Exploring Collection of Sign Language Datasets: Privacy, Participation, and Model Performance

Danielle Bragg
Microsoft Research
Cambridge, MA, USA

Oscar Koller
Microsoft
Munich, Germany

Naomi Caselli
Boston University
Boston, MA, USA

William Thies
Microsoft Research
Bangalore, India

ABSTRACT

As machine learning algorithms continue to improve, collecting training data becomes increasingly valuable. At the same time, increased focus on data collection may introduce compounding privacy concerns. Accessibility projects in particular may put vulnerable populations at risk, as disability status is sensitive, and collecting data from small populations limits anonymity. To help address privacy concerns while maintaining algorithmic performance on machine learning tasks, we propose privacy-enhancing distortions of training datasets. We explore this idea through the lens of sign language video collection, which is crucial for advancing sign language recognition and translation. We present a web study exploring signers' concerns in contributing to video corpora and their attitudes about using filters, and a computer vision experiment exploring sign language recognition performance with filtered data. Our results suggest that privacy concerns may exist in contributing to sign language corpora, that filters (especially expressive avatars and blurred faces) may impact willingness to participate, and that training on more filtered data may boost recognition accuracy in some cases.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility technologies**; *User studies*; • **Security and privacy** → **Human and societal aspects of security and privacy**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

data collection, privacy, machine learning, sign language

ACM Reference Format:
Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. Exploring Collection of Sign Language Datasets: Privacy, Participation, and Model Performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20), October 26–28, 2020, Virtual Event, Greece.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3373625.3417024

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ASSETS '20, October 26–28, 2020, Virtual Event, Greece
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7103-2/20/10…$15.00
https://doi.org/10.1145/3373625.3417024

1 INTRODUCTION

While powerful machine learning algorithms require large amounts of data, many application domains are still data-scarce. In particular, collecting sufficient data from small, underserved populations to build systems serving those groups is difficult, because the pool of potential contributors is greatly reduced. Furthermore, *human data* is typically required to train machine learning systems built to serve or assist humans, which introduces privacy concerns. While marginalized communities can greatly benefit from systems tailored to their needs, they can also be put at higher risk by contributing data to build those systems. Small group size makes personal identification easier, and marginalized status makes privacy breaches more dangerous. These privacy concerns may further inhibit data contributions from an already small pool. While privacy concerns are certainly not the only limitation in creating large corpora, and may not be the primary limitation in many domains, this work focuses on this particular barrier.

To help address such data scarcity problems *by addressing privacy concerns*, we present Privacy-Enhancing Data Filters (demonstrated in Figure 1). These filters obscure the identity of the contributor, for example by blurring videos of people signing, or scrambling pixels in the frame. As personal identification becomes more difficult, people's privacy concerns may be lessened, and people may become more willing to contribute to datasets, resulting in larger datasets. Filters can be used in a variety of domains, and a variety of filters can be designed for any type of data.

While filters may be useful for lessening privacy concerns, they may also be detrimental to models trained on the data. However, the increase in data quantity when privacy concerns are assuaged could in some cases more than compensate for weakened quality. In particular, in data-scarce domains, the increase in data may lead to *increased performance* for the resulting trained models. This technique can be particularly effective in building public datasets, which are powerful – they increase scalability, attract more diverse contributors, support broader research efforts, and help democratize data ownership – but compound privacy concerns. It could also be particularly useful for systems trained on data from small, vulnerable populations, where both data scarcity and privacy are problematic.

The problem of collecting data from small groups is exemplified in sign language data collection. Deaf signers form marginalized communities, comprising about 1% of the global population (totaling 70 million) [43]. Sign language recognition and translation software could enable many powerful developments for this community

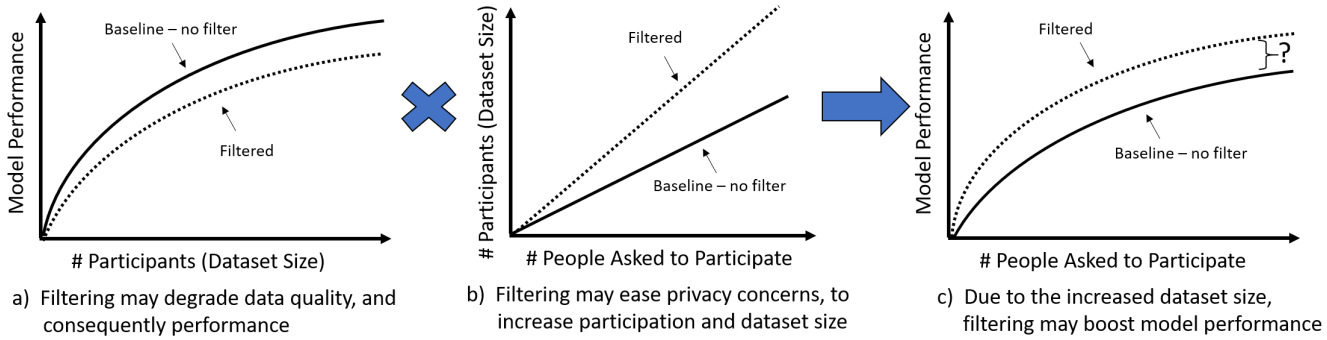


Figure 1: We envision that Privacy-Enhancing Data Filters may ease people's privacy concerns, thereby increasing dataset participation and boosting model performance due to larger training set size. Different filters may allow for varied levels of privacy, participation (dataset size), and model performance. The set of figures presents an abstract idea that can be applied to many domains, so the axes are not tied to specific metrics. For example, Model Performance could represent Word Error Rate for sign language recognition.

(e.g., digital assistants that respond to people signing; automatic translation, dictation, and transcription; and many educational applications). However, their development requires videos of people signing, which may introduce privacy concerns – video reveals personal identity, the Deaf community has a history of oppression which can make personal identification more dangerous, and small population size makes identification more likely.

This work explores privacy concern remediation as a tool to enable machine learning applications in data-scarce regimes, through the lens of sign language data collection, in two main studies. To shed light on the end-user experience of using filters, we conduct a web study that explores signers' privacy concerns, allows participants to experience some basic filters, and elicits recommendations for how videos might be modified in the future to increase dataset participation. To explore the potential impact of filtering on model performance, we also conduct an offline experiment on sign language recognition with varied amounts of filtered and unfiltered data. Our results suggest that privacy concerns may exist in contributing to sign language datasets, that filters may impact participation, and that increased filtered data may boost model performance in certain cases.

The main contributions of this work are:

- The idea of using Privacy-Enhancing Data Filters to address privacy concerns and thereby collect more data, which may ultimately improve machine learning performance. This idea may be particularly powerful for small, vulnerable populations where both privacy and data scarcity are problematic. We are the first to propose this idea, which introduces opportunities for exploring filters in many other domains (e.g., in building machine learning solutions for other disability communities, oppressed ethnic minorities, or victims of domestic abuse).
- An initial exploration of this idea, within the context of collecting sign language videos to train recognition and translation systems. We present two studies, one user study focused on signers' experience of contributing data and using filters, and an algorithmic experiment exploring filter impact on machine learning performance. We are the first to explore privacy issues related to contributing sign language videos to corpora, which introduces opportunities to further explore the large filter design space, as well as contributors' experiences and machine learning performance.

2 BACKGROUND AND RELATED WORK

To frame our exploration of Privacy-Enhancing Data Filters through the lens of sign language data, we provide background on Deaf culture and sign language and summarize related work on sign language recognition, privacy concerns with videos, and techniques for enhancing privacy in video and machine learning. To this foundation, our work contributes a demonstration of the potential of filters to increase data contributions, and for increased filtered data to improve model performance. We also provide an initial exploration of signers' privacy concerns.

2.1 Deaf Culture and Sign Language

Sign languages (e.g., American Sign Language i.e. ASL) are minority languages used primarily by Deaf people who often identify as members of a deaf cultural minority.[1] Many communication barriers result from using non-majority languages and modalities (e.g., exclusion from increasingly pervasive voice-controlled agents). Even the written form of the majority language can be inaccessible because of failures of Deaf Education systems to properly support acquisition of a first language [52, 53, 57] and subsequent low literacy rates for deaf children [58, 81]. Despite new technologies enabling Deaf people from across the world to connect more easily [87, 100], communities are still small, and Deaf people often know one another. Privacy concerns around video sharing may be increased, as people are more easily identified by others in the community.

"Audism" is the marginalization of Deaf people based on audiological status [6, 40, 62]. Like other forms of discrimination, audism

[1]We use uppercase "Deaf" to refer to members of a cultural minority group, and lowercase "deaf" to refer to audiological (hearing) status.

manifests in many different ways (e.g., state-mandated sterilization [79], doctors or companies rejecting Deaf patients or employees). Audism can also be embedded in technologies (e.g., technologies that inadvertently exclude Deaf people, such as voice assistants and airport PA announcements). Deaf people are often excluded from decisions that will profoundly affect them. A famous example is the Deaf President Now protest in which students demanded Deaf leadership at the only university designed for Deaf people, which had historically been run almost exclusively by hearing people [30]. The existence of audism can compound privacy concerns with sharing videos, by making personal identification and disclosure of Deafness more risky.

