

# Protest Activity Detection and Perceived Violence Estimation from Social Media Images

Donghyeon Won  
UCLA  
dh.won@ucla.edu

Zachary C. Steinert-Threlkeld  
UCLA  
zst@luskin.ucla.edu

Jungseock Joo  
UCLA  
jjoo@comm.ucla.edu

## ABSTRACT

We develop a novel visual model which can recognize protesters, describe their activities by visual attributes and estimate the level of perceived violence in an image. Studies of social media and protests use natural language processing to track how individuals use hashtags and links, often with a focus on those items' diffusion. These approaches, however, may not be effective in fully characterizing actual real-world protests (e.g., violent or peaceful) or estimating the demographics of participants (e.g., age, gender, and race) and their emotions. Our system characterizes protests along these dimensions. We have collected geotagged tweets and their images from 2013-2017 and analyzed multiple major protest events in that period. A multi-task convolutional neural network is employed in order to automatically classify the presence of protesters in an image and predict its visual attributes, perceived violence and exhibited emotions. We also release the **UCLA Protest Image Dataset**, our novel dataset of 40,764 images (11,659 protest images and hard negatives) with various annotations of visual attributes and sentiments. Using this dataset, we train our model and demonstrate its effectiveness. We also present experimental results from various analysis on geotagged image data in several prevalent protest events. Our dataset will be made accessible at <https://www.sscnet.ucla.edu/comm/jjoo/mm-protest/>.

## CCS CONCEPTS

- **Information systems** → **Multimedia information systems**;
- **Computing methodologies** → **Computer vision**; **Activity recognition and understanding**; *Scene understanding*;

## KEYWORDS

Protest; Action and Activity Recognition; Scene Understanding; Social Media Analysis; Visual Sentiment Analysis

## ACM Reference Format:

Donghyeon Won, Zachary C. Steinert-Threlkeld, and Jungseock Joo. 2017. Protest Activity Detection and Perceived Violence Estimation from Social Media Images. In *Proceedings of MM '17, Mountain View, CA, USA, October 23–27, 2017*, 9 pages. <https://doi.org/10.1145/3123266.3123282>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '17, October 23–27, 2017, Mountain View, CA, USA

© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

ACM ISBN 978-1-4503-4906-2/17/10...\$15.00

<https://doi.org/10.1145/3123266.3123282>

## 1 INTRODUCTION:

### IMAGES AND POLITICAL CONTENTION

Online social media have served as an open information channel which hosts public discussions on a number of social and political issues. Individuals respond to real world events in social media, and their responses can influence various social events and provoke public movements. As an alternative to traditional mass media outlets, social media, its users, and the content shared on them are often independent of government authorities. In particular, the possible impact of social media on *protests* has been analyzed in the context of the Arab Spring [42, 43], social movements in Europe [15], and election protests in Russia [10].

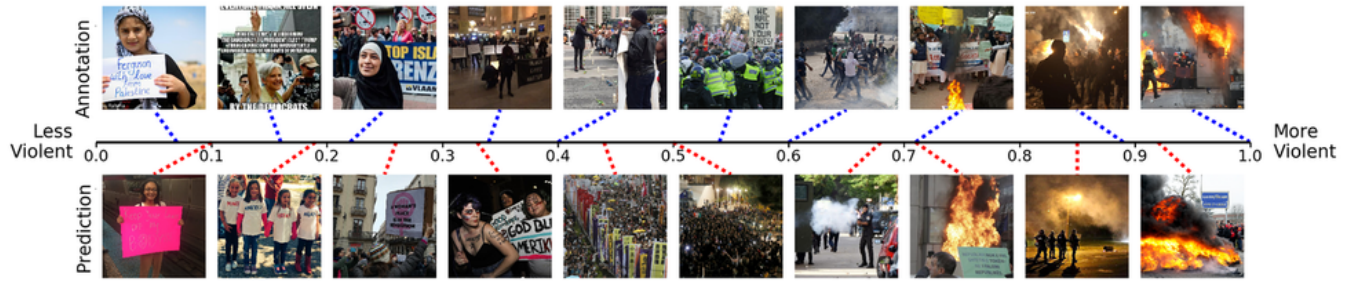
Social scientists have long studied protests, but the difficulty of acquiring and processing large-scale data has limited what questions are answerable. Work in this field is therefore dominated by formal and qualitative models [23, 27, 29, 45] or surveys of protest participants after a protest occurs [31, 44]. The spread of information and communications technologies (ICT), like the Internet and cell phones, has generated a new burst of theorizing and testing, with scholars modeling whether these technologies lead to more protest [25, 40] and using data on hundreds of thousands to millions of people across countries and times to test these models [3, 9, 15, 42].

With such data, scholars can now measure the behavior of various people across multiple cities and countries for weeks, months, or years. The advantage of these data has been that scholars can see what protesters say. In the last few years, however, accounts have started to share **images** with greater frequency, and scholars have yet to analyze what protesters show.

The objective of this paper is to develop an automated system that analyzes what images are shared during protests and how they change over space and time. In doing so, our visual approach identifies salient characteristics of protests – especially **violence** – by automatically assessing visual activities and attributes described in a given scene.

