Image Data in the Social Sciences

Elliott Ash,[*] Malka Guillot,[*] and **Philine Widmer**[**]

[*]ETH Zurich [**]University of St.Gallen

SICSS ETH Zurich, 18 June 2021

Images as Data

Using images in social science research is **not** new

- But: human annotation is costly
- Recent advances in computer vision: **potential for wider use?**
- Also: has visual material become more important in society?
  - "Media consumption today is highly nonverbal" (Boxell, 2021)[1]

[1] Boxell, L. (2021). Slanted Images: Measuring Nonverbal Media Bias During the 2016 Election.

Some Background

Why can image data be useful in social science research?

- "Instrumental use"
  - Unstructured content (archives, pdfs) $\rightarrow$ structured data
  - Gather data in absence of other evidence (on events, protesters, ...)
- Images as the (main) subject of study
  - Differences and biases in visual representations
  - Content classification
- Anything in between

Taking a step back: what is computer vision?

- How do **you** see?
  - How does a human distinguish a dog from a cat?
- Computer vision is also about "computer cognition"
  - Making sense of images requires higher-level "knowledge" (recognizing emotions, objects, background/foreground, etc.)
- For humans, images have semantic information
  - Computers just "see" a series of discrete numerical values
- Computers' vision tends to be sensitive to "trivial" variations (scale, perspective, light, intra-class differences)

Taking a step back: what is computer vision?

Figure 1: All the same?

What is an image?

- Simplest form (single-channel): two-dimensional function $f(x, y)$
  - Maps a coordinate pair to an integer/real value related to the intensity/color of the point
  - Single-channel: binary or mono-chrome, gray-scale, black/white images
- Each point called a **pixel** or pel
- Multiple channels possible, such as RGB:
  - Color represented using three channels
  - Pixel at point (x, y) represented as $(r_{x,y}, g_{x,y}, b_{x,y})$
- Channel-specific pixel value represented as an integer between 0 and 255 or a floating-point value in $[0, 1]$

What is an image?

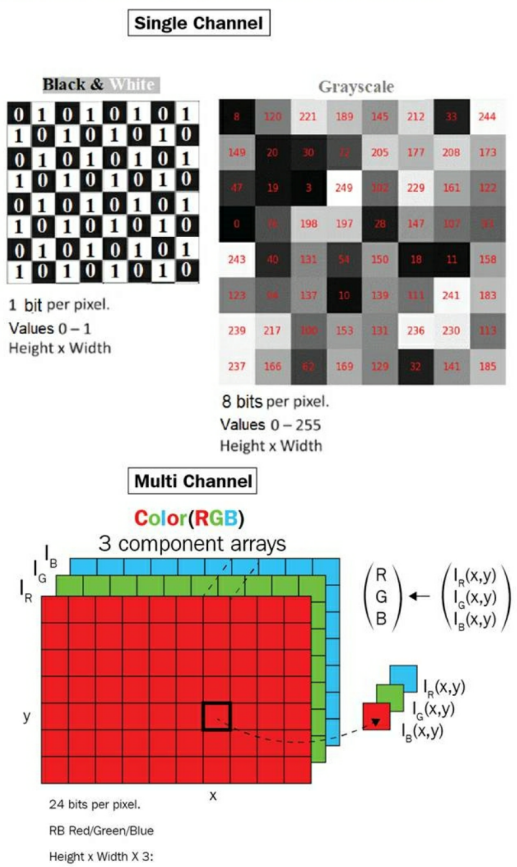


Figure 2: Dey (2018)[2]

[2]Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

Feature detection

- Imagine you search for images with a specific political leader
- Should you create a checklist of physical characteristics?
  - Shape of their nose, size of forehead?
  - Not robust (facing the camera? sunglasses?)
  - And infinite list of rules?
- Fairly general features can be helpful for various tasks
  - Corners
  - Blobs (dark/bright region)
- **Feature detection**: identification of points of interest
- Explicit feature detection in "traditional" machine learning

