Eye Blink Detection in Sign Language Data Using CNNs and Rule-Based Methods

Margaux Susman, Vadim Kimmelman
University of Bergen
Sydnesplassen 7, 5007 Bergen, Norway
{margaux.susman, vadim.kimmelman}@uib.no

Abstract

Eye blinks are used in a variety of sign languages as prosodic boundary markers. However, no cross-linguistic quantitative research on eye blinks exists. In order to facilitate such research in future, we develop and test different methods of automatic eyeblink identification, based on a linguistic definition of blinks, and in a dataset of a natural sign language (French Sign Language). We compare two main approaches to eye openness detection: calculating the Eye Aspect Ratio using MediaPipe, and training CNNs to detect openness directly based on images from the video recordings. For the CNN method, we train different models (with different numbers of signers in the training data, different frame crops and different numbers of epochs). We then combine the openness degree detection with a separate rule-based component in order to determine boundaries of blink events. We demonstrate that both methods perform relatively well, and discuss the practical implications of the methods.

Keywords: Blink detection, French Sign Language, Machine Learning

1. Introduction

Eye blinks are a natural physiological phenomenon which is independent of speech and language production, but which is also involved in language production in various ways. Crucially, in sign languages, eye blinks have been shown to serve as boundary prosodic markers (Sze, 2008). Some studies indicate that eye blinks are co-occurring with prosodic units, and that different sign languages employ eye blinks differently, that is, at different levels at the prosodic structure (ibid.). However, currently, no large quantitative research on eye blinks in sign languages exists, neither for specific sign languages, nor for the purposes of cross-linguistic comparison.

In order to make such research possible, it is necessary to have a reliable method of automatically identifying and annotating blinks in video recordings of signers communicating in a signed language. Due to recent advances in Computer Vision (CV) and Deep Learning, it is now possible to attempt this. In fact, the blink detection task has been pursued in many studies (Dewi et al., 2022; Fodor et al., 2023; Hong et al., 2024), but not specifically using sign language data or with sign languages in mind. In addition, the definition of blinks in the blink detection literature is quite different from the linguistic understanding of eye blinks in sign linguistics.

In this paper, we report a study in which we implemented and tested two proofs of concept of eye blink detection in a corpus of French Sign Language (LSF). We tested two main methods: a combination of a newly trained CNN associated with two different rule-based blink identification methods, and a combination of an existing CV solution, using MediaPipe (Grishchenko and Bazaresvky, 2020), with a simple eye aspect ratio (EAR) calculation, also followed by the rule-based algorithms. We specifically test how the number of signers in the dataset, as well as other specific methodological decisions influence the success of eye blink identification.

2. Background

2.1. Eyeblinks in communication

2.1.1. Physiology of blinks

In physiology, blinks have been defined as having three phases, that is a closing phase, a closed phase and a reopening phase. They have also been differentiated from closures as they last longer and do not carry the same meanings and functions as blinks do in communication. Stern and Skelly (1984) note that "for blinks, the time from initiation of lid movement to full eye closure is short, [...] less than 150ms, whereas for non-blink closures, the time taken to close the eyes is [...] generally greater than 250ms and frequently extends over a period of seconds." Blinks may exhibit an incomplete closure of the lids (Sforza et al., 2008).

As was noted by Ponder and Kennedy (1927) but also Hall (1945), Karson et al. (1981) and Bentivoglio et al. (1997), blink rates in conversations is higher than while resting or reading. Hall (1945) reports an average blink rate of 25.4 blinks per minute in conversation against an average blink rate of 3.29 blinks per minute while reading. Similar average blink rates while speaking are reported by Karson et al. (1981) and Bentivoglio et al. (1997). Hömke et al. (2017) suggested that blink events occur at turn taking points in conversations.

Finally, Descroix et al. (2022) recently investigated blinking in spoken communication. They found that in addressees, the blinking rate depends on the degree of interest in the communicated information; when presented with an interesting message, the blink rate of the addressee increases. On the other hand, the blink rate of the speaker is said to be higher than while being silent and alone, regardless of the interest to the shared information. The authors note that these findings give evidence to the "interactive communication function of Spontaneous Eye Blinks".

