Automated video analysis for social science research[1]

Dominic Nyhuis[2] / Tobias Ringwald / Oliver Rittmann / Thomas Gschwend / Rainer Stiefelhagen

[1] Chapter for „Handbook of computational social science, vol. 2: Data science, statistical modelling, and machine learning methods", edited by Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, and Lars Lyberg. We are grateful to Morten Harmening, Felix Münchow and Marie-Lou Sohnius for excellent research assistance. Funding for this project was provided by the *Deutsche Forschungsgemeinschaft* through the *Sonderforschungsbereich 884* „The Political Economy of Reforms".

[2] Corresponding author: Dominic Nyhuis, University of North Carolina at Chapel Hill, nyhuis@unc.edu.

1. Introduction: Moving beyond the analysis of digitized text

The widespread digitization has profoundly impacted research practices in the social sciences. The ubiquity of digital data has enabled research projects on scales that were almost inconceivable a mere thirty years ago. As human coding is often no longer a viable option to deal with the immense amounts of data, scholars have begun to embrace methodological innovations that aim to automate the transposition of digitized information into data points. Among the most successful and most widely employed techniques is the automated analysis of text, which has become a staple of social science research (Grimmer & Stewart, 2013; Lucas et al., 2015; Wilkerson & Casas, 2017). But even though text mining has yielded crucial insights in a number of disciplinary subfields, scholars have yet to appreciate the full potential of the universal digitization for the social sciences.

While some studies have recently started using techniques for the automated analysis of still images (Haim & Jungblut, forthcoming; Peng, 2018; Williams et al., 2020; Zhang & Pan, 2019), there are almost no social science contributions which have adopted tools for the automated analysis of other forms of digitized media, such as audio or video recordings. Not only do these data promise new and valuable insights, our perspective on some social phenomena clearly remains incomplete when we disregard visual information. The most obvious example is social media research. Our focus on textual information has severely hampered our understanding of these platforms, where ideas are frequently expressed as images and videos and where the visual cues arguably dominate the textual ones.

Studying the visual cues of social media posts underlines the need for embracing automated tools to make sense of the vast amounts of data that are continuously generated on these platforms. While previous studies have analyzed social media imagery using manual classification (Kharroub & Bas, 2016; Neumayer & Rossi, 2018; Rose et al., 2012), these efforts are limited due to the size of the data and the costs of human coders. These problems are compounded by the fact that we are typically interested in multiple dimensions of digitized media, further increasing the need for human coders, thus limiting the amount of material which can be covered.

While social scientists have so far resisted adopting the tools for studying images and video data, there has been tremendous progress in computer science to make sense of this type of data using automated systems. The aim of this chapter is to highlight some of these potentials and to encourage researchers to make greater use of the available techniques for analyzing digitized media. This chapter focuses on the automated analysis of video footage, but the underlying methods are similar across the different data types, not least due to the similarity of still images and videos. In either case, machine learning is used to classify images or video data based on a sample of human-coded training data.

To showcase the possibilities, section 2 begins with an overview of typical scenarios and tools for classifying video footage. Section 3 provides a summary of current research in the social sciences to give a sense of the type of research question which can be explored with video data.

Section 4 illustrates a case study in greater detail. Section 5 concludes with a discussion of some conceptual and methodological challenges and research perspectives.

2. The state of the art in automated video analysis

The core objective of machine learning is to automate the data analysis. Machine learning replaces the need for human coders who would otherwise have to perform mundane and repetitive tasks. This is especially true for high-bandwidth data such as videos which require human annotators to be attentive for long stretches of time while – depending on the specific coding task – paying attention to the audio and video tracks at once.

The tasks in automated video analysis are highly diverse and there are a multitude of sub-tasks depending on the desired characteristics and the intended usage of such systems. Common tasks are the detection of humans or generic objects (object detection), localization of body parts (pose estimation), detection of higher-level actions (activity recognition) or identification and tracking of humans (person re-identification) based on certain features such as face, gait or appearance.

Ever since Krizhevsky and colleagues (2012) have shown the enormous potential of deep neural networks for image classification, the research on automated video analysis has shifted towards deep convolutional neural networks (CNNs). We will introduce the above concepts in the context of CNNs as they usually outperform other approaches (Wang et al., 2011).

At a fundamental level, automated video analysis can be thought of as the analysis of a series of images. A basic classification task might consist of predicting a certain class label from a predefined list of possible classes for a single image or a series of images. Object detection extends this setup by also predicting the location of an object in the form of a rectangular bounding box. One of the earliest CNN-based methods in the area of object detection was the Region-based CNN (R-CNN; Girshick et al., 2014). R-CNN has a two-stage object detection pipeline: First, region proposals, i.e. a set of bounding boxes, are generated through a selective search algorithm (Uijlings et al., 2013). In the second stage, these regions are classified with a CNN as belonging to a class from a list of possible classes or as background in cases where the region does not contain an object.

Due to the slow inference speed of R-CNN, multiple improvements to the region proposal and classification stage have been proposed in Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2015). As even Faster R-CNN is too slow for real-time inference, recent research has shifted towards anchor box-based predictions instead of region proposals. Anchor box-based detection architectures directly regress coordinates of bounding boxes and do not require a separate region proposal stage. Research has also looked into more efficient CNN backbone architectures such as MobileNet (Howard et al., 2017) and SqueezeNet (Iandola et al., 2016), specifically designed for use on mobile devices. Examples of real time detection frameworks are the single-shot multibox detector (SSD; Liu et al., 2016) and YOLO (Redmon et al., 2016; Redmon & Farhadi, 2017).

