American Journal of Political Science

The Bag of Visual Words: Using Computer Vision to Understand Visual Frames and Political Communication

--Manuscript Draft--

Manuscript Number:	AJPS-43344
Full Title:	The Bag of Visual Words: Using Computer Vision to Understand Visual Frames and Political Communication
Article Type:	Article
Keywords:	Images-as-Data; Computer vision; Unsupervised visual analysis; Visual topic model; Political communication
Abstract:	This article introduces a tool to analyze the content of visual material in order to assess its relationship with political variables: the Bag of Visual Words. The article details 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, and that can be used to conduct exploratory and unsupervised analysis of images, in contrast to other popular classification tools that require training data. I illustrate the applicability of this approach by focusing on the identification of visual frames using a semi-supervised 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 that highlight the magnitude and size of the movement.

The Bag of Visual Words: Using Computer Vision to Understand Visual Frames and Political Communication

May 18, 2020

Abstract

This article introduces a tool to analyze the content of visual material in order to assess its relationship with political variables: the Bag of Visual Words. The article details 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, and that can be used to conduct exploratory and unsupervised analysis of images, in contrast to other popular classification tools that require training data. I illustrate the applicability of this approach by focusing on the identification of visual frames using a semi-supervised 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 that highlight the magnitude and size of the movement.

Word count: 9,957

1 Introduction

Citizens form their attitudes and act according to the information that their own experience and the sources surrounding them provide. Several studies use the analysis of political messages and information to provide not only a better understanding of political events, but also of the aforementioned attitude formation process (Dilliplane, Goldman, and Mutz 2013). However, with a few recent exceptions using audio-visual material (Bauer and Carpinella 2018; Casas and Williams 2019; Dietrich, Enos, and Sen 2019; Knox and Lucas Forthcoming; 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 visual material depicting the same political events in different ways. 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 in a systematic way? The purpose of this article is to offer 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. While some outlets illustrate this movement with images of the struggle and risks that the members endure to reach a better life in America (Figure A.1a), others focus on the "invasion" that the crowd represents and the risks that it poses to Americans (Figure A.1b). What information does an image provide about the way in which a communicator frames an event?

The large amount of images and the subjectivity of human coding are, among several others, two important challenges that complicate answering most of these questions. To

address this, I introduce a computer vision technique to political science that helps to summarize the content of visual material for its subsequent unsupervised and semi-supervised analysis: the Bag of Visual Words (BoVW), an approach that represents an image as a collection of features or "patches" that emulate words in a text. Further, I present a novel use of this dimension reduction technique in a semi-supervised setting, a structural topic model (STM), to improve the identification of topical dimensions in a corpus of images. While the computer vision field has extensively discussed several tools for the identification of topics in images, this manuscript contributes to the general literature on image analysis by applying the BoVW to a STM, and by providing a comprehensive set of diagnoses and analyses of the output from this method. Further, it is to my knowledge the first article to introduce a method for topic modeling of images to political science.

First, I present a survey of the literature regarding the importance of visual material when studying political questions, and the challenges that researchers face when quantifying pictures. Second, I introduce the Bag of Visual Words (BoVW) method as a tool to reduce the dimensionality of images into features that are useful for unsupervised analyses. 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. Then, I present a novel application of this method in a semi-supervised setting for the identification of meaningful political components of the images of the migrant caravan such as "crowds", "fences" and "individuals", 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), as well as practical challenges, solutions and recommendations when using this approach. Finally, I conclude with a list of steps for potential applications of the BoVW, as well as its impact on the social sciences field.

2 Beyond words: images, frames and political attitudes

The content and impact of political information by media and other actors have 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 these issues focuses solely on the verbal component of such information and does not consider the visual material that accompanies the text. There are several reasons to be concerned about this omission. First, we begin to respond emotionally to visual stimuli *before* we can even process them in a conscious manner (Zajonc 1984). 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 is imagery (Kress, Van Leeuwen et al. 1996). Images are everywhere and constantly flowing (Lyman and Varian 2001), and the reliance of organizations, parties, 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 not only helps to highlight a message, but that also influence the way in which recipients understand the message these visuals convey (Butz 2009; Mendelberg 1997, 2001; Valentino, Hutchings, and White 2002). 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).

To illustrate the existence, characteristics, and analysis of these *visual frames*, this article focuses on the depictions of the caravans of Central American migrants seeking refugee

in the United States. The caravans have intensified immigration debates. Citizens and public figures have 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 activities reflects the variability in the perceptions of the caravans, especially regarding the visuals used to illustrate news pieces. For example, on November 14-15 of 2018, several news outlet covered the arrival of the caravan groups at the U.S. border and used the exact same text from the Associated Press to inform their audiences about this event, but published different pictures (Figure 1). While the *Columbus Dispatch* illustrated its article with a photo of government officials and military personnel, the *Dayton Daily News* showed a group of individuals trying to cross the border. Most strikingly, the *Scranton Times-Tribune* chose a picture with a dense crowd of people arriving at the border. The portrayals of the same situation differ significantly between outlets and motivate the question: how can we quantify and explain these differences?

3 Quantifying images: the Bag of (Visual) Words

The first step to understand the origin and impact of underlying messages in images is to describe their content. For some studies the identification of a few elements in the picture provides enough information for its classification and study. For example, Torres and Cantú (2019) classify images in newspaper articles according to whether these show heavily armed police. Other studies require the identification of a broader concept depicted in an image. Won, Steinert-Threlkeld, and Joo (2017) measure the level of violence in pictures of protests using a training dataset with annotations of "perceived violence."

However, there are other instances where a broader analysis of visual content is more fitting. Researchers might be interested in discovering underlying topics in the content of images rather than assuming them, or like a distribution of different topics in an image rather than an indicator of whether a certain element exists in the picture (Feng and Lapata

Figure 1: One caravan, three perspectives: Pictures used in the October 5, 2018 coverage of the migrant caravan

(a) *Columbus Dispatch*

(b) *Dayton Daily News*

(c) *Scranton Times-Tribune*

(a) By: Joel Martinez/The Monitor via The Associated Press; (b) By: Gregory Bull via The Associated Press; (c) By: Marco Ugarte via The Associated Press

2010; Monay and Gatica-Perez 2007). We might be interested in measuring whether the ideology of news outlets determines the usage of the different "visual themes" in the photos they publish, or run an exploratory analysis of a new large corpus of political propaganda to identify the topics on which they focus through time. For these purposes, unsupervised and semi-supervised tools like topic models are appropriate.

Computer scientists have extensively discussed several topic models that a researcher can choose based on her objectives. These include parametric and non-parametric models such as Self Organizing Maps (El Agha and Ashour 2012, SOM), Latent Dirichlet Analysis (Xu et al. 2012, LDA), Probabilistic Latent Semantic Analysis (Hofmann 2001, PLSA), and others.[1] However, while these methods aid in the exploration and identification of themes and clusters in the images, they are not well tailored for the research needs of social scientists. First, the themes and objects that we are interested in exploring might not equate to those found by fully unsupervised clustering models. Second, social scientists are interested not only in the mere identification of topics but also in the role that relevant variables like the characteristics of the publisher/author, treatment assignment in an experimental setting, or details of an event have on the generation of such themes. The introduction of supervision through contextual and author-related variables helps to address these issues (Olaode, Naghdy, and Todd 2014). Thus, I suggest the use of structural topic models (Roberts et al. 2014, STM), a well-known and widely used tool in social sciences, for the identification of topics in images.

However, unlike text, the use of STMs and other clustering models is not trivial when images are the units of analysis. How can we transform raw pixel intensities into meaningful inputs of classification and topic models?

[1] For a comprehensive description and comparison of these methods, please refer to Olaode, Naghdy, and Todd (2014).

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 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 full objects but only some of their components. For example, we will have 10 pieces each showing a tire. A piece corresponding to a "tire" will therefore be a word in our "visual vocabulary". Then, during the classification of the images we will observe that the school bus and the car have four "tire" words each, the bicycle has two, and 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 techniques (Deselaers, Pimenidis, and Ney 2008; Yang et al. 2007; Zhang et al. 2009).

The BoVW involves a series of dimension reduction steps that ease the digestion of visual material that I detail in the following section (Csurka et al. 2004; Grauman and Darrell 2005; Sivic and Zisserman 2003; Sivic et al. 2005).

3.2 Step 1: Extracting and describing local key points

The first step consists of detecting local key points in the corpus of images 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. Once the key regions representing the content of the image are identified, we proceed to "describe" them through the extraction of their features. For the identification part we use a "locator", and for the feature extraction we use a "descriptor."

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

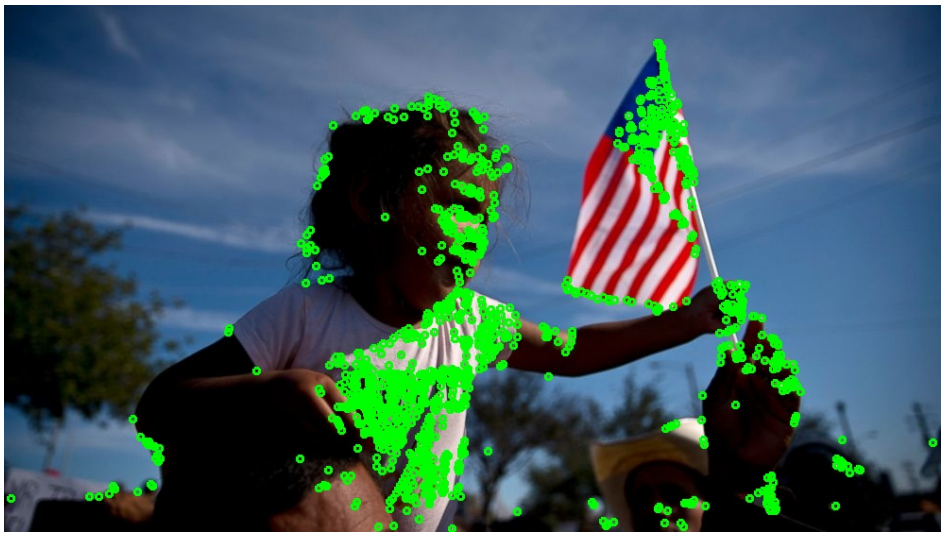
3.2.1 Detecting key points

The FAST Hessian detector is used to locate edges and corners in an image (Bay, Tuytelaars, and Van Gool 2006). This detector identifies the points and regions where significant changes in pixel intensity occur. 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 illustrates the key points identified in the photo with open circles. The points appear in salient regions of the image, and match lines, contours and edges of the most prominent elements of the picture.

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

Figure 2: Location of key points

(a) Original image



(b) Image with key points identified

Source: AP Photo/Ramon Espinosa.

