Eye Blinks in French Sign Language

Definition of eye blink types and automatic detection of eye blinks using computer vision, rule-based and machine learning-based methods.

Margaux Susman

Supervisor: Prof Dr. Vadim Kimmelman, Department of Linguistic, Literary and Aesthetic Studies

Department of Linguistic, Literary and Aesthetic Studies
University of Bergen
Norway
November 2022

Eye Blinks in French Sign Language

Definition of eye blink types and automatic detection of eye blinks using computer vision and machine learning methods.

Margaux Susman

Abstract

This study aims at defining the different types of eye blinks in French Sign Language as well as preparing a potential automatic eye blink types classification in creating a method that allows the detection of blinks in a systematic and reliable way.

In this thesis, we learn about the research done on non-manual markers in sign languages. We focus on eye blinks and mention studies which are interested in blinks in sign languages and then studies which have contributed to the creation of automatic eye blink detection methods in various fields, ranging from medicine to automatic engineering. Following, we present the different phases of annotation of our data. We define the different types of blinks, both linguistic and non-linguistic, that we have found in our dataset. We then go on presenting the methods that we use to detect blinks automatically. Finally, we report our results, proposing a proof of concept for automatic eye blink detection that combines a CNN regression model and logic rules. We analyze ours results and we show that the Eye Aspect Ratio calculation used in combination with a cascade classifier in most methods for eye blink detection though robust (as the EAR calculation relies on the reliability of the landmarks detector) might be outperformed by the combination of a CNN algorithm and logic rules. We note that our automatic eye blink detection method is only a proof of concept and that further development need to be introduced before it can be used reliably in the context of blink types automatic classification. These further developments notably include a more varied data used in the training of the CNN models.

Les clignements des yeux en Langue des Signes Française

Définition des types de clignements et leur détection automatique grâce à des méthodes de vision artificielle et de machine learning.

Margaux Susman

Résumé

Cette étude vise à définir les différents types de clignements des yeux en Langue des Signes Française ainsi qu'à préparer une éventuelle classification automatique en créant une méthode qui permette la détection des clignements de manière systématique et fiable.

Dans ce mémoire, nous en apprenons plus sur la recherche faite sur les signaux non-manuels dans les langues des signes. Nous nous concentrons sur les clignements des yeux et évoquons les études qui s'intéressent aux clignements dans une langue des signes et ensuite, celles qui ont visé à la détection automatique des clignements des yeux dans différents domaines, s'étendant du domaine médical au domaine de l'ingénieurie automatique. Ensuite, nous présentons les différentes phases d'annotation de nos données. Nous définissons les divers types de clignements, à la fois linguistiques et non-linguistiques que nous trouvons dans notre ensemble de données. Nous passons après ça à la présentation de la méthode que nous utilisons pour détecter les clignements des yeux de manière automatique. Finalement, nous rapportons nos résultats, proposons une démonstration de faisabilité pour l'analyse automatique des clignements des yeux qui allie un modèle de réseaux de neurones convolutionnel (CNN) et des règles de logiques. Nous analysons nos résultats et nous montrons que la mesure du rapport d'aspect de l'oeil (Eye Aspect Ratio, EAR), qui est largement utilisée dans les méthodes s'attaquant à la détection des clignements des yeux, souvent combiné à un classificateur en cascade, et qui se montre robuste (notamment parce que le calcul du EAR dépend de la fiabilité du détecteur de repères sur le visage) peut être dépassée par une méthode qui allie un algorithme utilisant un CNN avec une méthode basée sur des règles de logique. Nous notons également que notre méthode de détection automatique des clignements des yeux n'est qu'une démonstration de faisabilité et qu'elle nécessite un développement plus approfondi avant de pouvoir être utilisée de manière fiable dans l'élaboration de la classification automatique des différentes catégories de clignements des yeux. Ce développement plus approfondi inclu notamment

un entraînement des modèles CNNs qui incluerait des données plus variées.

Acknowledgments

I would like to start by expressing my deep gratitude to my professor and supervisor PhD. Vadim Kimmelman for the continuous support this past year and a half. Thank you for your patience, your knowledge and precious advice.