The recent Mask R-CNN (He et al., 2017) extends the idea of Faster R-CNN and predicts a fine-grained segmentation mask in addition to the coarse object bounding box. This results in more detailed object contours as generic rectangular regions often contain a large number of non-object pixels.

Regarding humans in video footage, fine-grained predictions are often done in the form of pose estimation (Cao et al., 2019; Wei et al., 2016). In this case, several key points on the human body, such as joints, eyes, or ears are predicted and merged into a full skeleton. This can be done in 2D (image space) or in 3D (world space) based on triangulation from multiple single views or depth cameras. Specialized pose estimation focuses on fine-grained detection of subparts of the human body, such as hands, which can deliver precise information about gestures.

All of these techniques can be applied to single frames of a video and tracked over time, i.e. an object can be tracked in a video by applying object detection at every frame and connecting the resulting bounding boxes over time based on their overlap between the individual frames. For a higher-level understanding of videos – such as activity recognition – still images are insufficient as they do not contain temporal context. In order to exploit the high performance of single image classification networks, Donahue and colleagues (2015) proposed extracting single image CNN features at specific time steps in videos and passing them to a long short-term memory (LSTM) module (Hochreiter & Schmidhuber, 1997). LSTMs are commonly used for sequence processing as they can encode and propagate states over multiple time steps and capture long-term dependencies within the input data.

CNNs are also commonly used for person (re-)identification. Here, the task is usually defined as retrieving similar images from a gallery given a single query image. For example, in a surveillance task the query could be the image of a suspect, while the gallery consists of footage from a surveillance camera. As there is an unknown number of identities in the gallery, (re-)identification differs from a normal classification task. Instead, models in this area try to predict the similarity between two inputs based on a metric such as Euclidean distance. One of the most popular works was published by Schroff and colleagues (2015), who use CNNs to learn a compact feature embedding for an input image. During training, embeddings from images showing the same person are optimized to be closer than embeddings from images showing different persons.

Given the increasing popularity of automated video analysis, several software solutions have been published in recent years. One of the most popular open source solutions for pose estimation is OpenPose,[3] which offers a multitude of different pose-related solutions such as body and hand keypoint estimation in 2D and 3D. For standard image classification tasks, many pretrained models are available for commonly used CNN architectures. Oftentimes, the

[3] Available at https://github.com/CMU-Perceptual-Computing-Lab/openpose

ImageNet dataset (Deng et al., 2009) is used for pretraining, as it offers a diverse set of 1,000 different classes that cover most objects found in everyday life. Similarly, deep learning frameworks such as PyTorch (Paszke et al., 2019) offer pretrained solutions for object detection, semantic segmentation, instance segmentation and action recognition.[4] Even if these models are not trained on a class required for a specialized prediction task, they can still be used as a starting point for fine-tuning, thus easing the training process and reducing the need for additional training data.

3. Current applications in the social sciences

Even though tools for the automated analysis of video data are fairly well established in computer science and even though there is great potential to advance social science research with video data in a number of areas, efforts to apply computer vision in empirical social research are few and far between. To provide an overview of this nascent research field, it is helpful to distinguish between the audio and visual components of video recordings. While we have seen some analyses of still images (Haim & Jungblut, forthcoming; Peng, 2018; Zhang & Pan, 2019), the analysis of moving images is virtually non-existent in the social sciences. By contrast, analyses of audio data have been somewhat more common.

In a series of papers, Dietrich and colleagues study audio recordings of political speech, where vocal pitch is taken as an indicator for the emotional intensity of speakers. Analyzing speeches in the US House of Representatives, Dietrich et al. (2019) find that female legislators exhibit greater variation in vocal pitch when they address women in their speeches. Dietrich and Juelich (2018) analyze the vocal pitch of Hillary Clinton and Donald Trump during the televised debates in the 2016 Presidential election campaign. The authors find that candidates' vocal pitch is related to the content of their speech, such that candidates addressing issues at the core of their parties' platform exhibit higher standardized vocal pitch and vice versa. Additionally, Dietrich and colleagues have analyzed the vocal pitch during oral arguments in the U.S. Supreme Court, which famously only publishes audio recordings of their deliberations, but does not allow video cameras into the court room. Dietrich et al. (2019) find that the Justices signal their eventual vote choice during oral argument. The Supreme Court deliberations are also analyzed by Knox and Lucas (forthcoming). Based on a general speech classifier, the authors train a model to detect skepticism in the utterances of the Justices (Knox and Lucas forthcoming).

While analyses of audio recordings have been – comparatively – more common in the recent social science literature, a few studies have also analyzed video recordings. For example, Dietrich presents analyses of video footage from the US House of Representatives (forthcoming, 2015). In these contributions, Dietrich is interested in how the increasing polarization in the US political system manifests in the behavior of legislators on the House floor. He argues that as the polarization has intensified, legislators have become less likely to

[4] Available in the torchvision package at https://pytorch.org/docs/stable/torchvision/index.html

interact with their counterparts from the opposing party – they are literally less likely to cross the aisle. To test this behavioral upshot of the changing political landscape, he studies the video footage from the House of Representatives that is published by C-SPAN. Dietrich specifically focuses on the video segments during the proceedings when legislators take a roll call vote. During voting, C-SPAN shows an overhead shot of the plenary floor. To proxy the interactions between legislators, Dietrich relies on a fairly simple technique that compares the pixels in the overhead shots, where more dissimilarity between two frames is taken as an indicator of more movement on the House floor. Assuming that crossing the aisle – moving across the screen to interact with members of the opposing party – creates greater dissimilarity between the frames, he confirms that the stability of the frames in the overheard shots has gone up over time.