Like spoken words, signs are made up of phonological features, originally thought to comprise handshape, location, and movement [106]. Current theories provide a more precise, comprehensive set of features [18, 96, 116]. Grammatical information can be produced non-manually (not with the hands) via eye gaze, eyebrow or mouth movement, or head/body posture (see [19, 119] for reviews). Multiple features occur simultaneously (e.g., the handshape and its location), unlike in speech where a single sound typically occurs at a time. The language complexity increases the quantity of data needed for accurate modeling, in an already data-scarce environment.

2.2 Sign Language Recognition Algorithms

Sign language recognition started out with glove-based approaches, dating back to 1983 [50]. The patent describes an electronic glove that recognized ASL fingerspelling based on a hardwired circuit. Since then, a lot of related work was built on "intrusive sign recognition", where users are required to use or wear intrusive gear [25, 42, 76, 84]. The first non-intrusive vision-based recognition system, presented in 1988, used skin color thresholding to recognize 14 isolated signs of Japanese sign language [109]. To address three-dimensionality, some vision-based approaches use depth cameras [113, 123], multiple cameras [16] or triangulation for 3D reconstruction [98, 99]. Some rely on colored gloves to facilitate tracking [31].

Continuous sign language recognition (CSLR) is a necessary extension to the works mentioned so far, which focused on isolated signs. CSLR deals with naturally produced language where different dialects, contextual references (e.g., shifting left and right to indicate who is talking in a story), and the effect of each sign's ending on the next's beginning make it a significantly more challenging, yet also more realistic, problem. The first work on CSLR used handcrafted features in traditional hidden markov models (HMMs) [105] to distinguish 40 signs. Many similar works followed, often inspired by improvements in speech recognition [44, 117].

Advances in deep learning and convolutional neural networks (CNNs) for image processing have reshaped the field. Embedded CNNs in HMMs [71, 72], long short term memory (LSTM) cells [70], 3D-CNNs [22, 60, 114] and a combination of 2D-CNNs with temporal 1D convolutions and optical flow [32] are the most promising directions. However, deep learning requires large amounts of data, and sign language video corpora are typically small. Privacy-Enhancing Data Filters offer a way to address such data scarcities by boosting people's willingness to contribute to datasets. We also show that with sufficient filtered data, recognition algorithms can outperform models trained on unfiltered data.

2.3 Privacy Concerns with Video

Privacy (and security) have been framed as social and cultural phenomena (e.g., [34]), Defined by society, privacy involves legal structures, as well as technical and social systems. Understanding what society deems acceptable can be difficult, especially as technology changes quickly, bringing with it a change of societal privacy expectations [110]. Understanding individuals' conceptualization of privacy can be difficult as well [61]. In this work, we deal with privacy in terms of the possible revelation of personal identity through video.

Privacy concerns with image and video data span many domains: video surveillance, which is increasingly used by businesses, police departments, and governments (e.g., [36]); robot assistants that may be used in sensitive environments like the home (e.g., [20]); health monitoring and assisted living systems, especially for the growing aging population (e.g., [21, 85]); biometrics, which reveal highly personal information and are difficult to recover if compromised (e.g., [90]); map applications containing images or video, (e.g., Google Street View [46]); social media and digital photography sharing (e.g., [56]), and cloud computing more generally (e.g., [125]). We are not aware of work on privacy concerns for people contributing to machine learning datasets, which we provide.

Privacy-enhancing techniques for visual data typically have a detrimental effect on the person viewing the data, for example a redacted video or image (e.g., [14, 51, 56, 75]). This past work highlights a privacy-utility trade-off, where techniques that enhance privacy often reduce the utility of the images or video. We are not aware of any work showing that addressing privacy concerns can be a tool to collect larger datasets, thereby boosting performance of machine learning applications. Our work demonstrates this possibility.

2.4 Privacy-Enhancing Techniques for Video

Privacy-enhancing techniques for video can be categorized in three main groups: methods that 1) prevent capture of sensitive information, 2) support computation on sensitive information without revealing the sensitive parts, and 3) obfuscate the video. (See [86] for a more complete review.)

Some methods preserve privacy by preventing sensitive data from being captured and stored in the first place. For example, systems have been developed to detect cameras and render them useless by directing bright light at them [55, 88, 112]. Such systems create *environments* where people can operate with reduced concern about undesirable data being collected.

Obfuscation methods target either select regions or entire frames. Targeting human faces or bodies may require face or body detection and tracking, often achieved by leveraging skin color [103], gait [118], or intentional cues (e.g., privacy-minded people wearing distinct clothes [97]). Once detected, people (or other content) are obscured by a variety of techniques, the most common listed below. Many can also be applied to entire frames (i.e., blur and pixelation), as can full-frame techniques (e.g., turning an image into a line-drawing or distorting colors).

- blur or pixelation (e.g., [2, 14, 63])
- masking or face swap - a person's face is modified or replaced with another face (e.g., [33, 111, 122])

- silhouette or skeleton - only the person's silhouette or skeleton is tracked (e.g., [85])
- avatar - the image of a real person is replaced by a cartoon-like character (e.g., [92])
- invisibility - a person or object is entirely removed, and the background is filled in (e.g., [89])

Methods that preserve privacy during computation provide privacy in a variety of ways. Cryptographic methods protect sensitive content from unwanted viewers, while allowing access for desired ones [12, 23, 36]. These methods typically blur sensitive parts of images or video using a cryptographic key, which can be used to unblur them. Other methods support sharing obfuscated data, or parameters learned by training on the data, without sharing the raw data itself [41]. Yet other methods support computation on humans without identifying the people, for example to estimate crowd size [24]. Many of these methods overlap with those in Section 2.5.

Algorithmic resilience to video/image distortion has been studied, but not for privacy-enhancing purposes. For example, the prevalence of low-quality cameras, compressed feeds, or out-of-focus picture motivates work on the effect of image quality on facial recognition or tracking [59, 74]. Protecting people from identity theft and biometric spoofing attacks similarly motivates work on detecting image tampering or face spoofing attempts [29, 80]. To the best of our knowledge, nobody has studied algorithmic resilience to distortion, in order to better support privacy enhancements and data collection, as we do.

2.5 Privacy-Enhancing Machine Learning

Work on preserving privacy in machine learning more generally falls into preventing unwanted exposure from three main sources: 1) training data, 2) the learned model, and 3) the model's outputs. Our work falls into the first category, preserving privacy of training data. However, we show that by doing so, we can actually increase dataset size and thereby potentially boost performance as well as ease privacy concerns.

Homomorphic encryption refers to encryption schemes that support computation on encrypted data without ever decrypting it [47]. A variety of machine learning models can be built on encrypted data: simple predictive analysis [10], decision trees, hyperplane decision, and Naive Bayes [11] low-degree polynomial classification [49], and neural networks [48]. Other encryption techniques allow target users to decrypt data, for use in machine learning applications [77, 124] (e.g., including examples in the previous subsection).

Differential privacy techniques have been applied to machine learning datasets and algorithms, to provide privacy guarantees. Differential privacy (roughly) is a guarantee that an adversary has negligible probability of identifying an individual datapoint based on data aggregates [37], and there are limits to such guarantees [8]. This framework has been applied to various classes of machine learning algorithms including linear and logistic regression [26], SVMs [93], PCA [27, 39], boosting [38], and deep learning [1, 101], sometimes coupled with network protocols [82].

Multi-party computation (MPC) is a domain with direct privacy implications. In MPC, multiple entities contribute to a single computation – for example, training a machine learning model on the aggregate of separate datasets owned by the entities. Techniques exist to aggregate high-dimensional data from multiple sources without revealing individual contributions [9], and to learn many types of models including decision trees [3, 78], linear regression [35] Naive Bayes classifiers [115], and k-means clustering [64]. Our work does not address MPC in particular, but rather provides a means for collecting data that contributors are comfortable sharing.

3 STUDY 1: END-USER EXPERIENCE OF FILTERS

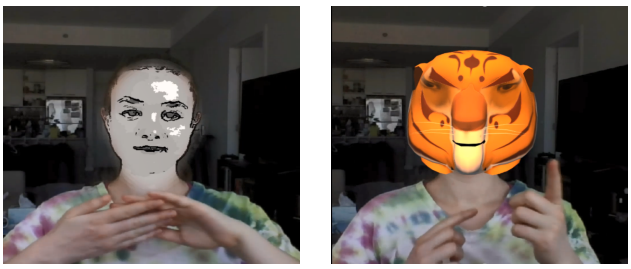
To explore the end-user experience of filters through the lens of sign language data collection, we ran a study with sign language users. The study was designed to explore 1) privacy concerns with contributing videos to sign language datasets, 2) whether filters might impact people's willingness to contribute, and 3) what video modification or filters users might want.

3.1 Procedure

The study was run as a web study, with IRB approval. We recruited American Sign Language (ASL) users to participate through relevant email lists and social media. The study took about ten minutes, and ran for two weeks. All questions were multiple-choice (some single-selection, some allowing multiple selections), or free response. All questions were available both in English and ASL video. The ASL translations were produced by a deaf native signer, with accuracy verified by two native signers (one deaf, one hearing).

