Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.1 (2022) 350-371
HTTPS://DOI.ORG/10.5117/CCR2022.1.010.BALD

ARTICLE

A Character Recognition Tool for Automatic Detection of Social Characters in Visual Media Content

Joshua Baldwin
Department of Communication, Michigan State University

Ralf Schmälzle
Department of Communication, Michigan State University

Abstract

Content analysis is the go-to method for understanding how social characters, such as public figures or movie characters, are portrayed in media messages. It is an indispensable method to investigate character-related media processes and effects. However, conducting large-scale content-analytic studies is a taxing and expensive endeavor that requires hours of coder training and incurs substantial costs. This problem is particularly acute for video-based media, where coders often have to exert extensive time and energy to watch and interpret dynamic content. Here we present a *Character-Recognition-Tool* (*CRT*) that enables communication scholars to quickly process large amounts of video data to identify occurrences of specific predefined characters using facial recognition and matching. This paper presents the *CRT* and provides evidence for its validity. The *CRT* can automate the coding process of on-screen characters while following recommendations that computational tools be scalable, adaptable for novice programmers, and open source to allow for replication.

Keywords: computer vision, computational communication, content analysis, face recognition, media

A Character Recognition Tool for Automatic Detection of Social Characters in Visual Media Content

Content analysis, the systematic characterization of meaning and patterns within media messages, is the backbone method of mass communication research (Lovejoy et al., 2014). In this context, televisual and filmic media

are a crucial area of interest. However, with the continuing rise of visual media specifically, the application of computational tools for video-based content analysis in mass communication research is limited. Furthermore, the social characters (e.g., on-screen fictional portrayals, public figures, politicians, etc.) depicted in televisual and filmic media are critical anchor elements central to their plot and widespread appeal. Therefore, character recognition is a central task for visual content analysis, much like named-entity-recognition (NER) is fundamental for text-based analysis (Goyal et al. 2018). Although sophisticated commercial systems enable the detection and recognition of faces, the proprietary nature and cost of these tools create barriers for researchers interested in analyzing visual media and raise the demand for open-source, free, and easy-to-use research tools (Trilling & Jonkman, 2018).

This paper introduces and validates a *Character-Recognition-Tool* (*CRT*) for automatic computational analysis that allows researchers to detect the presence of prespecified characters within video content. The structure is as follows: In the next section, we review evidence of why the detection of characters' faces is often central to visual content analyses, followed by a brief discussion of the challenges of human content analysis. Next, we discuss the history and state of the art of facial recognition to equip readers with the necessary technical background for understanding the methodology for computational character recognition. The following sections provide a practical introduction to the *Character-Recognition-Tool*, followed by a short description of a validation study. We end by discussing areas for application, limitations, potential expansions, and avenues for future research in this rapidly evolving area.

The Central Role of Characters in Visual Media Content

With the rise of computational communication research (Shah et al., 2015), many researchers have focused their attention on text-based content. Video content, however, should not be ignored because televisual or filmic media, like TV shows, online videos, and movies, are consumed very frequently and by large audiences. For instance, the average daily TV viewing time is almost three hours, and over two-thirds of the nearly 8 billion world population watch television (Krantz-Kent, 2018). Even on traditionally text-heavy social media sites, video content is on the rise, and video-focused sites like YouTube and TikTok range among the most popular platforms. In addition to frequency and amount of use, the higher immediacy of visual,

as opposed to text-based content, has been suggested to cause more potent audience effects (Draft & Lengel, 1984; Gibson & Zillmann, 2000; Messaris, 1997). This emphasizes the need for methods to analyze what content is on screen, its prevalence, and its impact.

A prominent feature of video-based media content is that it is replete with human characters, such as actors portraying fictional characters in entertainment content or public figures in news and advertisements. Humans are an inherently social species, and accordingly, they take a genuine interest when other humans are depicted on-screen (Giles, 2002). Indeed, human characters' actions, fate, and interactions with other characters are featured prominently in visual content. Such content is also central to the widespread appeal of entertainment (Raney, 2008) and non-entertainment media (Zillmann et al., 1998). As a result, most variables of interest in media content analyses, such as violence, morality, or interpersonal relationships, are also inherently social.

Critically, the human face is key to our ability to recognize those on screen, and therefore a critical anchor element in understanding visual media. For instance, in movies, audiences follow the protagonists by recognizing the faces of specific actors across scenes. Similarly, television viewers frequently encounter familiar persons such as news anchors or politicians. In each of these examples, the face is a crucial information source for *character recognition*, which is required to track individuals across scenes, infer their attributes, and understand their actions over time. Thus, the ability to recognize characters and track their appearance and behaviors over time is fundamental for the analysis of visual content (see Figure 1).

The Challenges of Human Coding and the Potential of Computational Analyses

Because characters are central to the social behaviors depicted in entertainment, news, and advertising media, their identification is central to the human-coding process in manual content analyses. However, this task is very laborious for human coders, especially when many characters appear on-screen simultaneously, or when a character appears infrequently. Also, if the amount of content to be analyzed is substantial, such as an entire multi-season TV show or a year's worth of news footage, it may even become prohibitively time-consuming and expensive. For instance, the 122 episodes of the popular show *Parks and Recreation* are each 22 minutes long, amounting to 161,040 seconds of content. Manual coding of such content