I would also like to thank Helen Hint, research fellow at Tartu University for taking the time to read through this thesis and provide insightful suggestions and comments.

I would like to thank all of the people who contributed to this thesis directly or indirectly, thank you to Annelies Braffort, research director of CNRS, who kindly provided access to the Dicta-Sign corpus. Thank you for Giorgia Zorzi, associate professor at HVL for her French Sign Language (LSF) reading suggestions. Thank you also to Ari Price, PhD. candidate at HVL for helping me decipher some LSF signs.

I am extremely grateful to Pierre, my partner, for his love, guidance and encouragement all throughout the months dedicated to this work. I would not have been able to finish this thesis without him. Thank you for helping me all along, for accepting to discuss this work over and over to help me keep my head clear and thank you for reminding me that I could do it.

Finally, I would like to thank my parents, family and friends who supported me throughout this journey, reminded me that I was doing great, and who pushed me to keep going.

Contents

Abbreviations

Sign Languages

LSF - French Sign Language
ASL - American Sign Language
DGS - German Sign Language
NGT - Dutch Sign Language
LIS - Italian Sign Language
TID - Turkish Sign Language
HKSL - Hong Kong Sign Language
LSE - Spanish Sign Language
CSL - Chinese Sign Language
BSL - British Sign Language
SSL - Swedish Sign Language
NS - Japanese Sign Language
IPSL - Indo-Pakistani Sign Language
LSC - Catalan Sign Language
LSB - Brazilian Sign Language
ISL - Irish Sign Language
RSL - Russian Sign Language
GSL - Greek Sign Language
LIU - Jordanian Sign Language
ÖGS - Austrian Sign Language
AUSLAN - Australian Sign Language

Other abbreviations

NMM - Non-Manual Marker
ELAN - EUDICO Linguistic Annotator
EAR - Eye Aspect Ratio
FACS - Facial Action Coding System
CNN - Convolutional Neural Network

Chapter 1

Introduction

Sign languages are natural languages. For this reason, they are numerous. They emerge and evolve within Deaf communities. A sign language possesses a grammar. It has its own structure that is not derived from that of the surrounding spoken language. Sign languages share both similarities and differences with spoken languages. Baker (2016, 11-12) give examples of such similarities:

> All languages make use of [consonants and vowels] **small meaningless elements**. From these small elements, all larger units are built and these in turn are combined to form sentences.
>
> In all languages, the users can express a negative statement, can ask a question, and can issue an order.

In her book, Millet (2020, chap.1) points out four major differences between sign languages and spoken languages. Unlike spoken languages which use the auditory-oral modality, sign languages make use of the visual-spatial modality. This modality offers a signer the possibility to articulate many *small meaningless elements* simultaneously, that is combine handshape, location and movement at once, while consonants and vowels in spoken languages are articulated sequentially. While in spoken languages, only few words have a form expressing their meaning (mostly limited to onomatopoeia), these languages are said to be arbitrary. Sign languages on the other hand, as they are articulated with the human body, allow for an imitation of reality and therefore, every sign language has iconic signs. The last difference Millet (2020, chap.1) indicates concerns the opposition of spatiality and temporality. Sign languages, because they are expressed spatially benefit from three more dimensions than spoken languages for which an utterance is articulated sequentially.

A sign has three phonological parameters: handshape, movement, and location. Some researchers add orientation or even non-manual components to the list. According to a basic understanding of sign language phonology, these parameters, by themselves, do no carry any meaning, they are the ‘small meaningless elements’ noted earlier. Combined, they form the signs.

Non-manual markers are features used more generally in coordination of the manual signs but which may occur on their own in some occasions. They consist of facial expressions and head and body movements. As this study aims at closely investigating a non-manual marker, namely eye blinks in French Sign Language (LSF), in section 2 and in particular in section 2.1, we take a closer look at what non-manual markers are and what their functions are in different sign languages.

Specifically, we want to be able to classify these blinks according to their types in an automatic manner. If this last goal will not be attained within this project as steps leading to the classification are more numerous than once expected, we present in this work the qualitative work elaborated to lead us toward this initial goal in section 3.1. Specifically, we describe the process of annotation of a LSF corpus and we give our definition of the various types of blinks used in sign language communication, be they linguistic or non-linguistic. Additionally, we report on the creation of an automatic eye blink detector, a step that brings us closer to achieving this classification task according to blink types.