Violence is a critical dimension of protest in understanding social mobilization, as violent protests typically generate a much higher level of media and public attention. There might be other cues that one could use to approximate the level of violence in a protest, such as police or government statements or the number of people who have been killed, injured, or arrested. However, this information can be often inaccurate or not provided at all to the public in an official channel. Therefore, the goal of our study is to take advantage of unfiltered stream of data in social media and to assess the level of perceived violence for protest events.

The key contributions of our paper can be summarized as follows.



**Figure 1: Sample images in our Protest Image dataset ordered by their perceived violence scores: (top) annotation (bottom) prediction.**

- We have collected, and will release, a novel dataset of protest images with human coded data of perceived violence and image sentiments. Our dataset is an order of magnitude larger than existing social event image datasets and provides fine-grained annotations which are directly relevant to protest research.
- We train a model based on a Convolutional Neural Network (CNN) that can automatically classify the content of images, especially perceived violence and sentiments, all of which are inferred jointly from the shared visual feature representation.
- We have collected geotagged tweets and associated images across the world from August 2013 to detect, track, and analyze various protests in different countries. In this paper, we analyze and compare five protest events including Black Lives Matter and Women’s March. Our analysis reveals that the degree of predicted perceived violence differs significantly across events and also across states within an event.

## 2 RELATED WORK

Although understanding protest and violence has been a critical topic of research in political science, there are few work in the fields of political science and media studies which attempts to automatically analyze visual or multimodal data due to the lack of proper methods and datasets.

While this paper is the first work that analyzes what images are shared during protests and how those change across events, recent studies in “**social**” **multimedia and computer vision** employ large-scale visual content analysis to tackle related research questions in political science, media studies and communication. For instance, facial attribute classification (gender, race, and age; [26]) has been used to examine the supporter demographics of major politicians in the U.S. using profile photographs of Twitter users [46]. Researchers have also analyzed photographs of politicians shared on social media [49] or their perceived personalities from facial appearance [21]. Public opinion about politicians has been also studied in relation to visual portrayals and persuasions in mass media [20] and presentations in social media [2]. These works all highlight the importance of the visual cue in human perception of media content and the advances in multimedia and computer vision have enabled to recognize subjective, perceived dimensions

of images or videos, e.g., creative [38], interesting [16], or sexually provocative [13].

Among more traditional work in **multimedia**, our paper is closely related to social event detection or classification [6, 35–37, 39, 48]. While these studies focus on identifying the same type of event (i.e., clustering) or classifying event type, we specifically concentrate on the protest event and investigate various ways to characterize them.

In addition, there have been a few works which propose to automatically classify violent activities from images or videos by static image or motion features [7, 8, 17, 32]. These existing studies on violence detection are mostly concerned with physical violence such as physical fights between players in sports games [32], detecting bloody frames in movies [7], or aggressive behavior in short videos [8]. Our work clearly differs from these works as we focus on perceived violence in protest activities of various types which include not only physical assaults, but also rallies, demonstrations, or even peaceful gatherings.

## 3 UCLA PROTEST IMAGE DATASET

In this section, we describe our dataset of social protest images. It is designed to support studies in protest activity detection, fine-grained attribute recognition, and visual sentiment analysis.<sup>1</sup>

Prior studies have proposed image datasets for general social event detection such as music concerts or birthday parties. However, our main research topic, protest and public mobilization, has not been sufficiently addressed in these datasets because they lack other rich annotations. The social event detection benchmark [39] has released an image dataset containing images of the protest category; however, a very small portion of the dataset was composed of protest images (800-1000 images) and the images do not have any other annotations than categorical information. Thus, this dataset is insufficient to conduct in-depth studies specifically aimed at analysis of protest images.

Our dataset contains 40,764 images which have been collected from Twitter and other online resources. 11,659 images are protest images identified by annotators and the rest are hard-negative images (e.g., crowd in stadium). A few negative examples are shown in Fig. 2 and positive examples with annotations and prediction scores of violence are presented in Fig. 1 and Fig. 3. Each positive

<sup>1</sup><https://www.sscnet.ucla.edu/comm/jjoo/mm-protest/>.



**Figure 2: Hard negative examples (non-protest) in our dataset. The images exhibit visual features common in protest images such as fire, a group of people, or a weapon.**

protest image was annotated for its visual attributes (e.g., children, fire, or large crowd) and image sentiment.

### 3.1 Image Collection

Our model should be able to distinguish between a protest crowd and other large gathering such as concerts or sporting events. It should also distinguish between non-violent and violent protests. In order to effectively capture diverse visual patterns of protests and train a robust model, we collect images from multiple sources by web search (Google Image search) and also our own Twitter data stream in a refining, active learning approach.

We first collected 10,000 images which may describe a protest scene by web search using a set of keywords. We manually selected some general keywords (e.g., “protest”, “riot”, “demonstration”) and also used the names of recent major protest events (e.g., “Black Lives Matter” or “Women’s March”) based on the Wikipedia page of the list of protests.<sup>2</sup> We trained our first, rough classification CNN, treating these noisy images as positive examples. For negative examples, we used keywords such as “concert” or “stadium.” Our model architecture is based on a 50-layer ResNet [18] (See Sec. 4). This model was applied back to the initial image set to filter out images of very low scores (i.e., easy negative) as they are unlikely to be protest scenes.