While the work by Dietrich is a good example of how video recordings can inform political research in diverse and unexpected ways, a more natural application is presented by Joo and colleagues (2019). The authors conduct an exploratory analysis to assess whether machine learning is useful for automatically classifying video recordings of political actors. Studying footage from the first televised debate between Hillary Clinton and Donald Trump during the 2016 Presidential election campaign, the authors manually code a variety of nonverbal candidate behaviors. Next, they extract facial and pose characteristics of the candidates from the footage using standard libraries, among them OpenPose, introduced in the previous section (Cao et al., 2019). Based on these features, the authors train a neural network to predict the manual codes. Overall, their model performs reasonably well, suggesting that automated video analysis may greatly decrease the need for manual labor in this research area, opening up enormous potentials for large-scale comparative work.

While these initial applications of automated video analysis in the social sciences show great promise, there is a lot of room for additional research. First, a lot of the work using audio and video data is exploratory and descriptive in nature, while these tools have yet to be incorporated into more conventional research programs with an explanatory interest. Second, applications of computer vision in the social sciences have not yet made full use of the tools which have been developed in recent years and which have become more easily accessible to interested researchers as highlighted in the previous section. Third, digitized video recordings are available in many areas which have not yet found their way into social science applications. What is more, the published studies have occasionally discarded relevant parts of the data. Specifically, none of the studies included in the overview have integrated audio and video data to make the most out of the available data. While the video or audio component are not always available or relevant for the research question, some of the studies on political speech could easily benefit from adding the video component of the footage. To further highlight the potentials of automated video analysis for empirical social research, the next section will discuss an application from legislative politics in greater detail.

4. A sample application: Analyzing parliamentary video footage

In a move to increase the transparency of parliamentary activities, many legislatures have allowed third-party cameras into the plenary chamber or have begun publishing video recordings of the plenary proceedings themselves (Ryle, 1991). Over time, these efforts have become more technologically sophisticated and, nowadays, digitized video footage of parliamentary proceedings is available on many parliamentary websites.

For researchers interested in legislative politics, these data hold an enormous potential. In addition to enabling analyses of parliamentary behaviors that are not recorded in the minutes of the plenary proceedings, such as legislator attendance or interactions, parliamentary video footage is characterized by several features that make it particularly suitable for an automated analysis. Most importantly, video footage of parliamentary proceedings is more uniform than footage generated in other contexts, say videos from traditional or social media. Knowing what is contained in a collection of videos greatly simplifies the classification task. First and foremost, we do not need to distinguish between relevant and irrelevant footage. For example, if we were interested in protest imagery on social media, a key task would be to identify the relevant footage before moving to the analysis. But even beyond selecting the relevant footage, videos of parliamentary proceedings are more uniform than video footage in other contexts, which means that building a classification scheme and applying it to the videos is easier and more robust. For example, in the application presented here, we gauge the dynamics of plenary attendance from the parliamentary video footage. To this end, we build a model to identify the legislators that are visible in the videos. This task is comparatively straightforward, as the number of legislators is small, such that collecting training footage for the identification is simple and the accuracy of the classification is high.

The video footage of the plenary proceedings that is published by many parliamentary administrations typically focuses on the speaker. Such data is well-suited to study questions about legislative speech. For example, analyses of the vocal pitch in audio recordings of political speech (Dietrich et al., 2019) could easily be supplemented with video footage, as the video track contains additional cues about the nonverbal characteristics of political speech, such as facial expressions or body movements. In addition to shots of the speaker, some parliamentary administrations intersperse their videos with shots of the full plenary, either to make the footage – marginally! – more captivating or to bridge sequences in the proceedings when no speaker addresses the plenary. This is exploited in the contribution by Dietrich (forthcoming) who studies the overhead shots from the US House of Representatives. While the work by Dietrich constitutes an excellent example of how these occasional shots of the plenary can inform research on legislative politics, the utility of this type of footage is nonetheless limited. Some of the most interesting questions on the behavior of legislators require a continuous record of the plenary proceedings and not just the occasional and often somewhat haphazardly inserted shots of the plenary chamber.