But before defining blink types exhibited in communication in LSF, we will learn more about the physiology of eye blinks and normal blink distribution in section 2.2. We will see that this blink distribution varies depending also on the context in which these blink events occur, depending on a person’s occupation and mental state.

Afterwards, we will look into the research that has been done on eye blinks in sign languages.

We will continue by looking at existing automatic eye blink detection methods developed within the field of computer science. The topic of blink detection has been active for these past twenty years, therefore we will go through the elements of methods used in several articles which undertake the matter of blink detection. We will see that the definition of ‘blink’ is not unique and that in the computer science field, blinks are portrayed in a simpler way than that found in the fields of physiology or linguistics. Partly for this reason, we propose in this thesis a new method approaching the detection of blinks as described in the fields of physiology and linguistics. We review the steps taken toward the elaboration of this method in section 3.2. Instead of using a combination of a cascade classifier and the Eye Aspect Ratio (EAR) calculation, we make use of CNN models and associate them with rules. We discuss our results in section 4. We go over what we were able to achieve along with the possible steps to take to improve our method which we present in terms of proof of concept in this thesis.

Chapter 2

Literature review

2.1 Non-manual markers in sign languages

Located on the upper part of the body of the signer and consisting specifically of the torso, the head, and the face, through which facial expressions and mouth movements are expressed, non-manual articulators carry meaningful information. These non-manual features are commonly used with the manual signs and may as well co-occur with other non-manuals. Non-manual markers (NMMs) can spread over a lexical sign or their span may extend to the whole prosodic or syntactic phrase. Finally, NMMs are not a natural grammatical class, the same way manual actions are not always lexical (Sandler, 2012), therefore they may bear both grammatical and prosodic functions. Before moving on and diving into the many roles these non-manual features have in the syntax and prosody of sign languages, it is important to differentiate grammatical and affective non-manual markers. Indeed, while the former have clear on- and off-sets and appear aligned with the congruous structures, the latter on the other hand, tend to surface gradually and be inconsistent: they do not emerge with specific structures and are not part of the syntax. They may, however, play a role in the language's prosody (Sandler, 2012).

2.1.1 Lexical and grammatical functions of non-manual markers

2.1.1.1 Lexical non-manuals

Finding examples of minimal pairs for non-manual markers is rare and the non-manuals tend not to be represented in phonological models of sign languages (Brentari, 2012), therefore, their place in a sign's lexical entry and them being phonological elements is debated. However, since we do find examples of such non-manual markers, we are going to talk about them here. Head and body movements, along with facial expressions and the mouth (may) play a lexical role in sign languages.Crasborn (2006) writes that this phonological role of non-manual markers appears as these NMMs are "obligatory formal features" meaningless on their own.

For example, in many sign languages, the verb SLEEP involves the palm of a hand toward which the head leans (Pfau and Quer, 2010). French Sign Language (LSF) as well as Spanish Sign Language (LSE) and Chinese Sign Language (CSL) show this head tilt. Head movements are also found to express negation in many sign languages. They may be specified in the lexical entry of a negative particle or be associated with a manual interjection sign. As explained by Wilbur (2013), negative headshakes used by signers are to be differenciated from those of non-signers, as the former coincide with syntactic constituents. In Turkish Sign Language (TID), Gökgöz (2013) notices that a backward head tilt accompanies the particle NOT. Pfau and Quer (2010) note that in American Sign Language (ASL) and in German Sign Language (DGS), a single lateral movement of the head co-occurs with the manual particle NO/NOT.

Facial expressions may sometimes be phonological non-manual articulators as well. We find examples in LSF for which the two signs SAD and SERIOUS are only distinguished by a change in facial expression (Millet, 2020, chap.3). In British Sign Language (BSL), the signs GOD and BOSS are distinguished by a different eye gaze direction, for GOD, the eye gaze is directed toward the sky while for BOSS, the signer has a straight eye gaze direction.