We then applied this classifier to random samples from our geo-tagged Twitter image collection (see Sec. 5.4 for detail), without any filtering by keywords, and obtained a set of images whose prediction scores were above a threshold. The optimal threshold was selected empirically from a set of hand-labeled images such that it prunes the majority of irrelevant images while keeping most of the positive examples. Therefore, this set contains many hard negative examples, such as flash mobs. These two sets of images were then merged and provided to the annotators from Amazon Mechanical Turk who labeled the presence or absence of a protester in each image.

<sup>2</sup>[http://en.wikipedia.org/wiki/List\\_of\\_riots](http://en.wikipedia.org/wiki/List_of_riots)

<sup>3</sup>[https://en.wikipedia.org/wiki/List\\_of\\_ongoing\\_protests](https://en.wikipedia.org/wiki/List_of_ongoing_protests)

### 3.2 Violence and Emotions in Protest Images

Each protest event is attended by different people who gather together for different purposes [12]. The degree of violence involved in a protest can also vary greatly with the participants and their demands [11]. Publicly shared photographs allow us to assess how a protest event is depicted: how **violent** it is and what kinds of **emotions** are expressed.

Many emotional dimensions, such as anger or fear, are highly correlated with perceived violence, but they sometimes capture different traits. The distributions of annotations also differ across dimensions, as shown in Fig. 4. For instance, protesters might be angry but still not violent. To reduce the annotation cost, we excluded the two emotional dimensions of surprise and disgust as they overlap with other dimensions.

In addition, we further identified common scene attributes associated with protest images and annotated them in order to analyze what kinds of visual attributes are correlated with the perceived violence or image sentiments. We first generated any related visual concepts for each image in a small subset of the image set and constructed the most common attributes among them which are shown in Table 1.

### 3.3 Image Annotation

We used Amazon Mechanical Turk (AMT) to obtain necessary annotations for each image in our dataset including (1) whether an image contains a protest activity or protesters, (2) visual attributes in the scene, (3) the level of perceived violence and other sentiments. The first two tasks require objective, binary annotations. We assigned two workers to each image for these tasks and confirmed the values when both workers agreed (if not, the image was sent to the third judge).

On the other hand, since our perceived violence is a subjective and continuous variable, we instead requested pairwise comparison-based annotations [33]. Specifically, we randomly sampled 58,295 image pairs among 11,659 protest images such that each image is paired up 10 times and assigned 10 workers to evaluate each pair. For each pair, the annotators were asked to choose an image which looks more violent than the other. We used the Bradley-Terry model [5] and estimated the global scores for images such that each image is assigned a real-number score of perceived violence.

The advantages of pairwise annotation method has been well understood in prior work [22, 33], but it requires much more annotations to be collected. To ease the burden of the overall annotation task, the remaining emotional sentiment annotations (angry, fearful, sad, happy) were obtained by individual evaluation (i.e., the annotator was given only one image at a time and asked to provide his response.) In both cases (violence and emotions), we obtained a scalar value in  $[0, 1]$ .

## 4 MODELS

We use two separate models to recognize protest activities in images. First, we train a CNN which takes a full image as input and outputs a series of prediction scores including the binary image category (i.e., protest or non-protest) (1), visual attributes (10), and perceived violence and image sentiment (1 + 4). Our model architecture is based on a 50-layer ResNet [18], which consists of 50 convolutional





Figure 3: Example images in our protest dataset with various scores obtained by our model: perceived violence, sentiments, and visual attributes.

Table 1: List of visual attributes.

Attribute	Description
Sign	A protester holding a visual sign (on paper, panel, or wood).
Photo	A protester holding a sign containing a photograph of a person (politicians or celebrities)
Fire	There is fire or smoke in the scene.
Law enf.	Police or troops are present in the scene.
Children	Children are in the scene.
Group 20	There are roughly more than 20 people in the scene.
Group 100	There are roughly more than 100 people in the scene.
Flag	There are flags in the scene
Night	It is at night.
Shout	One or more people shouting.

Table 2: The architecture of our model.

Layer	Output size	Building blocks			
conv1	112 × 112	7 × 7, 64, stride 2			
conv2	56 × 56	3 × 3 max pool, stride 2			
		1 × 1, 64	× 3		
		3 × 3, 64			
conv3	28 × 28	1 × 1, 128	× 4		
		3 × 3, 128			
		1 × 1, 512			
conv4	14 × 14	1 × 1, 256	× 6		
		3 × 3, 256			
		1 × 1, 1024			
conv5	7 × 7	1 × 1, 512	× 3		
		3 × 3, 512			
		1 × 1, 2048			
pooling	2048	average pooling			
classification	17	1-d fc (protest)	1-d fc (violence)	4-d fc (sentiment)	10-d fc (visual attribute)

layers with batch normalization and ReLU layers. The architecture of the model is briefly described in Table 2. The features computed through convolutional layers are all shared by linear layers for multiple classification tasks. We jointly train the model such that all parameters for 3 different tasks – protest classification, violence

and sentiment estimation, and visual attribute classification – are updated jointly. We use binary cross entropy loss to train our binary variables (protest and visual attributes) and mean squared error to train violence and sentiment dimensions.