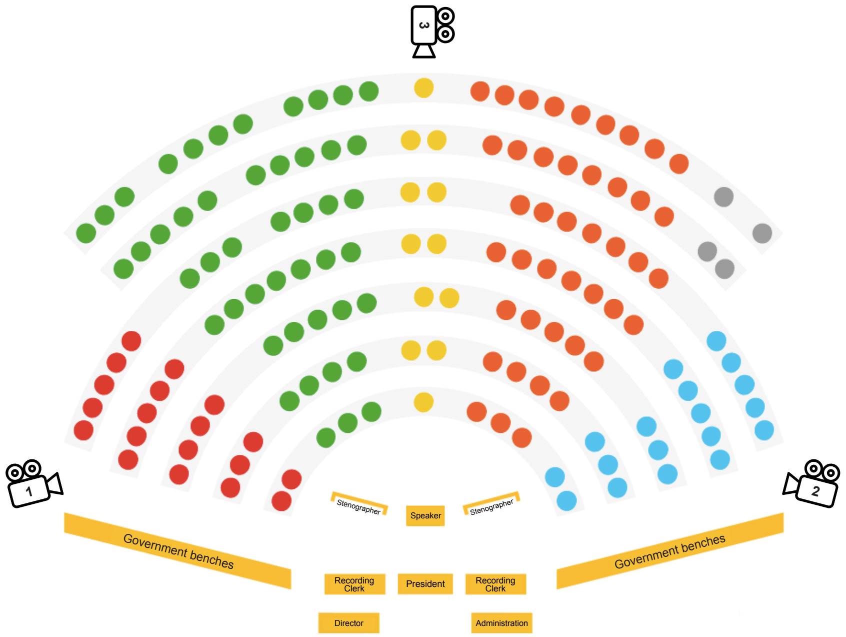

Figure 26.1: Camera locations in the *Landtag Baden-Württemberg*

Note: The layout of the plenary chamber was adopted from https://www.landtag-bw.de/home/der-landtag/sitzplan.html. The Figure reflects the composition of the *Landtag* in July of 2020.

To advance such a research program, we have collaborated with the *Landtag Baden-Württemberg*, a large German state-level parliament, to provide us with continuous footage of the plenary proceedings. The parliamentary administration records the plenary chamber with three cameras, as displayed in Figure 26.1, one focusing on the speaker (camera 3), two focusing on the plenary (cameras 1 and 2). The footage that is published by the administration predominantly consists of footage of the speaker from camera 3, along with occasional sweeps across the plenary. Hence, most of the footage from cameras 1 and 2 is never publicly released and typically deleted. To study interactions in the *Landtag*, we have been able to secure the footage from cameras 1 and 2 for one year between July 2018 and July 2019. Specifically, the following analysis is based on footage from 31 plenary sessions, ordinarily lasting from mid-morning to well in the evening. The sheer size of the data, 31 days with about 8-10 hours of video material per camera and day (31 x 2 x 9 = 558 hours or roughly 23 days), highlights the need for automating the analysis, as the cost for manually coding the material would be prohibitive – not to mention the human suffering caused by having to watch 23 straight days of plenary proceedings.

The footage allows studying a variety of questions about legislative behavior. For the sample application, we analyze which legislators are present in the plenary chamber at any given time. This is more of a validation exercise and does not make use of the full potential of this data for research on legislative politics. We briefly highlight some areas where future research could build on these preliminary analyses at the end of this section. It should be stressed, however, that even studying the ebbs and flows of plenary attendance is valuable in its own right and that this information is not available from the official parliamentary records otherwise. While some assemblies have legislators sign in when they enter the building or take attendance at the beginning of the session, to the best of our knowledge, no legislature in the world records whether legislators are actually present in the plenary chamber as the day progresses or whether they are off to attend a committee meeting, to meet with colleagues, or to work in their offices.

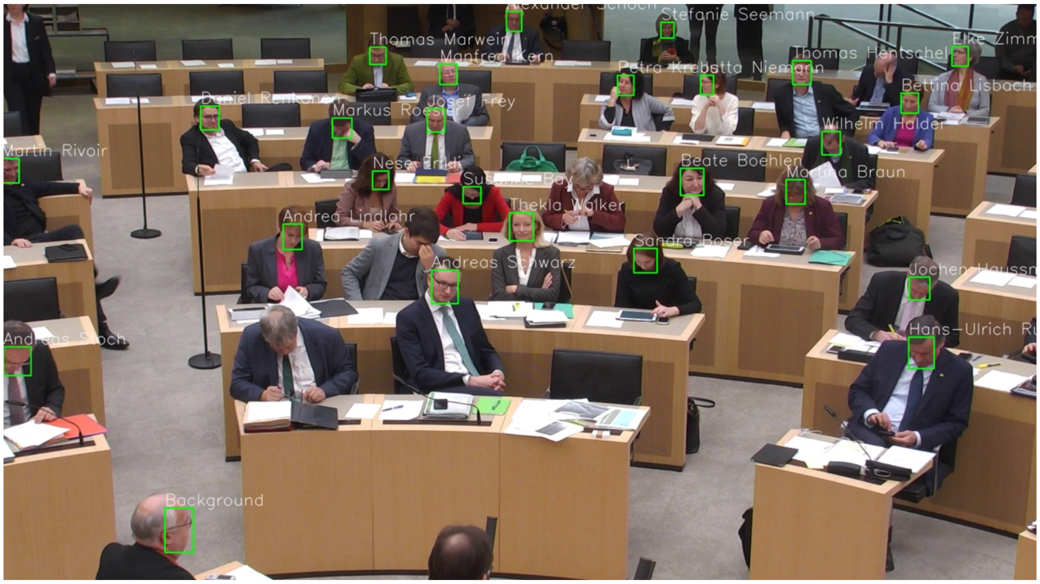

Figure 26.2: Screenshot from camera 2 of the *Landtag* plenary chamber focusing on the Green party

Note: The automatically detected and identified faces are highlighted in the screenshot.