"Traditional" machine learning in computer vision

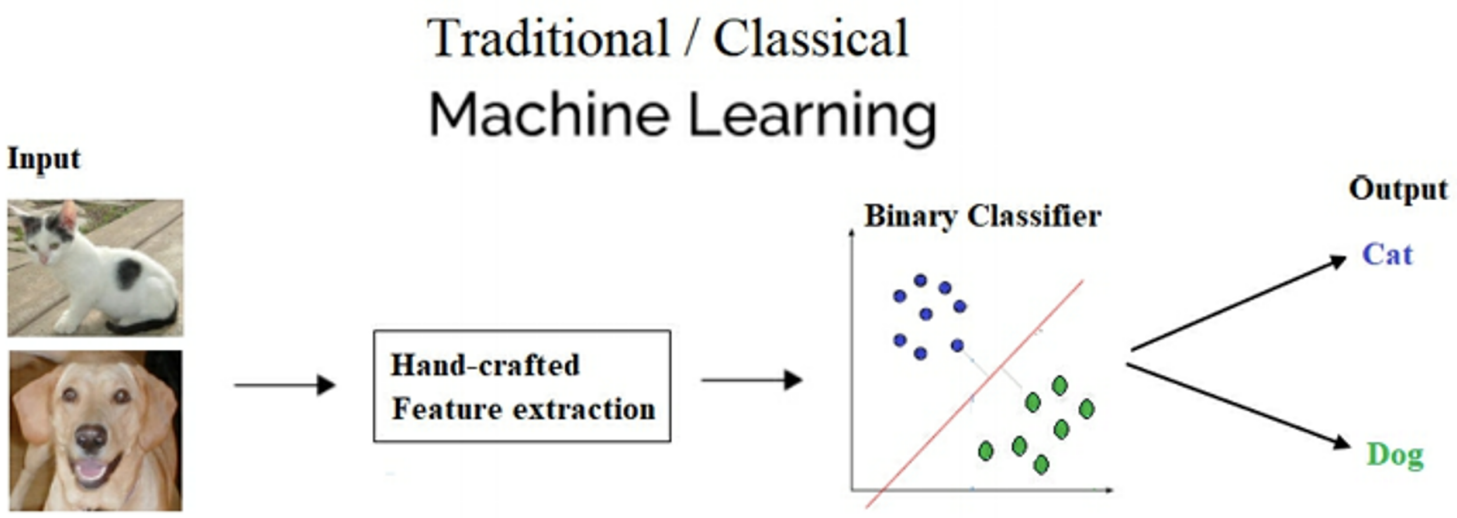


Figure 3: Dey (2018)[3]