3.2.2 Describing the key points

Next we need to extract features from these points. In texts, features are words, sentences, or n-grams describing each document. However, the identification of comparable features in images poses some challenges. Although intuitively it is easy to think of a "visual word" as a "piece" of an image (e.g. the "tire" in a car picture), in practice the actual quantification of this "patch" is problematic given the multi-dimensionality of a picture and the absence of semantic meaning for patches of a picture. Feature descriptors help in the task of representing image characteristics in mathematical forms. As in the case of detectors, there are multiple alternatives that vary in computational costs, efficiency, 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.[3]

In this project, I implement a RootSIFT descriptor which quantifies the region surrounding the key points. This 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 summary of those elements.

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 a grid with 4×4 pixel cells (Panel (a) of Figure 3). Then, the descriptor compares the intensity of a given pixel 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:

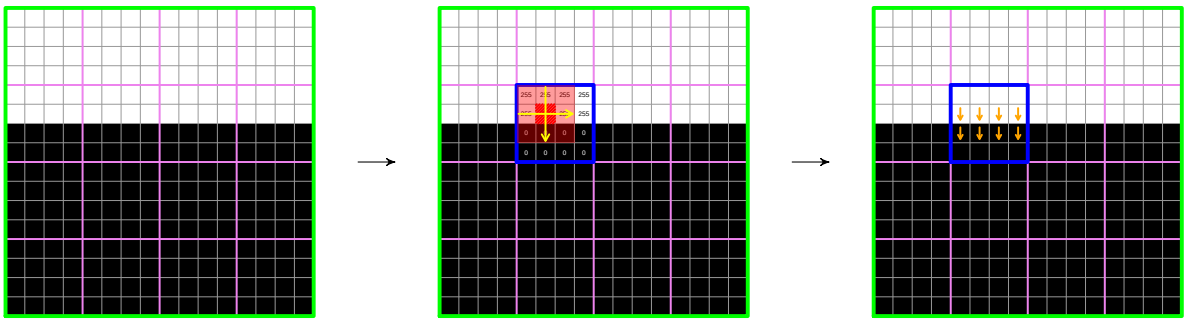
$$G_x = A(x, y) - A(x + 1, y) \qquad\qquad G_y = A(x, y) - A(x, y + 1)$$

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

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

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

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

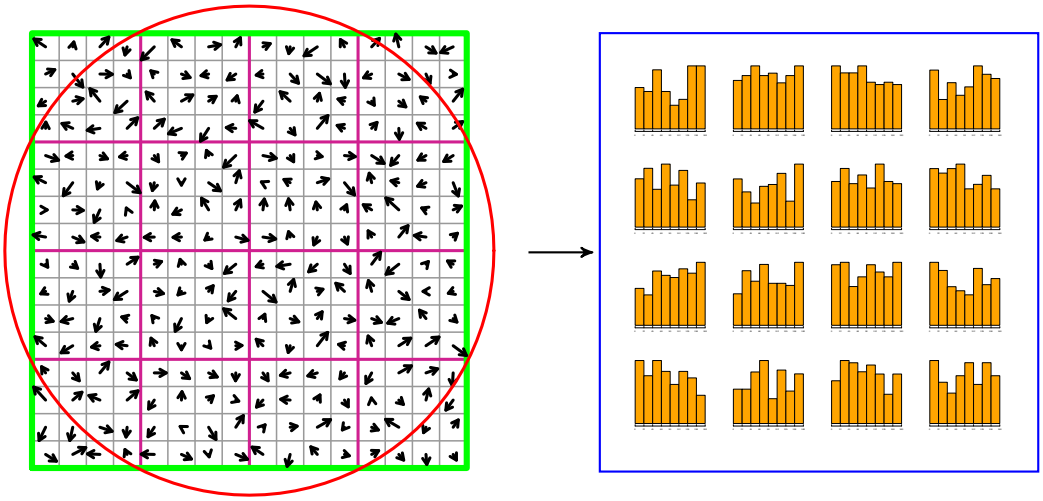
If we focus on a single cell out of the 16 that we defined in the first step, this process yields 16 gradients with their respective magnitude and orientation that we summarize using a weighted count. To do this, we first collapse all the potential gradient angles into 8 bins for the histograms. These angles are in the range of [0, 180] when unsigned[4] 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

[4]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.

each of the elements of this new "flattened" long vector. 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.

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

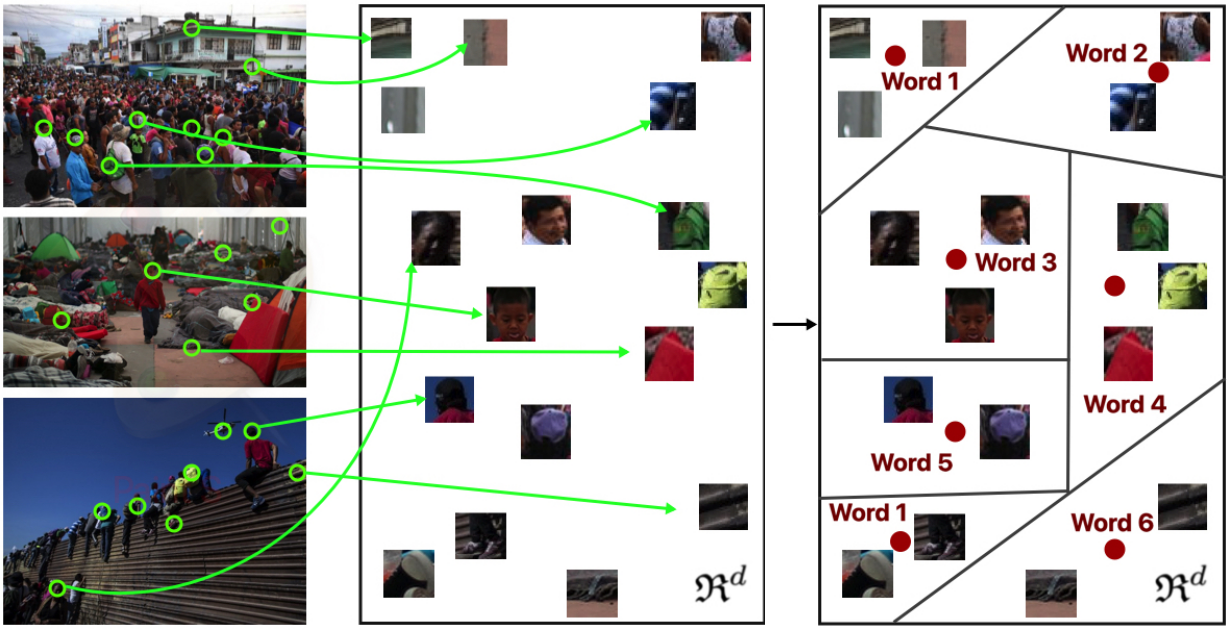


3.3 Step 2: Defining a vocabulary

The features found in images do not have a semantic meaning like words. Therefore, we must define our own codebook or "visual vocabulary". To do this, we will cluster a randomly selected sample of features extracted from the key points of the images in our pool. Once we identify the v clusters, the features associated with each cluster's centroid serve as the representation of a word. This process is illustrated in Figure 5.[5] Mathematically, a visual word is a vector with 128 elements, and graphically we can interpret it as a collection of the mini patches contained inside the cluster. Figure 6 shows a few examples of visual words that are generated by the clustering step.

[5]In the Appendix, I discuss the reasons for not using the full set of features.

Figure 5: Creating the visual vocabulary: clustering and centroids



For the clustering process, I use a mini batch k-means algorithm that optimizes the distance between the feature vectors. This method requires that the user specifies the number of clusters to be generated. That is, the size of the vocabulary V will be equal to this parameter. The diagnosis section and Appendix provide more guidance for the selection of this V and the impact it has on the outcome from the structural topic models.

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 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, and suppose that we identify 15 key points in it. This image is then 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

Figure 6: Examples of visual words

(a) Water (b) Large groups of people

(c) Wired fence

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 word appears in every photo. This constitutes our Image-Visual Word matrix.[6]

[6] An illustration is presented in the Appendix.

4 The BoVW in action: Feeding a semi-supervised model

To illustrate the process outlined in the previous section for topic discovery, I build a BoVW 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 the size, composition, mood, environment or central actors related to it, as well as a distribution of these components in each image. The content of an image illustrates what its creator thinks is worth highlighting of a given event. The creator can be the subject taking the picture but also the actor deciding what pictures to use to illustrate an event. Thus, for this application, I will equate the visual themes or topics in an image with "visual frames" and hereinafter refer to topics as such. Adapting the concept from the literature on framing, a visual frame is the imagery 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).

I compiled a dataset with around 6,500 images of the caravan from multiple sources including *Getty Images*, and 35 media outlets. This dataset includes photographs and metadata covering the author of the picture, source, caption, dimensions, and others.

4.1 Detecting visual frames

First, I build a visual vocabulary of 2,000 "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 biases

in the pictures and maximizing the number of frames.

Second, I build the IVWM of 688 images in 424 news articles covering the caravan of Central American migrants. The columns of this matrix are the 2,000 visual words generated from the *Getty* dataset. The news articles were published between October 3 and November 1, 2018. I compiled them both manually and with the `News API`.[7] I then feed the IVWM to a structural topic model (STM) which allows me to analyze patterns in the visual material under analysis.

Recall the pictures in Figure 1. The pictures that media outlets use to illustrate the same event show some variation in the use of visual elements: police, crowds, a fence with people climbing it. What factors explain this variation? There is evidence that media outlets define the coverage and content of the information they provide based on their audience's demands, marketing, 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.

To explore the composition and generation of such frames, I initialized a STM with 15 topics[8] and three prevalence covariates that account for the particularities of a given event and other characteristics of the actors responsible for choosing a picture: date, news outlet, and its ideology as measured by *All sides*, an organization that provides ratings of "media bias" (right, center-right, center, center-left, and left).[9] The inclusion of these political covariates allows the identification of topics that are more meaningful for social scientific research. Further, it also provides estimates of the effects that variables like ideology of the news outlet have in the generation of the visual frames.

[7] More information about this source is available in the Appendix.
[8] I provide more details regarding the choice of this parameter in the next section.
[9] For more information about this measurement and source, please see the Appendix.

Overall, the specified STM identifies coherent visual frames in the content of the images: border, groups walking, dense crowds, dark backgrounds, fields, indoor portrait, outdoor portrait, individuals, rally, and sky and sand. These frames offer information about the message and actors that the images portray: while some highlight individuals and officials in press conferences (portrait topics) others focus on the members of the caravan (crowds, groups walking). Further, topics like "sand and sky" or "border" also provide details about the place and time in which the evens are occurring.