Capturing the patterns of parliamentary attendance from the video recordings involves two steps. In the first step, we employ a neural network to detect the faces in the videos (Detection). In the second step, another neural network is used to assign the names of the legislators to the detected faces (Identification). For the identification step, we build a training data set containing photos and video clips of the legislators, so the model can learn the facial features of the legislators. Both steps are standard tasks in computer vision and robust tools have been developed that perform face detection and face identification with high levels of accuracy. For the detection part, the TinyFace architecture of Hu and Ramanan (2017) was used, as it offers accurate detections even for barely visible faces. The identification part was solved by an ImageNet pretrained ResNet-18 model (He et al., 2016) which we fine-tuned with our legislator dataset. Given the small number of legislators – currently, the *Landtag* has 143 seats – the classification is correct in 99.7% of cases based on a holdout validation set. To illustrate the footage and the analysis, Figure 26.2 provides a screenshot from camera 2 with all detected and identified faces highlighted. Prior to the visualization, detections and identifications with low confidence scores were filtered to prevent false positive detections or false assignments of identities.

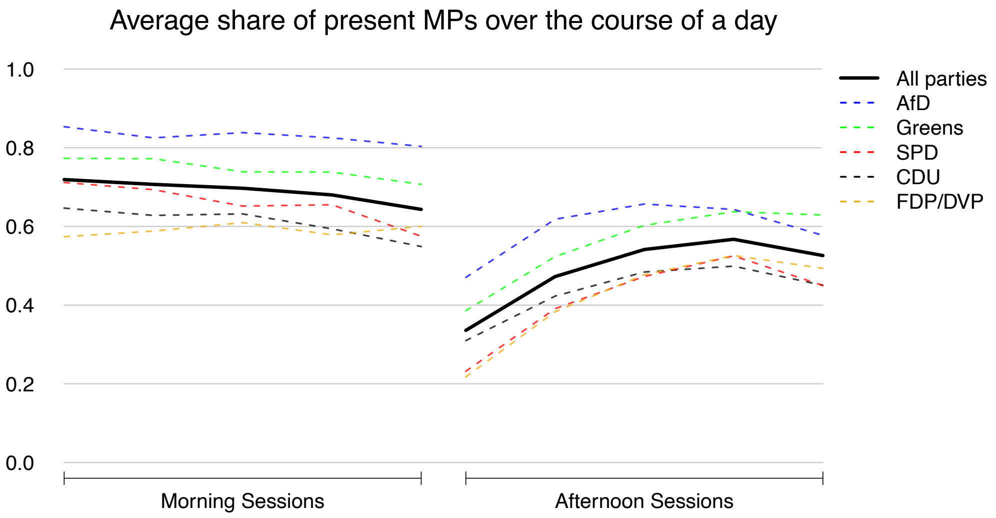

Figure 26.3: Average share of detected legislators in the *Landtag* over the course of the day

Note: The Figure displays the average share of legislators that attended the plenary sessions over the course of the day across all 31 sessions in the analysis. To combine the figures from the different sessions, the morning and afternoon sessions were split into five equally sized segments each. In addition to the overall average attendance shares in the ten segments, the Figure displays the values for the five parties that were represented in the *Landtag* at the time of the investigation.

For the validation step, we present two aggregate perspectives on the resulting attendance data. One, we study plenary attendance over the course of the day; two, we investigate plenary attendance by type of parliamentary procedure. For the first perspective, Figure 26.3 displays the average share of legislators that were detected in the video footage in all 31 sessions. The Figure distinguishes between a morning and an afternoon session, as the *Landtag* takes a lunch break for about one hour around 12:30am. To combine the attendance figures from the individual sessions, the morning and afternoon sessions were split into ten equally sized segments, five in the morning and five in the afternoon. In addition to the overall average, Figure 26.3 displays the attendance averages for the five parties with a parliamentary representation at the time of the investigation.

The results generally support conventional wisdom about legislative politics. Attendance is noticeably lower in the afternoon than in morning, which speaks to the quip that “there is no greater secret than the spoken word in the plenary after 2pm.” (Hohl, 2017, 18; translation by the authors) Interestingly, however, there is a clear uptick in attendance as the afternoon progresses, albeit not to the same levels as in the morning, which seems to belie the idea of an irrelevant parliamentary afternoon. Figure 26.3 suggests that while legislators tend to hang out at lunch too long, we do not find that attendance decreases as the session moves into the evening hours. Two factors may help explain this somewhat unexpected finding. One, compared with many national parliaments, where legislative sessions can easily run past midnight, the *Landtag* typically finishes its business in the early evening, making it less burdensome for legislators to stay until the end. Two, the *Landtag* meets comparatively infrequently, making attending the individual sessions more important. This proposition is supported by the high levels of attendance overall.

The party-level averages are similar to the grand mean. Notably, the new right-wing populist party AfD, which entered the *Landtag* for the first time in 2016, exhibits the highest levels of attendance. This observation nicely aligns with existing research, which has suggested that the AfD values public appeals in the plenary over substantive work in the committees (Ruhose, 2019; Schroeder et al., 2018).

The second way to assess the validity of the attendance data is to split the observations by type of parliamentary procedure. In this case, we should expect legislators to attend in greater numbers when the stakes of the debate are high. To examine this proposition, Figure 26.4 distinguishes between the six main plenary procedures in the *Landtag Baden-Württemberg*. In addition to government declarations, typically delivered by the prime minister, and ordinary legislative debates, the rules of procedure offer four plenary proceedings which can be subsumed under the heading of parliamentary control instruments: Urgent debates, major interpellations, oral questions, and the question hour. They differ in terms of the kind of issue that they address. Oral questions and the question hour are scheduled on a regular basis. They allow legislators to pose short oral questions on different topics which are answered by members of the executive. By contrast, major interpellations and urgent debates need to be explicitly requested and they dedicate an entire segment of the legislative session to just one issue when the topic is of general importance or urgency.