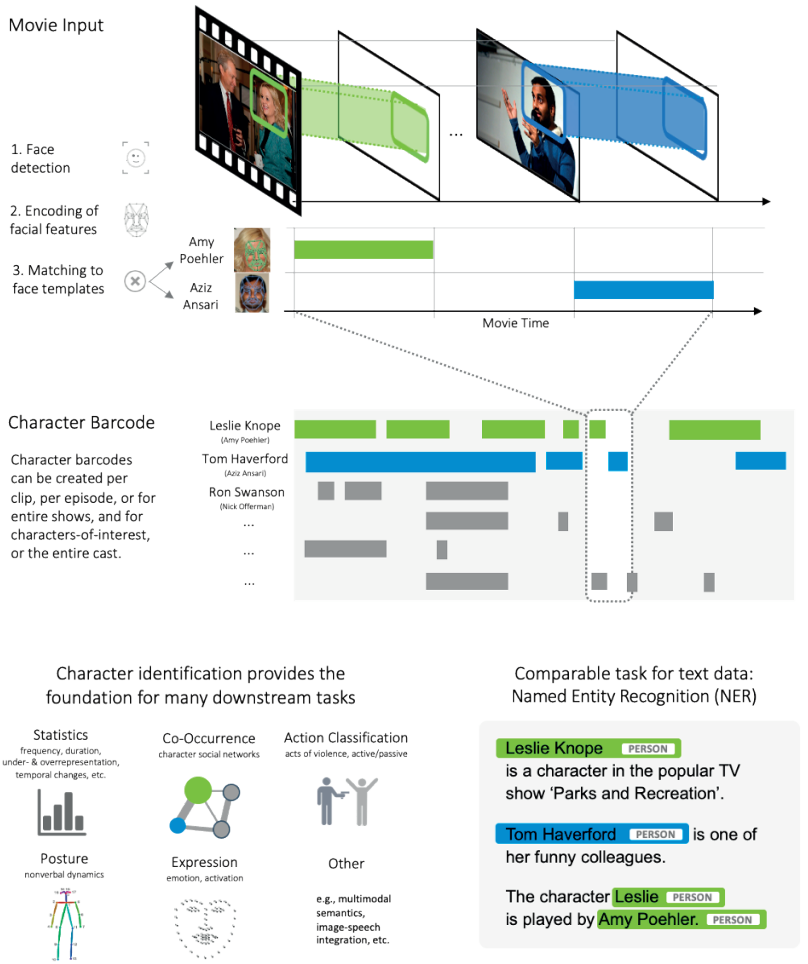

Figure 1 Illustration of the Concept of Character Recognition in Visual Data Streams

Note. Similar to how Name Entity Recognition (NER) can detect person names in text, the CRT can detect and match characters on screen using prespecified face templates (face encodings/vectors). This approach can create a character barcode, revealing when, for how long, and how often a character appears. Character recognition is a critical starting point for many downstream applications. Thus, in a media analysis pipeline, many subsequent tasks can take the video sequences in which a particular character appears as input data. Images are CC licensed and taken from https://commons.wikimedia.org/wiki/File:Aziz_Anzari_December_2011.jpg, Ed_L._Munson_and_Amy_Poehler,_May_2012.jpg, Amy_Poehler_by _David_Shankbone.jpg, Aziz_Ansari_2012.jpg.

poses operational and financial challenges, and the quality of the result may suffer from human capacity limitations, coder wear-out, and attentional fluctuations. This invites new methods for tracking specific characters' appearance over time, preferably ones with high automation potential.

As of today, most video-based content analytic research still relies on multiple human coders. Such time- and resource-consuming analyses are difficult to replicate, if not outright impossible when the scope of the project is large. For example, one of the most famous and respected content analyses, the *National Television Violence* Study (1996, 1997), took researchers from two universities and over 50 trained coders to code over 10,000 hours of television content over a three-year period. Granted, such large content analyses may not be the norm, but most content analytic researchers are well aware of the difficulties, time, and funds needed to train and manage human coders. From our own experiences, we know that maintaining intercoder and intracoder reliability with an ever-changing group of coders is a monumental task.

In sum, researchers have long known the difficulties involved in conducting a high-quality content analysis, especially for visual media (Krippendorff, 2004; Riff et al., 2014). Therefore, there is great interest in accessible methods that could make it easier to analyze media content (Trilling & Jonkman, 2018). In this paper, we propose a *Character-Recognition-Tool* (*CRT*) that leverages facial recognition technology to identify and track the appearance of specific characters in large streams of visual content. Critically, this tool is geared towards content analysis practitioners and thus requires minimal to no coding skill, while still offering flexibility to those who possess or want to develop it. Before introducing this tool, the next section reviews how characters can be detected automatically.

History and Present State of Facial Recognition Systems

Over the past decades, algorithms for automated face detection and analysis have become widespread (Datta et al., 2015). As of 2021, basic face detection is a standard feature of every cellphone, and facial recognition technology pervades our society, including social media, law enforcement, and various access control systems. Although facial recognition is now regularly employed in commercial settings, few studies have used these tools for visual-based content analyses. Furthermore, most existing studies have been published outside of communication (e.g., Zhu et al., 2013), and only very few have examined video content specifically (e.g., Joo et al., 2018; Guha et al., 2015), focusing instead on face recognition from photographs (e.g., Peng, 2018). This seems particularly notable since a handful of research advancing facial recognition and other computer vision tools have used the same type of video content that mass communication scholars are interested in studying

(i.e., movies and television) as training sets for large-scale machine-learning models (e.g., Nagrani & Zisserman, 2017; Patron-Perez et al., 2010). In sum, there exists a gap between existing technology and its use by content analytic researchers, and closing this gap is one of the main motivations of this paper.