2.1.2. Blinks in Sign Languages

Several researchers working on a variety of signed languages have argued that eye blinks have a linguistic function (Baker and Padden, 1978; Wilbur, 1994; Sze, 2008). For example, Wilbur (1994) argued that some eye blinks[1] in American Sign Language (ASL) occur at the end of Intonation Phrases, and thus serve as prosodic boundary markers, while other blinks occur on lexical signs and have lexical or emphatic functions.

Sze (2008) investigates eye blinks in Hong Kong Sign Language (HKSL). She finds both similarities and differences in the functioning of blinks in this language. Specifically, for prosodically aligned blinks ("boundary-sensitive" in her terminology), she argues that they do not necessarily align with Intonational Phrases. According to her, they occur at the potential Intonational Phrase boundaries in 55% of the cases, while in the rest of the cases they occur at boundaries of other and typically smaller prosodic/grammatical units. In addition, she demonstrates that eye gaze change and head movement can lead to the use of blinks, even in the absence of linguistic boundaries.

The issue of classifying the functions/types of blinks is thus very complicated. The classifications in Wilbur (1994) and Sze (2008) differ in the level of detail, and these authors would classify some of the blinks quite differently. In a recent study of LSF (Chételat-Pelé, 2010), yet another classification was applied.

To summarize, a few studies have shown that blinks have important linguistic functions in sign languages. Note however, the following limitations. First, only a handful of sign languages have been studied so far. Second, while all the researchers note that blinks often align with (prosodic) boundaries, more specific functions attributed to blinks vary between the different studies, and thus a comparison is not possible. Third, the datasets analyzed in the existing studies are quite small. It is clear that much more research is necessary on this issue, including using larger datasets and analyzing blinks across multiple sign languages, multiple genres, and across individual signers. This can be achieved if automatic blink detection is available.

2.2. Eyeblink Detection

Sign Language Recognition (SLR) is a task at the intersection with CV and Natural Language Processing (NLP). SLR is concerned with the automatic recognition of signs and their translation into written or spoken language. Over the years, SLR methods have improved and nonmanuals started to be integrated into such recognition algorithms but as reported by Koller (2020), eye gaze and eye blinks have never been taken into consideration. For this reason, we turned ourselves towards blink detection algorithms. Those algorithms have mostly been implemented to solve tasks such as driver drowsiness analysis, attention level measure and eye fatigue measure (Fodor et al., 2023).

Eye blink detection methods can be divided into two categories: sensor-based methods and vision-based methods, the latter having become more popular in recent years (Hong et al., 2024).

Soukupová and Cech (2016) introduced the Eye Aspect Ratio (EAR) as a measure of eye openness. They report that the EAR is an estimation of the degree of openness of the eye. The EAR is the calculation of the distances between the lower and upper lids (with two computations per eye) and of the distance between the left and right corners of each eye. The equation of the EAR measurement is presented in (1) and the placement of the points P is shown in figure 1. P_n are landmarks locations represented in $2D$. P_1 is the landmark denoting the outside part of the eye, P_4 denotes the inside part of the eye while P_2 and P_3 both denote point on the upper lid and P_5 and P_6 denote point on the lower eyelid.

$$EAR = \frac{|| P_2 - P_6 || + || P_3 - P_5 ||}{2 || P_1 - P_4 ||} \tag{1}$$

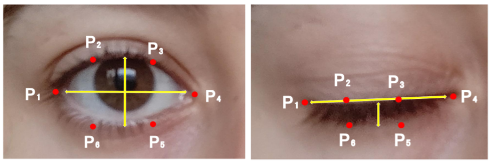

Figure 1: Eye landmarks position for the EAR calculation with open eye and with closed eye.

This EAR calculation has been widely used (Ibrahim et al., 2021; Dewi et al., 2022; Phuong et al., 2022).

[1] Clearly not all blinks have a linguistic function, as all the authors acknowledge, see also the previous section.

Recent blink detection studies include tasks addressing issues such as computation cost (Ibrahim et al., 2021), luminosity changes (Dewi et al., 2022), head movements (Hong et al., 2024). Most methods start by extracting the eye region using face detection and facial landmark detection methods or apply the EAR calculation combined with deep learning architectures (Hong et al., 2024). Another issue that we address in this present study is the lack of consideration for blinks with incomplete lid closure, which are considered blinks (Sforza et al., 2008) and which we encounter in our data.

