SoftwareX 20 (2022) 101236

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

Masked-Piper: Masking personal identities in visual recordings while preserving multimodal information

Babajide Owoyele [a], James Trujillo [b,c], Gerard de Melo [a], Wim Pouw [b,*]

[a] Hasso Plattner Institute, University of Potsdam, Germany
[b] Donders Center for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, Netherlands
[c] Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

A R T I C L E   I N F O

Article history:
Received 16 May 2022
Received in revised form 8 September 2022
Accepted 15 October 2022

Keywords:
Multimodal communication
Kinematic research
Data privacy
Open science
Masking
Research reproducibility

A B S T R A C T

In this increasingly data-rich world, visual recordings of human behavior are often unable to be shared due to concerns about privacy. Consequently, data sharing in fields such as behavioral science, multimodal communication, and human movement research is often limited. In addition, in legal and other non-scientific contexts, privacy-related concerns may preclude the sharing of video recordings and thus remove the rich multimodal context that humans recruit to communicate. Minimizing the risk of identity exposure while preserving critical behavioral information would maximize utility of public resources (e.g., research grants) and time invested in audio–visual research. Here we present an open-source computer vision tool that masks the identities of humans while maintaining rich information about communicative body movements. Furthermore, this masking tool can be easily applied to many videos, leveraging computational tools to augment the reproducibility and accessibility of behavioral research. The tool is designed for researchers and practitioners engaged in kinematic and affective research. Application areas include teaching/education, communication and human movement research, CCTV, and legal contexts.

Code metadata

Current code version	*v1*
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-22-00110
Permanent link to reproducible capsule	https://wimpouw.github.io/TowardsMultimodalOpenScience/Index
Legal code license	MIT License
Code versioning system used	git
Software code languages, tools and services used	python
Compilation requirements, operating environments and dependencies	Mediapipe Jupyter Notebooks
If available, link to developer documentation/manual	https://github.com/google/mediapipe
Support email for questions	wim.pouw@donders.ru.nl

1. Introduction

Recordings of human behavior are often unable to be shared due to concerns about protecting privacy [1]. As a result, there is often limited data sharing in areas such as behavioral science, multimodal communication, and human movement research. Nevertheless, data sharing is crucial in scientific contexts, as it allows for analyses on said data to be computationally reproducible [2,3]. For sensitive, quantitative data, such as medical records, new methods have been developed to share data without exposing personally identifiable information by creating synthetic data that preserve some statistical aspects of the original data [4]. For sensitive visual data of human behavior, there are currently no widely available comparable solutions [5].

Due to this lack of technical solutions, it is still common in research on multimodal communication and other kinematic research areas to forgo sharing the original video recordings

* Corresponding author at: Donders Center for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, Netherlands.
E-mail address: wim.pouw@donders.ru.nl (Wim Pouw).

https://doi.org/10.1016/j.softx.2022.101236

Fig. 1. Example of input frame (left) and masked frame (right) (video link: https://osf.io/4kj6a/).[1]
Source: Masterclass.com.

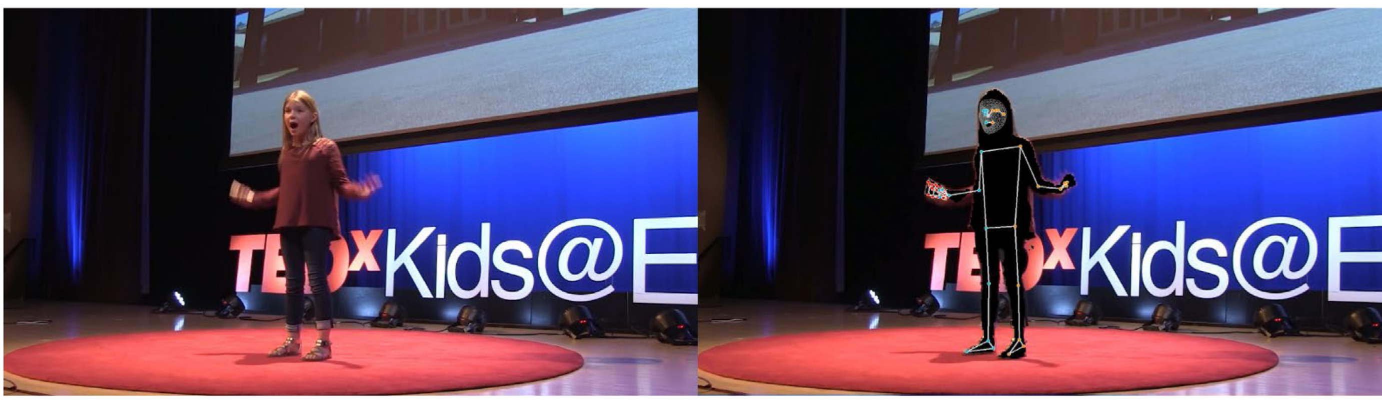

Fig. 2. Another example of input frame (left) and masked frame (right) for a non-adult speaker.[2]
Source: TedX.