Computers' ability to recognize faces builds on recent advances in computer vision, a subfield of artificial intelligence/machine learning (Taskiran et al., 2020; Bradski & Kaehler, 2008). In a nutshell, these algorithms involve first detecting whether a natural image contains a face (based on typical configural features), second analyzing the detected face in terms of its features (such as inter-eye distance, nose length, etc.; although features can be more complex and difficult to describe verbally), and finally matching these feature-based measurements with a named template, akin to fingerprint identification. Then, if the correlation between the facial measurements of the depicted image (e.g., *Leslie Knope* from the show *Parks and Recreation*) with the given template (e.g., a picture of the actress *Amy Poehler*) is high enough, a match-decision is made, and the face is assigned the template's name as a label (e.g., the character is measured as *Leslie Knope/Amy Poehler*, see Figure 1).

Over the past decades, this face recognition technology has matured to work at or above human-level performance. In recent benchmarks, such as the Labeled Faces in the Wild dataset (Huang et al., 2008), automated face recognition systems have achieved performance above 95% accuracy (e.g., Maze et al., 2018). A variety of open-sourced algorithms exist that differ in terms of the specific computer vision techniques they use. These range from HOG (histograms of oriented gradients) to modern deep learning (Chollet, 2018). Such methods are implemented, for instance, in the widely used dlib machine learning toolkit (King, 2009) and made accessible in dedicated packages (Geitgey, 2016).

The Character Recognition Tool (CRT): Principle and Design Philosophy

The potential use of facial recognition software for content analyses in communication science has been highlighted before (Evans, 2000), but communication researchers have not systematically leveraged the recent advances in computer vision. Outside of the communication discipline, there have been several projects that analyze media content, including facial recognition (e.g., Dhall et al., 2012), but these are geared towards different audiences, tend to have a relatively high technical threshold, and pursue

other goals than social scientific content analysis such as developing recommendation systems. As a result, performing content analysis still requires sifting manually through vast amounts of visual content - forwarding and replaying tracks until the critical scene is identified, information is extracted and coded by humans.

We introduce and validate a *Character-Recognition-Tool* (*CRT*) for automated visual content analysis to overcome this bottleneck. The *CRT* is built on the premise that characters play a central role in visual content and that advances in facial recognition technology make it relatively straightforward to detect a character's appearance computationally via template-matching (i.e., knowing that Actor X in this movie represents Character Y). Our simple yet effective solution is to use individual actors' faces as templates and use automated facial recognition software to detect matches between on-screen characters and the templates. Thus, by capitalizing on computerized face recognition systems' ability to quickly extract this meaningful information from large amounts of video data, we can simplify and substantially accelerate the content-analytic process. Moreover, identifying scenes in which a given character appears provides a springboard for more nuanced content analysis tasks for concepts that are still difficult to automatize (Figure 1).

Based on Trilling and Jonkman's (2018) recommendations, the *CRT* is free, open-source, and easy to adopt for visual content analysis. Over the past decade, several commercial platforms emerged for the study of visual content. For instance, *GoogleVision, Amazon Rekognition, Microsoft Azure,* and *Carifai* all provide interfaces that allow users to analyze content automatically, including tasks like face recognition. However, because these systems operate on a by-volume basis, they can become relatively expensive if large amounts of content are to be analyzed (e.g., a 90-minute movie with about 30 frames per second contains 162,000 frames). The algorithmically opaque nature of commercial systems is also incompatible with transparency and reproducibility. Overall, given that face recognition is a well-articulated task for which noncommercial methods exist, we decided to choose this route. This strategy also makes it possible to expand the system by performing analysis of facial expressions of actors and so forth. Again, successful models in the text-domain, such as the NLP-pipelines of NLTK, spacy, or stanza, provide good examples for this approach.

In the next section, we introduce the *CRT's* basic architecture and functionality. Then, we will report a validation test comparing human coders with the *CRT*.

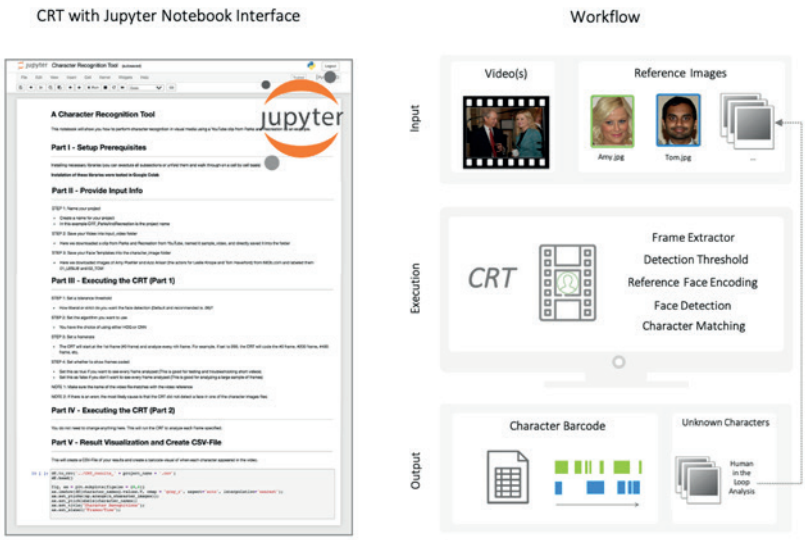

Figure 2 CRT User Interface and Principle

Note. Left Panel: Screenshot of the CRT, which is mainly executed as a Jupyter Notebook. Such notebooks provide narrative explanations along with the actual code, and the executed notebook can be saved for documentation purposes and to reproduce the results. Right Panel: Illustration of a typical CRT workflow in which a video and reference images are provided, and the CRT searches for occurrences of those images within the video. Faces that don't match the original reference images are marked and saved to disk for a human-in-the-loop coder to decide whether this face should be added to the list of reference images for the next iteration.