3. Methods

In the present research, we aim at detecting blinks automatically in sign language data in order to facilitate further linguistic analysis of blinks. We define a blink as a rapid closure and reopening of the lids, limited in duration and which can exhibit an incomplete closure. Would using a machine learning (ML)-based algorithm combined with a rule-based model improve the detection of blinks so defined?

Previous eye blink detection algorithms have failed to encompass the incompleteness of lids closure and the restriction on their duration. To address those shortcomings, we present a proof of concept that combines a ML-based classifier that determines the degree of openness of an eye (open, in-between, closed) with a rule-based model taking a set window of frames as input to determine whether a blink is occurring or not.

3.1. The Dataset

We work using sign language data. We use a subpart of the Dicta-Sign corpus (Matthes et al., 2010), namely the Dicta-Sign-LSF-v2 (Belissen et al., 2020). The dataset contains video recordings of discussion about European travel. The content of those recordings was loosely elicited. In the LSF subpart, nine dyads of signers are conversing. Each of the 18 signers performed between 3 and 9 tasks. Videos and the partial annotations of the data are available online. The annotated data includes glosses for the right and left hands as well as glosses for signs articulated with both hands. In the annotated files, the annotation of a gloss is represented by an ID which is linked to a gloss in a separate document. A subset of videos is annotated for loose translations of the signed utterances. For this study, we select a subset of the annotated data.

We use data from 5 different participants. Information about the signers is available in Table1.

Signer	G	Age	Learn LSF	Deaf fam.
A11	F	28	biling. school	no
B15	M	38	prim. school	yes
B14	F	28	kindergar.	yes
A9	F	28	birth	yes
B5	F	28	birth	yes

Table 1: Participants' metadata

3.2. Annotation

We annotated the blink occurrences using ELAN 6.2 software program (Sloetjes and Wittenburg, 2008). Videos were captured at 25fps. The shortest video consists of 5500 frames while the longest video contained over 16500 frames.The *.csv* files containing the original annotations of the corpus (Belissen et al., 2020) are transformed so that the frames are converted into time intervals using a Python script. A second Python script is used to connect the ID of the annotation to the gloss of its sign. As part of the current project, a total of 26 videos were annotated, that is a total of 2 hours and 59 minutes and 4342 blinks, giving an average of 24 blinks per minute. For the experiments conducted in this paper, we selected 9 videos, that is 60 minutes and 36 seconds and a total of 1565 annotated blinks, giving an average of 26 blinks per minute. 4 of the videos are used for the training of the various ML models while we apply the blink detection algorithm to the other 5.

Sze (2008) divides blinks into three phases, specifically the closing of the lid, the eyes closed and finally its reopening. On the other hand, (Chételat-Pelé, 2010) divides the blink into two phases, that is the closing of the lid and its reopening. She adds that the full closure of the lid should not exceed 40 milliseconds limit above which we observe a closed eye and not a simple blink anymore.

In this study, one annotation for a blink includes all three phases, and its duration covers the three phases, as motivated by the definition of blinks given by physiologists. In cases in which the lids reopen to squinted eyes for example, we stop the annotation of the blink at the frame where the lid is not opening further, while in regular cases, we stop the annotation when the lid excursion is back to what it was prior to the blink event. The annotation was conducted by the first author, with discussion of specific cases with the second author.

In the data of Chételat-Pelé (2010), the shortest recorded blink lasts 160 milliseconds, while the longest doesn't exceed 380 milliseconds. We obtain similar results with a mean blink duration across all signers of 230 milliseconds over our whole dataset and 233 milliseconds in the selection of 9 videos as shown in Table 2.

Video	Vid. duration	Blinks	Av. blink duration	Shortest blink	Longest blink
S2T1B15	11:05:000	229	0.215s	0.12s	0.60s
S9T1B5	10:35.240	282	0.204s	0.08s	0.48s
S5T9A9	05:21.823	206	0.236s	0.11s	0.39s
S4T4B14	06:14.560	204	0.250s	0.09s	0.63s
S2T2B15	03:51.000	100	0.240s	0.13s	0.68s
S9T2B5	04:07.520	156	0.243s	0.11s	0.61s
S5T3A9	05:47.680	159	0.239s	0.09s	0.50s
S4T7B14	09:41.040	166	0.219s	0.07s	0.55s
S2T3A11	04:28.000	63	0.259	0.13s	0.73s

Table 2: Blink annotation statistics

3.3. Automatic Blink Detection

In the field of automatic blink detection, blink events have rarely been defined and when it was done, the issue is described as a state of openness task rather than a blink detection task. Zeng et al. (2023) claim creating an eye blink detection model but compare their work to Phuong et al. (2022) who use "eye blink detection" in the title of their paper but keep noting that they are "propos[ing] a technique to detect the open/closed state of the eyes". Dewi et al. (2022) write: "We can assume that the eye is closed/blinked when: (1) Eyeball is not visible, (2) eyelid is closed, (3) the upper and lower eyelids are connected." Two problems arise from such a description of blinks: this definition (1) does not account for incomplete blinks (2) nor for closures which last typically longer than blinks.