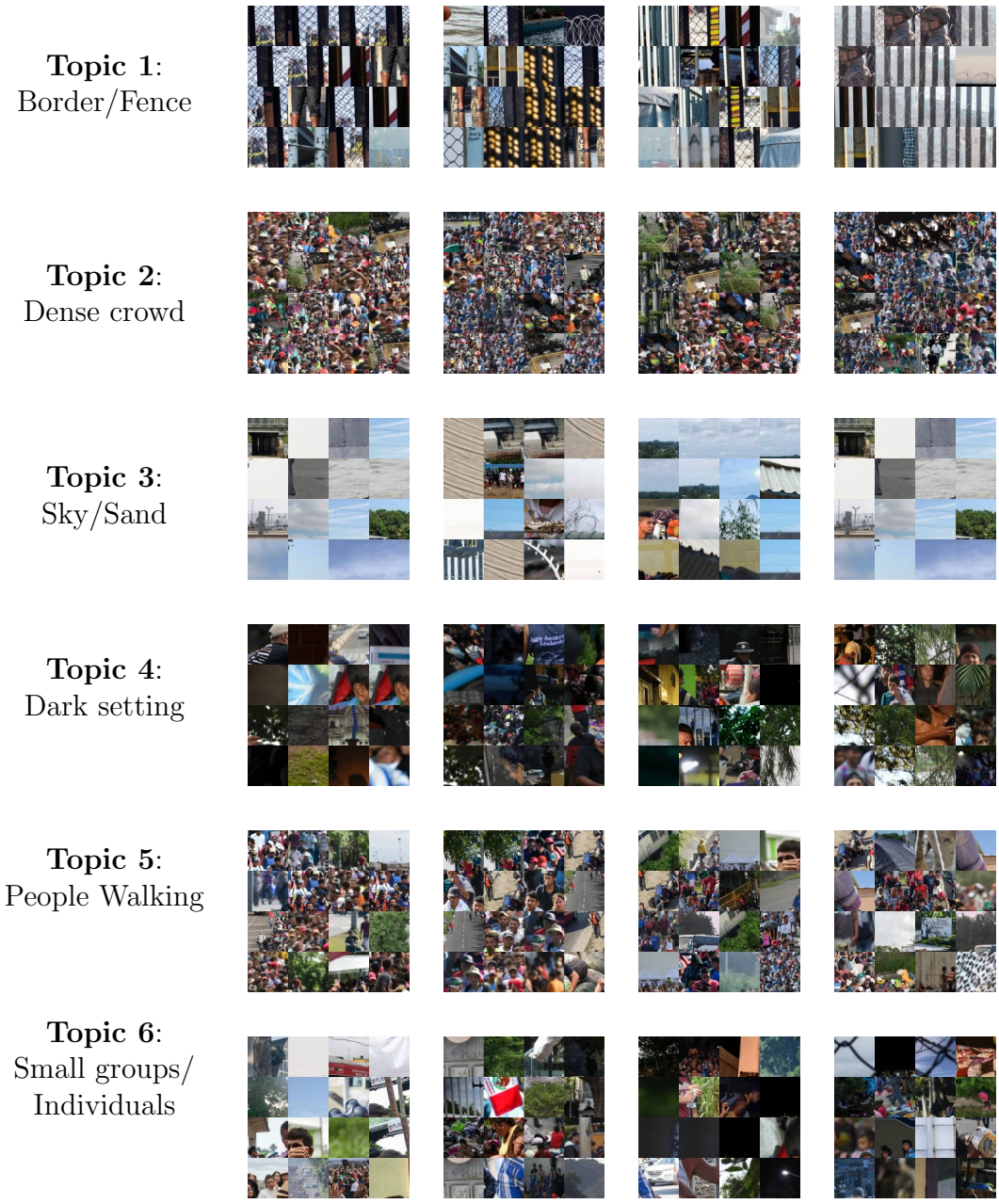
4.2 Exploring visual frames

To label the visual frames that the STM identified, we look at both the most frequent and exclusive visual words, and the most representative images per topic. Figure 7 shows four of the most frequent and exclusive visual words (FREX) from 6 of the 15 topics (the rest are in the Appendix). The replication code includes all the functions necessary to adapt the `stm` package and obtain this output.

Notice that the most representative visual words of the topics contain mini patches that represent components of the frame. For example, the topic "dense crowd" has visual words with patches showing large groups, dense conglomerations of people and granular textures, while the "border" frame has visual words with vertical lines and contrasting colors corresponding to the bars of a fence. Others contain less homogeneous patches given the composition of the frame. For example, "groups walking" reveals not only patches with close-up human figures and body parts, but also with pieces showing pavement, street, and sky suggesting an outdoor setting. Similarly, the "dark setting" topic shows visual words with body parts corresponding to the individuals in the pictures, but also patches with dark and solid colors corresponding to either nocturnal settings or dark backgrounds as in interviews.

The frames' labels become more obvious when we observe the most representative images per topic. The most representative images of a topic k are those photos with high proportions of such topic. Figure 8 presents examples of these. We can identify coherent

Figure 7: FREX visual words per topic



patterns in the data that contribute with our knowledge of the data at hand.

4.3 Framing a movement: factors behind the generation of visual frames

Although the characteristics of most of these frames are worth studying, for ease of exposition the analysis of the generation of frames focuses exclusively on the visual frame of magnitude of the caravans through "dense crowds."

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 an out-group has an effect on threat: "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 through the use of a "dense crowd" frame. On a 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. However, they also trigger other processes that occur in a less conscious manner. Brunyé, Howe, and Mahoney (2014) find that observers heavily rely on crowd size to estimate risk levels, while others highlight the ability of humans to detect anger and conflict elicited by facial expressions and body language of individuals in a crowd (Green

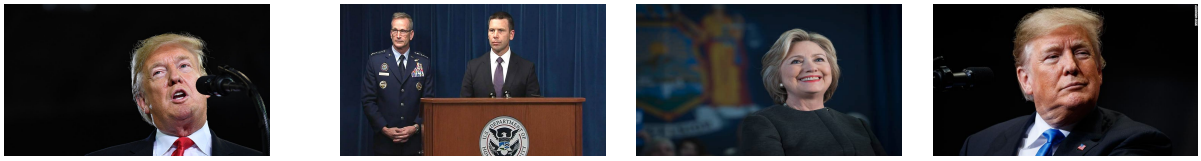
Figure 8: Most representative images per topic

Border/Fence

Dense crowd

Sky/Sand

Dark setting



People walking

Small groups/Individuals

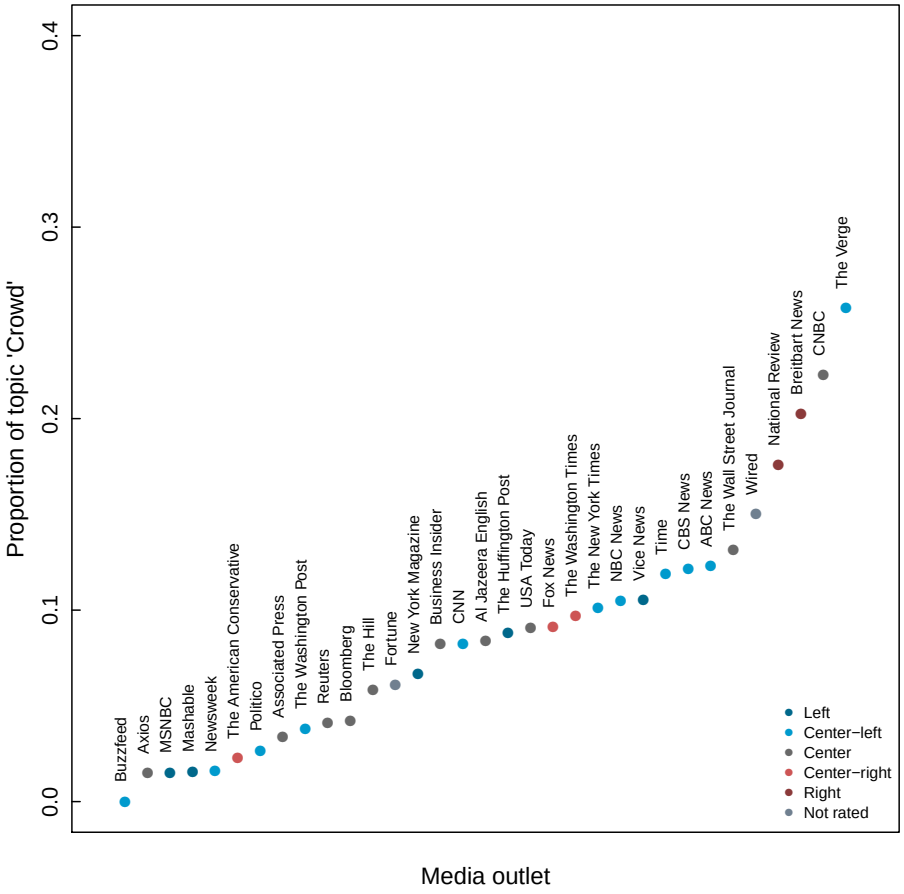
and Phillips 2004; McHugh et al. 2010; Öhman, Lundqvist, and Esteves 2001). Even in an early stage of visual processing, authors find that "feature congestion" and "display clutter," likely to be found in saturated images of crowds, have a negative effect on the attention and digestion of visual information (Rosenholtz et al. 2005).

The ideology of a news outlet determines their own biases with respect to immigration, and therefore influences the way in which they frame information (Kriesi 1995; Oliver and Myers 1999). This leads to the expectation that, for the case of the caravan, right leaning outlets will depict it in more threatening ways through the use of photos showing denser crowds than other outlets. This is line with the idea that conservatives and right-leaning actors are more likely to hold negative views about immigration (Abrajano and Hajnal 2017; Homola and Tavits 2018; Schemer 2012).

We can test this expectation by exploring the usage of the "dense crowd" frame by news outlet. Figure 9 illustrates the variation in the use of crowds in the images of the caravan (mean proportion of "dense crowd" along the y-axis) across the different outlets (x-axis). The color of each point indicates the ideological leaning of the outlet.

Is this variance associated with ideology? To study this question, I analyze the effect of the ideological leaning of the newspapers, the prevalence covariate of the STM, on the generation of the topic "dense crowd." Figure 10 shows the mean of this topic by ideological group and shows that the news outlets with right-leaning biases show significantly higher proportions of this topic than the other groups (all of these differences are positive and reliable). On average, right leaning outlets tend to publish images with 8.9 percentage points more content of the frame "dense crowd" than outlets in the center, and 13 percentage points more than left-leaning outlets. This suggests that right-leaning outlets tend to focus on the magnitude and size of the caravan when publishing news about it, which is in line with the perception of immigrants as a major threat that a large group of actors with such ideology hold.