Basic System Architecture and Installation

The *CRT* uses open-source Python packages and is available online at https://git.io/JDEBo. At the heart of the *CRT* lies the popular open-source Python library *face_recognition* (Geitgey, 2016), and we use the *pliers* package for extracting video frames (McNamara et al., 2017). In essence, *CRT* first converts a given video input file into individual image frames. Next, it searches whether a given frame contains a face that matches one of several template face images. If a match occurs, this is stored, resulting in an output matrix that indicates which frame contains which characters' faces.

The *CRT* uses Python, a free programming language that has become the most popular language for scientific computing, especially computer vision and machine learning. The *CRT* was programmed on Jupyter notebooks (Kluyver et al., 2016), but it can also be used via scripts. The Jupyter notebook is an open-source, interactive computing environment that runs in a web browser. Jupyter Notebooks combine narrative text and executable

software code, thereby making it very easy to explain computer programs to non-coders, and letting them execute code in a very accessible interface.

We opted to use this environment over standalone or web applications for several reasons. First, the *CRT* itself is open-source and uses only open-source software, which means no cost. Importantly, no further charges are incurred to execute projects because no proprietary software is used and no commercial APIs are called. While this already lowers the bar for adoption, the user interface is intuitive and easy to learn but still offers high flexibility for further development. Moreover, Jupyter notebooks are excellent tools for documenting analysis steps, thereby enhancing reproducibility.

We provide the *CRT* online for download and offline use along with installation instructions. Additionally, we provide a *one-click solution* that allows executing the *CRT* notebooks in free online environments, such as *Binder* and *Google Colab*. Thus, by simply clicking on the link on the website, users start a cloud-based environment to run their analyses and later download the results or save them to a cloud storage space, such as Google Drive. Again, these choices shall provide beginners with an easy-to-use tool that does not require complicated installations, while also giving more advanced users the ability to expand on the basic *CRT* routines without restriction.

Running the CRT: A Quick Walkthrough

The two necessary ingredients for running a visual content analysis via the *CRT* are 1) a video message and 2) one or more to-be-detected face templates. For instance, suppose one wanted to analyze a clip of the popular TV show *Parks and Recreations*, then one would need a video of the clip itself (as an MP4 file) and a photograph of at least one of the actors (e.g., a JPEG image of *Amy Poehler* for the character of *Leslie Knope,* which can easily be found on IMDb.com). Once these two prerequisites are placed in the subfolder *'input_data'*, everything is ready to run.

To run the analysis, the user simply executes the cells of the *CRT* notebook by pressing a *run*-button. We provide various explanations of the intermediary steps to inform users and enable even novice programmers to understand the relatively simple code. Users are also encouraged to change preconfigured settings. For instance, by analyzing only every 10th frame, one can speed up the computation at the cost of a coarser result. Once the tool is finished, the user receives a graphical output of which characters are detected at what time point of the input clip. This character-barcode

simply illustrates the *CRT*'s character-by-time output matrix, which is also saved out as a CSV file to the folder *'output_data.'*

Importantly, if the *CRT* detects a face, but does not find a match to any of the provided face templates, it saves a snapshot of the image to an interim folder *'unrecognized_images.'* After a first-pass run of the *CRT*, users can visit this output folder and judge whether they want to include this character in the analysis (e.g., a side character such as *Aziz Ansari*). If one chooses to do so, one can simply include this character as a face template in the *'input_data'* folder and rerun the *CRT*. This *human-in-the-loop* style procedure can be repeated iteratively until all relevant characters are included (Zamith & Lewis, 2015).

Example Study: Testing CRT's Validity

In this section, we demonstrate the *CRT* in practice by reporting a small example study wherein (a) we show how *CRT* can be used to detect characters in television shows and (b) provide evidence for the *CRT*'s validity. The motivation for this is that while there is no doubt that the *CRT* is superior to humans in terms of its low-cost, highly scalable, efficient, and objective nature, it remains to be demonstrated that computer-based character recognition matches human-level accuracy.

To provide this evidence of the tool's validity, we designed a test in which we compared results with human coders with the *CRT's* output. In brief, we drew a sample from different current TV shows and had human coders annotate the presence of faces and code their identity. The same procedures were also carried out by the CRT, enabling us to compare the overall performance. We find that the *CRT* is overall on par with human coders, but given its strong performance, low cost, and high scalability is superior to human coding of character appearance.

Methods and Materials

Sample. We devised a list of 154 television shows that aired during primetime on the major broadcasting networks in 2016. Although this sampling frame may not be exhaustive, we note that the purpose of the study was to establish measurement validity rather than generalizability. Thus, this is an appropriate sampling frame. From the list of 154 shows, a stratified random sample was drawn. Specifically, the shows were separated into three categories based on IMDb.com genre tags *comedy* (n = 31), *drama* (n = 64), and *reality shows* (n = 24), excluding mixed-genre shows. These groups were chosen to ensure that the *CRT* coded for characters who could be portrayed on screen

Table 1 Overview of Shows used for the Validation Study

Show	*Genre*	*Season*	*Episode*
Crazy Ex-Girlfriend	Comedy	2	*Where is Josh's Friend*
The Middle	Comedy	8	*The Core Group*
Quantico	Drama	2	*KUDOVE*
The Flash	Drama	3	*Flashpoint*
The Biggest Loser	Reality-Show	17	*Money Hungry*
Shark Tank	Reality-Show	8	*Episode #8.1*

in a variety of ways. For instance, a drama may have darker lighting and more special effects compared to a comedy or reality-show. Of note, animation shows were excluded from the sample since the *CRT* is not designed to pick up animated characters. From each genre, two shows were randomly chosen for the study. The chosen shows were *Crazy Ex-Girlfriend* and *The Middle* (comedy), *Quantico* and *The Flash* (drama), and *The Biggest Loser* and *Shark Tank* (reality; see Table 1). From each show, the first ten minutes of the pilot episode of the 2016 season were recorded and used for analysis.