Making the task a binary one, with *open* and *closed* classes is overseeing the *in-between* frames which exhibit an eye not completely closed nor completely open.

We use two methods for the detection of the eyes' degree of openness. We use Mediapipe to detect eyes landmarks on which we use the EAR measure on one hand and, on the other hand, we train a novel ML model.

3.3.1. State of Openness Detection

Before training the ML model, we create a dataset specifically for the task. We transform a subset of the annotated videos into images. We create two different crops of each frame: a face crop and an eyes crop. We use MediaPipe Face Landmarks (Grishchenko and Bazaresvky, 2020) to determine which region of the frames needs to be cropped. Depending on the frame, the crop varies in dimension. The images are divided into an *open* and *closed* folders based on our annotations (the frames overlapping with the blink annotations are placed in the closed folder). We create a third *in-between* folder and rearrange the data across those three folders image by image. Indeed, as all three phases of a blink are annotated as one event, in the *closed* folder, we have eyes half open. We apply this to 4 videos from 4 of our signers, namely B14, B15, A9, and B5. The *in-between* folder contains instances where the eyeball is not completely visible nor completely hidden, instances where the eye looks open but the signer keeps their head down, and instances where the eyes are hidden in cases where a sign is performed on the face. These observations reinforce the idea that a binary classification of eye openness is not ideal.

We use the EAR measurement to detect the eye openness degree. The EAR-based method includes extracting the relevant eye landmarks with MediaPipe and calculating the EAR value for each frame using the formula above. This is done in real time.

Another way of determining the eye openness degree can be done using ML techniques. We choose to use a Convolutional Neural Network (CNN) as we are working with images and CNNs are designed to treat such data. We create a CNN architecture inspired by the classic LeNet-5 architecture (Lecun et al., 1998). Our model consists of several blocks, each one includes a convolutional layer followed by a pooling layer to seize spatial correlation in the image at varying scales. The CNN ends with linear (or "fully-connected") layers. The model for the face crops is a bit more complicated and contains an extra convolutional layer to account for the larger spatial dimensions of input images (256x256 vs. 64x128). Specifically, the face crops model is made of four convolutional layers (against three for the eyes crops model). The size of the first layer also goes up from 2080 input features for the eyes to 9216 input features for the face crops. Aside from this, the models are the same: each convolutional layer is followed by a MaxPooling layer, followed by a flattening layer and two linear layers. All layers except the last are followed by the ReLu activation function to account for non-linearity. For both models, the last layer takes 80 nodes as input and has three output features, that is one per class (open, in-between, closed). The last layer of our networks is a softmax layer that outputs a vector of probabilities. We use the cross-entropy loss to calculate the distance between the probabilities given by the model and our groundtruths. Eventually, we use the Adam Optimizer to minimize complex linear functions.

3.3.2. Pipeline: State of Openness Detection Using Machine Learning

We create four models, each model is respectively trained on 1 signer, 2 signers, 3 signers, and 4 signers and we compare the results.

Once the data is ordered in the three folders (open, in-between, closed), we proceed and load the images. Using the PyTorch library (Paszke et al., 2019), we start developing our method. Our first step is to separate the images into a training, a validation and a test set. The preprocessing of the frames varies depending on whether those are in the training set or in the validation and test sets. We resize the frames in each sets to 256x256 for the face crops and 64x128 for the eyes crops and we convert those images into numerical values. For frames in the training set, we use the Trivial Augmentation Wide transform developed by Muller and Hutter (2021) and implemented in PyTorch. The frame distribution across our three classes is greatly unbalanced. Indeed, *open* received the vast majority of the data. If we take video S2T1 from signer B15 which we use in the training of all the models, we note that out of the 16298 frames distributed across the three categories, only 690 frames belong in the *closed* folder while the two remaining folders share the 15608 images evenly.