[3] Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

Limits of "traditional" machine learning and the promise of neural nets

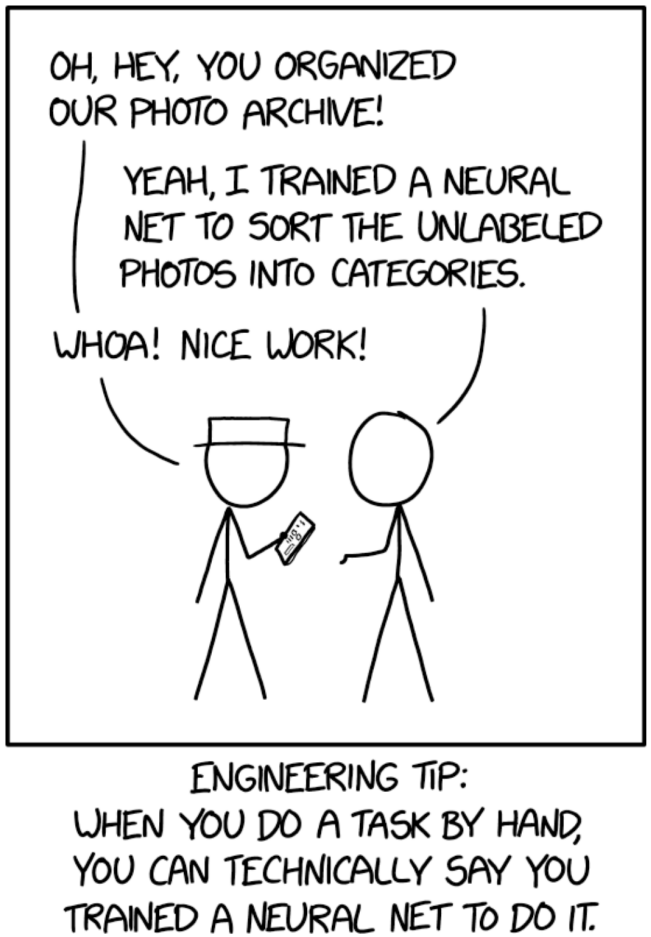


Figure 4: Source: XKCD

The promise of neural nets

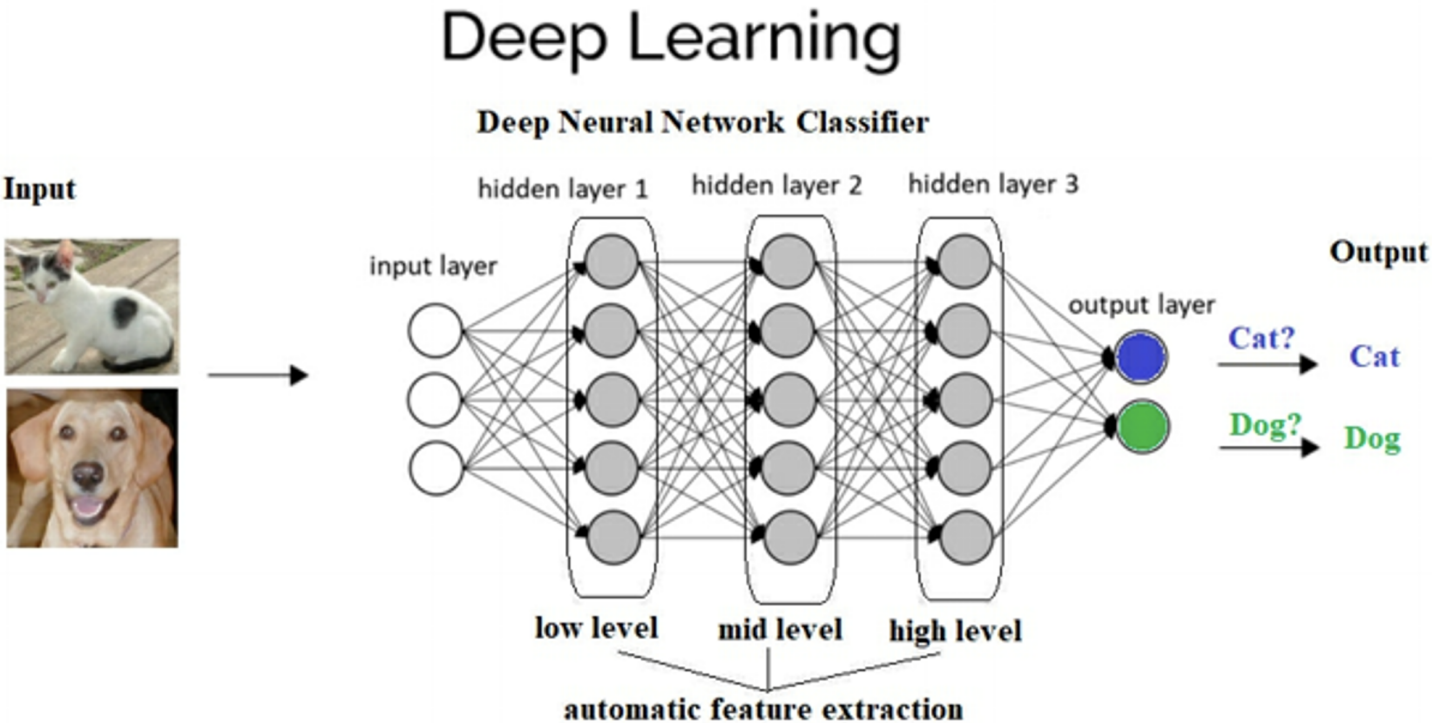


Figure 5: Dey (2018)[4]

[4] Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

Convolutional neural networks (CNN)

Figure 6: Yann Le Cun, a founding father of convolutional neural networks

- CNN are a class of deep neural networks (DNN)
- Developed for two-dimensional image data
  - Can be applied to one- and three-dimensional data
- Often reliable and cost-effective
- Past decade has seen huge advances in CV with CNN