In addition, another CNN captures various facial attributes from images. We use OpenFace [1] for our face model, which was developed for face recognition. We use the CelebA facial attribute dataset to train the attribute model. That model outputs gender, race, and

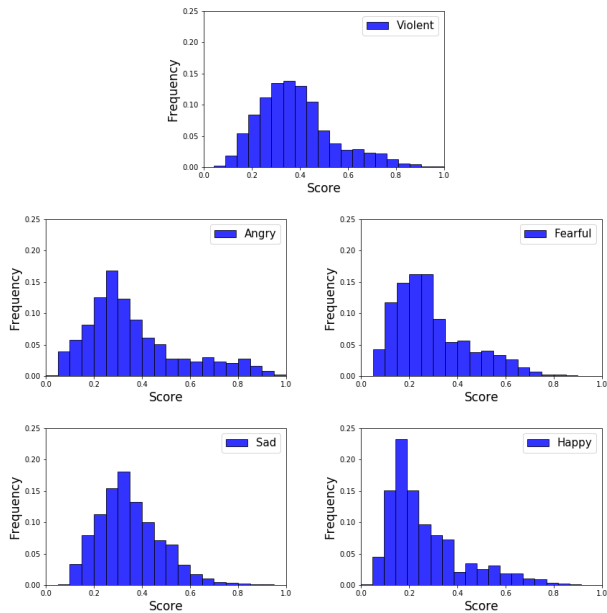


Figure 4: Distributions of perceived violence and image sentiment scores rated by annotators.

Table 3: Inter-rater reliability measured by Pearson’s correlation coefficients between two randomly split annotator groups.

	Violent	Angry	Fearful	Sad	Happy
Pearson’s $\rho$	.716	.568	.417	.362	.875

other expressions [26]. For each image, we use dlib’s face detection and alignment<sup>4</sup> and crop the internal facial region to feed into the facial CNN model. In our analysis, facial attributes are especially important because social scientists have theorized about the role of emotions in leading to and sustaining protests, but arguments have had to rely on qualitative models and case studies [30, 34, 47]. With our model, we can now test these theories with more precision than before.

## 5 RESULTS

This section presents various experimental results obtained from our analysis. We discuss the general performance of our model and then provide the results of actual analyses conducted on our geocoded tweet dataset of protests.

### 5.1 Inter-rater reliability

Table 3 reports the Pearson correlation coefficient for inter-rater reliability between two randomly split annotator groups. As we used Mechanical Turk to collect annotations, it is inappropriate to apply a standard test which typically assumes complete data. Therefore, we measure correlations between non-overlapping groups of

<sup>4</sup>dlib.net

Table 4: Pearson’s correlation coefficients between visual sentiments and visual attributes, measured from annotations. We only print fields which are statistically significant ( $p\text{-val} < 0.0001$ ).

	Violent	Angry	Fearful	Sad	Happy
Violent		.671	.575	.351	-.359
Angry			.795	.626	-.427
Fearful				.752	-.219
Sad					-.195
Sign	-.479	-.549	-.495	-.288	.225
Photo	-.047				
Fire	.567	.578	.504	.297	-.184
Law enf.	.367	.417	.399	.239	-.186
Group >100	.152	-.166	-.279	-.147	
Night	.206	.183	.143	.086	-.129
Shouting		.106			-.087

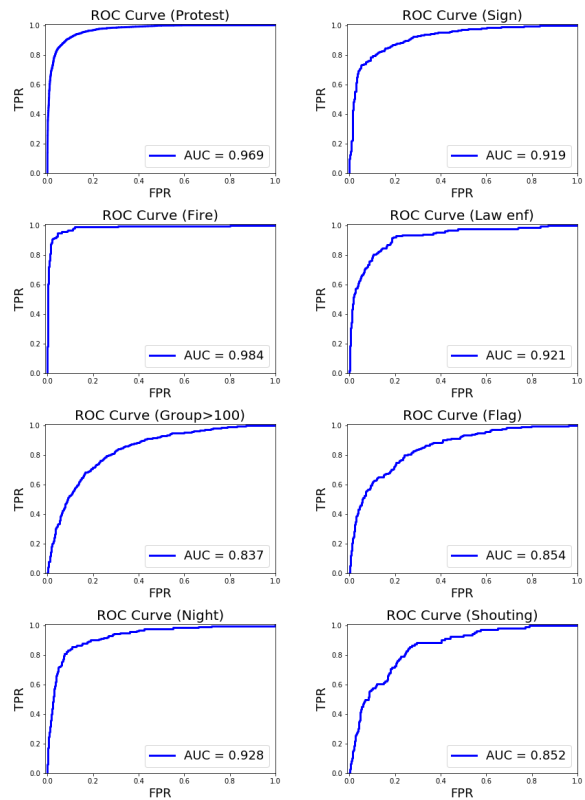


Figure 5: ROC curves for protest image and visual attribute classifications.

workers. This method has also been frequently used in the literature [19]. The results are all statistically significant.