We recreate balance in an artificial way as we fix the number of training images on a percentage of the minority class. We fix the percentage at 70% of the minority class. For example, 70% of the 690 images mentioned earlier are used in the training set for the 1 signer model. The training set therefore contains 482 images from each of the classes. The remaining 30% of the *closed* folder are divided into two: half of the frames goes to the validation set and the other half to the test set. The rest of the frames from the two other classes are also divided into a validation and test sets.

The training set is quite small due to the undersampling applied thus we use the virtual data augmentation method to modify the images within the training set randomly, that is, from one batch to another the images will appear differently. To this end, we use the TrivialAugment, an automatic augmentation method. The degree of transformation of an image fluctuates randomly but as noted by (Muller and Hutter, 2021), only one augmentation method is applied to the image at a time. The augmentation techniques applied to the images involve modifications of brightness, colors, contrast, blurring and sharpness along with image rotation and image flipping transformations.

All models are trained on the eyes crops for 100 and 200 epochs and on the face crops for 100 and 200 epochs as well.

3.3.3. Agglomeration Over Time Using Logic-Based Rules

Once we obtained our CNN results, we create the rules which will allow making a decision as to whether or not a blink is occurring.

We use the original groundtruths (data annotated with ELAN) as *.csv* files, one file for one video. As a blink occurs over a set of frames, a decision is made on a window of frames representing a time interval. We split the videos into non-overlapping windows of five frames each. We implement two different rules to detect whether we observe a blink event. Those rules are the high-low-value-difference rule (R1) and the curve rule (R2). Each one will be combined with the CNN outputs on one hand and with the EAR measurement on the other hand.

The high-low-value-difference rule looks at the maximum amplitude between the values within the selected window. According to our definition of a blink, the eye should still be somewhat open at the beginning and at the end of a blink, therefore we should observe low and large values within the window of frames when a blink happens. The difference in values between the frames of a unique window should be higher than the defined threshold when there is a blink event.

As we expect the CNN and EAR values to be lower in the middle of the window of frames and higher on the outskirts of this window when a blink is occurring, we implement the curve rule. We expect the values to form a U-shaped curve. We fit a second-degree polynomial using the polynomial regression model from Scikit-Learn (Pedregosa et al., 2011). A blink occurs when the curve goes down and up steeply, and we define the steepness with a threshold.

3.3.4. Pipeline: Blink or Not?

We want to make a decision as to whether a blink is happening or not based on a time interval lasting longer than the duration of a single frame. Therefore, we create windows of frames. The size of the window is set at 5 when no blink occurs and follows the length of the blink otherwise.We have a large class imbalance with more intervals without blinks than with blinks thus we use the $f1$-score as a our evaluation metric.

We compare the two rules, namely the high-low-value-difference rule with the curve rule within two methods (CNN and EAR). For the Convolutional Neural Networks, each rule is tested for the four trained models, noting that each of these four models has been trained for 100 and 200 epochs on the eyes crops and on the face crops. We combine the EAR measurement to each of the rules as well.

We test several thresholds which differ for the CNNs and for the EAR measurement. For the CNNs, we test 8 threshold values, specifically $0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. These thresholds represent the CNNs outputs probability of belonging into one of our three classes. For the EAR calculation, we test the following threshold values: $0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18$. Best thresholds for the EAR have been said to be contained between 0.18 and 0.20 (Soukupová and Cech, 2016) but others have criticized those thresholds and have mentioned that a greater variation can be observed (Dewi et al., 2022). The threshold represents the difference between the open and closed eyes needing to be observed for a blink to occur.

4. Results

4.1. CNN Training Results

In Table 3, we report the results obtained on the evaluation of the CNNs.

Strangely, the worst results are not obtained on the 1 signer model but rather on the 2 signers model and stay high with our lowest micro $f1$-score at 80.6% and the respective macro and weighted $f1$-score reaching 93.2% and 94.4% respectively.

The same way, for all eyes and face models, the best results are not exhibited for the 4 signers model but for the 3 signers model although the difference is slight. Our highest macro and weighted $f1$-scores each reach 97.3% and are obtained on the face model trained for 100 epochs.

4.2. Blink Detection Results

After the training of the CNN, we have seen that we obtained the best evaluation results on the 3 signers model. Combined with the rules, let us see whether the 3 signers model obtains the best results. We will also look at which of the CNN or the EAR combined with the rules is best suited for our problem.