Figure 26.4 strongly supports the conclusion that the stakes of the debate influence plenary attendance. The Figure shows the attendance distributions by the five parties under the six plenary proceedings, along with the proceeding-specific average values, indicated by the dashed red lines. Comparing the four control instruments, we find that urgent debates and major interpellations trigger average attendance values well over 60 percent, whereas oral questions and the question hour only result in an average attendance around 40 percent. Interestingly, there is a notable gap between the mean attendance values of the AfD and the other parties during oral questions and the question hour. Again, this observation nicely complements existing research which has highlighted how AfD members focus their efforts on government control and critique (Schroeder et al., 2018).

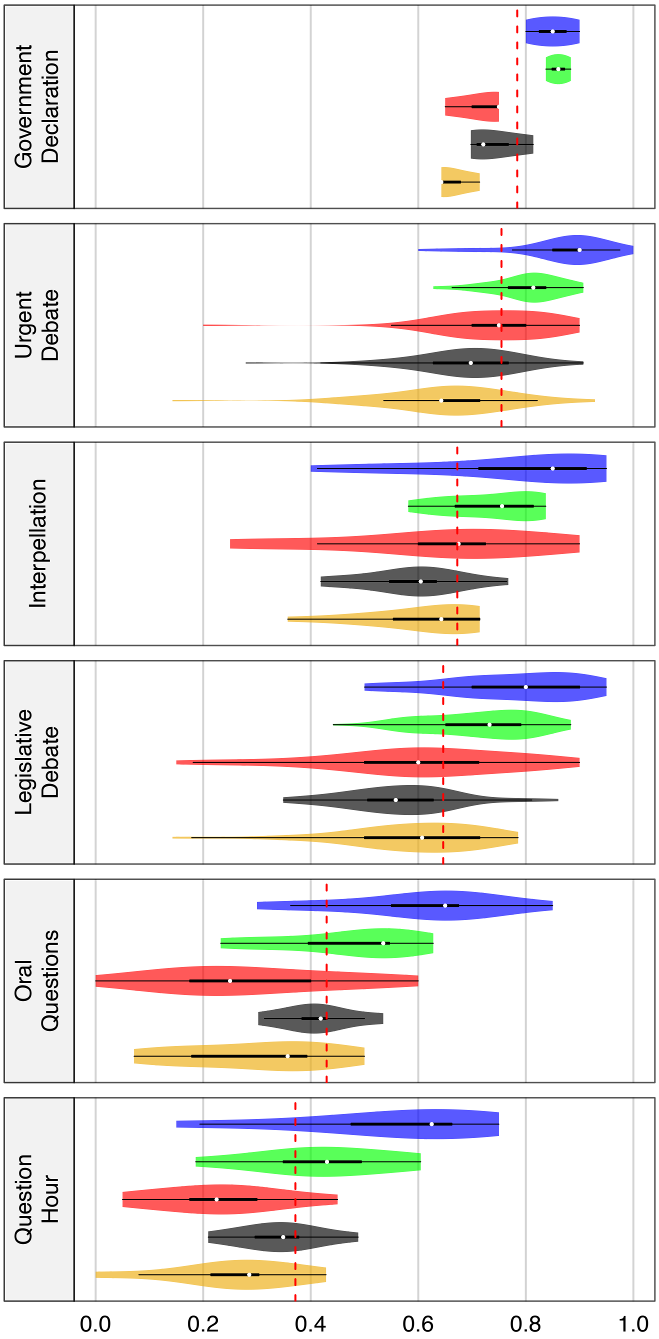

Figure 26.4: Share of detected legislators per party in the *Landtag* by type of parliamentary procedure

Note: The Figure displays the share of party representatives that were detected in the six types of plenary procedures across all 31 sessions. For example, between 60-100 percent of the AfD legislators attended the urgent debates. The colors represent the five parties: AfD (blue), Greens (green), SPD (red), CDU (black), FDP/DVP (yellow). The dashed red lines represent the mean attendance by parliamentary procedure. The six parliamentary procedures are Government declaration (*Regierungsinformation*), Urgent debate (*Aktuelle Debatte*), Interpellation (*Große Anfrage*), Legislative debate (*Beratung*), Oral questions (*Regierungsbefragung*), and Question hour (*Fragestunde*).

In summary, both perspectives on plenary attendance show high levels of face validity. Beyond highlighting how accurate computer vision has become for detecting and identifying faces, these preliminary analyses pave the way for future research on legislative politics using automated video analysis. Without going into unnecessary detail, there are at least four research perspectives that could be pursued with the footage introduced above. First, the aggregate figures have clearly shown how plenary attendance reflects political importance. Therefore, the analysis could be extended to study political importance attributions for specific issues among parties or individual legislators. Second, the data speaks to a research agenda on the determinants of legislator efforts with a particular emphasis on plenary attendance (Arnold et al., 2014; Bernecker, 2014; Besley & Larcinese, 2011). The video footage can provide a more nuanced perspective on plenary attendance by moving beyond the coarser measures based on absences during roll call votes. Third, a broad literature has analyzed the networks among legislators. Lacking a direct measure for legislator interactions, these contributions have typically focused on proxy measures, such as bill co-sponsorships (Bratton & Rouse, 2011; Cho & Fowler, 2010; Fowler, 2006). Building on the preliminary analyses presented here, it is possible to study interactions and networks in the parliamentary arena more directly. Fourth, the footage could also be re-analyzed to study questions beyond the presence or absence of legislators. For instance, one could train models to assess the focus of attention or the reactions of legislators in the plenary. In combination with the legislator IDs, such a model could inform a number of research questions, for example on how female/male legislators react to female/male speakers or on how members of the political mainstream react to speakers by the AfD and vice versa. Overall, we hope that these brief comments highlight how computer vision can advance our understanding of social phenomena and, in the present case, our understanding of legislative politics.