Figure 9: Crowd topic by media outlet

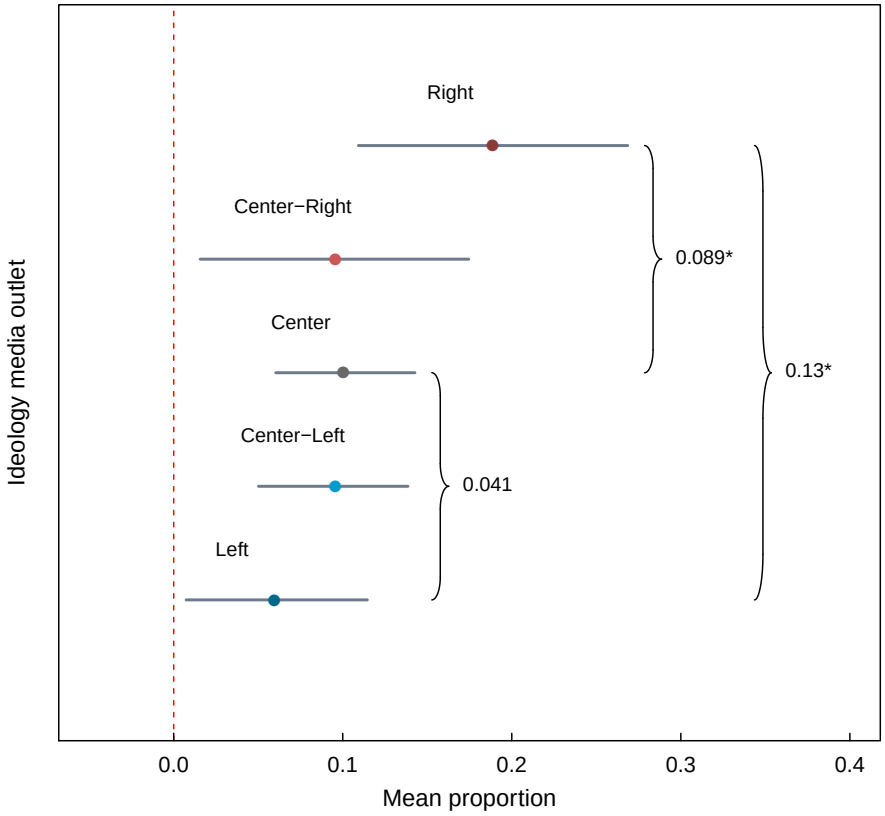


Note: Each point represents the mean "crowd" topic proportion among the images of each of the outlets in the sample. The points are ordered from lowest to highest proportion of topic "crowd". Colors indicate the ideological slant of the outlet.

5 Practical considerations: scope, challenges and diagnosis

Throughout this article, I have elaborated on the logic, implementation, applicability, and benefits of the BoVW. However, it is important to understand its limits and scope, as well as the impact of key decisions during its implementation.

Figure 10: 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 bias categories. Brackets indicate the differences between a few groups and the $*$ indicates that the 95% confidence interval of the difference does not cover 0.

5.1 Scope and strengths

A good way to understand the scope and limits of the BoVW is to discuss the differences between it and other popular 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 (Anastasopoulos et al. 2016; Cantú 2019; Lucas 2019; Won, Steinert-Threlkeld, and Joo 2017; Zhang and Pan 2019) that achieve high predictive power in image classification. Why should scholars use the BoVW then?

5.1.1 Supervised and unsupervised applications

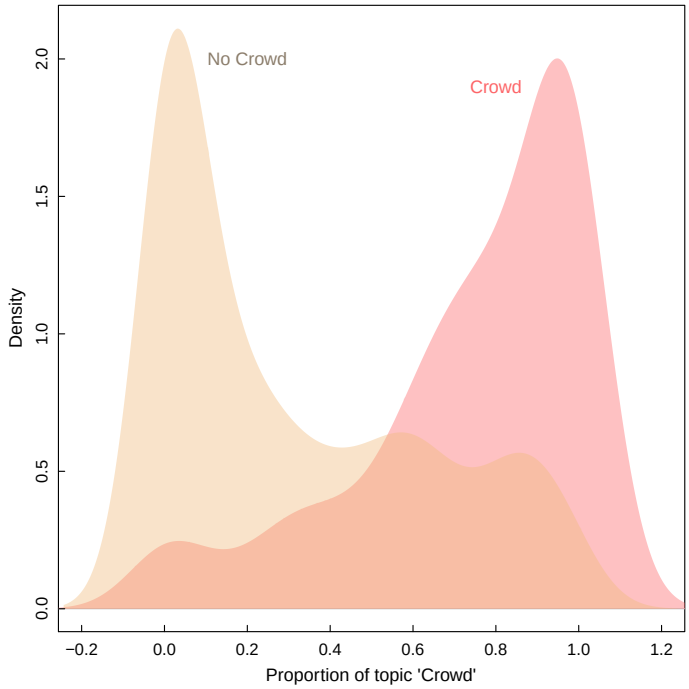
Although both methods share "feature extraction" as a crucial component of their processes, they have different final objectives. The BoVW provides a series of steps to reduce the dimensionality of images and returns a matrix that can serve as the basis of many classification algorithms, but that is particularly fitting for unsupervised and semi-supervised models that allow researchers to learn underlying content of the images without explicitly imposing categories of interest. CNNs are models with a directed graph structure composed of *layers* with nodes and connections that also detect and extract features from images in order to ultimately predict an outcome of interest. In order to reach these predictions, a process of error minimization takes place (Krizhevsky, Sutskever, and Hinton 2012; LeCun et al. 1998; LeCun, Bengio et al. 1995) that requires large amounts of labeled data to train and test the model.

Although the BoVW can also be used in supervised settings to classify, for example, whether a picture has a crowd or not, when it comes to predictive power, a CNN would perform better given that the complexity and amount of features it extracts from a picture are richer than those from a BoVW. This is in part due to a cyclical learning process that uses labeled data. In other words, the feature detection and extraction process of the BoVW follows a set of rules that are independent from any label; it occurs before any modeling strategy. In contrast, the CNN refines the extraction of features by learning about the images of interest from their associated labels. Thus, the BoVW relaxes the need of labeled training data, a step that in several applications is particularly hard to fulfill. Methods like transfer learning have reduced the need for large numbers of training data for CNNs, but they still rely on labeled data as well as on a clear target measure.

However, despite the differences in performance when it comes to prediction, the pairing of a BoVW with a tool like STM can still be used to aid classification tasks. As a validation exercise to illustrate this point, I hand-coded whether the pictures under study display a crowd. Figure 11 shows the distribution of the sum of proportions of all topics

that include crowds or large groups (not only "dense crowds") among those images labeled as 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. Third, although there are certain cases in the tails that are incorrect classifications and that I will discuss in the next subsection, the results suggest that the topics can help with the estimation of coefficients that enable classification.[10]

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



Note: The "No crowd" and "Crowd" labels are hand-coded. The density curves show the distribution of the topic "all crowd" (Dense Crowd + Groups walking + Rally + Outside crowd) in each group.

[10]Table **??** in the Appendix shows the results from a model predicting the hand-coded labels using the topic proportions, θ.

5.1.2 Different research objectives

Another feature of the BoVW is that it can fulfill research objectives beyond mere classification. For example, while a binary indicator of whether a picture shows a dense crowd might be useful for certain purposes, the proportion of such frame in different pictures might be informative for the question of how media frames the same topic using visuals. If we compare two images each showing a crowd of the same size but from different angles, the proportions of the "crowd" topic in those images may vary. Figure 12 illustrates this. A CNN would correctly indicate that both pictures contain a crowd, a binary indicator. However, the proportions of the "crowd" topic that a visual STM outputs are different and give a continuous measure instead. The photo on the left with a low proportion of this topic is focused on the display of the flags and contains other frames like "water/sky", while the picture on the right puts more emphasis on the individuals belonging to it. Thus, this richer and distinct measure of framing that differs from the mere identification of a particular object is important when studying the use of crowds as a potential driver of fear of immigration.

Figure 12: Comparison of different proportions of topic "crowd"

(a) Low proportion

(b) High proportion

(a) By: Sandra Cuffe/Al Jazeera; (b) By: Jesús Alvarado.

5.1.3 Resources, mechanisms and interpretation

There are other dimensions in which the BoVW and CNNs differ. First, it is possible to track and understand each step of the BoVW given that the data reduction process involves clear steps with basic mathematical foundations. The feature extraction of a CNN is less traceable given that it is prediction-oriented. In contrast, the construction of a visual vocabulary provides intuitive tokens that help with a better understanding of the role that different features of the images play in, for example, the discovery of topics.

Further, the computational costs of using this approach are low, especially compared to CNNs that require special infrastructure like graphics processing units (GPUs) or high performance computing clusters (HPC) if 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.

Overall, despite their distinct objectives and strengths, rather than being competitors, researchers should see these unsupervised and supervised methods as complementary (Grimmer and Stewart 2013). For example, while the BoVW in a semi-supervised setting could help to detect theoretically relevant but unknown dimensions in the data, CNNs can accurately classify images according to the newly discovered and more specific dimensions of interest.

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, in this section I provide a set of diagnosis tools and guidance for making some of those decisions. A detailed analysis of the impact of some of these parameters on topic discovery and estimation effects is included in the Appendix.