In fact, we note that the best results across four out of the five signers are obtained using the 4 signers model. The 3 signers model trained on eyes crops for 200 epochs gives the best results for the fifth signer, i.e. signer B15 with an $f1$-score of 97% with 0.5 as best threshold. Results for the four other signers include $f1$-scores spanning between 75% to 91.7% as can be seen in Table 4 where we report the results obtained with the four signers model and where E stands for Eyes and F for Face.

1 signer	**E 100**	**E 200**	**F 100**	**F 200**
test loss	0.145	0.210	0.188	0.213
test acc.	0.952	0.947	0.945	0.945
macro f1	0.848	0.835	0.840	0.833
micro f1	0.952	0.947	0.945	0.945
weighted f1	0.957	0.954	0.951	0.952
2 signers	**E 100**	**E 200**	**F 100**	**F 200**
test loss	0.162	0.262	0.211	0.173
test acc.	0.948	0.932	0.939	0.961
macro f1	0.840	0.806	0.819	0.869
micro f1	0.948	0.932	0.939	0.961
weighted f1	0.954	0.944	0.948	0.964
3 signers	**E 100**	**E 200**	**F 100**	**F 200**
test loss	**0.110**	**0.130**	**0.122**	**0.139**
test acc.	**0.966**	**0.968**	**0.973**	**0.970**
macro f1	**0.946**	**0.949**	**0.959**	**0.952**
micro f1	**0.966**	**0.968**	**0.973**	**0.970**
weighted f1	**0.966**	**0.969**	**0.973**	**0.971**
4 signers	**E 100**	**E 200**	**F 100**	**F 200**
test loss	0.160	0.191	0.137	0.173
test acc.	0.955	0.955	0.964	0.963
macro f1	0.937	0.937	0.951	0.951
micro f1	0.955	0.955	0.964	0.963
weighted f1	0.955	0.956	0.964	0.963

Table 3: CNNs evaluation results

For each signer, the eyes models are overall better than the face crops models. In addition, the best results are all obtained with rule 1 (R1), that is the high-low-value-difference rule, that is also true for the EAR calculations (Table 5). Concerning the EAR measurements, except for one signer, the CNN models combined with R1 gives better results than the EAR calculation combined with R1 as we see in Table 5 (where, in the parentheses of the last column, the number represents the signer model, E stands for eyes, 100 or 200 for the number of epochs the CNN has been trained and R1 stands for the high-low-value-difference rule). The difference is minimal except for signer B14 for whom we observe a 12 points difference.

Signer A11 is the only one whose data has not been used for training any of the models. In Table 6, we show the evolution of the results obtained on signer A11 across the four models. We note that for the face crops the best results are attained on the three signer model. This is in agreement with what we have seen of the evaluation of the training of the CNN models. Overall we see that the results for signer A11 are getting much better when the number of signers the CNN has been trained on increases.

We achieved the best results using the four signer CNN models combined with the high-low-value-difference rule, yet we note that the variation across signers is important and while we obtain

Signer	Eyes	Eyes	Face	Face
	100, R1	200, R2	100, R1	200, R2
B15	0.964 [0.5]	0.969 [0.6]	0.953 [0.5]	0.953 [0.6]
B5	0.874 [0.5]	**0.917** [0.5]	0.874 [0.5]	**0.917** [0.5]
B14	**0.758** [0.9]	0.724 [0.9]	0.743 [0.7]	0.728 [0.9]
A9	0.870 [0.8]	0.822 [0.8]	**0.888** [0.7]	0.863 [0.8]
A11	**0.751** [0.8]	0.727 [0.8]	0.629 [0.5]	0.636 [0.8]

Table 4: Results of the 4 signer models

Signer	Rule 1	Rule 2	CNN best
B15	0.943	0.877	**0.970** (3, E, 200, R1)
B5	**0.944**	0.884	0.917 (4, E, 200, R1)
B14	0.638	0.650	**0.758** (4, E, 100, R1)
A9	0.874	0.806	**0.888** (4, F, 100, R1)
A11	0.723	0.650	**0.751** (4, E, 100, R1)

Table 5: Results of EAR combined with R1 and R2.