### 5.2 Model Performance

5.2.1 *Protest Scene and Attributes Classification.* We randomly split our entire dataset into a training set (80%) and the test set

**Table 5: Performance of protest scene and attributes classification. AUC is area-under-curve in a ROC curve.**

Fields	Protest	Sign	Photo	Fire	Law	Children
Pos. rate	.286	.829	.036	.057	.067	0.030
AUC	.969	.919	.738	.984	.921	.813
Fields		Group > 20	Group > 100	Flag	Night	Shout
Pos. rate		.730	.252	.083	.084	.047
AUC		.795	.837	.854	.928	.852

**Table 6: Perceived violence and image sentiment prediction accuracy of our model measured by Pearson’s correlation coefficients and  $r^2$  values.**

	Violent	Angry	Fearful	Sad	Happy
Pearson’s $\rho$	.900	.753	.626	.340	.382
$r^2$	.809	.566	.392	.116	.146

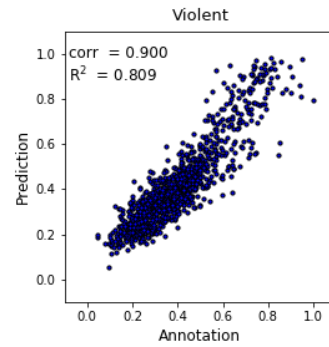
(20%). Table 5 shows the classification accuracies for protest scene classification and visual attribute classification, measured on the test set. The ROC curves for selected variables are also shown in Fig. 5. Some variables such as ‘children’ or ‘shouting’ only have a very small number of positive examples, but our model in general achieved reasonable accuracies for most variables.

**5.2.2 Violence and Sentiment Estimation.** We also measure the accuracy of our model in estimating perceived violence and image sentiments on emotional dimensions. Table 6 reports the Pearson’s linear correlation coefficients and the coefficients of determination ( $r^2$ ) between human annotations and our model’s predictions, measured on the test set. Fig. 6 shows the scatter plot of annotations and predictions.

We found that our model performs very well in predicting image violence. It is less accurate for emotional sentiments; we believe this is at least partly because the individual annotation scheme (vs. pairwise) sometimes led to less consistent annotations across annotators (e.g., due to the lack of a reference scale).

Fig. 3 shows qualitative examples in our test dataset. While our model successfully predicts violence and image sentiment, there are some difficult cases where the prediction does not match annotators’ ratings. We found the most important factor that our model does not address very well is a semantic relation between uncommon visual feature (symbolic gestures such as “die-in” in Black Lives Matter) and their meanings and associated emotions. For instance, in the bottom-left image, people demonstrate at the protest by pretending that they are casualties. However, the model might have treated this group gesture as people who are actually wounded, and some images in our dataset contain actually wounded people.

**5.2.3 What makes a protest image violent?** We now analyze visual attributes common in images which annotators rate as violent. We identify these features by measuring correlations between visual attributes and perceived sentiments (from annotations). As shown in Table 4, annotators find images with more dangerous (or potentially more dangerous) physical activities (‘fire’ and ‘law enforcement’) as more violent, angry and fearful than images with a ‘big group’ of people holding ‘signs.’



**Figure 6: A scatter plot of perceived violence in annotations and predictions.**

**Table 7: Correlation coefficients between facial attributes and perceived violence and image sentiment. (p-values < 0.0001)**

	Violent	Angry	Fearful	Sad	Happy
Male	.120	.146	.145	.137	-.151
White	-.166	-.172	-.171	-.160	.158
Black	.189	.193	.194	.180	-.143
Smile	-.151	-.196	-.186	-.145	.229
Frown	.181	.223	.210	.175	-.237

Another important set of visual features are human attributes, including expressions or demographic information about human subjects in a scene. We assess these features from their faces. The correlations between the facial attributes and sentiment predictions are presented in Table. 7.

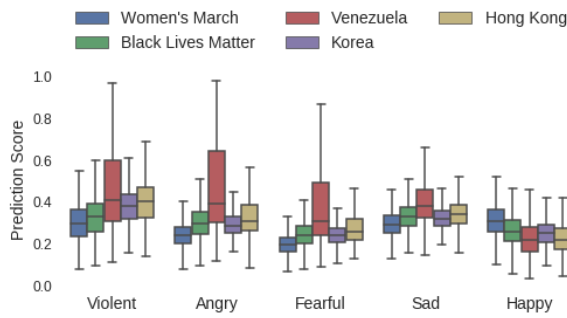
From this analysis, we find smiling faces negatively correlate with perceived violence and other emotional dimensions such as angry or fearful (but not happy). Perceived violence also differs across gender and race groups. This could arise if some demographic sub-groups are involved in a more violent protest activity captured in each photograph, or the level of violence varies in different underlying protest events which have different groups of participants (but the violence is exogenous to the participants).

### 5.3 Protest Event Analysis

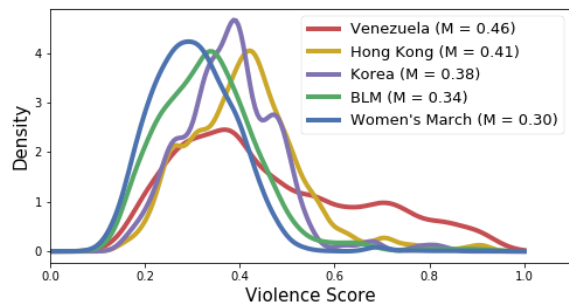
The key advantage of using photographs in our analysis is to assess various non-verbal characteristics of protest events. The following subsections present the results obtained by applying the trained model to the tweet data stream collected over the past three years.