5.2.1 Detecting key points

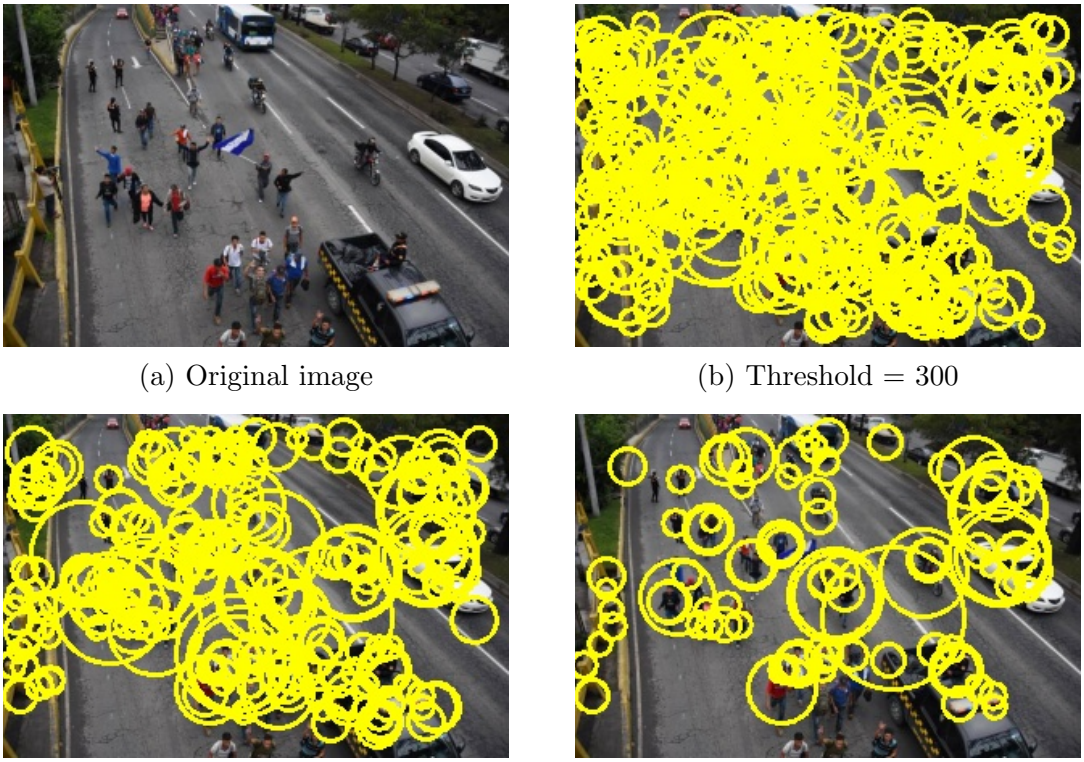
There are two important things to consider during the key point detection: 1) the type of salient regions to identify, and 2) the precision of this identification. The first one relates to the definition of what constitutes a salient region. For example, certain applications require an accurate identification of "corners" in an image (e.g. identification of buildings and houses, captured with rectangular shapes). Others, however, rely more heavily on "blobs" (e.g. classification of texture). 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 (Olaode, Naghdy, and Todd 2014). 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 impact on the number of key points that the detector identifies in each image. The decision should consider what features are substantively relevant to the objective of the 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. A visual inspection of a small sample of images and the key points detected in each of them using different thresholds is helpful to address this issue. Consider Figure 13 where a low threshold of 300 yields a very large number of key points detecting finer features such as lines on the pavement. In contrast, a higher threshold of 3000 leads to the identification of more prominent features like the people and cars in the picture.

5.2.2 Building a vocabulary

For the visual vocabulary, the first consideration should be from what images it will be constructed. As explained above, each of the feature vectors in an image is associated with a visual word. Thus, it could be the case that a given feature is linked to a visual word that

Figure 13: Comparison of key point detection outputs with different thresholds



(a) Original image (b) Threshold = 300

(c) Threshold = 1000 (d) Threshold = 3000

does not properly represent it if there are no better candidates.[11] Therefore, it is crucial to build a vocabulary with images relevant to the target pool under study.

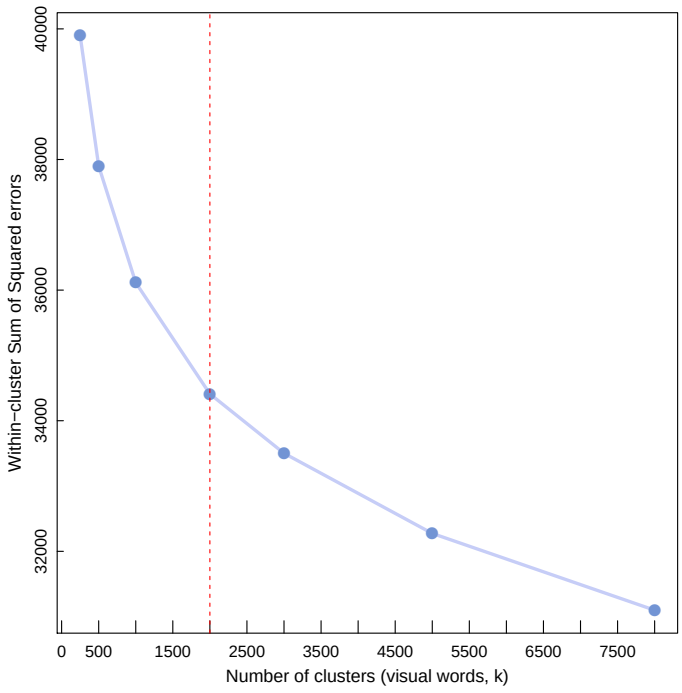
Another consideration is the number of clusters or "visual words" to extract. A richer vocabulary has more power to discriminate and distinguish features, but a more parsimonious one focuses less on the details that each visual word is capturing. The decision to have a more fine-grained vocabulary depends on the substantive motivation of a project: is it relevant to keep features of a given object with light and dark backgrounds as separate visual words? If only the object itself is a relevant component of the visual theme then it might be adequate to keep them together. If instead the background is relevant as a time-space indicator, then the separation is more useful.

Researchers can also rely on commonly used tests providing an "optimal" number of clusters. Figure 14 illustrates the "elbow" method. The x-axis of the figure shows the

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

number of potential clusters/visual words, and the y-axis the within-cluster sum of squared errors (WSSE). A lower WSSE is desirable suggesting more cohesiveness of a cluster (i.e. smaller distances between the features and the cluster centroid). The elbow method consists of determining a "break point" or "elbow" in the plot where the WSSE decreases sharply and starts flattening afterwards. In Figure 14 we observe a strong jump between 1000 and 2000 visual words but a less intense change from 2000 to 3000. Thus, this test informs our decision to select an appropriate number of visual words in our vocabulary (in this case, 2000 visual words).

Figure 14: Within-cluster sum of squared errors by number of clusters



The length of the visual vocabulary has an impact on the discovery and characteristics of the topics in a corpus of images. For example, a BoVW based on a 500 visual word vocabulary yields slightly different topics than those from a model with 2,000 visual words. Quantities like semantic coherence, exclusivity, and likelihood of the topics also differ between designs. After the comparison of multiple models and designs, the results for the caravan application indicate that models based on smaller vocabularies have a higher

semantic coherence than the rest, but lower scores of exclusivity of words to topic (Figure A.8). Overall, while vocabulary length has a substantive impact on the frames that an STM can identify in a corpus of images, the differences in the estimates of prevalence covariates on topics that are clearly identified in the different set-ups are minor (see Appendix).

5.2.3 Determining the number of visual frames

Another parameter to consider is the number of topics to extract from the visual STM. Here, I follow the practices recommended by Roberts, Stewart, and Tingley (2014) involving the evaluation of aspects like semantic coherence and exclusivity. Using the functions embedded in the STM package paired with the functions provided in the replication code of this article, researchers can obtain average measures of mean held-out likelihood, semantic coherence, exclusivity, and residuals from models with a varying number of topics. Visual inspection of plots with this information such as those in Figure A.8 allows researchers to select an "optimal" number of topics that preserve parsimony, and maximizes semantic coherence, held-out likelihood and exclusivity. Researchers can take these statistics as guidance for their decision but should also complement them with considerations of their research needs and qualitative inspection of the topics using the FREX visual words and most representative topics.

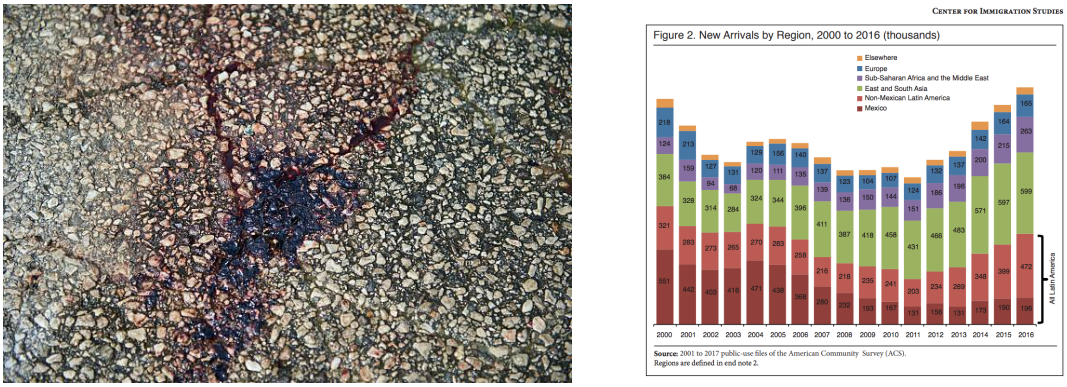
5.2.4 Inspect and visualize

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 of the data, and the absence of concrete tokens 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. The replication code covers the construction of visual words as well as their visualization. The visual inspection of these

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

Some of the potential errors and pitfalls become obvious in a post-BoVW stage. For example, the picture in the left panel of Figure 15 has a high percentage of topic "crowd" although it is just a shot of pavement. The granularity and texture of the pavement resembles that of a big crowd in terms of pixel intensity changes. Further, the picture on the right is clustered with the "border/fence" pictures due to the vertical lines of the graph. Thus, the manual analysis of those pictures or the removal of customized "visual words" (like those with pavement) are alternatives that help to improve the study. 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, digest, and explore visual material. 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). Further studies should consider the inclusion of "color statistics" to the BoVW routine.

Further, the applications of this method to visual framing should be extended to include text and other relevant covariates. 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.

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 political issues such as the way in which leaders in each country visually present foreign interactions to their constituencies, or the different frames of protests across countries becomes more viable.

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.

Anastasopoulos, L Jason, Dhruvil Badani, Crystal Lee, Shiry Ginosar, and Jake Williams. 2016. "Photographic home styles in Congress: a computer vision approach." Working paper.

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. 2019. "The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election." *American Political Science Review* 113(3): 710–726.

Casas, Andreu, and Nora Webb Williams. 2019. "Images that matter: Online protests and the mobilizing role of pictures." *Political Research Quarterly* 72(2): 360–375.

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 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.

El Agha, Mohammed, and Wesam M Ashour. 2012. "Efficient and fast initialization algorithm for k-means clustering." *Efficient and fast initialization algorithm for k-means clustering* 4(1).

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.

Hofmann, Thomas. 2001. "Unsupervised learning by probabilistic latent semantic analysis." *Machine learning* 42(1-2): 177–196.

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. Forthcoming. "A Dynamic Model of Speech for the Social Sciences." *American Political Science Review* .

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* 4(4): 237–243.

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

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.

Olaode, Abass, Golshah Naghdy, and Catherine Todd. 2014. "Unsupervised classification of images: A review." *International Journal of Image Processing* 8(5): 325–342.

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, and Dustin Tingley. 2014. "stm: R package for structural topic models." *Journal of Statistical Software* 10(2): 1–40.

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.

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.

Sivic, Josef, and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*. Vol. 2 pp. 1470–1477.

Sivic, Josef, Bryan C Russell, Alexei A Efros, Andrew Zisserman, and William T Freeman. 2005. Discovering objects and their location in images. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*. Vol. 1 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.