Character Selection. Seventy-seven characters were selected to code for their presence or absence based on the credit list of each episode's IMBd.com page. Images of the actor or actress playing the character were used as the reference images for the CRT and the human coders, respectively.

Computer Coding. The *CRT* was provided with the videos and reference images to code for the presence or absence of a character. The *face_recognition* package was set to the default tolerance of .6. Lower tolerance increases strictness while higher numbers reduce strictness. Each ten minutes video was cut into two five-minutes videos for easier processing. After processing 108,000 frames from the 12 five-minute videos, the units were collapsed into 15-second intervals (1800 frames) and then collapsed as percentage scores (i.e., the percentage of frames the character was coded in the 1800 frames).

Human Coding. In addition to the computer coding, a human-based content analysis was conducted with the same variables and sample. Two human coders were provided the same reference images of the actors/actresses who played the seventy-seven characters in the shows. The coders then coded for the presence and absence of characters within 15-second segments. After the coding was completed, a third referee coder was recruited to resolve any disagreements between the two coders. The average Krippendorff alpha for character presence was .62. After removing characters with low reliabilities (i.e., those with $\alpha < .50$; $n = 16$) the average Krippendorff alpha was .82. The final analysis is thus based on a total of sixty-one characters.

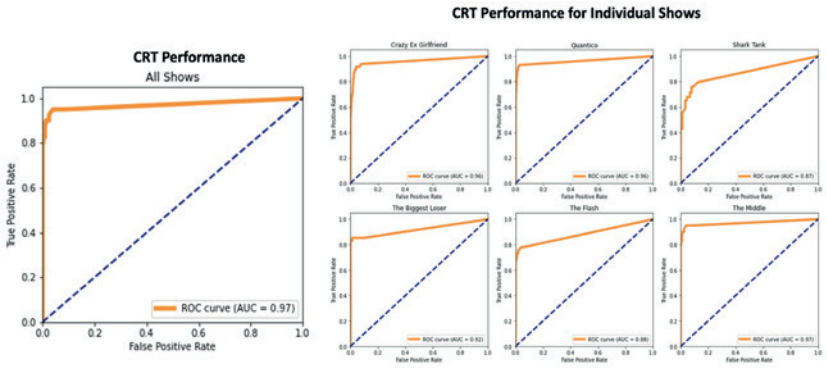

Figure 3 ROC Curve of CRT Performance

Note. Human coders are treated as ground truth. Left panel: ROC for all shows. Right panel: ROC curves for individual shows.

Validation Study: Results and Discussion

To test the accuracy of the *CRT*, we compared its performance against the performance of human coders. Specifically, we plotted Receiver Operating Characteristic (ROC) curves across all shows as well as for each show individually (Figure 3). A ROC curve plots the false positive fraction (i.e., 1 - specificity) against the true positive fraction (i.e., sensitivity). In the context of the character recognition task, true positives represent detections of a character when the character was actually on-screen (using human raters as ground truth), whereas false negatives would represent mistakenly detected characters who were not actually depicted. Practically, the closer the ROC curve is to the upper left corner and away from the chance line along the main diagonal, the higher the discrimination, and therefore, the higher the accuracy of the tool (Zweig & Campbell, 1993). A summary statistic for ROC curves is the area under the curve (AUC), whereby a ROC AUC of 1 represents a perfect score and 0.5 indicates performance at chance level.

We assumed that the human coders were perfectly correct in discerning presence and non-presence, providing so-called 'ground truth' against which the *CRT* results can be compared. Figure 3 shows the result of the ROC analysis for all shows (i.e., 2440 potential character detection cases). The overall performance across shows was at *AUC* (n = 2440) = .93, 95% *CI* = [.92, .94]. This demonstrates that the *CRT* has the ability to detect faces well above chance and at a level close to human performance. All analysis and data can be found on the study's Github repository: https://git.io/JDE07.

While previous research has shown that facial recognition can have high accuracy for detecting the static faces of people in individual photographs, our small study demonstrates how the *CRT* is able to detect a set of dynamic characters on screen that are not a part of an initial training-testing dataset and correctly match them to prespecified templates. Unlike the still photos that most facial recognition models are built upon, the characters in our six shows are constantly changing in their positions, head direction, how they move, how they are lighted, and whether or not there are obstructions. In other words, the way the characters are depicted in the videos is a lot messier. Thus, this study provides support for the validity of the *CRT* for content analytic research. Together with the fact that the CRT is immensely scalable, cost-free, and can sift through large amounts of video footage without ever getting tired, this suggests that the tool can be used profitably to aid visual content analysis of video-based media.

Example Applications across Communication Science

Having demonstrated the *CRT* and affirmed its validity in a small validation study, we next discuss its potential for content analysis projects across different domains of mass communication research. Specifically, we illustrate promising use-cases in entertainment studies, political communication, health communication, and communication neuroscience. We note here that the *CRT* may be especially useful when combined with other computational tools in order to go beyond its basic capabilities - face/character recognition - and to approach higher-level socially latent variables, such as the emotional expressions, character co-occurrences, or face-object pairings and visual sentiment analysis. In its current form, the *CRT* only performs the basic task of visual character recognition, which is conceptually equivalent to the task of named entity recognition in NLP. In the following examples, we highlight some ideas on how to combine multiple computational tools in a pipeline to capture variables of interest in contexts relevant to communication researchers.