A very basic primer on how CNN work

- Typically: input layer $\rightarrow$ hidden layers $\rightarrow$ output layer
- What do different convolutional layers do?
  - Layers close to the input layer learn low-level features (e.g., lines)
  - Middle layers learn complex abstract features (combining lower level features)
  - Layers closer to the output interpret the extracted features in the light of the classification task

Why CNN have been a revolution in computer vision

- Impressive results for many computer vision tasks
- End-to-end approach
  - Feature extraction no longer required
  - Neural nets extract their own features
  - Few pre-processing steps required
- Often more general solutions ("off the shelf")

Challenge: dealing with data requirements of CNN

- They require enormous amounts of data – annotated data!
- ImageNet has been instrumental to the advance of CV
- For end users: sometimes pre-trained models are good enough "as-is"

ImageNet

- $>$ 14 million images; classified in 22,000 "synonym sets"

"ImageNet is an image dataset organized according to the WordNet hierarchy. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". [...] In ImageNet, we aim to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. In its completion, we hope ImageNet will offer tens of millions of cleanly labeled and sorted images for most of the concepts in the WordNet hierarchy."

ImageNet (2021)

Challenge: dealing with data requirements of CNN

- Transfer learning
  - Also called domain adaptation
  - Minimizes data and computing power requirements of end user
  - No model training from scratch, but re-training of pre-trained models
  - To mitigate generalization errors
  - Technically speaking: output of the model from a layer prior to the (final) output layer of the model taken as input to a new classifier model
  - Relevant question: how similar is the task?

Transfer learning

"*[Transfer learning] is typically understood in a supervised learning context, where the input is the same but the target may be of a different nature. For example, we may learn about one set of visual categories, such as cats and dogs, in the first setting, then learn about a different set of visual categories, such as ants and wasps, in the second setting.*" (Goodfellow et al., 2016, p. 536)

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). Deep Learning (Vol. 1, No. 2). Cambridge: MIT Press.

Challenge: dealing with the opacity of CNN

- CNN are opaque in how they link model inputs and outputs
- Good practice: do not identify latent dimensions or ambiguous features in the data[5]
- Set transparent goals: tasks that are (in principle) accomplishable by humans[6]
  - Importance of human validation

[5] Torres, M., and Cantú, F. (2021). Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data. Political Analysis, 1-19.

[6] Lipton, Z. C. (2016). The Mythos of Model Interpretability. In 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York.

Some Applications

Collective Smile: Measuring Societal Happiness from Geolocated Images (Abdullah et al., 2015)

- Research goal: offer a new index of wellbeing
  - Based on share of images with smiling face
  - Share out of **all** images
- Data: 9m geolocated tweets
  - Random tweets with images 2012-2013
- Approach: large-scale sentiment assessment based on smiles detected in images shared on social media
- Key assumption: smiles as generally genuine display of emotion

Collective Smile: Measuring Societal Happiness from Geolocated Images (2015)