(e.g., [6]) or to conditionally share them upon request as often required by journal policies [7,8]. Unfortunately, researchers seldom ratify sharing requests, as previous research shows [9]. One way of circumventing certain privacy concerns is to share only the quantified kinematic data. While this may allow other researchers to perform statistical analyses, it obscures the data source and makes it impossible to observe the original recordings. In contrast, the ability to inspect the original video and audio can help researchers understand how quantified results relate to the real-world context observable in the raw video and audio data. It also enables third parties to assess the quality of the recordings, the requirements for which may differ depending on the specific research questions, e.g., see Pouw and colleagues [10,11] as well as Rasenberg and colleagues [12] to determine the particular level of multimodal communication one needs to consider [12]. Furthermore, presenting the actual video recordings rather than just plots and graphs is important for optimally communicating research to peers, such as an academic conference or public lecture. It is, therefore, crucial to find an effective middle ground that allows data to be as anonymous as possible while maintaining pertinent visual information about the context in relation to dynamic human behavior.

Building on advancements in computer vision and deep learning techniques to track full-body kinematic information instantiated in MediaPipe [13], we present *Masked-Piper* as a tool to address the above anonymity-related challenges. Our tool makes use of MediaPipe's convenient light-weight CPU-based processing pipeline, which we leverage to:

- track hand, body, and facial kinematics, and store the quantitative information as a frame-by-frame time series,
- distinguish a human body from background information,
- mask the human body in the original video while retaining background information,
- project kinematics onto the masked video.

Figs. 1 and 2 provide two examples of input frames and the resulting outputs. Researchers can apply Masked-Piper (code here) on a large number of videos by placing the original folders into a processing folder. Masked-Piper will iteratively process all videos storing masked videos and kinematic time series. The kinematic time series files produced contain time stamp information (based on the frame rate of the original video) next to the available key points.

For the body pose information, 33 key points are available with 3D position coordinates and additional visibility variables, which helps judge the reliability of position estimates. For hand kinematics, 42 position key points are tracked in 3D; for facial kinematics, 3D position coordinates are provided for a face-mesh, containing 478 key points corresponding to designated areas of the face mesh (for more information see https://google.github.io/mediapipe/solutions/holistic.html). Masked-Piper utilizes MediaPipe's kinematics drawing module to project the kinematic information on top of the masked video. Information about the face mesh coordinates, body pose, and hand kinematics is maintained in the video.

1 https://www.masterclass.com/classes/neil-degrasse-tyson-teaches-scientific-thinking-and-communication

2 https://www.youtube.com/watch?v=OMbNoo4mCcI

In addition, MediaPipe's drawing module renders a full-body pose in a 2D space that aligns with the original video at each frame. Thus, Masked-Piper masks original visual information and reinstates key bodily information in the video in de-identified form.

The current tool has applications beyond the behavioral sciences. Firstly, it can solve the problem of collecting unnecessarily superfluous information about human behavior. Consider, for example, that many (audio)visual surveillance systems might not require recording a person's identity but still record such information. Using our masking tool, such systems might be employed to monitor certain activities (e.g., running) or levels of activities (e.g., amount of people), which do not require amassing identifiable information. The current tool could resolve this issue of superfluous information that inflates privacy risks by only maintaining *relevant* information about human behavior while mitigating privacy risks. We envision many other applications of the current tool, such as in legal contexts, where hearings with witnesses can be recorded to reduce the risk of identity exposure while maximizing the embodied communicative information inherent to human communication.

2. Implementation

The tool is currently implemented in Python, with all required code provided in the supplementary material. Users need to install the required modules, copy the files to be processed into a local directory, and run the provided notebook (or batch file). The open nature of the code provides additional transparency into how the tool works and allows customizability on the user's part. Nonetheless, the notebook is fully functioning in its current form and thus does not require much technical expertise.