The study consisted of three main parts (after consent):

(1) **Demographics:** We asked basic demographic questions spanning age, gender, audiological status, and ASL level.
(2) **Filter experience:** We allowed participants to experience three different filters while signing requested content, and asked if they would be willing to contribute the resulting videos to datasets with different owners, to advance sign language recognition and translation. We asked about willingness to contribute, as this procedure (consent) aligns with IRB ethics standards for any data-collection effort. The process went as follows (three times):
   (a) The participant viewed a canvas displaying their webcam stream. The feed was either unmodified, or modified with one of two filters (described below). The participant was told to execute the sign HELLO to the camera.
   (b) The participant was asked "Your video was NOT recorded, but imagine that it WAS, exactly as you just saw it. Would you give permission to use your recording for making apps respond to ASL to.."
   – "a Deaf advocacy group at [a company]? The video would only be accessible within the company."
   – "the Deaf Studies group at [a university]? The video would only be accessible within the university"
   – "the general public? The video would be available on a public website'
   Participants selected from Yes (definitely or probably) and No (definitely or probably). We chose specific recipients, because willingness to contribute data to different entities can differ greatly.



(a) Frame cel shading (b) Tiger face

Figure 2: Stillframes of the two filters that user study participants experienced, which track the face in real-time. The filters were implemented starting with open-source code from Jeeliz (see https://jeeliz.com/).

(3) **Concerns and requested solutions:** We asked several questions about privacy concerns with contributing sign language videos of themselves, which solutions might affect their willingness to contribute, and open feedback.

To help ensure that participant's relative privacy concerns are reflected in their responses, we do not use the words "privacy" or "security" prior to or while asking about their willingness to contribute videos with different filters. (according to privacy guidelines [17]).

3.2 Filters

For this initial exploration, we sought a small, diverse, mainstream set of filters that were technically integratable in our web study. To meet these criteria, we chose two filters from Jeeliz's jeelizFaceFilter library [65], an open-source Javascript and WebGL-based library for real-time face detection and modification. The library provides a set of augmented reality experiences through webcams. The chosen filters cover the two main methods of obfuscation in the literature (blur and masking), and the two main regions of interest (face and full frame). We also compared against the baseline of the original unmodified video.

- **Frame cel shading**[2] - the full frame is replaced with a flattened greyscale version. Jeeliz's filter applies this transformation to the face only, and we extended it to cover the entire frame.
- **Tiger face**[3] - the face is detected, and replaced with a tiger avatar head, which emits blue bubbles when the mouth opens. This filter mimics smartphone Animoji filters, which have gained attention within the Deaf community (e.g., [121]).

The filters were experienced in order of increasing novelty, to help ensure that ratings of simple solutions are not artificially lowered by the perceived availability of more complex solutions.

3.3 Results

Our study results suggest that privacy concerns may exist in contributing sign language videos to corpora (especially concern of misuse), and that filters and data owners may impact people's willingness to contribute videos. They also shed light on the types of filters that sign language dataset participants may want (especially expressive avatars and face blur). We recognize that these results are preliminary, and follow-up studies are needed.

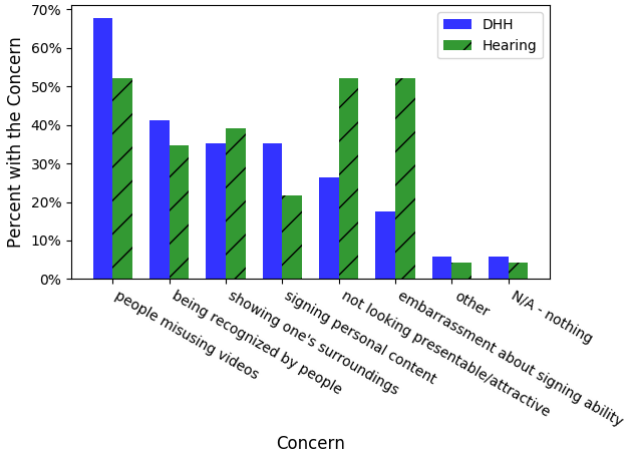


Figure 3: Privacy concerns with contributing sign language videos to help develop sign language recognition and translation, for DHH and hearing participants. Prompt: "What are your concerns with sharing videos of yourself signing, if any? (Check all that apply.)"

3.3.1 Participants. We had 61 participants. Two participants left the site before completing the final questions about privacy concerns, but completed the demographics and filter experience portions. Basic demographics for the group were: **gender:** 40 male, 18 female, 3 other; **age:** 19-75, mean 33.52, SD 12.71; **audiological status:** 23 deaf, 11 hard of hearing, 25 hearing, 2 other; **ASL level:** on a scale of 1 (fluent) to 7 (does not use ASL), range 1-6 (all ASL users), mean 3.36, SD 1.63; **age when started learning ASL:** range 0-51, mean 13.07, SD 10.58

3.3.2 Privacy Concerns. To better understand privacy concerns that people might have with contributing videos to sign language corpora, we asked participants about any concerns they may have (see Figure 3).

The most common concern for both deaf and hard-of-hearing (DHH) and hearing participants was video misuse (61% overall, 68% DHH, 52% hearing). This concern is particularly relevant to public datasets that anyone can access, and correspondingly a lower percent of participants were willing to contribute to these compared to private datasets (discussed below). The filters we explored decrease the desirability of videos for misuses like creating online memes or fake social media profiles, and correspondingly increased participant willingness to contribute publicly (also discussed below). The next-most-common concerns for DHH participants were about revealing sensitive information – being recognized by people (39% overall, 41% DHH, 35% hearing), showing one's surroundings (36% overall, 35% DHH, 39% hearing), and signing personal content (29% overall, 35% DHH, 22% hearing). Filters may similarly help address

[2]based on https://jeeliz.com/demos/faceFilter/demos/threejs/celFace/
[3]taken from https://jeeliz.com/demos/faceFilter/demos/threejs/tiger/

Table 1: Ordinal logistic regression predicting willingness to contribute. Abbreviations:Est. estimate, SE standard error. Significance codes: * < .001, ** < .01, * < .05**

Coefficient	Est.	SE	t	p	
filter:frame shading	0.288	0.365	0.790	0.430	
filter:tiger face	-0.732	0.320	-2.289	0.022	*
entity:company	1.573	0.317	4.957	<.001	***
entity:university	1.537	0.319	4.826	<.001	***

Table 2: Percent of participants willing to contribute videos of themselves to help develop sign language recognition and translation, with different recipients (columns) and different filters applied (rows).

	Company	University	Public
Baseline	90.16%	88.52%	36.07%
Frame cel shading	86.89%	86.89%	55.74%
Tiger face	63.93%	63.93%	45.90%

these concerns by obscuring or disguising the person and/or video background.

Concerns differed substantially between DHH and hearing participants. Hearing participants were much more concerned with embarrassment about their signing abilities, and not looking presentable/attractive. The difference in embarrassment may be due to differences in fluency between groups – average fluency rating of 1.82 DHH vs. 3.36 hearing on a 7-point scale (with lower meaning more fluent). Furthermore, sharing sign language videos publicly on social media is already fairly common within the Deaf community, suggesting that many DHH people may have already overcome concerns about contributing videos to the public domain, while hearing people who primarily use written text on such platforms have not.

Overall, the vast majority of participants reported some privacy concerns. Only 7% of participants (6% DHH, 4% hearing) reported having no privacy concerns. Though asking people about concerns may result in over-reporting, our results still suggest the existence of privacy concerns, which may be a barrier to collecting scalable sign language video datasets from this already small community.

3.3.3 Filter Experience. To shed light on participants' experience of our demoed filters, we analyzed participant responses about willingness to contribute data with the filters to different data owners. First, we examine which variables (filter and/or entity) may have impacted participants' reported willingness to contribute, through a factorial analysis via logistic ordinal regression (see Table 1). Second, to gain some intuition about these interactions, we tallied the percent of participants willing to contribute data with each filter to each entity (see Table 2, with strong and weak responses are grouped together for interpretability.).

Our regression model reveals a significant impact of both filter and data owner (entity), with the entity being more significant. Overall, relative to the baseline, the frame shading filter is expected to increase willingness to contribute, while the tiger face filter is expected to decrease willingness (as seen in the sign of the coefficients). Similarly, compared to contributing publicly, contributing to the company or university is expected to increase willingness to contribute.

Our tallied table shows that the most participants reported willingness to contribute publicly *with filters*, and to the private datasets *without filters*. Participant responses suggest that this result for the public domain may relate to the filters (which resemble participants' top-requested modifications) addressing their top concerns (which primarily pertain to public corpora – i.e., misuse and being recognized). Their feedback also suggests that the high willingness to contribute privately without filters may relate to counterbalancing concerns about the utility of filtered videos for research and development. For example, one person explained, "For training purposes... I suggest a mask/distortion feature that leaves a part of the face visible."

The tallied table also shows that across data owners, a higher percent of participants reported willingness to contribute with frame cel shading than with tiger face. Participant feedback suggests that this difference may stem from concerns about data quality (described above) pertaining more to tiger face than frame cel shading. In particular, participants expressed concern that the tiger face would not sufficiently capture facial expressions, which are grammatically meaningful, for example commenting, "The problem with putting a tiger over a person's face is that the face is used for contextualization of the sign."

3.3.4 Requested Filters. To explore more generally how videos could be modified to address people's concerns and boost contributions, we asked participants for their input (see Figure 4).