5. Conclusion: Potentials and challenges for the application of computer vision in social research

Computer vision promises to be a key tool for empirical social research that could impact research in a number of subfields. Current applications have barely scratched the surface of what is possible with automated video analysis. Not only have many of the available tools not yet been adopted for social science research, there is also a lot of untapped potential on the data side. Digitized video footage is created and published in a variety of settings, and there are numerous applications beyond the obvious use cases of video data from social and traditional media. For example, political scholars could learn a lot about deliberative practices from video recordings in committee settings. Existing research on this question has relied on video recordings of committee deliberations which were analyzed with painstaking manual coding procedures (e.g., Nullmeier et al., 2008; Weihe et al., 2008). Automated video analysis could support and expand such efforts by enabling a comparative perspective with high reliability at low cost. Another major area for the application of automated video analysis in the social sciences is in the study of video footage that is explicitly collected for the purpose of that research. Qualitative research on video data frequently does not study footage "from the wild", but analyzes videos that were shot for the purpose of coding actor behaviors in a second step (Heath et al., 2010). Automated video analysis promises to be a disruptive force in this field by enabling research projects on much larger scales.

Despite the undeniable value of computer vision for social science research, scholars face a number of challenges in bringing the available tools to real-world applications. First, in some research applications, the first step is choosing which material to include in the study. Particularly when studying footage from social or traditional media, only a – potentially small – subset of the footage might be relevant for the research problem. In principle, the same tools we use for classifying the data can also be used for distinguishing between the relevant and the irrelevant material. For example, in a research application on political protests, we can use neural networks to classify whether a segment contains protest footage or not.

Second, real-world applications of computer vision in the social sciences are often more complex than the case studies in computer science. Whereas applications in computer science tend to zero in on one particular classification problem, real-world applications need to deal with multiple problems at once. Following up on the case study outlined in the previous section, we might not only be interested in detecting and identifying legislators, but we might also be interested in classifying their parliamentary behaviors, which would require additional classification steps. While this is technically possible, the ambiguities associated with a research project increase as more classification steps are added, in our case from face detection to face recognition to activity recognition.

Third, the simplest way to conduct automated video analysis is to rely on pre-trained models to assign the footage to common categories of interest. The downside of using out-of-the-box solutions is that the categories are often too blunt to speak to our research questions. Therefore, to improve the results, it is often preferable to train a model for the specific research problem, which means having to manually code a portion of the material to ensure that the categories for the classification fit the research problem.

Fourth, privacy considerations and formal privacy rules and regulations are an important factor to consider when analyzing real-world video footage using automated video analysis. In light of the many applications of computer vision for the purposes of law enforcement, the general public has become increasingly aware of these technologies and wary of the possibilities for their use and abuse. Therefore, along with stricter rules on other forms of big data analysis, it is likely that computer vision will continue to be a controversial technology, potentially resulting in stricter limitations on its use. At the very least, privacy laws, such as the European General Data Protection Regulation, frequently mandate the storage of the resulting data on secured servers, creating added levels of difficulty when engaging with video data.

These challenges notwithstanding, we hope that this chapter has highlighted the potentials of automated video analysis for the social sciences and how the available computer vision tools might support research in diverse areas.

Bibliography

Arnold, F., Kauder, B., & Potrafke, N. (2014). Outside earnings, absence, and activity: Evidence from German parliamentarians. *European Journal of Political Economy*, 36, 147–57.

Bernecker, A. (2014). Do politicians shirk when reelection is certain? Evidence from the German parliament. *European Journal of Political Economy*, 36, 55–70.

Besley, T., & Larcinese, V. (2011). Working of shirking? Expenses and attendance in the UK parliament. *Public Choice*, 146(3), 291–317.

Bratton, K., & Rouse, S. M. (2011). Networks in the legislative arena: How group dynamics affect cosponsorship. *Legislative Studies Quarterly*, 36(3), 423–60.

Cao, Z., Hidalgo, G. M., Simon, T., Wei, S.-E., & Shekh, Y. (2019). *OpenPose: Realtime multi-person 2D pose estimation using part affinity fields*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Cho, W. K. T., & Fowler, J. H. (2010). Legislative success in a small world: Social network analysis and the dynamics of Congressional legislation. *Journal of Politics*, 72(1), 124–35.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *Imagenet: A large-scale hierarchical image database*. IEEE Conference on Computer Vision and Pattern Recognition.

Dietrich, B. J. (2015). If a picture is worth a thousand words, what is a video worth? In R. X. Browning (Ed.), *Exploring the C-SPAN archives: Advancing the research agenda* (pp. 241–63). Purdue University Press.

Dietrich, B. J. (forthcoming). Using motion detection to measure social polarization in the U.S. House of Representatives. *Political Analysis*.