As described above, Masked-Piper carries out several processing steps for each video provided. First, MediaPipe motion tracking is applied to the video, such that hand, body, and facial key points are located on each frame, providing time series for each of the key points, in x,y coordinates (given in pixels, local to the still frame). These are collected and provided as output in CSV format for further processing, sharing, etc. In parallel to collecting the key point-based tracking data, the *holistic* module from MediaPipe automatically detects a *silhouette* of the person in the video frame and extracts it from the background. This silhouette we subsquentely repurpose to serve as the mask.

To accomplish this, Masked-Piper draws this silhouette in black on top of the current video frame. The key point positions are then drawn onto the same silhouetted frame using the MediaPipe drawing module. Finally, the silhouetted (i.e., masked) marked frames are saved as a new video file using OpenCV. This new video file is the pseudo-anonymized output video, which preserves the holistic context of the video (e.g., background, human subject within the background) and provides more fine-grained information about the position of the limbs and fingers, facial expression, mouth movement, etc.

2.1. Choice of framework

The MediaPipe framework affords analysis of complex and dynamic bodily behaviors and is more easily installed than more heavy-duty GPU-based approaches to tracking human poses such as FrankMocap [14]. The value of using MediaPipe is (1) a good balance of resource consumption, (2) incremental and iterative processing, and (3) a broad library of supporting toolkits/libraries for developers and researchers to select from and customize [13].

2.2. Modification of underlying tool

Our modification of the MediaPipe tool consists of re-using the silhouette to determine background and body to create a mask. We then use MediaPipes body tracking to overlay the kinematics back on the mask. We further modify the original code so that time series data provide all kinematic information per frame over time. This tool is thus convenient for researchers to mask videos and extract kinematic time series for their research using a next-generation body tracker going beyond slower 2D tracking systems such as OpenPose [15].

3. Discussion

Any tool that risks exposing personally identifiable information needs to be used carefully. Several paths may lead to the exposure of someone's identity. Since we preserve information about communicative body movements and speech content, it is clear that any identifiable audio information remains unmasked for audio–visual recordings. Further, we can imagine cases where what is said or how one moves can still be sufficient to retrieve someone's identity [16] in cases where you know the person in question well, and know their potentially unique ways of communicating. The masking tool should therefore be seen as considerably *reducing* rather than completely abolishing risks of identity exposure. In behavioral science, identity exposure risks due to familiarity are not generally applicable though, as the researchers using the video recordings often have no connection with the study sample. Indeed, body movement data are not considered identifiable by any legal standard (e.g., GDPR guidelines). A limitation of the current tool is that only one body per frame can be detected. Thus, further iterations of the masking tool will need to be developed to mask multiple persons in one video. Finally, any automated computer vision-based tracking may be insufficiently precise[3] depending on your research questions (but see [7,17] for comparisons of video-based tracking versus device-based trackers). Fortunately, researchers can easily verify the quality of the videos and tracking performance produced by Masked-Piper.

The careful use of Masked-Piper has the potential to improve ethical research practices as well as maximize open science practices. We believe that these types of masking tools will become an important part of audio–visual research. Interesting parallel developments are currently underway that will further promote the use of masking tools, such as the Red Hen Anonymizer (https://sites.google.com/case.edu/techne-public-site/red-hen-anonymizer). In similar vein, the Deep Privacy Face Anonymization platform by Hukkelås and colleagues [18] masks people, thought it loses all information about facial expression. However, the Red Hen Anonymizer, for example, does leave a more human-realistic mask. These tools can together tailor different kind of research needs, and together they will help researchers to mask videos in large quantities, which they can then easily share with their peers. The current tool is especially suitable for its ease of use and maximal preservation of kinematic information (hands, body, face).

To conclude, we are convinced masking tools such as these will indirectly improve the core of the scientific process itself, because research reproducibility increases by allowing other

[3] https://google.github.io/mediapipe/solutions/holistic.html

researchers easy access to the original research context as contained in the video recordings.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

We wish to thank Linda Drijvers, Judith Holler, Asli Ozyurek and their affiliated research groups for their helpful comments on the masking tool. We are also grateful to the multimodal community on Twitter that encouraged us to explore such a tool and its value. We acknowledge continued support from Max Planck Institute, Donders Institute, the Hasso Plattner Design Thinking Research Program, particularly Jonathan Edelman and Joaquin Santuber. The collaborative nudges from Victor Omolaoye and colleagues at the HPI Chair for Artificial Intelligence and Intelligent Systems, and funding from the Hasso Plattner Foundation is also acknowledged.