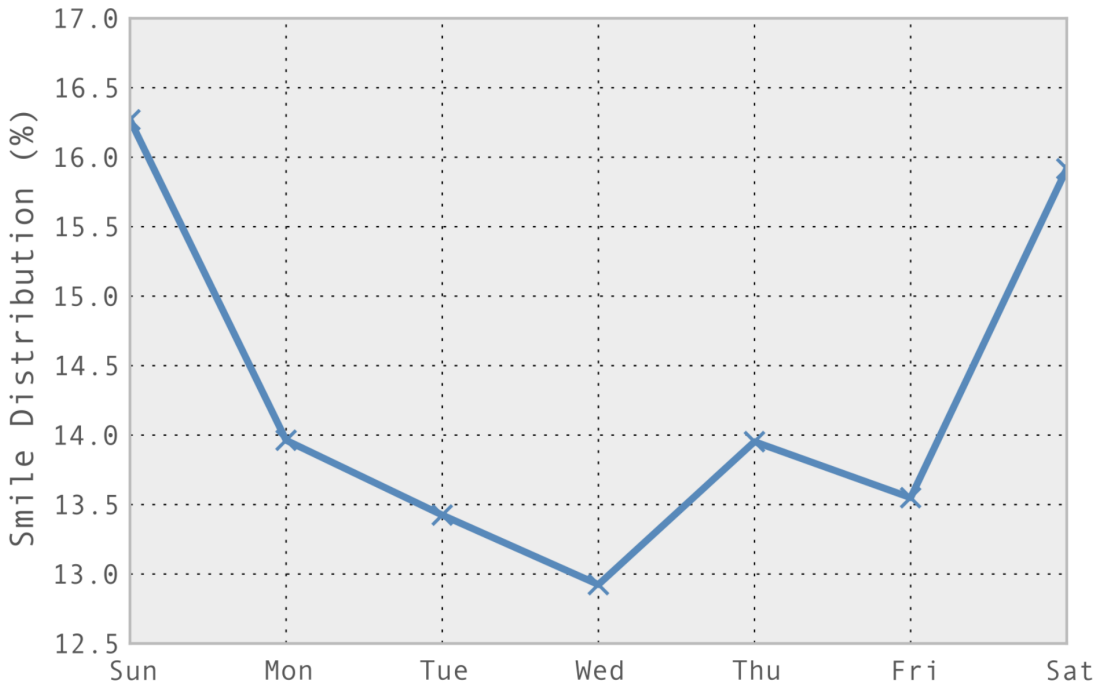
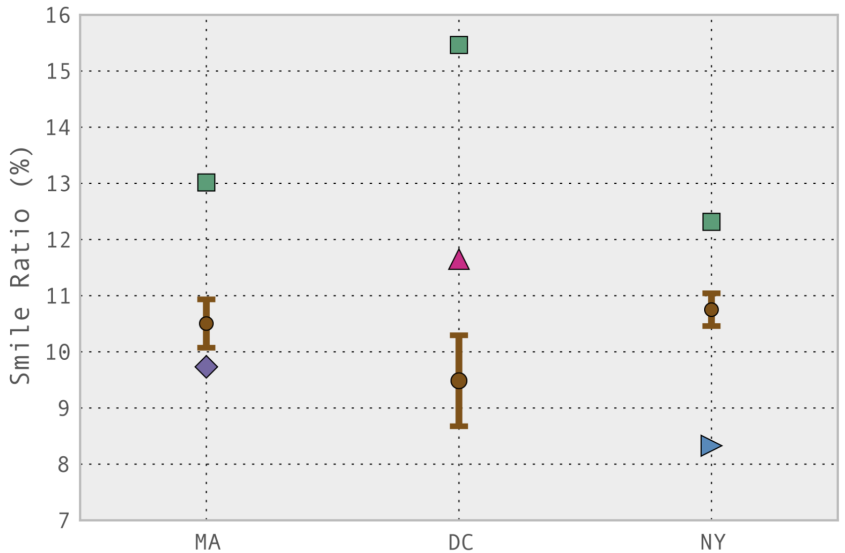


Figure 7: Abdullah et al. (2015): Distribution of smiles in U.S. tweets over days of the week

Collective Smile: Measuring Societal Happiness from Geolocated Images (Abdullah et al., 2015)



Smile Index (SI) of indicated events. • shows the mean for each of the three geographical regions, and 99% confidence interval is included. ♦ indicates SI in Massachusetts during Boston Marathon Bombings and subsequent manhunt period; ▲ marks SI during Presidential Inauguration in DC; ► indicates SI during Hurricane Sandy in NY; ■ represents SI in the corresponding areas during New Year's Eve

Figure 8: Abdullah et al. (2015)

Collective Smile: Measuring Societal Happiness from Geolocated Images (Abdullah et al., 2015)

