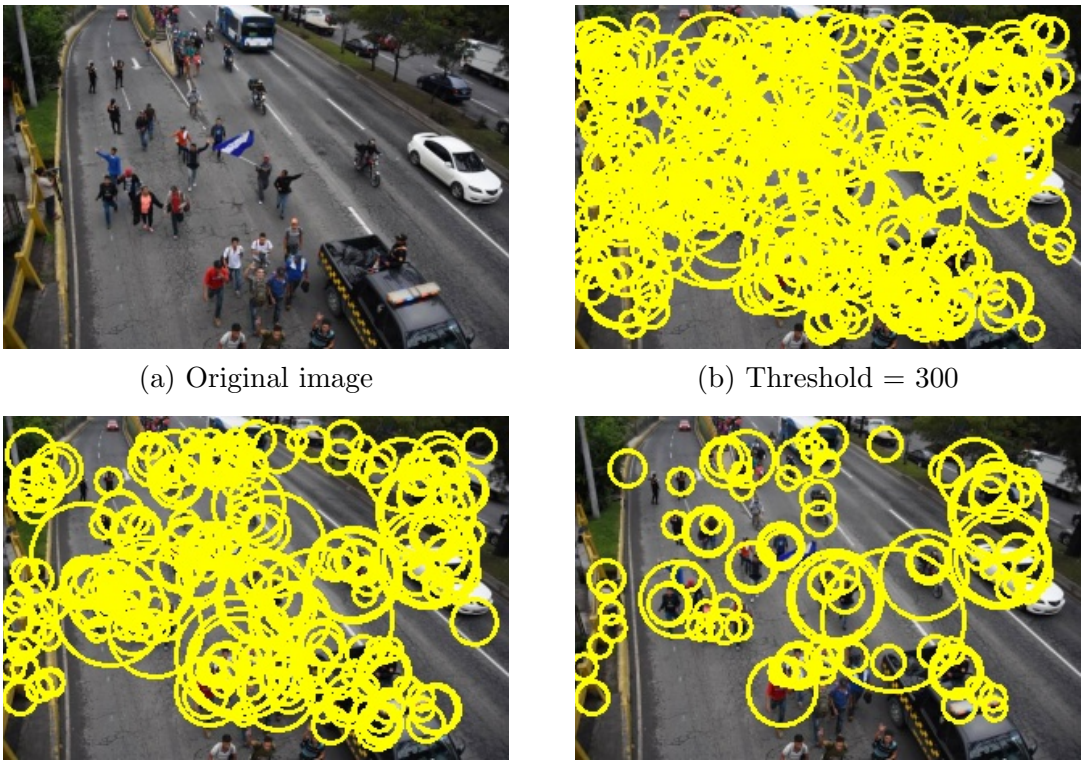
5.2.1 Detecting key points

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: 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 14 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 14: Comparison of key point detection outputs with different thresholds



(a) Original image

(b) Threshold = 300

(c) Threshold = 1000

(d) Threshold = 3000

5.2.2 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.[13] 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 (such as "elbow" and "silhouette" tests), and use visual inspection to assess the composition and formation of visual words.

5.2.3 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 the text covers the construction of visual words as well as their visualization. The visual

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

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 15 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.

Figure 15: Visualizing mistakes

(a) High proportion topic of "crowd" (b) High proportion of topic "border"

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.

6 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.

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: their 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. Further, there are also 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 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.

References

Abrajano, Marisa, and Zoltan L Hajnal. 2017. *White backlash: immigration, race, and American politics*. Princeton, NJ: Princeton University Press.

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.

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.

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.

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.

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* .

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. 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 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.

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.

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

Deselaers, Thomas, Lexi Pimenidis, and Hermann Ney. 2008. Bag-of-visual-words models for adult image classification and filtering. In *2008 19th International Conference on Pattern Recognition*. IEEE pp. 1–4.

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. 2019. "Emotional arousal predicts voting on the US supreme court." *Political Analysis* 27(2): 237–243.

Dilliplane, Susanna, Seth K Goldman, and Diana C Mutz. 2013. "Televised exposure to politics: New measures for a fragmented media environment." *American Journal of Political Science* 57(1): 236–248.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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* .