Fig. 7 shows that our image analysis reveals two interesting results when comparing the Women’s March, Black Lives Matter, protests in South Korea, Hong Kong, and Venezuela.<sup>5</sup> First, protests in Venezuela are more violent and angrier than the other protests; the standard errors are large, but the effect is persistent across the

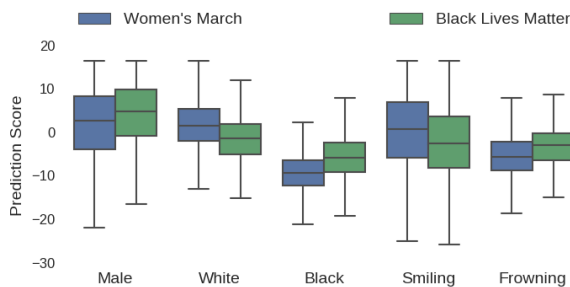
<sup>5</sup>We obtained event specific tweets in the following way. For Black Lives Matter, which has spanned for 3 years across the country, we simply filtered tweets by the hashtag of #BlackLivesMatter. For other events, we specified the region and date for which we know the protests happened and classified images to obtain protest related tweets.



**Figure 7: Predicted violence and sentiments of tweet images in different protest events. The box represents the range between the first and third quartiles.**



**Figure 8: Distribution of predicted violence scores of images in different protest events.**



**Figure 9: Predicted face attributes in tweet images from the Women's March and Black Lives Matter. They are separate dimensions, so one cannot directly compare the absolute score of White and that of Black. The box represents the range between the first and third quartiles.**

five emotions. This result matches our prior beliefs, as Venezuelan protests, especially recently, have been violently repressed [28]. Second, the Women's March is the least violent and angry, and this pattern holds across the five emotions as well. The statistics of these two events are statistically significant (p-val < 0.00001) when

compared to the other protests. The distributions of violence scores for each event are shown in Fig. 8.

From Fig. 9, we can observe that more females appeared in Women's March protest images, and the proportion of African-Americans in images was higher in Black Lives Matter images. This result also matches our beliefs. We manually verified the gender ratio of these two events by counting randomly sampled 500 detected faces. In case of Women's March, the gender ratio between male and female was 0.22 : 0.78. On the other hand, in Black Lives Matter images, the gender ratio was 0.46 : 0.54.

Fig. 10 shows the difference of predicted visual attributes in different events. Venezuelan protesters used less protest signs than protesters in other countries. The event that had most protest signs was Women's March. In contrast, fire was detected in Venezuelan protest images most frequently. Also, images related to law enforcements appeared most frequently in Venezuelan protests. These are all the indicators of a higher level of violence as discussed above. Another interesting result to note is that images with large groups appeared most frequently in Korean protest images since the recent Korean protests were very effectively organized on every Saturday for a few months.

### 5.4 Geo-coded Tweet Analysis

We have collected tweets from August 26, 2013 to the present, asking Twitter's streaming API for only tweets containing GPS coordinates. Twitter returns all tweets with GPS coordinates up to 1% of the total volume of tweets at a given time; since approximately 2% of tweets contain GPS coordinates [4, 24], we estimate we have collected half of all tweets with GPS coordinates (approximately 6.4 billion). We therefore have precise location information for all tweets and their images in our dataset, allowing us to track the spatial distribution of the protest event coverage in Twitter.

Fig. 11 shows the spatial distributions of frequencies of the #BlackLivesMatter hashtag and violent protest images in 2014-2016. Note as well that our classifier detects more violence in Missouri, Maryland, and New York. Each state was the site of major protests after the deaths of Michael Brown (Ferguson, Missouri), Freddie Gray (Baltimore, Maryland) and Eric Garner (New York City, New York). Investigating the temporal variation of the predictions of our classifier will provide more insight, as the classifier's violence predictions should spike during the protests, but the initial correlation between the classifier's prediction and our understanding of events is encouraging.

### 5.5 Multimodal Cues: Visual vs. Textual

In order to examine the alignment between visual and textual cues in tweets, we measured the correlation between the text sentiment and the predicted image sentiments. The result is shown in Table 8 and Fig. 12 shows two sample tweets. We used python's VADER (Valence Aware Dictionary and sEntiment Reasoner) package[14] which provides the sentiment measure from text. 12,055 tweets with protest images from Black Lives Matter and 10,566 tweets with protest images from Women's March were used to calculate the correlation. As expected, the visually inferred violence correlates negatively with positive text sentiment. However, the strength



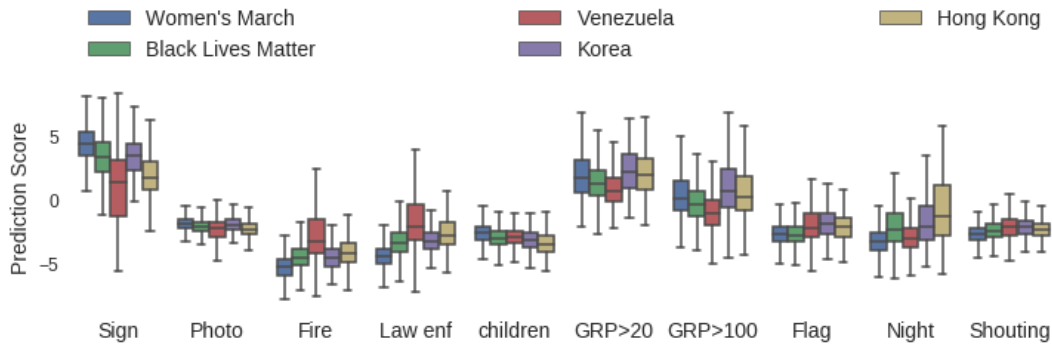


Figure 10: Predicted visual attributes in tweet images in different protest events. The box represents the range between the First and Third quartiles.