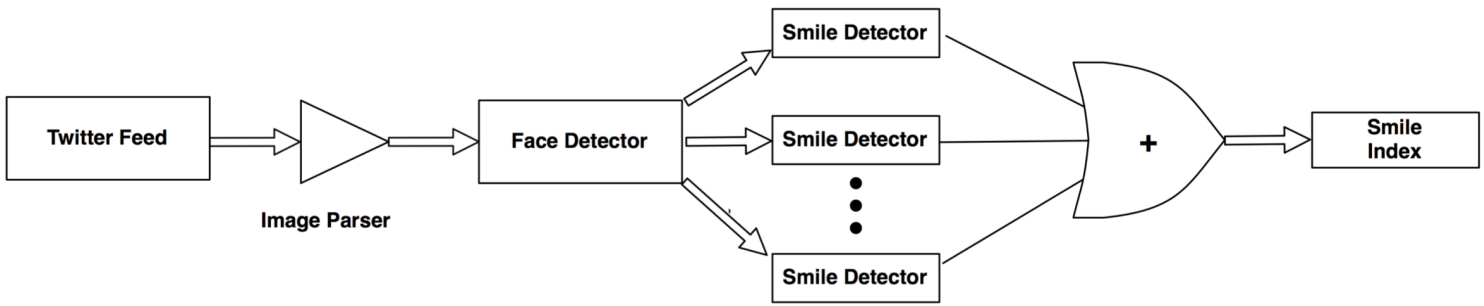


Figure 9: Abdullah et al. (2015): Framework for assessing happiness through Smile Index

How to identify smiling faces in images: the methodological gist of Abdullah et al. (2015)

- Cascade of boosted classifiers with Haar-like features
  - "Traditional" machine learning
- Based on the Viola-Jones object detection technique
  - Regularities in human faces matched using Haar features

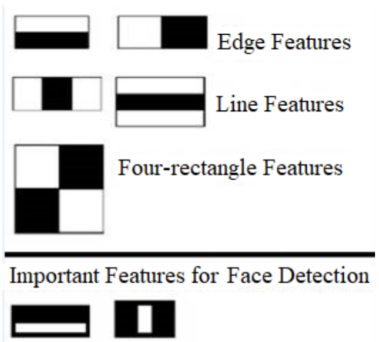


Figure 10: Haar-like features

How to identify smiling faces in images: the methodological gist of Abdullah et al. (2015)