Finally, Crasborn (2006) explains that the most common lexical non-manual is the mouth. Mouth movements are divided into two categories: mouth gestures and mouthings.

Mouth gestures are to be differentiated from mouthing. Unlike mouthing, mouth gestures are not influenced by the surrounding spoken languages. During a sign's articulation, these gestures might be static or may change along with the manual sign (Pfau and Quer, 2010).

Crasborn (2006, 4) demonstrates the difference between the two uses of the mouth movements, taking examples from Dutch Sign Language (NGT). He notices the appearance of "rounded and pursed lips" and the production of an "egressive airstream [...] resembling IPA [ʃ]" with the sign BE-PRESENT, which are clear examples of mouth gestures, as they do not resemble any relevant spoken Dutch word.

Lewin and Schembri (2011) explain that echo phonology is a type of mouth gesture that “echoes” the manual movement and is a mandatory part of a sign’s citation form. As an example, they present the BSL sign SUCCEED/FINALLY which is produced with a burst of air and the articulation of /pu:/ as the hands abruptly change orientation.

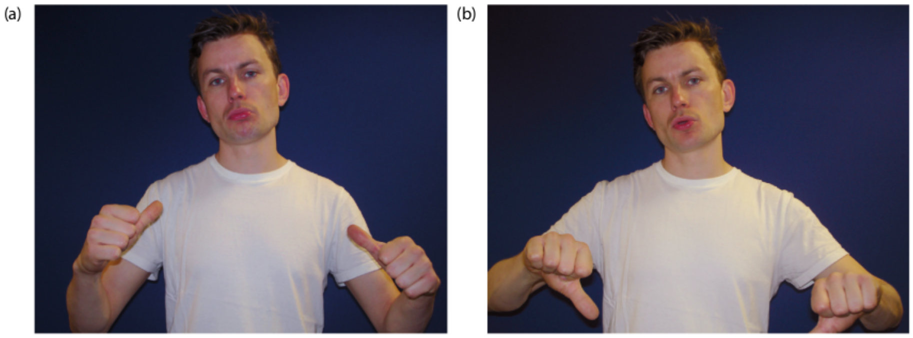


Figure 2.1: BSL example of SUCCEED/FINALLY, from Lewin and Schembri (2011)

Pfau and Quer (2010, 3) define mouthing as being “silent articulations of (a part of) a corresponding spoken word of the surrounding language”. Whether or not mouthings are part of sign languages is still discussed. What is seen, however, is that they may be redundant in some occurrences but they may sometimes add meaning to a sign or disambiguate it.

Examples of mouthing include the addition of meaning on signs such as ÉPOUX (‘spouse’) in LSF where the words ‘woman’ or ‘man’ may be silently uttered. The same can be said of the noun phrase MALADIE MÉMOIRE (‘memory disease’) precised by a labialization ‘Alzheimer disease’ (Millet, 2020, chap.6).

We have seen that some non-manual markers may carry a phonological function. This feature of NMMs is sometimes contested. In the following section, we learn more about the morphological functions of NMMs.

2.1.1.2 Morphological non-manuals

Morphological non-manuals are non-manual markers that carry their own separate meaning and may modify lexical items or carry an adjectival or adverbial function. Morphological non-manuals may co-occur with lexical manual signs. They may also co-occur with the various elements of a verb or adjectival phrase. Eventually, these NMMs may accompany signs of a whole clause (Reilly, 2005).

When occurring with a single manual sign, the morphological non-manual has the effect of changing the original meaning of the manual sign. Morphological non-manuals may also carry an adjectival or adverbial function (Pfau and Quer, 2010).

Non-manuals bearing an **adjectival** or an **adverbial** function tend to be located on the lower part of the face of the signer. Adjectives such as SMALL or BIG, for example, can be produced by sucked-in or puffed cheeks in many sign languages. Baker (2016) notably cites Inuit Sign Language for which an example is provided in figure 2.2.