Example 1: Entertainment Studies

Characters often take on an important role in audiences' lives when people develop parasocial relationships with them (Giles, 2002). A researcher interested in how audiences react emotionally to their favorite characters over time may conduct a diary study in which participants record their media

use as well as their emotional states over a set time period. The researcher could then use the *CRT* to content analyze the consumed media content mentioned in the diary to extract the timepoints when the participant's favorite characters were on screen. From this, one could then correlate the times when favorite characters were on-screen and participants' emotional state. Furthermore, once the *CRT* timestamps a person's favorite character, human coders could conduct a follow-up content analysis to evaluate the character's situation in each scene (e.g., was the character in a positive or negative situation).

Example 2: Political Communication

Political communication scholars have long been interested in the manner in which political figures and events are visually portrayed (e.g., Casas & Williams, 2019). Of particular interest is how political figures set the public agenda (McCombs & Shaw, 1972) and how the news frames their behaviors (Masters et al., 1991). With the use of the *CRT*, it would be easy to extract the exact frequency and length a political figure appears on the 24-hour news cycle. If we want to know how the news coverage frames a particular political figure, we could utilize additional computational tools to also extract the speech of news anchors right before and after the political figure is shown, or quantify qualities of the depicted images, such as the share of screen covered, the presentation angle (from below vs. above), or whether there are national symbols in the background (Schill, 2012).

Example 3: Health Communication

The prevalence of alcohol and tobacco products in television and movies is a longstanding public health concern (e.g., Bergamini et al., 2013). A related topic is how frequently minority groups are represented using alcohol and tobacco products on-screen (e.g., Lee et al., 2013). By combining the *CRT* with additional object detection tools, such as the Alcoholic Beverage Identification Deep Learning Algorithm (Kuntsche et al., 2020), one can explore the frequency in which minority groups are depicted with alcohol and tobacco products. Specifically, researchers could use the *CRT* to detect faces from a pre-existing list of minority actors and then use additional object detection tools to examine whether alcohol or tobacco products are presented at the same frame-by-frame time points at which the actors appear. For actors that are not in the a-priori list, the *CRT* can save the unknown detections for human coders to manually inspect for ethnicity and feed this information back into the *CRT* using the human-in-the-loop technique.

Example 4: Communication Neuroscience

The health communication example provides a good transition for another use case for the *CRT*. Wagner et al., (2011) measured the brain activity of smokers and nonsmokers exposed to cues of smoking in movies. Such studies could be scaled up and refined by using the *CRT* to detect when particular characters of interest are on-screen with great specificity. More broadly, the emerging field of communication and media neuroscience (e.g., Weber et al., 2015; Schmälzle & Meshi, 2020) strives to complement media effects research with insights into media mechanisms. However, this endeavor hinges critically on our ability to quantify the content elements that evoke specific neural responses. Since changes in brain activation happen within milliseconds (Schmälzle et al., 2011; Luck & Kappenman, 2011), which is too fast for human coders, the *CRT* can analyze data at the frame level to get the precise moments when characters and their attributes are on and off-screen.

Future Directions, Expansions, and Current Limitations

Having outlined how the *CRT* provides an easy and cost-free way to measure character prevalence and how it could be used across different domains, we next turn to avenues for future research and expansion. The *CRT* can be a starting point to select relevant scenes that can be displayed to human coders for further analysis. For instance, many content analyses are designed to capture information that computers currently cannot identify automatically (e.g., higher-level latent qualities such as humor). For such cases, the *CRT* can be used essentially as a cutting-tool to identify the scenes in which a relevant character occurs and save them out as a 'character-trailer.' This takes away the burden from human coders of having to identify the character within potentially hours of video footage (Zamith & Lewis, 2015).

In addition to describing character prevalence or preparing data for human coders, the *CRT* can also provide the input for further computational analyses. As discussed above, the field of natural language processing has embraced a pipeline-based approach in which projects are broken down into subtasks, such as tokenization, named-entity-recognition, coreference resolution, etc. A similar strategy will be beneficial for visual data, and character recognition will be central to it. For instance, based on the identification of scenes featuring the character *Leslie Knope*, it becomes possible to ask further questions like how her facial expression changes over time, in which settings she appears, or how surrounding characters respond. Thus, the *CRT*

could again be used to save out relevant clips or frames, which can then be further analyzed using custom tools for facial expression analysis, action recognition, or analyses of character co-appearance.

Beyond these immediate possibilities, we anticipate that the use of computer vision methods for large-scale analyses of media content is about to explode (e.g., Zellers et al., 2019). Content analytic researchers have long understood the potential for computational tools to reliably automate the coding process involved in video-based content analyses. For instance, twenty years ago Evans (2000) imagined a situation wherein computational tools could "permit social scientists to automate many common content analysis procedures" that would allow researchers to measure relevant higher-level variables such as acts of violence within videos (p. 247). While early computer vision tools were neither accessible nor powerful enough to code such variables, nowadays, this is becoming a possibility (Monfort, et al., 2019; Zellers et al., 2019). With the introduction of computational communication science (Shah et al., 2015), advances in deep learning and computer vision (Chollet, 2018), and the rise in code-sharing websites (e.g., github.com), powerful and accessible computational tools and methods are readily becoming more available. Although it is certainly not yet the case that computers can 'understand' what they see, their vision becomes increasingly refined (e.g., Levesque, 2019; Mitchell, 2020).

For instance, researchers may soon want to build so-called "visual dictionaries" to code for the presence of variables of interest in a similar manner to text-based content analysis methods (Pennebaker, 2001). One may create a "violence dictionary" where a computer can code for the presence of objects and actions such as a gun or a punch as indicators of violence. Combined with facial recognition, researchers could code not only for the prevalence of violence but also *who* are the victims and perpetrators of that violence. While we are unaware of any such visual dictionary in use, some scholars have begun exploring how a particular combination of visual cues may indicate specific social phenomena. For example, Schmälzle and Grall (2020) used object detection to find movie scenes that were particularly suspenseful (e.g., close-up shots of a gun).