Mod.	Eyes	Eyes	Face	Face
	100, R1	200, R2	100, R1	200, R2
1	0.611	0.661	0.594	0.617
2	0.561	0.561	0.309	0.521
3	0.705	0.681	**0.723**	**0.688**
4	**0.751**	**0.727**	0.629	0.636

Table 6: Evolution of the results across the signer models for signer A11

$f1$-scores in the 90% for some signers, we also get $f1$-scores around 75% for other signers. Let us try to understand why.

When we introduced the dataset, we mentioned that the data had been loosely elicited. The signers had access to screens placed between the signers at a low height, therefore in some videos, the signers spend part of the time with their heads down, directed towards that screen. The blinks are noticeable but with difficulty, and while they have been annotated manually, the difference between the open eye and the closed one might not be enough for the models to detect it.

5. Discussion and Outlook

We have seen that using data from different signers in the training of the CNN allows us to obtain better results. However, we noted that the best evaluation results were obtained on the 3 signers model. We can ask ourselves whether there is a limit in terms of number of signers before the models start having less performing results. Training the CNN on more signers would allow us to test this hypothesis.

We have demonstrated that both a CNN-based approach and an EAR-based approach (which uses an existing CV solution, MediaPipe), perform the task of eye blink identification in sign language data reasonably well, but only if supplemented by specific rules that take into account the temporal structure of eye blinks. However, we have also observed that there is a quite strong variation between individual videos/signers, so the solutions achieve very high results only when certain circumstances are favorable.

In most cases, the proposed CNN-based solution is performing slightly better than the EAR-based solution. Within the parameters of the CNN-based solution, using the eyes crops and training the CNN for 200 epochs, on data from 4 signers produces the best results. This can be taken into account in future studies.

Interestingly, of the two rules we proposed to account for the temporal structure of eye blinks, the simpler Rule 1 always performs best. It might be the case that the U-shape from Rule 2 is not an appropriate representation of the actual dynamics of eye lid movements, or that the CV/ML-based measurements are not precise enough to allow for this method to fully apply. Another explanation might lay in the chosen size of the window of frames which might not capture the full extent of a blink.

Note that both approaches of eye blink identification were tested with different threshold values for the CNN outputs or EAR, and the best results are reported. We also found that the optimal threshold values differ for the different videos and the different models. This can be explained for example by the fact that signers are holding their head down, therefore the threshold at which a blink may be observed is reduced, or by physiological differences between different people. This presents a complication for the practical use of these approaches for full automatic eye blink identification in novel data: for such an approach, specific threshold values must be provided to the model, and it might not be easy to determine in advance how to choose the value.

As discussed in Section 2.2, currently several other methods have been proposed for eye blink detection, but not specifically for sign language data, or with a linguistic definition of blinks in mind. We intend to test and adapt these approaches for further application to detecting blinks across sign languages.

6. Data availability

Codes and data are available on the OSF online repository: https://osf.io/dth2q/?view_only=79f372c1776849258261c26f0f0b6ca6

Author Contributions

Margaux Susman: Conceptualization, Data Curation, Methodology, Formal Analysis, Investigation, Software, Writing. **Vadim Kimmelman**: Conceptualization, Funding Acquisition, Writing

Acknowledgements

Funded by the European Union (ERC, NONMANUAL, project number 101039378). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

7. Bibliographical References

C Baker and C Padden. 1978. Focusing on the nonmanual components of American Sign Language. In *Understanding language through sign language research*, p. siple edition, pages 27–57. Academic Press, New York.

Valentin Belissen, Annelies Braffort, and Michèle Gouiffès. 2020. Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing. In *12th conference on Language Resources and Evaluation*, Marseille, France.

Anna Rita Bentivoglio, Susan B. Bressman, Emanuele Cassetta, Donatella Carretta, Pietro Tonali, and Alberto Albanese. 1997. Analysis of blink rate patterns in normal subjects. *Movement Disorders*, 12(6):1028–1034.

Emilie Chételat-Pelé. 2010. *Les Gestes Non Manuels en Langue des Signes Françaises ; Annotation, analyse et formalisation : application aux mouvements des sourcils et aux clignements des yeux.* Theses, Université de Provence - Aix-Marseille.

Emmanuel Descroix, Wojciech Świątkowski, and Christian Graff. 2022. Blinking While Speaking and Talking, Hearing, and Listening: Communication or Individual Underlying Process? *Journal of Nonverbal Behavior*, 46(1):19–44.