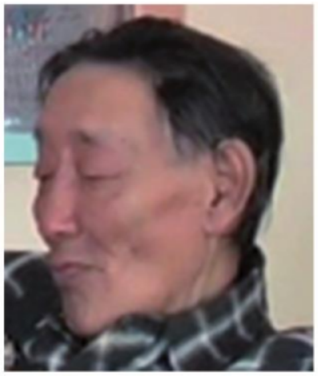


Figure 2.2: SMALL with sucked-in cheeks in Inuit Sign Language, from (Baker, 2016)

Pfau and Quer (2010) note that these non-manuals bearing an adjectival function may in fact carry an intensifier function when they appear along with their manual equivalent.

For BSL, Lewin and Schembri (2011) show that mouth gestures may play an adverbial function by accompanying a verb, thus modifying it. They give the example of DRIVE, which, associated with a mouth gesture glossed as 'mm' (bearing the meanings 'average', 'effortless') takes the meaning 'to drive in a relaxed manner'. Other mouth gestures include a 'th' (Liddell 1978, 1980 in Wilbur (2013)) carrying the meaning of 'carelessness' or 'incorectness'. Facial expressions that give information about the manner an action is carried out (e.g.: 'quickly', 'meticulously'), exist in Israeli Sign Language as noted by Meir (2012).

Meir (2012) adds that mouthing in Austrian Sign Language (ÖGS) and Australian Sign Language (Auslan) disambiguate nouns from verbs. The former are more frequently accompanied by mouthing than the latter.

Non-manual cues may also be markers of **plural**. This is the case in Italian Sign Language (LIS) where body anchored signs for nouns co-occur with a sideward movement of the head, often repeated three times (Steinbach, 2012).

In addition, in ASL **agreement** may optionally be marked by a change in eye gaze direction for object agreement which is accompanied by a head tilt appearing on the verb phrase for subject agreement (Thompson et al., 2006).

In LIS, morphological non-manuals may be **tense markers** (Zucchi, 2009). Signers of LIS use shoulder position to present information of when the action is placed relative to the utterance. A backward lean of the shoulder would therefore indicate an action taking place before the time of discourse while a forward lean of the shoulder would indicate that the action is to happen in the future. A neutral shoulder position places the utterance in the present. Zucchi (2009) adds that in cases where an adverb denoting time is used, the non-manuals cannot appear or else the utterance would be considered ungrammatical.

Pfau et al. (2012) explain that in some sign languages, **aspect** may be marked non-manually. They note that to mark continuative aspect, sign languages usually use puffed cheeks, pinched lips along with an air blow while simultaneously using reduplication of the signs concerned. They also give the example of Swedish Sign Language (SSL) for which it has been observed that durative aspect is marked by "cyclical arc movement" of the head.

Finally, Grose (2003, in Pfau et al. (2012)) notes that perfective aspect in ASL may be marked with a head nod as seen in example 1 from Grose. This non-manual can either co-occur with an aspectual marker or be alone in marking the perfective, appearing with a lexical sign or at the end of a clause.

(1) INDEX$_1$ PAST WALK $\overset{\text{hn}}{\overline{\text{SCHOOL}}}$

'I have walked to school / I used to walk to school.'

Non-manual markers can in part mark **modality** accompanying modal verbs and auxiliaries (Pfau et al., 2012). Common non-manual markers of modality are furrowed eyebrows (fe), but pursed lips and head nods may also encode modality. In ASL, the deontic modality of necessity is expressed with the modals SHOULD or MUST marked by furrowed eyebrows (Wilcox and Shaffer, 2006). An example from Wilcox and Shaffer (2006) is shown in example 2.

(2) (leaning back) SHOULD COOPERATE, WORK TOGETHER, INTERACT FORGET (gesture) PAST PUSH-AWAY NEW LIFE FROM-NOW-ON $\overset{\text{fe}}{\overline{\text{SHOULD}}}$

> ‘They (deaf community) should cooperate and work together, they should forget about the past and start anew.’

Concerning epistemic modality, as it conveys a great degree of certainty, the researchers explain that SHOULD is often used. The modal co-occurs with furrowed brows along with a head nod. The two non-manual markers also come along with the following words, which, in ASL connote epistemic modality: FEEL, OBVIOUS and SEEM are marked the same way (Wilcox and Shaffer, 2006).

Non-manual markers have many different roles in sign language morphology ranging from their adjectival and adverbial functions to the marking of tense and aspect. Let us see whether these roles are as varied in sign languages’ syntax.