Most participants (68%) reported that videos of themselves signing could be modified in some way to make them more willing to contribute. This suggests that adding various filter options to data collection mechanisms could increase contributions. Compared to deaf and hard-of-hearing participants, hearing participants more frequently reported that no possible changes would increase their willingness to contribute (43% hearing vs. 21% DHH). Hearing people may be less likely to contribute videos no matter the accommodations because they are generally less invested in the end result (i.e., sign language recognition and translation). Hearing people who sign may also be less comfortable creating and sharing videos of themselves signing because they do so less frequently than their DHH peers.

Replacing the contributor with a cartoon character was the most popular solution for DHH participants, followed by blurring the face, which was the most popular solution for hearing participants. Participants' feedback made it clear that such cartoon characters or avatars should capture facial expressions, which are semantically meaningful in ASL (and other sign languages). For instance, participants specifically requested new avatar filters that "allow for facial expression changes in the avatar."

3.3.5 General Feedback. In the open feedback, participants expressed support for the idea of collecting data from sign language users to build a digital assistant for signers, but also re-iterated privacy concerns. Many expressed support for the end-goal of empowering sign language technologies, one stating that they'd had "enough of hearing access – where is sign access!" and another, "it's obvious the technology will revolutionize access for the Deaf". One participant also commented on the diversity of signers that this

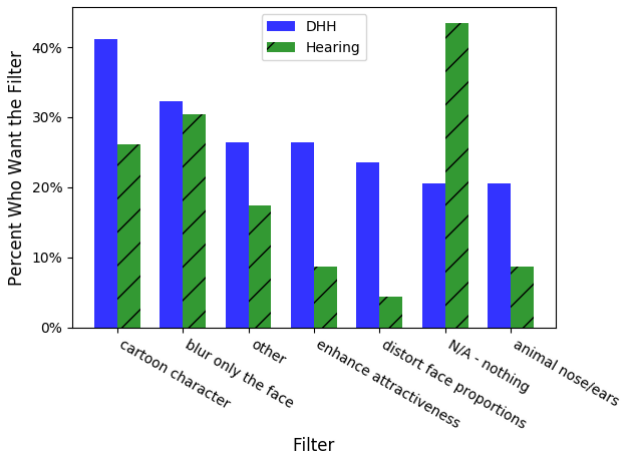


Figure 4: Filters requested by our participants. Question prompt: "Are there any other ways your video could be changed to make you more willing to share for ASL research? (Check all that apply.)"

type of data collection would elicit, that they "like the idea of showing a diversity of faces that this study seems to suggest". While the feedback on the project motivation was entirely positive, several participants also re-iterated concerns about people misusing their videos, especially if released in a public dataset.

4 STUDY 2: MACHINE LEARNING WITH FILTERED DATA

To explore the impact that filters may have on machine learning performance in the context of sign language, we ran a set of experiments using state-of-the-art computer vision for sign language recognition. To probe the effects of filtering and potential related changes in dataset size, we compared training on filtered and unfiltered videos, and varied training set size.

4.1 Dataset

For our experiments, we chose the RWTH-PHOENIX-Weather 2014 [45] dataset, a computer vision benchmark used in many papers to compare progress in the field. The dataset comes partitioned into train, dev, and test sets. Train was used for training, dev for parameter tuning, and test was held out for testing (as standard in computer vision on such datasets).

The dataset contains German Sign Language interpretations of weather forecasts from public television. We used the signer independent set [70], where one signer is reserved for the test set. It includes nine signers – eight hearing interpreters in the train and dev partitions, and one Coda (child of Deaf adults) in the test set exclusively. Table 3 provides the dataset statistics. Each signed sentence is aligned with a gloss transcript (a written representation which retains sign order and grammar), created and reviewed by multiple Deaf transcribers. Content is unscripted, which results in fast movements, co-articulation (where neighboring signs affect execution), and misarticulated signs. Due to recording quality, motion blur is present in many videos, and resolution is limited to 210 x 260 px. These factors result in a realistic, challenging recognition problem.

Table 3: PHOENIX 2014 Signer Independent SI5 dataset stats.

	Train	Dev	Test
Signers	8	1	1
Duration [hours]	6.80	0.18	0.30
Frames	612,027	16,460	26,891
Sentences	4,376	111	180
Running glosses	49,966	1,167	1,901
Vocabulary	1,080	239	294

Though this dataset is state-of-the-art, it still has limitations which impact our experiments. In particular, it is relatively small for this type of task, and does not allow for cross-validation due to highly imbalanced signer proportions (see the distribution plot in [69]). However, these types of limitations characterise the field; the lack of large, labeled, continuous datasets is a primary barrier to progress in sign language recognition and translation [15].

4.2 Filters

For this initial exploration, we sought to conduct a systematic comparison with a basic filter type. To meet these criteria, we focus exclusively on blur, which is a primary way to enhance video/image privacy (see related work), and can be applied to various regions of the frame (vs. Animoji filters, which only apply to the face). This choice allowed us to explore the two main ways that videos are redacted to enhance privacy in related work (targeted and entire-frame changes) as more of the frame is blurred, unconfounded by different types of filtering. We did *not* attempt to use optimal filters, but rather to explore reasonable baseline filters in this initial work.

Specifically, we compared the unmodified PHOENIX dataset to two filtered variants (see Figure 5) – cel shading of the face, and of the full frame. As in the web study, these filters are variants of Jeeliz's [65] Face cel shading filter.[4]

- **face cel shading** - the face of the signer is replaced with a flattened greyscale version. The original implementation was modified to fit our offline use-case by removing face tracking (unreliable offline), and fixing the 3D modeling window in the frame over the signer's face.
- **frame cel shading** - the full frame is replaced with a flattened greyscale version. This filter was implemented by removing the face tracking functionality, fixing the 3D modeling window in the center of the frame, and stretching it to span the full frame.

4.3 Language Recognition Framework and Implementation

Sign language recognition is a sequence learning task, which means we want to predict a sequence of output symbols w_1^N. The symbols are typically sign glosses, written words representing signs. Given an input video as a sequence of images $X_1^T = X_1, \ldots, X_T$ and

[4] https://jeeliz.com/demos/faceFilter/demos/threejs/celFace/

(a) baseline (b) face cel shading (c) frame cel shading

Figure 5: A stillframe of a signer from the Phoenix dataset (see [45]), with the filters we compared: (a) the unchanged baseline, (b) cel shading applied to the face, and (c) cel shading applied to the full frame. The filters were implemented starting with open-source code from Jeeliz (see https://jeeliz.com/).

the resulting mean-normalized images $x_1^T = x_1, \ldots, x_T$, automatic continuous sign language recognition tries to find an unknown sequence of glosses w_1^N for which x_1^T best fits the learned models. We assume the video (image sequence) and gloss transcription share the same sign ordering and grammar. This clearly distinguishes sign language recognition from translation, which requires re-ordering.

To solve this problem, we use a state-of-the-art continuous sign language recognition algorithm [70, 72]. The framework models sign language by embedding a deep convolutional neural network (CNN) followed by a recurrent long short term memory (LSTM) in a hidden Markov model (HMM). It treats the outputs of the neural network as true Bayesian posteriors and trains the system as a hybrid CNN-LSTM-HMM in an end-to-end fashion. The iterative training approach that includes frequent re-alignments has been shown to be very successful on different sign language datasets [67, 70]. Our implementation ran on a modified GoogleNet CNN architecture [108], with a depth of 22 layers. To boost performance, we pre-trained on the 1.4M images of the Imagenet dataset [94] and performed scrambled re-alignment [68].

To compare the unmodified PHOENIX dataset to our two filtered versions with datasets of various sizes, we performed independent training runs for each of the three input video conditions, and three training set sizes (the full dataset, 50% and 75% of the data). The smaller training sets were chosen at random, by randomly selecting signed sentences. Entire sentences were chosen for inclusion, because they were the unit of alignment with the glossed transcript. For consistency in comparing results, the 50% set was chosen to be a subset of the 75% training set. We test on videos from the same filter condition. A single training instance took about 6 days.

For the full framework, algorithmic derivations and explanations, and implementation details, see Appendix A.

4.4 Metrics

We follow the standard metrics of the PHOENIX dataset: word error rate (WER). This error measure is suitable for comparing sequences of different length. It is a common metric used to evaluate automatic speech and sign recognition. The metric is computed on aligned reference and hypothesis sequences (in our case, glossed sentences). It divides the minimal edit distance (summing substitutions, deletions and insertions) by the total reference token count (in our case, the number of signs in the correct gloss transcription), as follows:

$$\text{WER} = \frac{\#\text{deletions} + \#\text{insertions} + \#\text{substitutions}}{\#\text{symbols in reference}} \quad (1)$$

4.5 Results

Our results suggest that while filters may degrade model performance for a fixed training set size, models trained on larger filtered datasets may outperform those trained on smaller unfiltered ones in some cases. Figure 6 shows recognition accuracy for the unchanged baseline and our two filters, with varied training set size. (See Appendix A for full results from training and test phases.)

For reference, state-of-the-art continuous sign language recognition trained on the employed benchmark dataset, achieves ~ 40% WER [15]. As we rely only on RGB input and do not use temporal convolutions, our WER is slightly worse. Nevertheless, low state-of-the-art performance highlights the difficulty of the problem, underscored by a lack of sufficient training data.