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

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. 2019. "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.

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.

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

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.

Mutz, Diana C. 1998. *Impersonal influence: How perceptions of mass collectives affect political attitudes*. Cambridge, UK: Cambridge University Press.

Mutz, Diana C. 2007. "Effects of "in-your-face" television discourse on perceptions of a legitimate opposition." *American Political Science Review* 101(4): 621–635.

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

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.

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.

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.

Robertson, Ronald E, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. "Auditing partisan audience bias within google search." *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 148.

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.

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.

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.

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.

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.

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.

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.

Yang, Jun, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM pp. 197–206.

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

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

Zhang, Shiliang, Qi Tian, Gang Hua, Qingming Huang, and Shipeng Li. 2009. Descriptive visual words and visual phrases for image applications. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM pp. 75–84.

7 Appendix

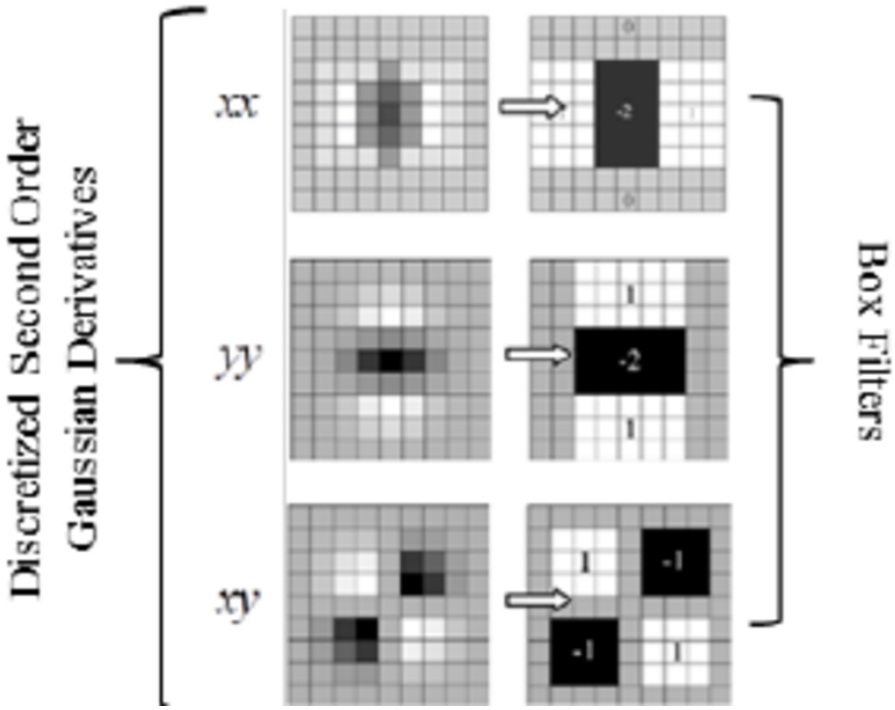
7.1 Key-point detection

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 $\mathbf{x}$.[14] 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 16 illustrates the original and approximated filters.

Figure 16: 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

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

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 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 3×3×3 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 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.

7.2 Emulating the Document-Term matrix: the Image-Visual word Matrix

The Image-Visual Word matrix (IVWM) emulates the Document-Term matrix (DTM) in text analysis. Their underlying logic and structure is similar: the units of analysis are in rows, while each column has an element contained in the full sample. A cell in row i and column j is a number indicating the number of times that the element in column j appears in observation i. This can be a count or proportion, either weighted or unweighted.

In the case of a DTM, each row represents a text under analysis, while the columns are generally words, word stems, sentences, n-grams, etc. that appear in the full pool of texts. In the IVWM, the rows are images, and the columns are visual words. Figure 17 illustrates both.