Torres, Michelle, and Francisco Cantú. 2019. "Learning to see: Convolutional neural networks for the analysis of social science data." Working paper.

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.

Xu, Kan, Wen Yang, Gang Liu, and Hong Sun. 2012. "Unsupervised satellite image classification using Markov field topic model." *IEEE Geoscience and Remote Sensing Letters* 10(1): 130–134.

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." *Sociological Methodology* 49(1): 1–57.

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*. pp. 75–84.

A Appendix for "The Bag of Visual Words: Using Computer Vision to Understand Visual Frames and Political Communication"

This Appendix provides additional analyses, diagnoses, and details regarding the application and plots presented in the main text. It is organized into 8 sections.

A.1 Motivation

A.1.1 Visual frames (example)

Images used in two different news outlets reporting on the caravan of migrants from Central America. The first one was published in the article "The migrants risking it all on the deadly Rio Grande" in *The Independent*, while the second one was featured in the article published by *Fox News* "Caravan migrants begin to breach border as frustration with slow asylum process grows" on the *Fox News* website.

(a) Photo by John B. Moore, published by *The Independent*

(b) Photo published by *Fox News*

A.1.2 Same facts, different visual perspectives

The pictures in Figure 1 in the main text were used by different newspapers to illustrate an article about the Caravan written by staff members of the Associated Press. The text is identical but the headline changes slightly. The article itself was written by Elliot Spagat and Maria Verza from the Associated Press and its text can be found in this link: `https://apnews.com/566cc181c2734d30ba1a03133d1b3304`

A.2 Technical details: detection and description of key points

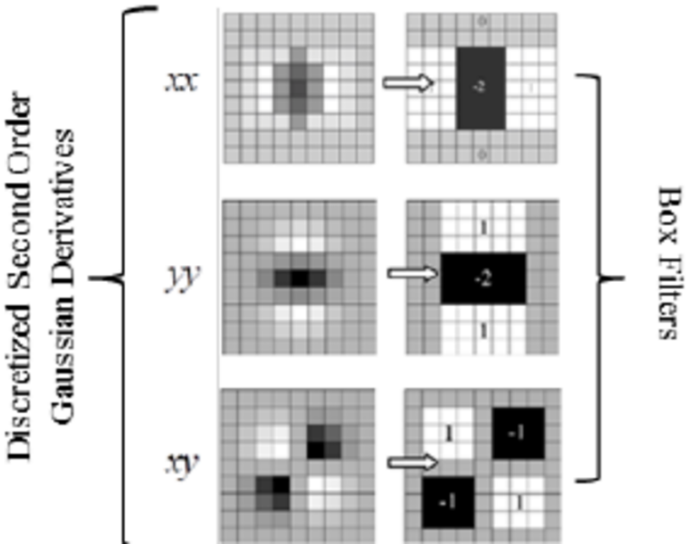
A.2.1 Key-point detection

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

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

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

Figure A.2: 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 allow us to estimate the determinant of the

[1] 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 in the main text 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.

In this article we use the RootSIFT descriptor. This 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.

The RootSIFT steps are not applied to the original image, I, but to a "blurred" version of it, A, using a Gaussian-smoothing filter. This processing step helps to clean the image by decreasing the sharpness of irrelevant elements like irregular blobs.

A.3 Building a vocabulary: sample of features and clustering

The BoVW requires the researcher to select and cluster a sample of the total features identified in the images under analysis. Why do we lump together a subset of the features instead of using the full set? 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 observation 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".

In order to achieve higher levels of speed and efficiency we form this vocabulary based on a random sample of the feature vectors. 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.

A.4 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 indicates 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 A.3 illustrates both.

Figure A.3: 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

A.5 Media outlets application

The data used in this article was collected using the `News API` and manually curated by the author. The `News API` is a tool that allows users to 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.

The data was used to feed a structural topic model with 15 topics and three prevalence covariates: news outlet, date, and ideology of the news outlet. The scores for ideology are provided by *All Sides* and 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 are 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 424 articles published by 33 sites with different ideological groups: left leaning (center-left and left, n=16), center (n=10), and right leaning (center-right and right, n=5). The other two have not been rated. 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.

A.5.1 STM Results (cont.)

The most representative images per topic, and most frequent and exclusive words for topics 7-15 are presented below in Figure A.4 and Figure A.5.

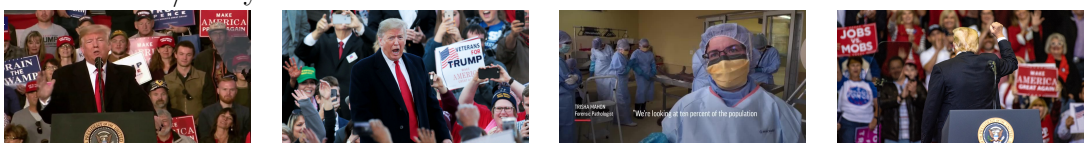
Although most of the topics are sensible and meaningful for the study of visual framing, we also find a few that are less relevant than the others. For example, there is a topic whose most representative images are pictures with banners and ribbons that typically appear at the bottom of news shows. Although this "TV screenshots" topic is not a politically relevant dimension for the study of portrayals of immigration, it makes sense from the computational viewpoint: there are elements of pictures with high proportions of this frame such as figures with text at the bottom that make them look similar.

Figure A.4: Most representative images per topic

Indoor Portraits

TV Screenshots

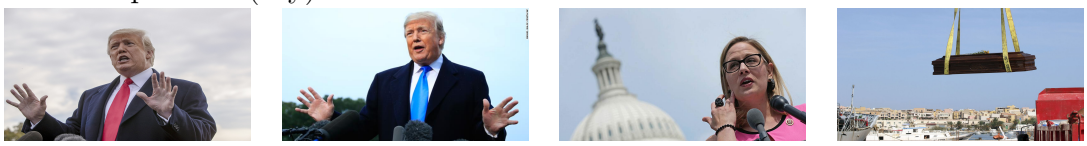
Small crowd/Rally



Outdoor crowd

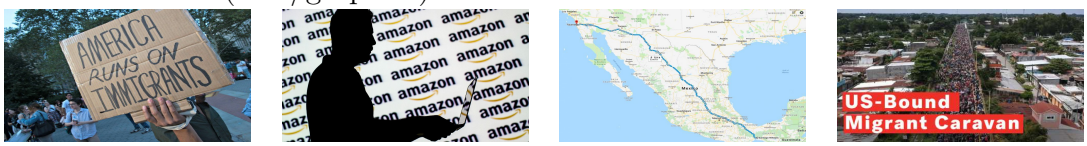
Open field/Grass and trees

Outdoor portrait (sky)



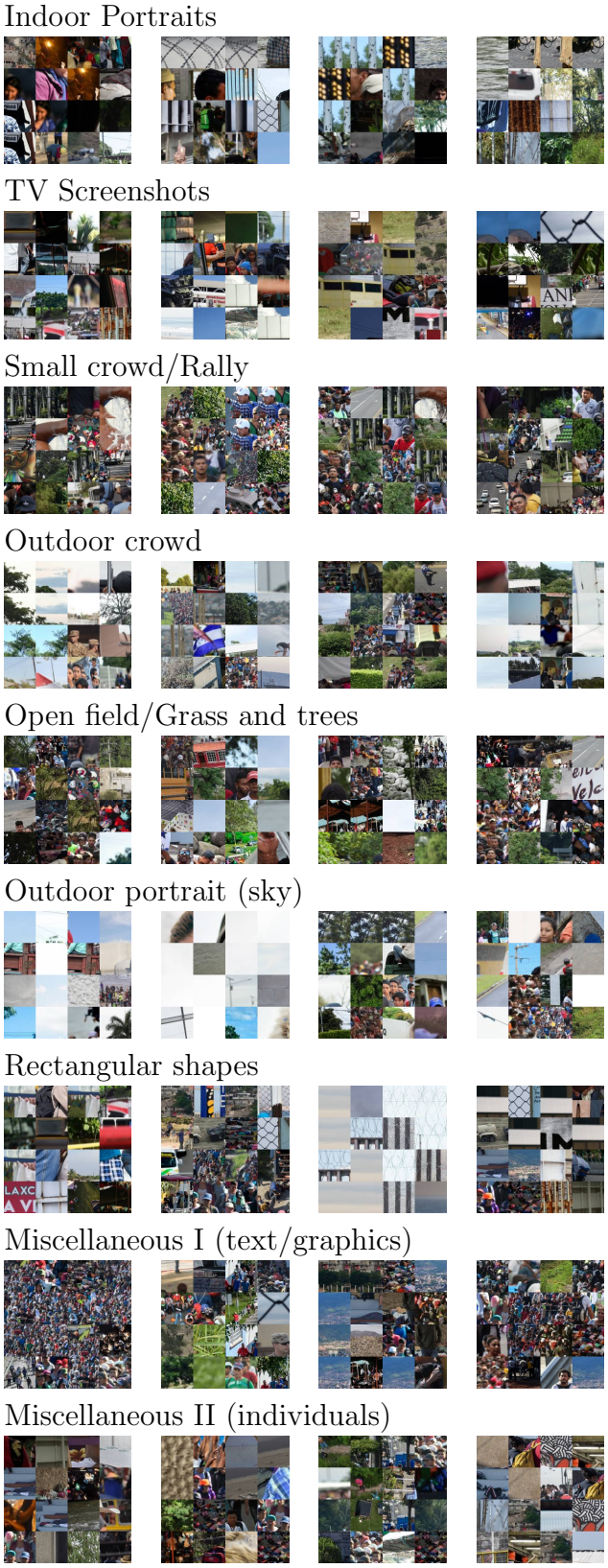
Rectangular shapes

Miscellaneous I (text/graphics)



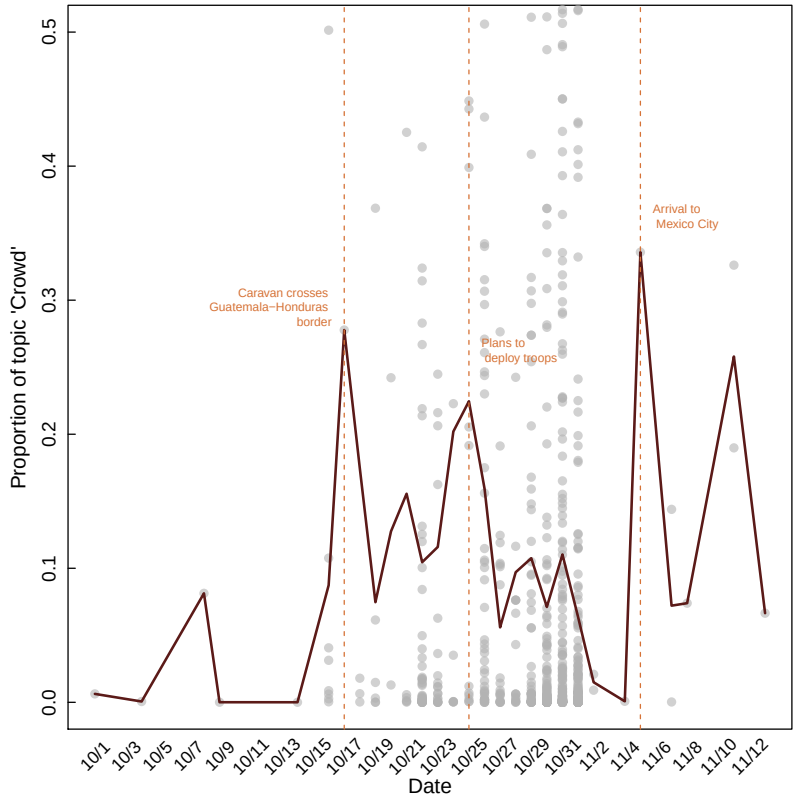
Miscellaneous II (individuals)

Figure A.5: FREX Visual Words per Topic



We can analyze the use of the topic "crowd" over time to identify whether there is any variation in the coverage of the magnitude dimension of the caravan. Figure A.6 shows on the y-axis the proportion of topic "dense 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 Mexico City. 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 A.6: Use of topic "crowd" over time (2018)



Note: The line shows the trend of the topic "crowd" from October to November 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.