4.5.1 Impact of Training Set Size. Overall, the performance of the baseline and two filters improved as training dataset size increased, indicating that we are operating in a data-scarce domain. The performance improvements do not plateau, signifying that the model is not yet saturated with data. As the dataset we used is one of the largest continuous sign language corpora, this means that sign language recognition and translation is a data-scare environment where we expect performance to improve further with more data. Privacy-enhancing filters aim to help address this problem by expanding the pool of willing contributors.

Given sufficient data, both the face cel shading and frame cel shading filters outperformed the baseline. WER is lowered by moving from the 25k baseline dataset to the 37.5k frame cel shading dataset (and beyond) or the 50k face cel shading dataset. Similarly, WER is lowered by moving from the 37.5k baseline dataset to the 50k face cel shading dataset. While it is natural to ask how much larger the training dataset must be to compensate for filtered data, it is not possible to answer this question in absolute terms. Performance depends on a set of factors – baseline dataset size, baseline

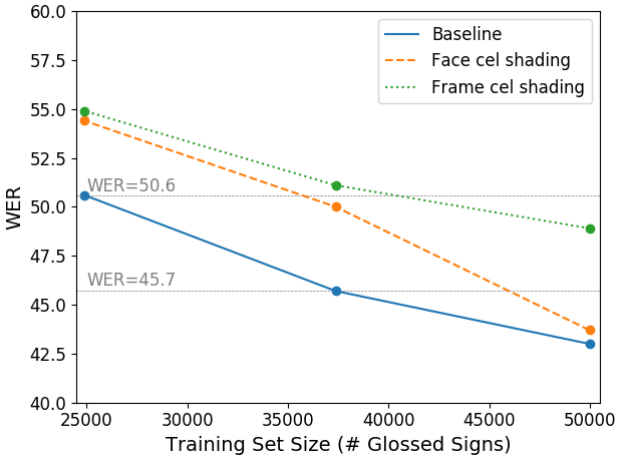


Figure 6: Recognition performance on the PHOENIX SI05 test dataset, with the filters and training set sizes we compared. Lower WER is better. Models trained on larger filtered datasets may outperform the baseline trained on less data in some cases (e.g., baseline 25k vs. frame cel shading 37.5k and 50k).

data quality, the filter design, and the complexity of the task and model.

4.5.2 Impact of Filter. For fixed training set sizes, the baseline outperformed the two filters, as demonstrated by its lower values in Figure 6. This lower error rate is to be expected, as the filters we explored degrade video quality by removing linguistically meaningful information. The resulting models are limited in power because they do not have access to this information. In contrast, the baseline model has access to the full original data, and can leverage more information such as color.

Comparing the two filters, face cel shading generally outperformed frame cel shading (as seen in face cel shading's lower WER across training set sizes). For the largest training set size, face cel shading performed comparably to the baseline (43.8% baseline, 44.1% face WER), while frame cel shading has a higher error rate (48.0% WER). This difference is likely due to the fact that face cel shading preserves details outside of the face, whereas frame cel shading does not. When trained on a small amount of data, a complex model like ours might learn meaningless patterns in the external details while placing less weight on meaningful patterns in the low-resolution face, but given enough data it learns stronger signals.

5 DISCUSSION

While participants in our first study expressed a higher willingness to contribute filtered data exclusively to public datasets, such datasets can be particularly powerful. Public datasets support broad scientific research and public advancement, compared to privately owned datasets that benefit individual organizations. As a result, people may be motivated to contribute to public datasets for free, reducing monetary constraints that limit dataset size. Voluntary contributions also result in more diversity [91], which is essential for training systems to recognize diverse signers (not currently possible). Organizations also monetize datasets, which can be particularly offensive to the Deaf community if they are not the primary beneficiaries. Despite democratization, increasing the control Deaf people have over these datasets may be vital, as a fully democratic system may leave Deaf people far outnumbered despite having the most at stake.

While this work explored filters that degrade model performance due to information loss, it may be possible to design filters that do not degrade model performance. For example, a deepfake filter that swaps one face for another (e.g., a dataset-wide volunteer) while preserving facial expressions would provide privacy without impacting data quality. Such deepfakes can be so convincing that they are generally undetectable by humans and computers [73]. Other lossless filters such as shuffling video pixels would not be expected to degrade model performance. To the human eye, the filtered data would be unrecognizable, but to a computer the data quality would be identical. Furthermore, it may even be possible to design filters that *increase* model accuracy by removing irrelevant content that models may latch onto in data-scarce environments (e.g., filters that remove the video background or superimpose uniform clothing on all contributors).

There are many use cases for filtering sign language videos, besides encouraging corpora participation. Because sign languages are not typically written, signed content is often shared via video. The lack of anonymity in video makes many interactions impossible, e.g. posting content anonymously, collaboratively authoring content, sharing ideas in the abstract (vs. through a particular body, with social connotations of gender, race, etc.), and submitting work for anonymized review. It is also often impossible to edit signed content to create a cohesive piece, especially if the signer differs across clips. Filters can enable people to share signed content anonymously, and the homogeneity of filters may newly enable editing and piecing together video clips into a single cohesive result. This is all in contrast to written languages, which already support anonymous (or near-anonymous) communication, sharing ideas in the abstract, and easy editing and collaboration.

While this work explores privacy-preserving filters to help address privacy concerns in dataset participants, it is possible that other privacy enhancements may be beneficial. Video misuse was the most common privacy concern among both DHH and hearing participants in our web study (e.g., using videos to create memes or fake profiles, distorting or photo-shopping them, or viewing them for amusement). While the filters we explored might reduce misuse by obscuring people's identities, they do not target misuse. Other possible solutions that target misuse include watermarking dataset videos, or creating strict usage terms coupled with methods for checking for violations. We also note that ethical data collection encompasses more than privacy (e.g. consent), and that privacy concerns are not the only limiting factor in collection. To understand how to best support potential contributors, further research on concerns and possible solutions is needed.

We note that sign language recognition and translation software is controversial in some Deaf communities. As addressed in this work, collecting training data (videos) may introduce privacy concerns. There may also be pressure to use translation software in

lieu of more expensive and effective accommodations (e.g., sign language interpreters in complex medical situations). These concerns may be heightened by the community's history of audism, but also may be reduced if Deaf people take leadership in developing these technologies, which have the potential to profoundly affect their lives [54, 102]. They may also be addressed by simply improving the accuracy of recognition and translation, which requires addressing data scarcity problems, as our filters attempt to do.

6 LIMITATIONS AND FUTURE WORK

We want to clarify the limitations of this work. First, the application of filters to human faces or other parts of videos is not novel (see Section 2 for prior examples). However, the application to sign language datasets, and the idea that such filters might increase dataset participation, are new. Similarly, we do not mean to suggest that privacy is the main or only barrier to curating corpora of sign language or other types of data. Many other types of barriers may exist, for example lack of funding or data-collection infrastructure, difficulty labeling, and lack of deaf community support for resulting applications. Finally, the idea that filters may increase the volume of datasets and thus boost machine learning model performance requires further work to understand the domains and communities in which this may or may not hold.

Nonetheless, our vision of privacy-enhancing filters enabling machine learning applications in data-scarce regimes introduces many opportunities for future exploration. First, there are many other domains besides sign language datasets where this idea could be explored. These include: medical data (where filters might obscure personally identifying information such as name and date-of-birth); audio recorded by digital personal assistants (where filters might anonymize the user's voice or distort other sounds); and video from surveillance cameras (where filters may anonymize people's faces, bodies, or sensitive data like credit card numbers).

As the first project to explore sign language filters, we introduce a large design space for exploration and optimization. While a small, basic set of filters was appropriate for an initial exploration, a wide range of filters are possible, including methods that lose *no* information (e.g., scrambling pixels), and those that are lossy (e.g., artistic styling, various blurs). In particular, avatars that capture facial expressions would be interesting to examine, as our web study participants requested. New filters also technically challenging to create and integrate into applications, often involving real-time face or body tracking and complex graphics operations. This line of future work also includes developing privacy guarantees, for example proving that the original video cannot be recovered with certain filters.

Relatedly, it would be valuable to study the effect of a wider spectrum of filters on recognition accuracy. Because the filter's effect is not independent of the model and data complexity, it would also be informative to study a variety of recognition models and datasets. Because our privacy study revealed a variety of filter preferences, it is possible that contributors need a variety of filters from which to choose, in order to maximize data contributions. The resulting dataset would be highly heterogeneous, requiring the development of training methods that efficiently leverage both filtered and unfiltered data, and an understanding of the effect of training (and testing) on such mixed datasets. Of course, as the field of sign language recognition and translation evolves, so too should the techniques for incorporating filtered data.

In making initial steps into this space, the two exploratory studies we present have several limitations. In our first study, we note that people's responses about willingness to contribute may not equate to real-world actions (as in any study). However, running a study allowed us to collect qualitative feedback, and to explore the space prior to deploying a large data-collection initiative. In addition, we can expect relative reported willingness to reflect relative real-world willingness to contribute [17]. Our computer vision experiments were also limited due to the small size of existing sign language datasets (a problem which also motivated our work). Future work includes running larger experiments on both the end-user experience and machine learning capabilities, including deploying real-world data-collection initiatives.

7 CONCLUSION

In this work, we present the idea of *Privacy-Enhancing Data Filters*, data modifications designed to increase contributions towards machine learning corpora by addressing participants' privacy concerns. We do *not* claim to have designed or evaluated optimal filters, but have hopefully demonstrated the possibility that privacy enhancements may in some cases encourage dataset participation and subsequent model improvements.