Figure 17: DTM and IVWM

(a) Document-Term MAtrix

Document/Term	President	elections	⋯	migrants	troops	Central
President Donald Trump has focused heavily on issues related to immigration in the run-up to the midterm elections, warning of an "invasion" of Central American migrants, and sending thousands of troops to the border.	1	1	⋯	1	1	1
President Donald Trump is trying to frame the upcoming midterm elections as a national referendum on immigration issues. The President complains that Mexico is not doing enough to stop the caravan of migrants.	2	1	⋯	1	0	0
Thousands of Central American migrants have again resumed their trek through southern Mexico after failing find buses to carry them. President Donald Trump said Wednesday that the deployment of active troops to the southern U.S. border could increase dramatically.	1	1	⋯	1	1	1

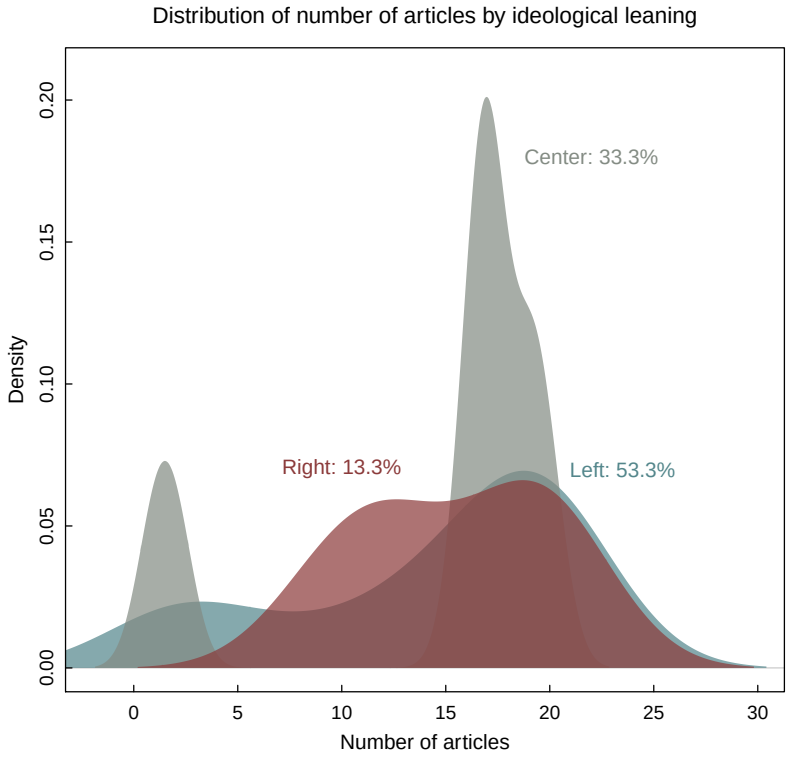
(b) Image-Visual Word matrix

Image/Visual Word			⋯	
	20	0	⋯	3
	0	7	⋯	5
	12	9	⋯	0

7.3 Media outlets dataset

7.3.1 Descriptives

Figure 18: Number of images by ideological group



Note: The plot shows the distribution of number of articles by ideological group. The gray bimodal distribution corresponds to the "center", the red one to the "right" and the blue one to the "left". The numbers next to each distribution show the percentages of media outlets within that category. Despite the differences (53.3% vs. 13.3%), the number of articles is similar between categories.

7.3.2 STM Results

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 19: FREX Visual Words per Topic