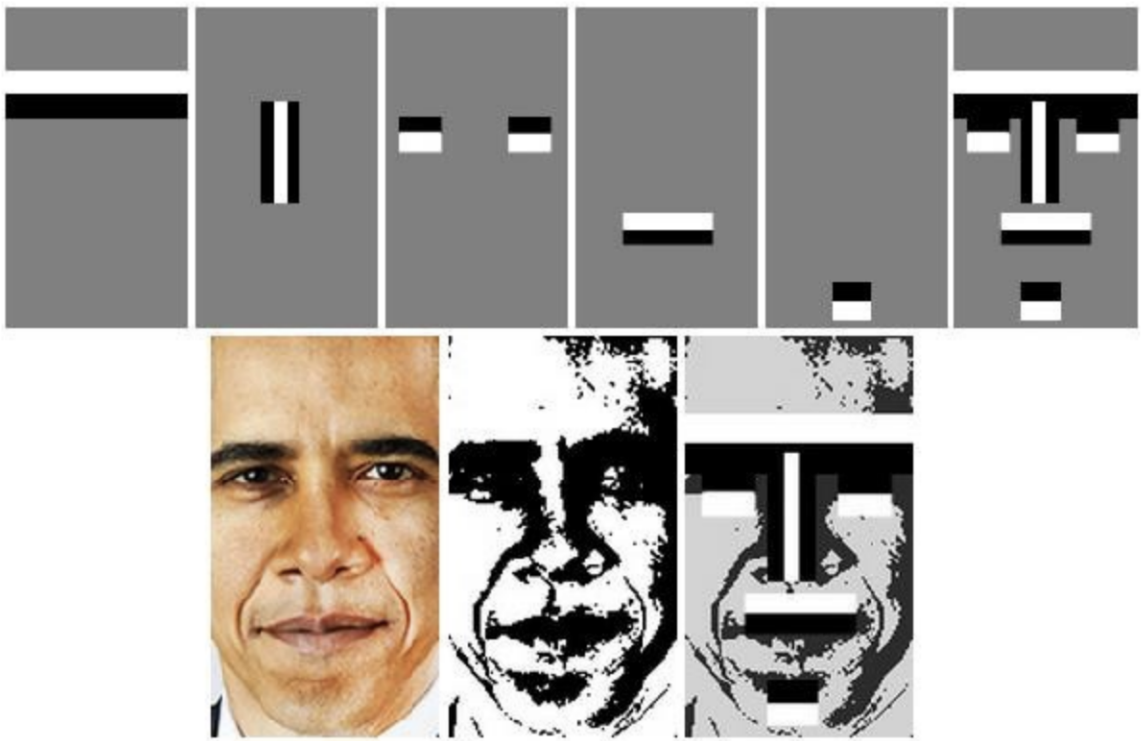


Figure 11: Kadir et al. (2014): an application of Haar-like features[7]

[7] Kadir, K., Kamaruddin, M. K., Nasir, H., Safie, S. I., and Bakti, Z. A. K. (2014). A Comparative Study between LBP and Haar-like Features for Face Detection using OpenCV. 4th International Conference on Engineering Technology and Technopreneuship (ICE2T) (pp. 335-339). IEEE.

Slanted Images: Measuring Nonverbal Media Bias during the 2016 Election (Boxell, 2021)

- Research goal: quantify "media slant" in images
  - Who is shown? Democrat or Republican politician? With what emotion?
  - Focus on Trump and Clinton
- Motivation: nonverbal information in news is persuasive
- Data: 1m images from news front pages
  - Around the 2016 U.S. election
- Approach: large-scale emotion detection (happiness, sadness, surprise, fear, anger, disgust, and contempt)

Slanted Images: Measuring Nonverbal Media Bias during the 2016 Election (Boxell, 2021)

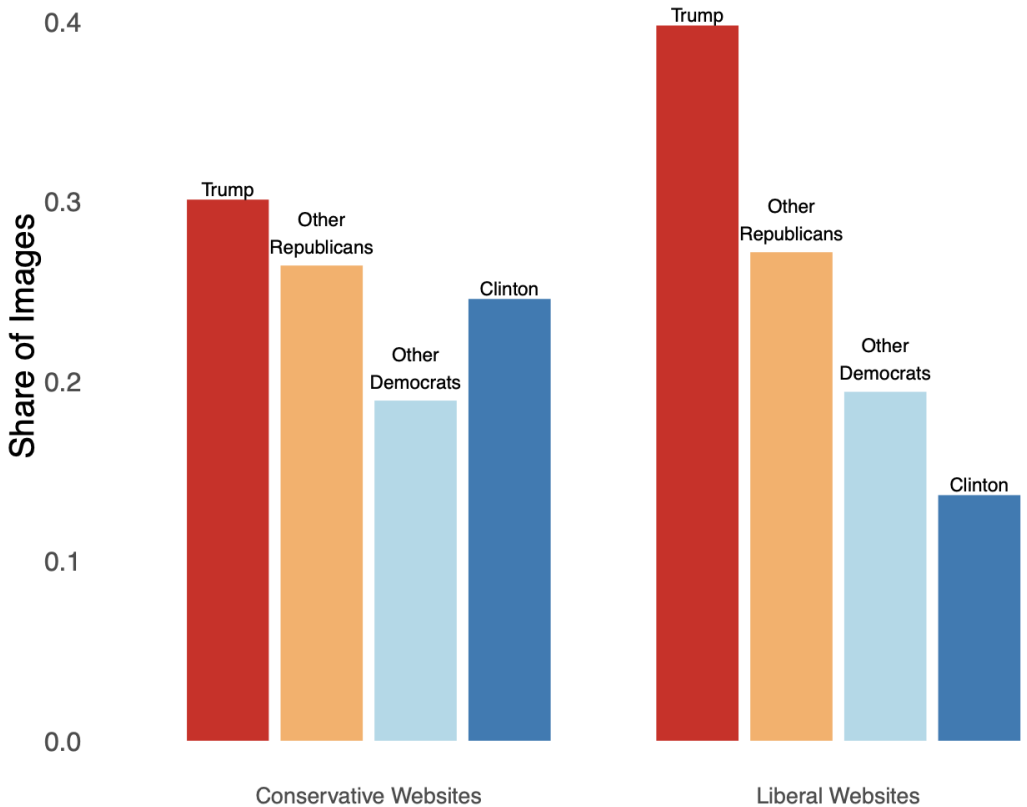


Figure 12: Boxell (2021): Bias in who to cover

Slanted Images: Measuring Nonverbal Media Bias during the 2016 Election (Boxell, 2021)