We explored the idea of privacy-enhancing filters through the lens of sign language data in two studies. To investigate the end-user experience of contributing to sign language corpora and using privacy-enhancing filters to do so, we ran a web study that asked participants about their concerns, allowed them to experience filters, and elicited filter requests. Our results suggest that privacy concerns may be pervasive in the community, and that filters may impact willingness to participate. They also shed light on the types of filters the community may want (in particular, expressive avatars). To explore how filtering may impact machine learning model performance, we also ran a set of computer vision experiments comparing state-of-the-art sign language recognition accuracy on filtered and unfiltered videos with varied data quantity. Our results suggest that a higher quantity of filtered data may improve recognition accuracy in some cases for data-scarce environments.

To the best of our knowledge, our work is novel in several ways: 1) we provide a vision that addressing privacy concerns may help overcome data scarcity problems in building machine learning systems for underserved, vulnerable minority populations; 2) we provide the first exploration of the sign language community's privacy concerns with contributing videos of themselves signing; and 3) we provide the first exploration of the effect of privacy-enhancing video filters on sign language recognition. Similar work might benefit other small, vulnerable populations for whom it is difficult to build powerful machine learning solutions due to data scarcity (e.g., building tools for other disability communities or language models for oppressed ethnic minorities).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications*

Security. ACM, 308–318.

[2] Prachi Agrawal and PJ Narayanan. 2011. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 3 (2011), 299–310.

[3] Rakesh Agrawal and Ramakrishnan Srikant. 2000. Privacy-preserving data mining. In *ACM Sigmod Record*, Vol. 29. ACM, 439–450.

[4] L. R. Bahl, F. Jelinek, and R. L. Mercer. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5, 2 (March 1983), 179–190. https://doi.org/10.1109/TPAMI.1983.4767370

[5] Raimo Bakis. 1976. Continuous Speech Recognition via Centisecond Acoustic States. *The Journal of the Acoustical Society of America* 59, S1 (1976), S97–S97.

[6] H-Dirksen L Bauman. 2004. Audism: Exploring the metaphysics of oppression. *Journal of deaf studies and deaf education* 9, 2 (2004), 239–246.

[7] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Transactions on Neural Networks* 5, 2 (1994), 157–166.

[8] Avrim Blum, Katrina Ligett, and Aaron Roth. 2013. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)* 60, 2 (2013), 12.

[9] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1175–1191.

[10] Joppe W Bos, Kristin Lauter, and Michael Naehrig. 2014. Private predictive analysis on encrypted medical data. *Journal of biomedical informatics* 50 (2014), 234–243.

[11] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. 2015. Machine learning classification over encrypted data.. In *NDSS*, Vol. 4324. 4325.

[12] Terrance Edward Boult. 2005. PICO: Privacy through invertible cryptographic obscuration. In *Computer Vision for Interactive and Intelligent Environment (CVIIE'05)*. IEEE, 27–38.

[13] Herve A. Bourlard and Nelson Morgan. 1993. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, Norwell, MA.

[14] Michael Boyle, Christopher Edwards, and Saul Greenberg. 2000. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 1–10.

[15] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudrealt, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. *arXiv preprint arXiv:1908.08597* (2019).

[16] Helene Brashear, Thad Starner, Paul Lukowicz, and Holger Junker. 2003. Using Multiple Sensors for Mobile Sign Language Recognition. In *7th IEEE International Symposium on Wearable Computers*. IEEE, White Plains, NY, USA, 45–52.

[17] Alex Braunstein, Laura Granka, and Jessica Staddon. 2011. Indirect content privacy surveys: measuring privacy without asking about it. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*. ACM, 15.

[18] Diane Brentari. 1996. Trilled movement: Phonetic realization and formal representation. *Lingua* 98, 1-3 (1996), 43–71.

[19] Diane Brentari. 2018. Representing Handshapes in Sign Languages Using Morphological Templates1. *Gebärdensprachen: Struktur, Erwerb, Verwendung* 13 (2018), 145.

[20] Daniel J Butler, Justin Huang, Franziska Roesner, and Maya Cakmak. 2015. The privacy-utility tradeoff for remotely teleoperated robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 27–34.

[21] Kelly E Caine, Arthur D Fisk, and Wendy A Rogers. 2006. Benefits and privacy concerns of a home equipped with a visual sensing system: A perspective from older adults. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications Sage CA: Los Angeles, CA, 180–184.

[22] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2016. Using Convolutional 3D Neural Networks for User-Independent Continuous Gesture Recognition. In *Proc. Int. Conf. on Pattern Recognition Workshops (ICPRW)*. Cancun, Mexico, 49–54.

[23] Paula Carrillo, Hari Kalva, and Spyros Magliveras. 2010. Compression independent reversible encryption for privacy in video surveillance. *EURASIP Journal on Information Security* 2009, 1 (2010), 429581.

[24] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–7.

[25] C. Charayaphan and A. E. Marble. 1992. Image Processing System for Interpreting Motion in American Sign Language. *Journal of Biomedical Engineering* 14, 5 (Sept. 1992), 419–425. https://doi.org/10.1016/0141-5425(92)90088-3

[26] Kamalika Chaudhuri and Claire Monteleoni. 2009. Privacy-preserving logistic regression. In *Advances in neural information processing systems*. 289–296.

[27] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. 2013. A near-optimal algorithm for differentially-private principal components. *The Journal of Machine Learning Research* 14, 1 (2013), 2905–2943.

[28] S. Chen and J. Goodman. 1998. *An Empirical Study of Smoothing Techniques for Language Modeling*. Technical Report TR-10-98. Computer Science Group, Harvard University. 1–63 pages.

[29] Ivana Chingovska, André Anjos, and Sébastien Marcel. 2012. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*. IEEE, 1–7.

[30] John B Christiansen and Sharon N Barnartt. 2003. *Deaf president now!: The 1988 revolution at Gallaudet University*. Gallaudet University Press.

[31] Helen Cooper and Richard Bowden. 2010. Sign Language Recognition Using Linguistically Derived Sub-Units. In *LREC Workshop on the Representation and Processing of Sign Languages*. Valletta, Malta, 57–61.

[32] R. Cui, H. Liu, and C. Zhang. 2019. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia* 0 (2019), 1–1. https://doi.org/10.1109/TMM.2018.2889563

[33] Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video face replacement. In *ACM Transactions on Graphics (TOG)*, Vol. 30. ACM, 130.

[34] Paul Dourish and Ken Anderson. 2006. Collective information practice: Exploring privacy and security as social and cultural phenomena. *Human-computer interaction* 21, 3 (2006), 319–342.

[35] Wenliang Du, Yunghsiang S Han, and Shigang Chen. 2004. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, 222–233.

[36] Frederic Dufaux and Touradj Ebrahimi. 2006. Scrambling for video surveillance with privacy. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 160–160.

[37] Cynthia Dwork. 2011. Differential privacy. *Encyclopedia of Cryptography and Security* (2011), 338–340.

[38] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. 2010. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 51–60.

[39] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. 2014. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM, 11–20.

[40] Richard Clark Eckert and Amy June Rowley. 2013. Audism: A theory and practice of audiocentric privilege. *Humanity & Society* 37, 2 (2013), 101–130.

[41] Jianping Fan, Hangzai Luo, Mohand-Said Hacid, and Elisa Bertino. 2005. A novel approach for privacy-preserving video sharing. In *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 609–616.

[42] S. S. Fels and G. E. Hinton. 1993. Glove-Talk: A Neural Network Interface between a Data-Glove and a Speech Synthesizer. *IEEE Transactions on Neural Networks* 4, 1 (Jan. 1993), 2–8. https://doi.org/10.1109/72.182690

[43] World Federation fo the Deaf. 2018. *Our Work*. http://wfdeaf.org/our-work/

[44] Jens Forster, Christian Oberdörfer, Oscar Koller, and Hermann Ney. 2013. Modality Combination Techniques for Continuous Sign Language Recognition. In *Iberian Conference on Pattern Recognition and Image Analysis (Lecture Notes in Computer Science 7887)*. Springer, Madeira, Portugal, 89–99.

[45] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*. Reykjavik, Island, 1911–1916.

[46] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. 2009. Large-scale privacy protection in google street view. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2373–2380.

[47] Craig Gentry et al. 2009. Fully homomorphic encryption using ideal lattices.. In *Stoc*, Vol. 9. 169–178.

[48] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*. 201–210.

[49] Thore Graepel, Kristin Lauter, and Michael Naehrig. 2012. ML confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*. Springer, 1–21.

[50] Gary J. Grimes. 1983. Digital Data Entry Glove Interface Device. US Patent.

[51] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. 2005. Integrating utility into face de-identification. In *International Workshop on Privacy Enhancing Technologies*. Springer, 227–242.

[52] Wyatte C. Hall. 2017. What You Don't Know Can Hurt You: The Risk of Language Deprivation by Impairing Sign Language Development in Deaf Children. *Maternal and child health journal* 21, 5 (May 2017), 961–965. https://doi.org/10.1007/s10995-017-2287-y

[53] Wyatte C Hall, Leonard L Levin, and Melissa L Anderson. 2017. Language deprivation syndrome: a possible neurodevelopmental disorder with sociocultural origins. *Social psychiatry and psychiatric epidemiology* 52, 6 (2017), 761–776.