Funding

This research has been co-funded by a VENI grant (VI.Veni.201G.047) awarded by the Dutch Research Council (NWO) to Wim Pouw (PI).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.softx.2022.101236.

References

[1] Narayanan A, Huey J, Felten EW. In: Gutwirth S, Leenes R, De Hert P, editors. Data Protection on the Move. Law, Governance and Technology Series. 24, Dordrecht: Springer; 2016, http://dx.doi.org/10.1007/978-94-017-7376-8_13.

[2] Buchanan EM, Crain SE, Cunningham AL, Johnson HR, Stash H, Papadatou-Pastou M, et al. Getting started creating data dictionaries: How to create a shareable data set. Adv Methods Pract Psychol Sci 2021;4(1). http://dx.doi.org/10.1177/2515245920928007.

[3] Gilmore RO, Kennedy JL, Adolph KE. Practical solutions for sharing data and materials from psychological research. Adv Methods Pract Psychol Sci 2018;1(1):121–30. http://dx.doi.org/10.1177/2515245917746500.

[4] Abay NC, Zhou Y, Kantarcioglu M, Thuraisingham B, Sweeney L. Privacy preserving synthetic data release using deep learning. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 11051 LNAI. 2019, p. 510–26. http://dx.doi.org/10.1007/978-3-030-10925-7_31.

[5] Joel S, Eastwick PW, Finkel EJ. Open sharing of data on close relationships and other sensitive social psychological topics: challenges, tools, and future directions. Adv Methods Pract Psychol Sci 2018;1(1):86–94. http://dx.doi.org/10.1177/2515245917744281.

[6] Gawne L, Krajcik C, Andreassen HN, Berez-Kroeker AL, Kelly BF. Data transparency and citation in the journal gesture. Gesture 2019;18(1):83–109. http://dx.doi.org/10.1075/GEST.00034.GAW/CITE/REFWORKS.

[7] Meyer MN. Practical tips for ethical data sharing. Adv Methods Pract Psychol Sci 2018;1(1):131–44. http://dx.doi.org/10.1177/2515245917747656.

[8] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3. http://dx.doi.org/10.1038/SDATA.2016.18.

[9] Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in plos journals. PLOS ONE 2009;4(9):e7078. http://dx.doi.org/10.1371/JOURNAL.PONE.0007078.

[10] Pouw W, Trujillo JP, Dixon JA. The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. Behav Res Methods 2020;52(2):723–40. http://dx.doi.org/10.3758/S13428-019-01271-9/FIGURES/4.

[11] Pouw W, Dingemanse M, Motamedi Y, Özyürek A. A systematic investigation of gesture kinematics in evolving manual languages in the lab. Cogn Sci 2021;45(7):e13014. http://dx.doi.org/10.1111/COGS.13014.

[12] Rasenberg M, Özyürek A, Dingemanse M. Alignment in multimodal interaction: An integrative framework. Cogn Sci 2020;44(11). http://dx.doi.org/10.1111/COGS.12911.

[13] Lugaresi C, Tang J, Nash H, Mcclanahan C, Uboweja E, Hays M, et al. MediaPipe: A framework for building perception pipelines. 2019, http://dx.doi.org/10.48550/arxiv.1906.08172.

[14] Rong Y, Shiratori T, Joo H. FrankMocap: Fast monocular 3D hand and body motion capture by regression and integration. 2020, http://dx.doi.org/10.48550/arxiv.2008.08324.

[15] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings - 30th IEEE conference on computer vision and pattern recognition, CVPR 2017, 2017-January. 2017, p. 1302–10. http://dx.doi.org/10.1109/CVPR.2017.143.

[16] Runeson S, Frykholm G. Kinematic specification of dynamics as an informational basis for person-and-action perception: expectation, gender recognition, and deceptive intention. Journal of experimental psychology: general 1983;112(4):585.

[17] Kosourikhina Veronika, Diarmuid Kavanagh, Michael J. Richardson, David M. Kaplan. Validation of deep learning-based markerless 3D pose estimation. Plos one 2022;17:10. http://dx.doi.org/10.1371/journal.pone.0276258.

[18] Hukkelås H, Mester R, Lindseth F. DeepPrivacy: A generative adversarial network for face anonymization. 2019, CoRR abs/1909.04538, http://arxiv.org/abs/1909.04538.