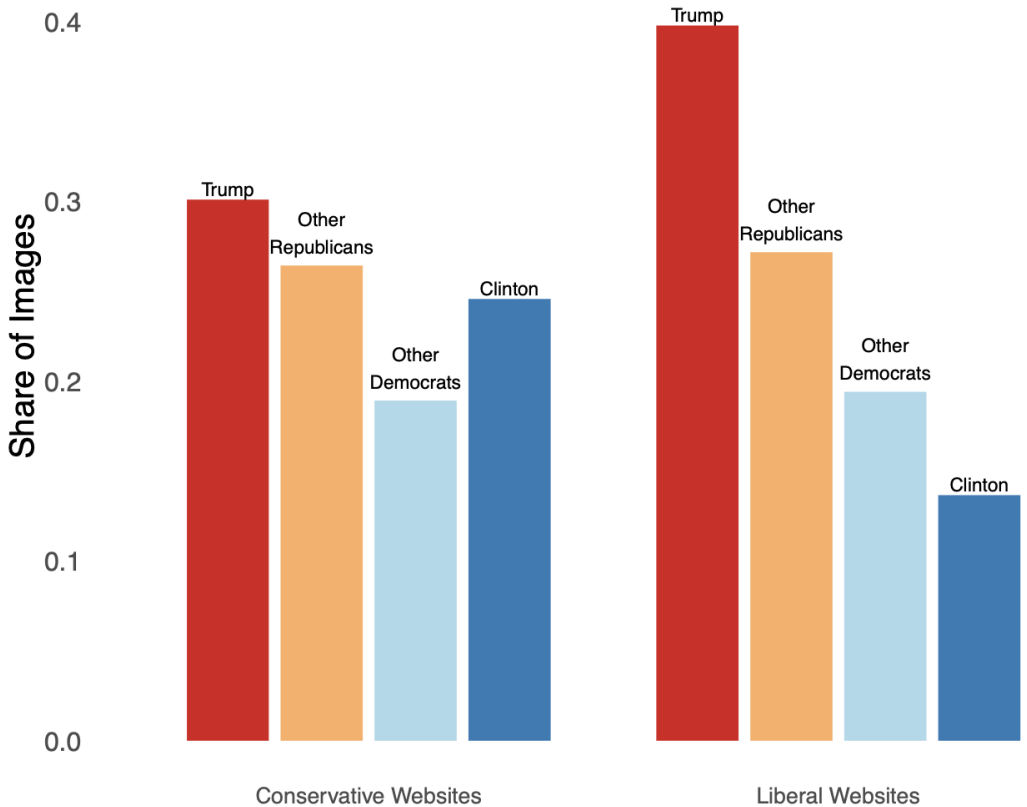


Figure 13: Boxell (2021): Bias in how to cover

How to identify politicians and emotions: the methodological gist of Boxell (2021)

- Step 1: Pre-select images with eye detector (algorithm readily available in Matlab)
- Step 2: Create a manually labeled set of politician images
- Step 3: Based on (2), identify politicians in pre-selected, unlabelled images (Microsoft's Face API)
- Step 4: Annotation of emotions (Microsoft Emotion API)
- Step 5: Human validation
  - High correlation for happiness and neutral emotions (0.87/0.67)
  - Low correlation for *distinct* negative emotions

Visual Stereotypes in News Media (Ash et al., mimeo)

- Research goal: Study gender and, in particular, ethnic stereotypes in news media images
  - Uncover verbal associations
  - Link representations to "ground truths" (population shares)
  - Ethnicity typically not conveyed in language
- Data: Images and articles from Fox News and the New York Times
  - 2010-2020
- Approach: standard feed-forward neural net architecture trained on publicly available data

Visual Stereotypes in News Media (Ash et al., mimeo)

Figure 14: Ash et al. (mimeo): Country associations of different ethnicities