[54] Raychelle Harris, Heidi M Holmes, and Donna M Mertens. 2009. Research ethics in sign language communities. *Sign Language Studies* 9, 2 (2009), 104–131.

[55] A Harvey. 2010. Camoflash-anti-paparazzi clutch. *URL http://ahprojects. com/projects/camoflash/Accessed* 8 (2010), 2014.

[56] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. 2018. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 47.

[57] Jon Henner, Rama Novogrodsky, Jeanne Reis, and Robert Hoffmeister. 2018. Recent issues in the use of signed language assessments for diagnosis of language disorders in signing deaf and hard of hearing children. *The Journal of Deaf Studies and Deaf Education* 23, 4 (2018), 307–316.

[58] Judith A Holt. 1993. Stanford Achievement Test—8th edition: Reading comprehension subgroup results. *American Annals of the Deaf* 138, 2 (1993), 172–175.

[59] Fang Hua, Peter Johnson, Nadezhda Sazonova, Paulo Lopez-Meyer, and Stephanie Schuckers. 2012. Impact of out-of-focus blur on face recognition performance based on modular transfer function. In *2012 5th IAPR International Conference on Biometrics (ICB)*. IEEE, 85–90.

[60] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2015. Sign Language Recognition Using 3D Convolutional Neural Networks. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2015.7177428

[61] Lee Humphreys. 2011. Who's watching whom? A study of interactive technology and surveillance. *Journal of Communication* 61, 4 (2011), 575–595.

[62] Tom Humphries. 1975. Audism: The making of a word. *Unpublished essay* (1975).

[63] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. 2015. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 781–792.

[64] Geetha Jagannathan and Rebecca N Wright. 2005. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 593–599.

[65] Jeeliz. [n.d.]. *jeelizFaceFilter*. https://github.com/jeeliz/jeelizFaceFilter

[66] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *CoRR* abs/1408.5093 (2014).

[67] Hamid Reza Vaezi Joze and Oscar Koller. 2018. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. *arXiv:1812.01053 [cs]* (Dec. 2018). arXiv:1812.01053 [cs]

[68] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. 2019. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* accepted for publication (2019), 15.

[69] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding (CVIU)* 141 (Dec. 2015), 108–125. https://doi.org/10.1016/j.cviu.2015.09.013

[70] Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. Re-Sign: Re-Aligned End-To-End Sequence Modelling With Deep Recurrent CNN-HMMs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, 4297–4305.

[71] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2016. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proc. British Machine Vision Conference (BMVC)*. York, UK, 1–12. https://doi.org/10.5244/C.30.136

[72] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision (IJCV)* 126, 12 (Dec. 2018), 1311–1325. https://doi.org/10.1007/s11263-018-1121-3

[73] Pavel Korshunov and Sébastien Marcel. 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685* (2018).

[74] Pavel Korshunov and Wei Tsang Ooi. 2011. Video quality for face detection, recognition, and tracking. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 7, 3 (2011), 14.

[75] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 67.

[76] Rung-Huei Liang and Ming Ouhyoung. 1998. A Real-Time Continuous Gesture Recognition System for Sign Language. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG)*. Nara, Japan, 558–567.

[77] Chen-Yi Lin. 2016. A reversible data transform algorithm using integer transform for privacy-preserving data mining. *Journal of Systems and Software* 117 (2016), 104–112.

[78] Yehuda Lindell and Benny Pinkas. 2000. Privacy preserving data mining. In *Annual International Cryptology Conference*. Springer, 36–54.

[79] Paul A Lombardo. 2008. *Three generations, no imbeciles: Eugenics, the Supreme Court, and Buck v. Bell*. JHU Press.

[80] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. 2012. Face spoofing detection from single images using texture and local shape analysis. *IET biometrics* 1, 1 (2012), 3–10.

[81] Rachel I Mayberry, Alex A Del Giudice, and Amy M Lieberman. 2011. Reading achievement in relation to phonological coding and awareness in deaf readers: A meta-analysis. *The Journal of Deaf Studies and Deaf Education* 16, 2 (2011), 164–188.

[82] Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 19–38.

[83] Hermann Ney and Stefan Ortmanns. 2000. Progress in Dynamic Programming Search for LVCSR. *Proc. IEEE* 88, 8 (2000), 1224–1240.

[84] Cemil Oz and Ming C. Leu. 2011. American Sign Language Word Recognition with a Sensory Glove Using Artificial Neural Networks. *Engineering Applications of Artificial Intelligence* 24, 7 (Oct. 2011), 1204–1213. https://doi.org/10.1016/j.engappai.2011.06.015

[85] José Padilla-López, Alexandros Chaaraoui, Feng Gu, and Francisco Flórez-Revuelta. 2015. Visual privacy by context: proposal and evaluation of a level-based visualisation scheme. *Sensors* 15, 6 (2015), 12959–12982.

[86] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. 2015. Visual privacy protection methods: A survey. *Expert Systems with Applications* 42, 9 (2015), 4177–4195.

[87] Jeffrey Levi Palmer, Wanette Reynolds, and Rebecca Minor. 2012. "You Want What on Your Pizza!?" Videophone and Video-Relay Service as Potential Influences on the Lexical Standardization of American Sign Language. *Sign Language Studies* 12, 3 (2012), 371–397.

[88] Shwetak N Patel, Jay W Summet, and Khai N Truong. 2009. Blindspot: Creating capture-resistant spaces. In *Protecting Privacy in Video Surveillance*. Springer, 185–201.

[89] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2536–2544.

[90] Salil Prabhakar, Sharath Pankanti, and Anil K Jain. 2003. Biometric recognition: Security and privacy concerns. *IEEE security & privacy* 2 (2003), 33–42.

[91] Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 1364–1378.

[92] Chi-Hyoung Rhee and C LEE. 2013. Cartoon-like avatar generation using facial component matching. *Int. J. of Multimedia and Ubiquitous Engineering* 8, 4 (2013), 69–78.

[93] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. 2009. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *arXiv preprint arXiv:0911.5708* (2009).

[94] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (Dec. 2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[95] David Rybach, Stefan Hahn, Patrick Lehnen, David Nolden, Martin Sundermeyer, Zoltán Tüske, Simon Wiesler, Ralf Schlüter, and Hermann Ney. 2011. RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Waikoloa, HI, USA.

[96] Wendy Sandler. 1989. *Phonological representation of the sign: Linearity and nonlinearity in American Sign Language*. Vol. 32. Walter de Gruyter.

[97] Jeremy Schiff, Marci Meingast, Deirdre K Mulligan, Shankar Sastry, and Ken Goldberg. 2009. Respectful cameras: Detecting visual markers in real-time to address privacy concerns. In *Protecting Privacy in Video Surveillance*. Springer, 65–89.

[98] Christoph Schmidt, Oscar Koller, Hermann Ney, Thomas Hoyoux, and Justus Piater. 2013. Enhancing Gloss-Based Corpora with Facial Features Using Active Appearance Models. In *International Symposium on Sign Language Translation and Avatar Technology*, Vol. 2. Chicago, IL, USA.

[99] Christoph Schmidt, Oscar Koller, Hermann Ney, Thomas Hoyoux, and Justus Piater. 2013. Using Viseme Recognition to Improve a Sign Language Translation System. In *International Workshop on Spoken Language Translation*. Heidelberg, Germany, 197–203.

[100] Sherry Shaw and Len Roberson. 2013. Social connectedness of Deaf retirees. *Educational Gerontology* 39, 10 (2013), 750–760.

[101] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. ACM, 1310–1321.

[102] J Singleton, Amber Martin, and Gary Morgan. 2015. Ethics, deaf-friendly research, and good practice when studying sign languages. *Research methods in sign language studies: A practical guide* (2015), 7–20.