Building on this idea further, Araujo et al. (2020) recently demonstrated that it is possible to use supervised machine learning using data retrieved from object detection models to detect socially latent variables (e.g., different types of environmental frames in this case) in images. Applied to video content, researchers could use a combination of character recognition, object detection, and speech recognition together as data to train a model using supervised learning without the need for a specific visual dictionary.

With that said, the application of this technique with video content may be a worthy goal for future research.

Despite our optimism for the application of facial recognition in content analytic research, there are several limitations it brings that need to be noted. First, many computer vision models are biased in one way or another due to the training data used during supervised learning (Zou & Schiebinger, 2018). This is also apparent in facial recognition where reports have shown that minority groups are underrepresented, or that methods are biased to relevant variables, such as skin tone. In a similar vein, the face_recognition library underlying the *CRT* has not yet been shown to work well with children's faces or animated characters reliably (Geitgey, 2016). Although we tried with our validation study to sample from a broad range of shows depicting a diverse pool of characters, researchers should be aware of these limitations. Thus, if the goal is to specifically study minority representation or examine children's media, researchers should devote special attention to validate results with a subsample coded by human raters.

Another more specific limitation arises in situations when a character's face is obstructed (e.g., when looking away, hiding behind something, or wearing a mask). This will cause the tool to fail, even though a human coder could still identify the person based on clothing or other cues. Although complete facial obstruction still poses problems for the tool, we noticed that oblique face orientations may initially fail to be matched to reference images, but this can often be fixed by simply adding the corresponding shot from the *'unrecognized_images'* folder to the set of reference images and rerunning the *CRT*. Lastly, we note that facial recognition can sometimes be 'too good.' For instance, the *CRT* also tracks unimportant faces, such as those in a crowd of people, background photographs, or pieces of art, such as the Mona Lisa portrait. However, these false positives can also be screened out easily via the *human-in-the-loop* technique.

Within this context, it is also important to point to the broader societal debate about responsible AI, privacy, and potential side effects of technologies, particularly facial recognition. In fact, face recognition technology can be an instrument for social control if used for the purpose of surveillance (e.g., at public places), applicant selection (e.g., by insurance companies), or access control, and recently there have been several calls for regulating the use of such technology. Although these concerns do not directly apply to the use of facial recognition in computational media research, social scientists should be particularly aware of these rapid technological developments and ongoing discussions about their implications.

Conclusion

In summary, character recognition in visual media is a task equivalent to that of named-entity-recognition in text analysis. This paper introduces a *Character-Recognition-Tool* (*CRT*) using Python open-source software to accomplish this task. The *CRT* is designed to be free, easy-to-use, scalable, and expandible to downstream applications. We have demonstrated the validity of the tool and provided examples for its use. We anticipate that the *CRT* will be useful to content analysts across different domains of communication science.

References

Araujo, T., Lock, I., & van de Velde, B. (2020). Automated visual content analysis (AVCA) in communication research: A protocol for large scale image classification with pre-trained computer vision models. *Communication Methods and Measures,* 14(4), 239-265. https://doi.org/10.1080/19312458.2020.1810648

Bergamini, E., Demidenko, E., & Sargent, J. D. (2013). Trends in tobacco and alcohol brand placements in popular US movies, 1996 through 2009. *JAMA Pediatrics*, 167(7), 634-639. https://doi:10.1001/jamapediatrics.2013.393

Bradski, G., & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc.

Datta, A. K., Datta, M., & Banerjee, P. K. (2015). *Face detection and recognition: Theory and practice*. CRC Press.

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Annals of the History of Computing*, 19(3), 34-41. https://doi.ieeecomputersociety.org/10.1109/MMUL.2012.26

Casas, A., & Williams, N. W. (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, 72(2), 360-375. https://doi.org/10.1177/1065912918786805

Chollet, F. (2018). *Deeping learning with Python*. Manning Publications Co.

Evans, W. (2000). Teaching computers to watch television: Content-based image retrieval for content analysis. *Social Science Computer Review*, 18(3), 246-257. https://doi.org/10.1177/089443930001800302

Geitgey (2016). *Machine learning is fun! Part 4: Modern face recognition with deep learning*. Medium. https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

Gibson, R., & Zillmann, D. (2000). Reading between the photographs: The influence of incidental pictorial information on issue perception. *Journalism & Mass Communication Quarterly*, 77(2), 355-366. https://doi.org/10.1177/107769900007700209

Giles, D. C. (2002). Parasocial interaction: A review of the literature and a model for future research. *Media Psychology*, 4(3), 279–305. https://doi.org/10.1207/s1532785xmep0403_04

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, 29, 21-43. https://doi.org/10.1016/j.cosrev.2018.06.001

Guha, T., Huang, C. W., Kumar, N., Zhu, Y., & Narayanan, S. S. (2015). Gender representation in cinematic content: A multimodal approach. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 31-34.

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Workshop on faces in 'Real-Life' Images: Detection, alignment, and recognition*.

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, 4044–4066. https://ijoc.org/index.php/ijoc/article/view/10725

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10, 1755-1758.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., ... et al., Jupyter development team. (2016). Jupyter Notebooks – A publishing format for reproducible computational workflows. In F. Loizides & B. Scmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (pp. 87–90). IOS Press.

Krantz-Kent, R. (2018). Television, capturing America's attention at prime time and beyond. *Beyond the Numbers: Special Studies & Research*, 7(14), 1-11. https://www.bls.gov/opub/btn/volume-7/television-capturing-americas-attention.htm

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.