Dietrich, B. J., Enos, R. D., & Sen, M. (2019). Emotional arousal predicts voting on the U.S. Supreme Court. *Political Analysis*, 27(2), 237–43.

Dietrich, B. J., Hayes, M., & O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of Congressional speech. *American Political Science Review*, 113(4), 941–62.

Dietrich, B, J., & Juelich, C. L. (2018). When presidential candidates voice party issues, does Twitter listen? *Journal of Elections, Public Opinion and Parties*, 28(2), 208–24.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). *Long-term recurrent convolutional networks for visual recognition and description*. IEEE Conference on Computer Vision and Pattern Recognition.

Fowler, J. H. (2006). Connecting the Congress: A study of cosponsorship networks. *Political Analysis*, 14(4), 456–87.

Girshick, R. (2015). *Fast R-CNN*. IEEE International Conference on Computer Vision.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation*. IEEE Conference on Computer Vision and Pattern Recognition.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–97.

Haim, M., & Jungblut, M. (forthcoming). Politicians' self-depiction and their news portrayal: Evidence from 28 countries using visual computational analysis. *Political Communication*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. IEEE Conference on Computer Vision and Pattern Recognition.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask R-CNN*. IEEE International Conference on Computer Vision.

Heath, C., Hindermarsh, C. J., & Luff, P. (2010). *Video in qualitative research: Analysing social interaction in everyday life*. Sage.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–80.

Hohl, K. (2017). *Agenda Politics im Parlament: Das Themen- und Tagesordnungsmanagement der Opposition im Landtag von NRW*. Springer VS.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *Mobilenets: Efficient convolutional neural networks for mobile vision applications*. arXiv:1704.04861.

Hu, P., & Ramanan, D. (2017). *Finding tiny faces*. IEEE Conference on Computer Vision and Pattern Recognition.

Iandola, F., Hang, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size*. arXiv:1602.07360.

Joo, J., Bucy, E. P., & Seidel, C. (2019). Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal of Communication*, 13, 4044–66.

Kharroub, T., & Bas, O. (2016). Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. *New Media and Society*, 18(9), 1973–92.

Knox, D., & Lucas, C. (forthcoming). A dynamic model of speech for the social sciences. *American Political Science Review*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. Advances in Neural Information Processing.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., & Fu, C.-Y. (2016). *SSD: Single shot multibox detector*. European Conference on Computer Vision.

Lucas, C., Nielsen, R., Roberts, M., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2), 254–77.

Neumayer, C., & Rossi, L. (2018). Images of protest in social media: Struggle over visibility and visual narratives. *New Media and Society*, 20(11), 4293–310.

Nullmeier, F., Pritzlaff, T., Weihe, A. C., & Baumgarten, B. (2008). *Entscheiden in Gremien: Von der Videoaufzeichnung zur Prozessanalyse*. VS Verlag für Sozialwissenschaften.

Pasczke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). *Pytorch: An imperative style, high-performance deep learning library*. Advances in Neural Information Processing.

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of Presidential candidates with computer vision. *Journal of Communication*, 68(5), 920–41.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). *You only look once: Unified, real-time object detection*. IEEE Conference on Computer Vision and Pattern Recognition.

Redmon, J., & Farhadi, A. (2017). *YOLO9000: Better, faster, stronger*. IEEE Conference on Computer Vision and Pattern Recognition.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). *Faster R-CNN: Towards real-time object detection with regional proposal networks*. Advances in Neural Information Processing Systems.

Rose, J., Mackey-Kallis, S., Styles, L., Barry, K., Biagini, D., Hart, C., & Jack, L. (2012). Face it: The impact of gender on social media images. *Communication Quarterly*, 60(5), 588–607.

Ruhose, F. (2019). *Die AfD im Deutschen Bundestag: Zum Umgang mit einem neuen politischen Akteur*. Springer.

Ryle, M. (1991). Televising the House of Commons. *Parliamentary Affairs*, 44(2), 185–207.

Schroeder, W., Weßels, B., & Berzel, A. (2018). Die AfD in den Landtagen: Bipolarität als Struktur und Strategie: Zwischen Parlaments und 'Bewegungs'-Orientierung. *Zeitschrift für Parlamentsfragen*, 49(1), 91–110.

Schroff, F., Klenichenko, D., & Philbin, J. (2015). *Facenet: A unified embedding for face recognition and clustering*. IEEE Conference on Computer Vision and Pattern Recognition.

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). *Hand keypoint detection in single images using multiview bootstrapping*. IEEE Conference on Computer Vision and Pattern Recognition.

Uijlings, J. R. R., Van de Sande, K., E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, 104(2), 154–71.

Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). *Action recognition by dense trajectories*. IEEE Conference on Computer Vision and Pattern Recognition.

Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). *Convolutional pose machines*. IEEE Conference on Computer Vision and Pattern Recognition.

Weihe, A. C., Pritzlaff, T., Nullmeier, F., Felgenhauer, T., & Baumgarten, B. (2008). Wie wird in politischen Gremien entschieden? Konzeptionelle und methodische Grundlagen der Gremienanalyse. *Politische Vierteljahresschrift*, 49(2), 339–59.

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analsis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20, 529–44.

Williams, N. W., Casas, A., & Wilkerson, J. D. (2020). *Images as data for social science research.* Cambridge University Press.

Zhang, H., & Pan, J. (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.