Christine Dewi, Rung-Ching Chen, Xiaoyi Jiang, and Hui Yu. 2022. Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks. *PeerJ Computer Science*, 8:e943.

Ádám Fodor, Kristian Fenech, and András Lőrincz. 2023. BlinkLinMulT: Transformer-Based Eye Blink Detection. *Journal of Imaging*, 9(10):196.

Grishchenko and V Bazaresvky. 2020. MediaPipe Holistic - Simultaneous Face, Hand and Pose Prediction, on Device.

A. Hall. 1945. THE ORIGIN AND PURPOSES OF BLINKING. *British Journal of Ophthalmology*, 29(9):445–467.

Jeongmin Hong, Joseph Shin, Juhee Choi, and Minsam Ko. 2024. Robust Eye Blink Detection Using Dual Embedding Video Vision Transformer. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6274–6384.

Paul Hömke, Judith Holler, and Stephen C. Levinson. 2017. Eye Blinking as Addressee Feedback in Face-To-Face Conversation. *Research on Language and Social Interaction*, 50(1):54–70.

Bishar R. Ibrahim, Farhad M. Khalifa, Subhi R. M. Zeebaree, Nashwan A. Othman, Ahmed Alkhayyat, Rizgar R. Zebari, and Mohammed A. M. Sadeeq. 2021. Embedded System for Eye Blink Detection Using Machine Learning Technique. In *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, pages 58–62, Babil, Iraq. IEEE.

Craig N. Karson, Karen Faith Berman, Edward F. Donnelly, Wallace B. Mendelson, Joel E. Kleinman, and Richard Jed Wyatt. 1981. Speaking, thinking, and blinking. *Psychiatry Research*, 5(3):243–246.

Oscar Koller. 2020. Quantitative Survey of the State of the Art in Sign Language Recognition. Publisher: arXiv Version Number: 2.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Silke Matthes, Thomas Hanke, Jakob Storz, Eleni Efthimiou, Athanasia-Lida Dimou, Panagiotis Karioris, Annelies Braffort, Annick Choisier, Julia Pelhate, and Eva Safar. 2010. Elicitation tasks and materials designed for dicta-sign's multi-lingual corpus. In *sign-lang@ LREC 2010*, pages 158–163. European Language Resources Association (ELRA).

Samuel G. Muller and Frank Hutter. 2021. TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 754–762, Montreal, QC, Canada. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tran Thanh Phuong, Lam Thanh Hien, Do Nang Toan, and Ngo Duc Vinh. 2022. An Eye Blink detection technique in video surveillance based on Eye Aspect Ratio. In *2022 24th International Conference on Advanced Communication Technology (ICACT)*, pages 534–538, PyeongChang Kwangwoon_Do, Korea, Republic of. IEEE.

Eric Ponder and W. P. Kennedy. 1927. ON THE ACT OF BLINKING. *Quarterly Journal of Experimental Physiology*, 18(2):89–110.

Chiarella Sforza, Mario Rango, Domenico Galante, Nereo Bresolin, and Virgilio F. Ferrario. 2008. Spontaneous blinking in healthy persons: an optoelectronic study of eyelid motion. *Ophthalmic and Physiological Optics*, 28(4):345–353.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-elan and iso dcr. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

Terezq Soukupová and Jan Cech. 2016. Real-Time Eye Blink Detection using Facial Landmarks. Rimske Toplice, Slovenia. Luka Cehovin, Rok Mandeljc, Vitomir Struc (eds.).

John A Stern and June J Skelly. 1984. The eye blink and workload considerations. In *Proceedings of the human factors society annual meeting*, volume 28, pages 942–944. SAGE Publications Sage CA: Los Angeles, CA.

F Sze. 2008. Blinks and intonational phrasing in hong kong sign language. In *Signs of the Time*, j. quer edition, pages 83–107. Signum, Hamburg.

Ronnie Wilbur. 1994. Eyeblinks & ASL Phrase Structure. *Sign Language Studies*, 84(1):221–240.

Wenzheng Zeng, Yang Xiao, Sicheng Wei, Jinfang Gan, Xintao Zhang, Zhiguo Cao, Zhiwen Fang, and Joey Tianyi Zhou. 2023. Real-time Multi-person Eyeblink Detection in the Wild for Untrimmed Video. ArXiv:2303.16053 [cs].