A.6 Diagnosis – the impact of parameters on topic discovery and estimation

In this section, I present an analysis of the impact that three elements of the BoVW have on 1) topic quality, and 2) the estimates of the effect that prevalence covariates, like news outlet's ideology, have on the generation of visual frames. For the former, I use statistics like semantic coherence, exclusivity, held-out likelihood and residuals as Roberts, Steward, and Tingley (2014) suggest. For the latter, I estimate the differences in mean "dense crowd" proportions between right leaning outlets and left leaning outlets. Recall that Figure 10 in the main text shows a positive and reliable difference between these two types of outlets indicating that right leaning outlets tend to publish higher proportions of the "dense crowds" frame in the articles they publish about the migrant caravan.

The three parameters I explore and compare are a) Hessian threshold, which regulates the accuracy of key point detection in an image, b) number of visual words in the vocabulary, and c) number of topics in the visual STM.

A.6.1 Hessian threshold for key point detection

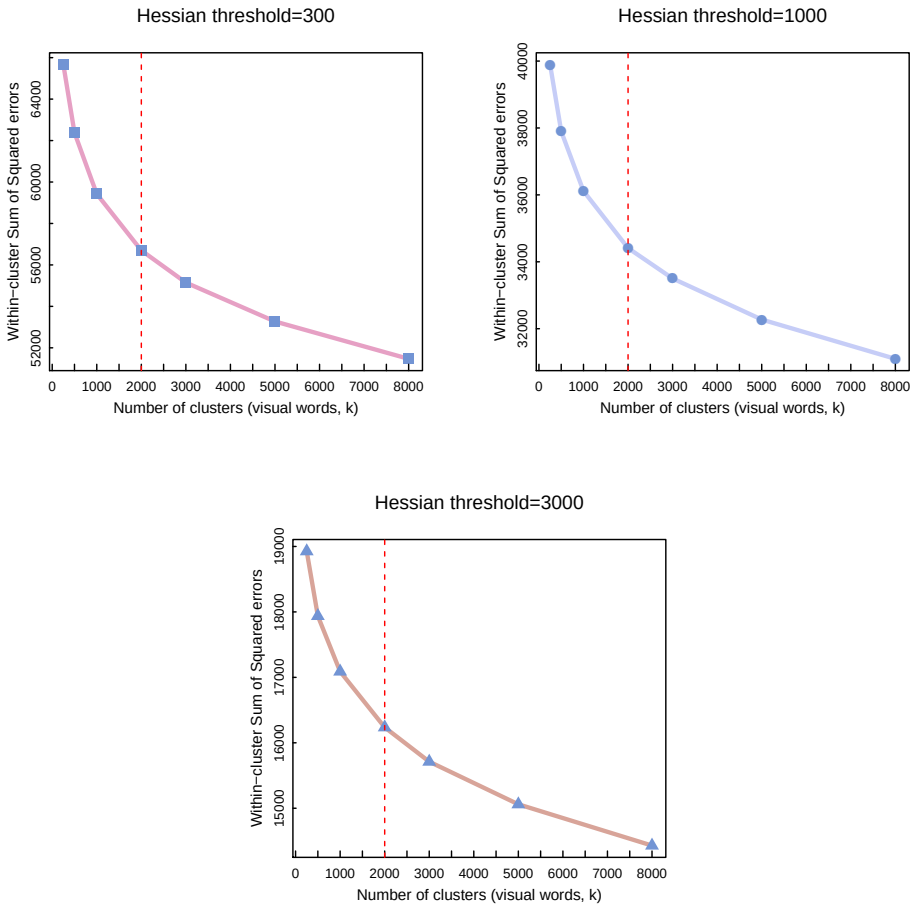
The Hessian threshold is a parameter that regulates the accuracy of the key point detection. In other words, only features whose Hessian threshold is larger than the defined value are retained by the SURF detector. This means that the larger the threshold, the lower the number of key points identified. Figure 13 in the main text illustrates the points that a SURF detector identifies when this threshold changes. It is clear from these pictures that a low threshold of 300 yields a high number of points detected. Substantively this means that the detector considers even minor changes in pixel intensity as salient. In contrast, a detector with a large threshold focuses on the most prominent pixel intensity changes and therefore retains only the most salient key points in an image. The decision to have a higher or lower threshold depends on the research objectives of the user as well as on the nature of the images under study. For example, images in canonical computer vision datasets like CALTECH 101 tend to depict one or a few close-up objects with low complexity. In this case, a threshold between 300 and 500 would retain an adequate number of features to make subsequent classifications. However, as Figure 13 shows, the complexity in the composition of the images under analysis is high, which leads to a very high number of key points identified when using a low threshold. Several of these key points do not contribute with meaningful information about the image (e.g. points along the lines on the pavement). However, a high threshold misses a few key points that provide information about the environment and set up of the event depicted in the picture. Thus, a number in the middle is an adequate option.

How would the number of key points impact the construction and application of the BoVW? In a nutshell, more key points represent more information about the picture. This extra information is represented by a larger number of features that especially impact the clustering process. First, the process is more computationally expensive due to the higher volume of features. Second, for a fixed number of clusters (visual words) v, the within-cluster sum of squared errors (WSSE) will be larger as the number of key points increases. Thus, if the objective is to achieve low WSSE while also keeping a low number of visual words, a lower number of key points is preferred.

Consider Figure A.7. For the same number of clusters v (on the x-axis), the WSSE is the lowest when using a threshold of 3,000 (around 19,000), and the highest with one of 300 (around 66,000). However, it is interesting to notice that while in absolute terms, as expected, the model in the bottom panel minimizes the WSSE, all models show a similar trend in the reduction of WSSE as the number of clusters increases. From the plots and using the "elbow" method, the chosen number of clusters would be 2,000 in all cases.

The effects on topic quality are illustrated in Figure A.8, where the different colors correspond to the different thresholds. For held out likelihood, the pattern is remarkably similar between the models with the different thresholds as the number of topics increases. Although the differences in exclusivity between the models are not stark, it is worth highlighting that this indicator is higher when there are more key points. For semantic coherence, the differences are larger even when comparing models with the same number of visual words. In this case, those with lower thresholds show more semantic coherence. Given that more key points retain more information about the picture, it is not surprising to find that the topics are better described when using additional features. However, this is also conditional on increasing the number of visual words in the vocabulary to preserve the cohesiveness of the clusters identified.

Figure A.7: Comparison of WSSE by Hessian threshold



A.6.2 Number of visual words in the vocabulary

The second step in the process of building a BoVW consists of clustering features to create a codebook composed of visual words. Recall that unlike text, images do not contain easily identifiable tokens such as words or sentences. Thus, this step is necessary to create "references" and tokens that serve as the columns of an IVWM. To achieve this, we simply apply a clustering technique to a sample of features extracted from a reference dataset. Once the clusters are identified, the centroid of each of these will represent a visual word. Researchers have flexibility to choose the clustering algorithm. However, this article uses and recommends a mini-batch k-means which achieves good levels of accuracy while also being computationally efficient. Given the number of features and the large number of images that researchers generally have under analysis, efficiency and speed are two elements that researchers might favor.

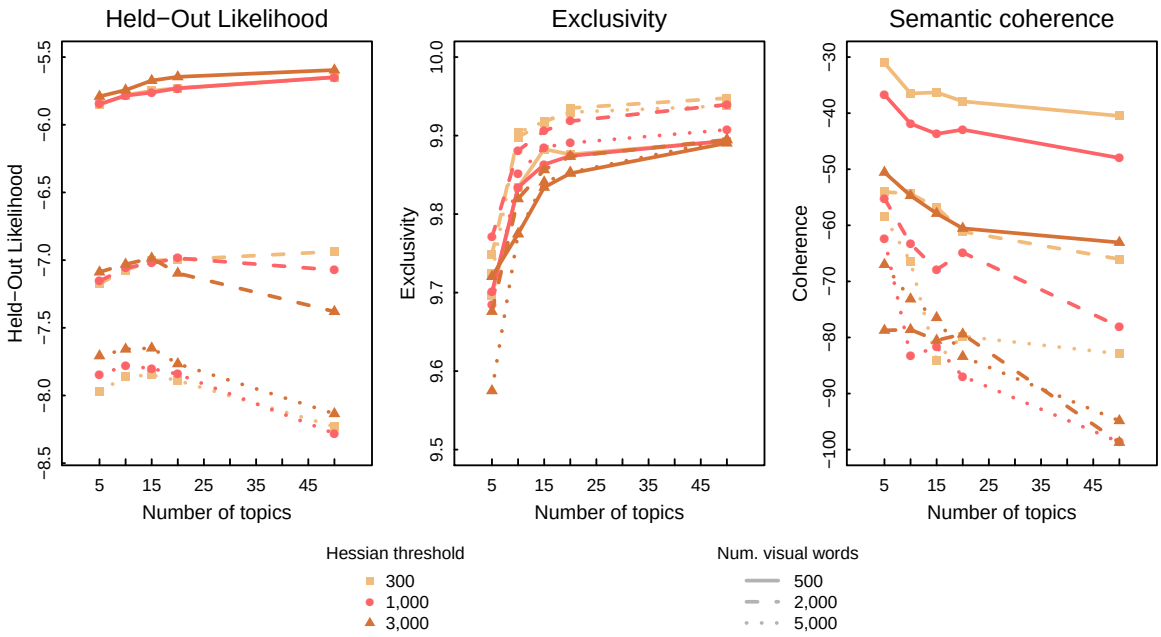
Given that the most commonly used clustering methods require the specification of a "number of clusters" parameter, the question about the optimal number of clusters to choose is a natural one. While the answer is not straightforward and depends on issues like the substantive motivation and objectives of a project, as well as the number of features extracted from the images under study, there are certain practices that researchers can follow to inform the decision of how many clusters to choose. In the main text we suggest diagnosis tools like the "elbow" method which is based on the assessment of within-cluster sums of squared errors, and also more qualitative alternatives like visual inspection.