2.1.1.3 Syntactic non-manuals

While lower parts of the face are usually involved in the non-manual marking of morphological constituents in SLs, upper parts of the face tend to mark higher syntactic constituents in sign languages. These upper parts include the eyes, eyebrows, head positions/movements and they intervene to mark clauses and sentences (Wilbur, 2013). More specifically, syntactic non-manuals play a role in marking sentence types, expressing agreement, negating sentences, marking relative and conditional clauses as well as marking topics or even serving an argument pronominalization (Cecchetto, 2012).

As sentence type markers, non-manual markers notably mark **interrogative sentences** in sign languages. Specific non-manuals have been found to mark polar questions (or yes/no questions) and content questions (or wh-questions) quite similarly across sign languages.

To mark **yes/no questions**, many non-manuals may be articulated simultaneously, however, raised eyebrows are the most prominent marking of such questions. In some sign languages, the only features differentiating polar interrogatives from declarative sentences are non-manual components. These non-manual components may therefore appear on their own, in which case their presence tends to be mandatory but they may as well co-occur with manual features. The scope of these non-manual components may be restricted or they may spread to the whole clause or sentence depending on the language.

In ASL, non-manual markers of polar questions involve raised eyebrows, widened eyes accompanied by a forward body lean and a tucked chin. They are not optional and extend to the whole question, excluding topics. Sometimes a question particle may co-occur with the non-manuals to put emphasis on the sentence type. This question particle is also used in questions for which the signer omits the normally mandatory non-manual articulators (Fischer, 2006).

In Japanese Sign Language (NS), polar questions are only differentiated from declarative sentences as one can see in example 3 from Morgan (2006) where pol-q refers to an agglomeration of non-manuals, namely raised eyebrows, a slight head nod and tucked chin appearing on the last sign .

(3) INDEX$_2$ BOOK BUY

‘You bought a book.’

pol-q
(4) $\overline{\text{BOOK BUY INDEX}_2}$

‘Did you buy the book?’

When it comes to **content questions**, lowered or furrowed eyebrows are the most outstanding characteristics (Cecchetto, 2012). These furrowed eyebrows are notably the non-manuals used to mark content questions in ASL. In other sign languages, content questions are marked by a variety of non-manual markers. In Japanese Sign Language (NS), content questions may be marked by raised or lowered eyebrows (example 5)that may be accompanied with a chin thrust, a sideward inclination of the head as well as a “side-to-side tremolo wag” (Morgan, 2006, 102). The author precises that different combinations of these non-manuals occur.

cont-q (lowered eyebrows)
(5) $\overline{\text{LECTURE WHO INDEX}}$

‘Who was it giving the lecture?’

cont-q (raised eyebrows)
(6) $\overline{\text{LECTURE WHO INDEX}}$

‘Who is giving the lecture (they’re not there yet)?’

Pfau and Bos (2016) note that Indopakistani Sign Language (IPSL) has an unusual way to mark content questions. In IPSL, content questions are marked by a brow raise and a backward head tilt. Cecchetto (2012) also note that only one wh-sign is used to mark content question in IPSL (example 7 from Cecchetto (2012)). This sign appears sentence finally and is marked non-manually. Polar questions in IPSL are marked by wide opened eyes and a forward tilt of the head.

(7) FATHER INDEX$_3$ SEARCH $\overset{\text{wh}}{\overline{\text{G-WH}}}$

‘What is/was father searching?’

(8) INDEX$_3$ COME $\overset{\text{wh}}{\overline{\text{G-WH}}}$

‘Who is coming?’

Zeshan (2006b) oberves that in TID, a head movement is also involved in the non-manual marking of content questions. In this language however, we find a sideward headshake that can co-occur with an eye contact and the head in forward position. According to Fisher (Fischer, 2006), ASL content questions are marked by narrowed eyes and furrowed and raised eyebrows. In ASL, non-manual components in content questions need to spread over the whole clause at least in cases where the wh-phrase appear at the beginning of the sentence (Fischer, 2006) We provide illustrations of possible wh-questions in ASL in examples 9, 10 and 11 taken from Cecchetto (2012).