[103] Franc Solina, Peter Peer, Borut Batagelj, Samo Juvan, and Jure Kovač. 2003. Color-based face detection in the" 15 seconds of fame" art installation. (2003).
[104] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15 (2014), 1929–1958.
[105] T. Starner and A. Pentland. 1995. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. In *International Symposium on Computer Vision*. Coral Gables, Florida, USA, 265–270.
[106] William C Stokoe. 1960. Sign language structure (Studies in Linguistics. *Occasional paper* 8 (1960).
[107] A. Stolcke. 2002. SRILM- an Extensible Language Modeling Toolkit. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*. Denver, Colorado, 901–904.
[108] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Boston, Ma, USA, 1–9.
[109] Shinichi Tamura and Shingo Kawasaki. 1988. Recognition of Sign Language Motion Images. *Pattern Recognition* 21, 4 (1988), 343–353. https://doi.org/10.1016/0031-3203(88)90048-9
[110] Omer Tene and Jules Polonetsky. 2013. A theory of creepy: technology, privacy and shifting social norms. *Yale JL & Tech.* 16 (2013), 59.
[111] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2387–2395.
[112] Khai N Truong, Shwetak N Patel, Jay W Summet, and Gregory D Abowd. 2005. Preventing camera recording by designing a capture-resistant environment. In *International Conference on Ubiquitous Computing*. Springer, 73–86.
[113] D. Uebersax, J. Gall, M. Van den Bergh, and L. Van Gool. 2011. Real-Time Sign Language Letter and Word Recognition from Depth Data. In *Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*. 383–390. https://doi.org/10.1109/ICCVW.2011.6130267
[114] Hamid Vaezi Joze and Oscar Koller. 2019. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. In *Proc. British Machine Vision Conference (BMVC)*. Cardiff, UK.
[115] Jaideep Vaidya, Murat Kantarcıoğlu, and Chris Clifton. 2008. Privacy-preserving naive bayes classification. *The VLDB Journal* 17, 4 (2008), 879–898.
[116] Els Van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. Netherlands Graduate School of Linguistics.
[117] U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K. F Kraiss. 2008. Recent Developments in Visual Sign Language Recognition. *Universal Access in the Information Society* 6, 4 (2008), 323–362.
[118] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. 2003. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence* 25, 12 (2003), 1505–1518.
[119] Ronnie B Wilbur. 2000. Phonological and prosodic layering of nonmanuals in American Sign Language. *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima* (2000), 215–244.
[120] Ronald J. Williams and David Zipser. 1995. Gradient-Based Learning Algorithms for Recurrent Networks and Their Computational Complexity. *Back-propagation: Theory, architectures and applications* 1 (1995), 433–486.
[121] YouTube. 2018. *#SIGNanimojichallenge [Interview With Bilal Chinoy]*. https://www.youtube.com/watch?v=0av-wKimT9Q
[122] Xiaoyi Yu, Kenta Chinomi, Takashi Koshimizu, Naoko Nitta, Yoshimichi Ito, and Noboru Babaguchi. 2008. Privacy protecting visual processing for secure video surveillance. In *2008 15th IEEE International Conference on Image Processing*. IEEE, 1672–1675.
[123] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. 2011. American Sign Language Recognition with the Kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI '11)*. ACM, Alicante, Spain, 279–286. https://doi.org/10.1145/2070481.2070532
[124] Jiantao Zhou, Weiwei Sun, Li Dong, Xianming Liu, Oscar C Au, and Yuan Yan Tang. 2016. Secure reversible image data hiding over encrypted domain via key modulation. *IEEE transactions on circuits and systems for video technology* 26, 3 (2016), 441–452.
[125] Minqi Zhou, Rong Zhang, Wei Xie, Weining Qian, and Aoying Zhou. 2010. Security and privacy in cloud computing: A survey. In *2010 Sixth International Conference on Semantics, Knowledge and Grids*. IEEE, 105–112.

A CONTINUOUS SIGN LANGUAGE RECOGNITION EXPERIMENT DETAILS

This appendix provides additional details on our sign language recognition experiment, including those of interest for computer vision readers, and those needed for replication.

A.1 Hybrid CNN-LSTM-HMM Framework for Sign Language Modeling

We base our experiments on a state-of-the-art continuous sign language recognition implementation as published in [68, 70, 72]. The framework models sign language by embedding a deep convolutional neural network (CNN) followed by a recurrent long short term memory (LSTM) in a hidden Markov model (HMM) while treating the outputs of the neural network as true Bayesian posteriors and training the system as a hybrid CNN-LSTM-HMM in an end-to-end fashion. The iterative training approach that includes frequent re-alignments has been shown to be very successful on different sign language data sets.

Sign language recognition is a sequence learning task, which means we want to predict a sequence of output symbols w_1^N. The symbols are typically sign glosses, representing the semantics of the described sign. Given an input video as a sequence of full images $X_1^T = X_1, \ldots, X_T$ and the resulting mean-normalized images $x_1^T = x_1, \ldots, x_T$, automatic continuous sign language recognition tries to find an unknown sequence of glosses w_1^N for which x_1^T best fit the learned models.

To find the best fitting sequence, we follow the statistical paradigm [4] using the maximum-a-posteriori simplification of Bayes' decision rule. We maximize the class posterior probability distribution $Pr(w_1^N|x_1^T)$ over the whole utterance, which we can split up into separate probabilities. Those can then be modelled by different information sources: a language model, a visual model and a transition model.

To replace the generative with modern and powerful discriminative CNNs, we follow the hybrid approach known from automatic speech recognition [13] to convert the posterior probability of the CNN-LSTM to scaled likelihoods. We define the sub-gloss label $\alpha := s, w_1^N$ which represents the hidden state s that belong to the gloss sequence w_1^N. We apply Bayesian inference, converting the posteriors to class-conditional likelihoods following Bayes' rule, where the prior probability $p(\alpha)$ can be approximated by the relative state label frequencies in the frame-state-alignment used to train the CNN-LSTM.

After applying the viterbi approximation, which considers only the most likely alignment path and adding several hyper-parameters to the implementation we get Equation (2). This is what we optimize to find the best output sequence. The hyperparameters allow us to control the effect of the language model (γ) and the neural network label prior (β).

$$\left[w_1^N\right]_{\text{opt}} = \underset{w}{\operatorname{argmax}} \Big\{ \underbrace{p\left(w_1^N\right)^{\gamma}}_{\text{lang. model}} \cdot \max_{s_1^T} \Big\{ \prod_{t=1}^{T} \times \underbrace{\frac{p(\alpha|x_t)}{p(\alpha)^{\beta}}}_{\text{visual model}} \cdot \underbrace{p\left(s_t|s_{t-1}, w_1^N\right)}_{\text{transition model}} \Big\} \Big\}$$

A.2 Implementation Details

In this work, we use the 22 layer deep GoogleNet [108] CNN architecture, which we initially pre-train on the 1.4M images from the

Table 4: Recognition results in WER [%] (the lower the better) on PHOENIX 2014 SI05 Signerindependent set. Using full frame inputs with no preprocessing (Baseline), Face cel shading and Frame cel shading. Best results in boldface.

Running Glosses	Corpus Fraction in [%]	Baseline	Face cel	Frame cel	Dev	Test
49,966	100	x			**43.8**	**43.0**
			x		44.1	43.7
				x	48.0	48.9
37,368	75	x			**48.0**	**45.7**
			x		50.7	50.0
				x	51.7	51.1
24,887	50	x			**52.0**	**50.6**
			x		54.7	54.4
				x	54.2	54.9

Imagenet large-scale visual recognition challenge [94]. GoogLeNet makes use of two auxiliary classifiers which help to propagate the gradient in lower layers of the network. Their losses contribute with a weight of 0.3 to the final loss. The network uses rectified linear units as non-linearity. To prevent over-fitting dropout [104] is applied. We set the threshold to 70% dropout ratio on the auxiliary classifiers and 40% on the final classifier. LSTMs are recurrent neural networks. Their intrinsic structure helps to overcome the vanishing gradient problem [7]. We attach two bi-directional LSTM layers with 1024 units on top of the last pooling layer of GoogLeNet and followed by a final softmax classifier. We train the recurrent network with truncated back propagation through time [120] and limit the temporal context of the LSTMs to 32 frames. We use stochastic gradient descent with a momentum $\mu = 0.9$ and an initial learning rate $\lambda_0 = 0.001$ for CNN-LSTM architectures and $\lambda_0 = 0.01$ for CNN networks. We employ a polynomial scheme to decrease the learning rate λ_i for iteration i as the training advances while reaching $\lambda_i = 0$ for the maximum number of iterations being equivalent to 4 epochs.

$$\lambda_i = \lambda_0 \cdot \left(1 - \frac{i}{i_{max}}\right)^{0.5} \tag{2}$$

Our CNN-LSTM implementation is based on [66]. The language model (LM) is estimated as 4-gram with modified Kneser-Ney discounting [28] using the SRILM toolkit by [107]. We estimate it on the available training annotations.

We use RASR [95] for the HMM and search implementation. It is a freely available and open-sourced speech recognition framework. We employ language model (maximum 4000 hypotheses), histogram (maximum 20000 hypotheses) and threshold pruning (maximum difference in log-likelihood score of 2000) of the search space for better performance and memory consumption. The prior-scaling-factor β is set to 0.3 and not optimized. The HMM is employed in bakis structure [5]. This is a standard left-to-right structure with forwards, loops and skips across at most one state. Additionally, two subsequent states share the same class probabilities. We use a topology of 6 states per gloss and a single state to account for background/garbage frames. The transition model is pooled across all glosses. Only the garbage class is modeled as an ergodic state with separate transition penalties to add flexibility, such that it can always be inserted between sequences of sign-words. For recognition, we perform a grid search over possible hyper parameters for γ and the transition model, which acts in log-domain and is composed of forward, loop, skip and exit transition penalties. They are optimised on the development set in order to minimise the WER. RASR provides an efficient implementation of the word conditioned tree search [83], which is used for this work.

Following [70], we perform iterative training with re-alignments starting from a linearly segmentation without any dependency on externally generated alignments. We train for 8 re-alignment iterations using a the CNN only. Each iteration comprises training of 4 epochs. We finetune from the CNN weights of the previous iteration. After 4 re-alignment iterations and different to [70], we restart the CNN training from the Imagenet pretrained network and also resegment the alignment path performing a linear segmentation under gloss start and end constraints given by the previous iteration. Finally, we train the full CNN+LSTM network by finetuning from the CNN only weights for additional 5 re-alignment iterations. We perform independent training runs for each of three input image conditions and additionally for the full data set, 50% and 75% of the data.

A.3 Results Details

Table 4 contains the exact results of our recognition experiments, on both sets: dev (used for parameter tuning) and test (held out throughout training).