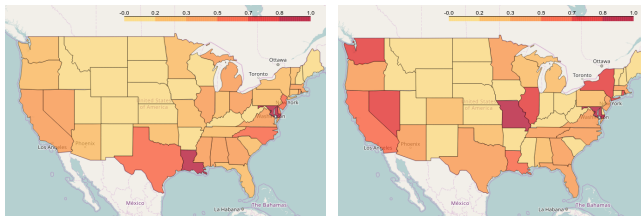


Figure 11: Spatial distributions of statistics related to Black Lives Matter movement. (left) The frequency of the hashtag of BlackLivesMatter. (right) The frequency of violent protest images. The statistics are normalized by the number of users in each state.



(a) Text:-.599, Violence: .948 (b) Text:-.946, Violence: .251

Figure 12: Two example tweets with text sentiment scores and image violence scores: (left) “#BlackLivesMatter okay #WhiteLivesMatter okay but #DemsLivesMatter UGH not so much when this happens.” (right) “protesting ignorance and fear that lead to hate and violence. #blacklivesmatter end #policeshooting”

of the correlation is very weak, although they are statistically significant. This might be due the fact that tweet texts are very short or strong texts do not necessarily accompany strong images, and vice-versa.

## 6 CONCLUSION

We have presented new approaches to estimate violence and protest dynamics from social media images. As our method is primarily

Table 8: Correlation coefficient between predicted text sentiment and perceived violence and image sentiment

	Violent	Angry	Fearful	Sad	Happy
Pearson’s $\rho$	-.080	-.085	-.088	-.090	.047
P-Value	$9.3 \times 10^{-34}$	$1.3 \times 10^{-37}$	$7.5 \times 10^{-40}$	$5.4 \times 10^{-45}$	$1.3 \times 10^{-12}$

based on visual analysis, it can generalize easier than textual analysis so long as visual language is more universal than spoken language.

We constructed a large-scale novel dataset, UCLA Protest Image Dataset, which contains more than 10k protest images with their perceived violence values manually annotated. We will release the dataset with all the annotations collected for perceived violence and attributes. Using this data and a model trained on it, we have presented the results of our analysis on various past and on-going protest events in the world.

Research in media studies and political science has suggested that the visual dimension of human communication can play a significant “persuasive” role in shaping public opinions [20, 41]. Our study demonstrates that the advances in computer vision and multimedia enable to systematically and automatically measure the impacts of visual media content to major social events in our society.

Multimedia research has long investigated human emotion processing by computational approaches on large scale multimodal data. While its applications have reached out to a number of different disciplines, understanding social and political activities and their meanings and implications have been relatively overlooked. Therefore, our paper suggests a novel collaborative area of research between multimedia and political science.

## 7 ACKNOWLEDGEMENTS

This research is supported by UCLA TSG Program, “Visual Big Data: Using Images to Understand Protests.” We also acknowledge the support of NVIDIA Corporation for their donation of hardware used in this research.