Kuntsche, E., Bonela, A. A., Caluzzi, G., Miller, M., & He, Z. (2020). How much are we exposed to alcohol in electronic media? Development of the Alcoholic Beverage Identification Deep Learning Algorithm (ABIDLA). *Drug and Alcohol Dependence*, 208, 107841. https://doi.org/10.1016/j.drugalcdep.2020.107841

Lee, J. G., Agnew-Brune, C. B., Clapp, J. A., & Blosnich, J. R. (2014). Out smoking on the big screen: Tobacco use in LGBT movies, 2000–2011. *Tobacco Control*, 23(e2), e156-e158. http://dx.doi.org/10.1136/tobaccocontrol-2013-051288

Levesque, H. J. (2019). *Common sense, the Turing test, and the quest for real AI*. MIT Press.

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2014). Assessing the reporting of reliability in published content analyses: 1985–2010. *Communication Methods and Measures*, 8(3), 207-221. https://doi.org/10.1080/19312458.2014.937528

Luck, S. J., & Kappenman, E. S. (Eds.). (2011). *The Oxford handbook of event-related potential components*. Oxford University Press.

Masters, R. D., Frey, S., & Bente, G. (1991). Dominance & attention: Images of leaders in German, French, & American TV news. *Polity*, 23(3), 373-394. https://doi.org/10.2307/3235132

Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., ... & Grother, P. (2018, February). Iarpa janus benchmark-c: Face dataset and protocol. *2018 International Conference on Biometrics* (*ICB*), 158-165.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2), 176-187. https://doi.org/10.1086/267990

McNamara, Q., de la Vega, A., & Yarkoni, T. (2017). *Developing a comprehensive framework for multimodal feature extraction. arXiv* [*cs.CV*]. Retrieved from http://arxiv.org/abs/1702.06151

Messaris, P. (1997). *Visual persuasion: The role of images in advertising*. Sage.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. MIT Press.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., ... & Oliva, A. (2019). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 502-508. https://doi.org/10.1109/TPAMI.2019.2901464

Nagrani, A., & Zisserman, A. (2018). *From Benedict Cumberbatch to Sherlock Holmes: Character identification in tv series without a script*. arXiv. https://arxiv.org/abs/1801.10442

National Television Violence Study. (1996). *National television violence study* (*Vol. 1*). Thousand Oaks, CA: Sage.

National Television Violence Study. (1997). *National television violence study* (*Vol. 2*). Studio City, CA: Sage.

Patron-Perez, A., Marszalek, M., Zisserman, A., & Reid, I. (2010). High five: Recognising human interactions in TV shows. *Proceedings of the British Machine Vision Conference*, 1(2), 1-11. https://doi:10.5244/C.24.50

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5), 920–941. https://doi.org/10.1093/joc/jqy041

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count* (*LIWC*)*: LIWC2001*. Mahway: Lawrence Erlbaum Associates.

Raney, A. A. (2008). Affective disposition theories. In *The International Encyclopedia of Communication*. Wiley. https://doi.org/10.1002/9781405186407.wbieca031.pub2

Riff, D., Lacy, S., & Fico, F. (2014). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Routledge.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90. https://doi.org/10.1145/3065386

Schmälzle, R., & Grall, C. (2020). The coupled brains of captivated audiences: An investigation of the collective brain dynamics of an audience watching a suspenseful film. *Journal of Media Psychology*, 32, 187-199. https://doi.org/10.1027/1864-1105/a000271

Schmälzle, R., Schupp, H. T., Barth, A., & Renner, B. (2011). Implicit and explicit processes in risk perception: neural antecedents of perceived HIV risk. *Frontiers in Human Neuroscience*, 5(43), 1-10. https://doi.org/10.3389/fnhum.2011.00043

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 6-13. https://doi.org/10.1177/0002716215572084

Schill, D. (2012). The visual image and the political image: A review of visual communication research in the field of political communication. *Review of Communication*, 12(2), 118-142. https://doi.org/10.1080/15358593.2011.653504

Taskiran, M., Kahraman, N., & Erdem, C. E. (2020). Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106, 1-28. https://doi.org/10.1016/j.dsp.2020.102809

Trilling, D., & Jonkman, J. G. F. (2018). Scaling up content analysis. *Communication Methods and Measures*, 12(2-3), 158–174. https://doi.org/10.1080/19312458.2018.1447655

Wagner, D. D., Dal Cin, S., Sargent, J. D., Kelley, W. M., & Heatherton, T. F. (2011). Spontaneous action representation in smokers when watching movie characters smoke. *Journal of Neuroscience*, 31(3), 894-898. https://doi.org/10.1523/JNEUROSCI.5174-10.2011

Weber, R., Eden, A., Huskey, R., Mangus, J. M., & Falk, E. (2015). Bridging media psychology and cognitive neuroscience. *Journal of Media Psychology*, 27, 146-156. https://doi.org/10.1027/1864-1105/a000163

Zamith, R., & Lewis, S. C. (2015). Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 307-318. https://doi.org/10.1177/0002716215570576

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6720-6731.

Zhu, J., Luo, J., You, Q., & Smith, J. R. (2013). Towards understanding the effectiveness of election related images in social media. *2013 IEEE 13th International Conference on Data Mining Workshops*, 421-425.

Zillmann, D., Taylor, K., & Lewis, K. (1998). News as nonfiction theater: How dispositions toward the public cast of characters affect reactions. *Journal of Broadcasting & Electronic Media*, 42(2), 153-169. https://doi.org/10.1080/08838159809364441

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature*, 559, 324-326. https://doi.org/10.1038/d41586-018-05707-8

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4), 561-577. https://doi.org/10.1093/clinchem/39.4.561