Using the latter, I find that the number of clusters have an effect on the discovery of topics. While there are topics and frames that can be clearly identified regardless of the parameter definition, even in small models like a STM with 5 topics, there are still differences in the composition of the topics across the different vocabularies. In this particular case, while topics like "dense crowd", "sky" or "dark background" where constant across the three visual vocabulary categories, the model with the smallest vocabulary yielded a topic whose distinctive features corresponded to light backgrounds and "sand" like texture. In contrast, the longer vocabularies containing pieces with text and drawings on similar backgrounds contributed to the formation of a topic with "maps and infographics" rather than just "light backgrounds". When the number of key points and visual words increase, the information retained from the pictures also increase and lead to a finer distinction between pictures. This might be something desirable depending on the characteristics of the study and the research objectives of the user.

Here we analyze the impact that this number might have on topic quality as well as on the results from effect estimation of prevalence covariates in topic models. Figure A.8 shows the result for the former. As we can see, the first panel shows that models with the "simpler" vocabulary with 500 words have a higher held-out likelihood than those with 2,000 words, and these in turn, higher than the ones with 5,000 visual words. This indicator does not vary dramatically within Hessian threshold categories. With respect to exclusivity, the lowest levels come from the models with the smallest number of visual words. The relationship between this parameter and exclusivity does not seem to be linear. While the simpler vocabulary has the lowest levels of exclusivity, the models with 2,000 words (middle category) evidence the highest levels of exclusivity. In contrast, the semantic coherence seems to be the highest among models with the shortest vocabularies, and it decreases as

the number of visual words grows. This is in line with the intuition that more words, and the added complexity that these represent, make topics less cohesive given the increased number of components to consider. However, it is important to highlight that these statistics also interact with other parameters such as key point detection. For example, the differences in semantic coherence are not large between models with 500 visual words and Hessian thresholds of 300 or 1,000. However, this indicator decreases sharply when the Hessian threshold is 3,000. The lowest levels of semantic coherence are found in models with low number of key points but high levels of clustering. Overall, these findings highlight the importance or running multiple diagnosis exercises and comparisons between parameters to understand the particular effects of certain choices on indicators like the ones discussed in this section.

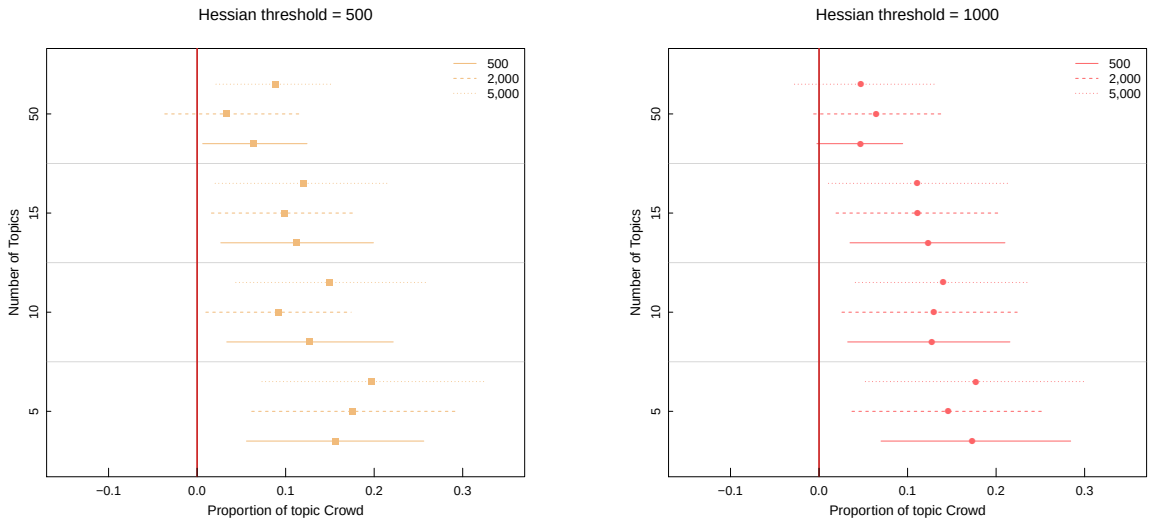
Figure A.8: Quality indicators across BoVW designs



With respect to the impact that the number of clusters has on the estimation of an effect of a prevalence covariate on a topic of interest, the results are remarkably similar within number of topic and Hessian threshold groups. Figure A.9 shows three plots corresponding to different Hessian thresholds. The lowest with 300 contains the highest number of key points while the one with 3,000 has the lowest. The rows in each plot correspond to a STM initialized with k topics. The estimates with their respective confidence intervals correspond to the difference in "dense crowd" proportions used in images published by right leaning and left leaning outlets. While there are strong differences between the results of models with high and low numbers of topics, within each of these categories the differences between number of visual words are very small. In a few cases, the results and their implications are different depending on the length of the vocabulary (based on statistical significance), but for the most part the results and findings are the same across categories. This is also due to

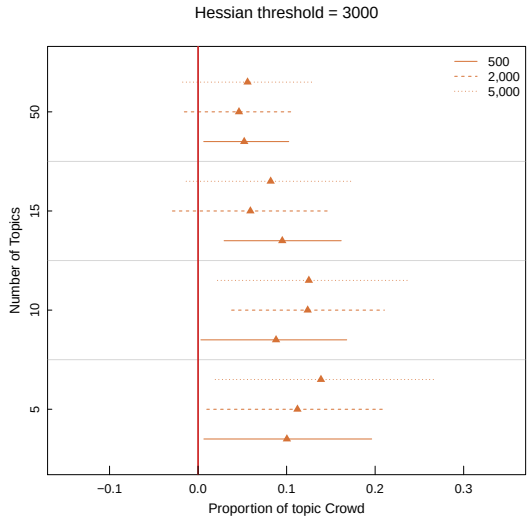
the distinctiveness of the topic under analysis. "Dense crowd" was a topic that was clearly identified in all the topic models regardless of the specification of the parameters. However, these results might differ in the cases where the parameters substantially affect the topic structure or composition. Further investigation regarding these effects is required.

Figure A.9: Comparison of "ideology" effects by number of topics, vocabulary, and Hessian threshold



(a) Hessian threshold = 500

(b) Hessian threshold = 1,000



(c) Hessian threshold = 3,000

A.6.3 Number of topics

Finally, I implemented a series of topic models varying the number of topics, and across different BoVW specifications with respect to number of key points detected (Hessian threshold) and number of visual words in the vocabulary. From visual inspection of the topic composition, there are substantive differences between the models. While smaller models with 5 topics cluster many different pictures into a "crowd" topic containing visual words with small human bodies and granular textures, increasing the number of topics allows for a more careful differentiation of topics like "crowd". For example, setting the STM to 15 topics allows for the identification of different types of crowds: "groups walking" that not only contain visual words of people but also of pavement, or "outside crowds" where the focus is more on environmental factors like the sky than the people. The differences between these frames and the role they play in the research projects should help with the user's decision regarding the selection of topics. Other methods aimed to assess topic quality, such as semantic coherence, held-out likelihood, and exclusivity are also suggested as part of the STM routine to evaluate the optimal number of topics to keep as illustrated in Figure A.8.

Finally, beyond the differences that we observe in topic quality and composition, it is also important to assess how the number of topics affects some of the potential inferences that we can make about the effect of prevalence covariates on the generation of visual frames. In this case, if we focus on whether the ideology of the news outlet (right vs. left) has an effect on the proportion of topic "dense crowds" in the pictures that such outlets use, and compare these differences between topic models, we get some interesting results. First, although the results are different between model set ups, in most of them the substantive finding that right leaning outlets publish pictures with higher proportions of dense crowds than left leaning outlets does not change. Second, the effect size decreases as the number of topics increases. This is a natural consequence of increasing the number of topics: the proportions of a single topic go down as the possibility of containing other similar but new themes arises. Thus, finding a difference distinguishable from zero as the proportions of a given topic become smaller gets more complicated as the number of topics increase. In Figure A.9, we can observe that although we still find a positive and reliable difference between right and left outlets, this differences becomes indistinguishable from zero when we have 50 or more topics. This is solely based on the estimation of the proportion "dense crowd" topic and does not consider summing up the proportions of topics associated with "crowds".

A.7 Continuous vs. Binary classifications

Binary quantifiers of a picture such as "Presence of a crowd" might be fitting for certain projects. However, as in the case of visual framing, if we do not only care about the mere presence but also the way in which a given object or theme is included in a picture, then other measures are more adequate. In the following table, I regress the proportions of the "dense crowd" topic per image (Column 1 of Table 1) and the indicator of whether there is a crowd (manually coded, Column 2 of the same table) on the ideology of the news outlet. The use of a continuous measure, the proportion of a frame, might not only be useful for research objectives but may also uncover interesting results. In this case, the proportion of crowds that left leaning outlets use is significantly lower than the proportion used by outlets in the center. Although the model with a binary indicator also shows a negative coefficient for left-leaning outlets, this coefficient is not distinguishable from zero at conventional levels.

Table 1: Binary vs. continuous measures of "crowds"

	Dense Crowd proportion.	Presence of a crowd
	OLS	Logistic
	(1)	(2)
Left	**−0.044**	−0.685
	(0.021)	(0.356)
Center left	−0.004	0.050
	(0.017)	(0.271)
Center right	−0.017	0.687
	(0.031)	(0.424)
Right	**0.092**	**0.784**
	(0.029)	(0.397)
Not rated	−0.013	1.440
	(0.064)	(0.864)
Constant	0.023	−17.254
	(0.176)	(2,399.545)
N	688	688
R^2	0.073	
Adjusted R^2	0.027	
Log Likelihood		−301.910
AIC		671.820

Bolded coefficients indicate $p<0.05$

A.8 References used in the Appendix

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.

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

Lowe, David G. 1999. Object recognition from local scale-invariant features. In *Proccedings of the International Conference on Computer Vision.* IEEE Corfu: p. 1150-1157.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2014. "stm: R package for structural topic models." Journal of Statistical Software 10(2): 1-40.

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.