Topic 1:
Outdoor/Portraits

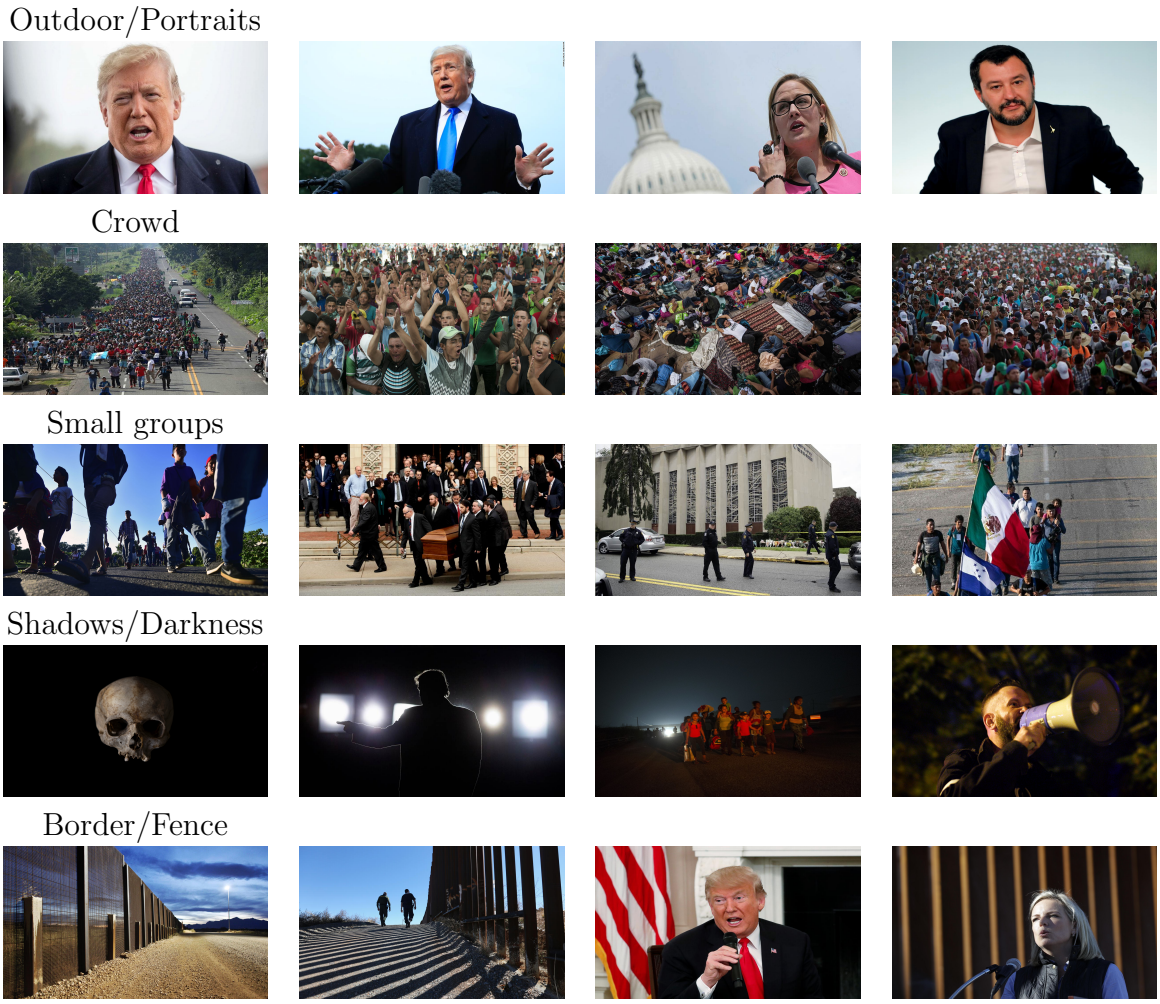
Topic 2:
Crowd

Topic 3:
Small Groups

Topic 4:
Shadows/
Darkness

Topic 5:
Border/Fence

Figure 20: Most representative images per topic



7.4 Results from Latent Dirichlet Allocation (LDA) analysis

To ameliorate concerns regarding the potential sensitivity of the topic distribution to the prevalence covariates in the structural model setting, I conduct a regular Latent Dirichlet Allocation (LDA) analysis and a subsequent regression of the identified "crowd" proportions on the ideological leaning of the media outlets. The topics, most frequent words and most representative images are very similar to those from the STM.

The table below shows two columnds: the first one using a categorical "ideological slant" variable which recreates the results presented in Figure 13 in the main text, and a second one that uses ideological slant as a continuous variable ranging from left to right. The results are in line with the findings in the main text. Right leaning outlets tend to use a higher proportion of topic "crowd" in the pictures the publish of the caravan.

Table 1: Association between ideological slant and topic "crowd"

	Proportion "Crowd"	
	(1)	(2)
Center-right	−0.189*	
	(0.064)	
Center	−0.166*	
	(0.049)	
Center-left	−0.186*	
	(0.048)	
Left	−0.178*	
	(0.053)	
Continuous ideological slant		0.030*
		(0.011)
Constant	0.403*	0.162*
N	437	437
R^2	0.035	0.017
Adjusted R^2	0.026	0.014

* $p < 0.05$. Standard errors in parentheses.