## REFERENCES

- [1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *Open-Face: A general-purpose face recognition library with mobile applications*. Technical Report. CMU-CS-16-118, CMU School of Computer Science.
- [2] Lefteris Anastasopoulos and Jake Williams. 2016. Identifying violent protest activity with scalable machine learning \*. (2016). <http://scholar.harvard.edu/janastas>
- [3] Pablo Barberá, Ning Wang, Richard Bonneau, John T. Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. 2015. The Critical Periphery in the Growth of Social Protests. *PLoS ONE* 10, 11 (2015), 1–15. DOI : <https://doi.org/10.7910/DVN/WCXK3Z>. Funding
- [4] Marco Bastos, Raquel Recuero, and Gabriela Zago. 2014. Taking tweets to the streets: A spatial analysis of the Vinegar Protests in Brazil. *First Monday* 19, 3 (2014), 1–27.
- [5] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.
- [6] Markus Brenner and Ebroul Izquierdo. 2012. Social event detection and retrieval in collaborative photo collections. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. ACM, 21.
- [7] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. 2011. Violence detection in movies. In *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*. IEEE, 119–124.
- [8] Filipe DM De Souza, Guillermo C Chavez, Eduardo A do Valle Jr, and Arnaldo de A Araújo. 2010. Violence detection in video using spatio-temporal features. In *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*. IEEE, 224–230.
- [9] Jesse Driscoll and Zachary C. Steinert-Threlkeld. 2017. Structure, Agency, Hegemony, and Action: Ukrainian Nationalism in East Ukraine. (2017).
- [10] Ruben Enikolopov, Alexey Makarin, and Maria Petrova. 2016. Social Media and Protest Participation: Evidence from Russia. (2016).
- [11] Matthew Feinberg, Robb Willer, and Chlose Kovacheff. 2017. Extreme Protest Tactics Reduce Popular Support for Social Movements. (2017). DOI : <https://doi.org/10.1007/s10551-015-2769-z>. For
- [12] Dana R. Fisher. 2014. Studying Large-Scale Protest: Understanding Mobilization and Participation at the People's Climate March. (2014).
- [13] Debashis Ganguly, Mohammad H Mofrad, and Adriana Kovashka. 2017. Detecting Sexually Provocative Images. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 660–668.
- [14] CJ Hutto Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) <http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>.
- [15] Sandra Gonzalez-Bailon, Javier Borge-Holthoefer, and Yamir Moreno. 2013. Broadcasters and Hidden Influentials in Online Protest Diffusion. *American Behavioral Scientist* 57, 7 (mar 2013), 943–965. DOI : <https://doi.org/10.1177/0002764213479371>
- [16] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. 2013. Visual interestingness in image sequences. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 1017–1026.
- [17] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. 2012. Violent flows: Real-time detection of violent crowd behavior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 1–6.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [19] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 145–152.
- [20] Jungseock Joo, Weixin Li, Francis Steen, and Song-Chun Zhu. 2014. Visual Persuasion: Inferring Communicative Intent of Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 216–223.
- [21] Jungseock Joo, Francis F Steen, and Song-Chun Zhu. 2015. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision*. 3712–3720.
- [22] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2012. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2973–2980.
- [23] Timur Kuran. 1989. Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution. *Public Choice* 61, 1 (1989), 41–74.
- [24] Kalev H. Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18, 5-6 (2013), 1–33.
- [25] Andrew T. Little. 2015. Communication Technology and Protest. *Journal of Politics* 78, 1 (2015), 152–166.
- [26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*. 3730–3738.
- [27] Susanne Lohmann. 1994. The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-91. *World Politics* 47, 1 (1994), 42–101.
- [28] Virginia Lopez and Jonathan Watts. 2017. Deaths and injuries reported amid 'mother of all marches' in Venezuela. (apr 2017).
- [29] Doug McAdam. 1986. Recruitment to High-Risk Activism: The Case of Freedom Summer. *Amer. J. Sociology* 92, 1 (1986), 64–90.
- [30] Jonathan Mercer. 2010. Emotional Beliefs. *International Organization* 64, 01 (jan 2010), 1. DOI : <https://doi.org/10.1017/S0020818309990221>
- [31] Edward N. Muller and Karl-Dieter Opp. 1986. Rational Choice and Rebellious Collective Action. *The American Political Science Review* 80, 2 (1986), 471–488.
- [32] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. 2011. Violence detection in video using computer vision techniques. In *International conference on Computer analysis of images and patterns*. Springer, 332–339.
- [33] Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 503–510.
- [34] Wendy Pearlman. 2013. Emotions and the Microfoundations of the Arab Uprisings. *Perspectives on Politics* 11, 02 (may 2013), 387–409. DOI : <https://doi.org/10.1017/S1537592713001072>
- [35] Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2012. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. ACM, 23.
- [36] Georgios Petkos, Symeon Papadopoulos, Emmanouil Schinas, and Yiannis Kompatsiaris. 2014. Graph-based multimodal clustering for social event detection in large collections of images. In *International Conference on Multimedia Modeling*. Springer, 146–158.
- [37] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and M Shamim Hossain. 2015. Social event classification via boosted multimodal supervised latent dirichlet allocation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11, 2 (2015), 27.
- [38] Miriam Redi, Neil O'Hare, Rossano Schifanella, Michele Trevisiol, and Alejandro Jaimes. 2014. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4272–4279.
- [39] Timo Reuter, Symeon Papadopoulos, Giorgos Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher de Vries, and Shlomo Geva. 2013. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013*.
- [40] Jacob N Shapiro and David A Siegel. 2015. Coordination and security: How mobile communications affect insurgency. *Journal of Peace Research* 52, 3 (feb 2015), 1–11. DOI : <https://doi.org/10.1177/0022343314559624>
- [41] Stuart Soroka, Peter Loewen, Patrick Fournier, and Daniel Rubenson. 2016. The Impact of News Photos on Support for Military Action. *Political Communication* 33, 4 (2016), 563–582.
- [42] Zachary C. Steinert-Threlkeld. 2017. Spontaneous Collective Action: Peripheral Mobilization During the Arab Spring. *American Political Science Review* 111, 02 (2017).
- [43] Zachary C. Steinert-Threlkeld, Delia Mocanu, Alessandro Vespignani, and James Fowler. 2015. Online social networks and offline protest. *EPJ Data Science* 4, 1 (2015), 19. DOI : <https://doi.org/10.1140/epjds/s13688-015-0056-y>
- [44] Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls Pre-print. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*. Ann Arbor.
- [45] Gordon Tullock. 1971. The Paradox of Revolution. *Public Choice* 11 (1971), 89–99.
- [46] Yu Wang, Yuncheng Li, and Jiebo Luo. 2016. Deciphering the 2016 US Presidential Campaign in the Twitter Sphere: A Comparison of the Trumpists and Clintonists. In *Tenth International AAAI Conference on Web and Social Media*.
- [47] Guobin Yang. 2000. Achieving Emotions in Collective Action: Emotional Processes and Movement Mobilization in the 1989 Chinese Student Movement. *The Sociological Quarterly* 41, 4 (sep 2000), 593–614. DOI : <https://doi.org/10.1111/j.1533-8525.2000.tb00075.x>
- [48] Xiaoshan Yang, Tianzhu Zhang, and Changsheng Xu. 2015. Cross-domain feature learning in multimedia. *IEEE Transactions on Multimedia* 17, 1 (2015), 64–78.
- [49] Quanzeng You, Liangliang Cao, Yang Cong, Xianchao Zhang, and Jiebo Luo. 2015. A multifaceted approach to social multimedia-based prediction of elections. *IEEE Transactions on Multimedia* 17, 12 (2015), 2271–2280.