(9) $\overset{\text{wh}}{\overline{\text{WHAT JOHN BOUGHT YESTERDAY WHAT}}}$

‘What did John buy yesterday?’

(10) $\overset{\text{wh}}{\overline{\text{JOHN BUY WHAT YESTERDAY}}}$

‘What did John buy yesterday?’

(11) $\overset{\text{wh}}{\overline{\text{JOHN BUY YESTERDAY WHAT}}}$

‘What did John buy yesterday?’

Another sentence type that non-manual markers affect is **imperative sentences**. According to Baker and Cokely (1980), imperative sentences in ASL are expressed with emphasis on the verb through the use of sharp and fast movements combined with an eye gaze directed at the addressee. We also find this eye gaze toward the addressee in Australian Sign Language (Auslan) which is also accompanied by a frown and a special stress on the utterance’s signs (Johnston and Schembri, 2007).

Besides sentence types, non-manual components affect and mark **negation**. Non-manual markers of negation originate from affective gestures and facial expressions commonly found in the surrounding spoken languages. As sign languages allow different kinds of marking of negative sentences, a distinction has been made between manual-dominant languages and non-manual dominant languages (Zeshan, 2006a). When negation is only marked by non-manual markers, these non-manuals are obligatory. Examples of such non-manual dominant languages include ASL, DGS and Catalan Sign Language (LSC). LIS, or TID on the other hand are manual dominant languages, i.e. the non-manual markers cannot appear by themselves and manual components are needed to mark the negation. Non-manuals present in negative constructions of manual dominant languages co-occur with a manual negator and have a scope restricted to the manual negation. In non-manual dominant languages, the non-manual components are at present grammaticalized and are controlled by the grammatical rules of the language (Quer, 2012).

The most commonly found non-manual markers of negation are headshakes, head turns and finally head tilts (Quer, 2012). Headshakes may co-occur with a manual negator. A head turn occurring in negative utterances has been reported in BSL, CSL, Russian Sign Language (RSL) but also Greek Sign Language (GSL) and even Jordanian Sign Language (LIU). In Eastern Mediterranean sign languages, a head tilt is found to mark negation. This marker appears in GSL and LIU and also in TID. In TID, the head tilt is realized backward. Sometimes, a side-to-side headshake appears to mark negation in TID. Zeshan (2006b) notes that the two NMMs occur with specific manual components in TID. Besides head movements, facial expressions may be non-manual markers of negation. Examples of such expressions are wrinkling, frowning, squinted eyes, and in some cases lips positions. Quer (2012) gives the example of Brazilian Sign Language (LSB) which uses a combination of a headshake and facial markers, notably lips corners down of an O-like shape of the mouth. In Israeli Sign Language as well as in LIU, it is the facial feature itself that negates the sentence.

The scope of the non-manual markers of negation is restricted. Headshakes can extend over the clause but in TID, for example, the scope of the head tilt is limited to the manual negator it comes with and sometimes a preceding or following unstressed sign. In ASL, spreading of the non-manual feature is mandatory when no manual negator is used. However, in cases in which manual negation is employed, the spreading of the negator is optional. In DGS, the rule is different, the spreading of the non-manual negator seems to always be optional, regardless of the presence of a manual sign but if the non-manual features spreads, then it must do so over whole constituents, the sentence would otherwise be ungrammatical (Quer, 2012).

Non-manual elements may also mark **relative clauses**. Pfau and Quer (2010) explain that in many sign languages, "relative clauses are marked by raised eyebrows". In some languages, raised upper lips or pursed lips and tensed eyes can also be found to mark relative clauses (Tang and Lau, 2012). Pfau and Quer (2010) give the example of LIS and DGS for which this non-manual appears to encode relative clauses. In LIS, the non-manual components also include facial expressions: tensed eyes and pursed lips (example 12 from Branchini and Donati (2009)). The facial expressions' scope extends to the noun as well as the adverb, while the raised eyebrows cover the entire relative clause. In DGS, a body lean can also mark relativization and the non-manuals, both the raised eyebrows and the body lean, only co-occur with the relative pronoun (Tang and Lau, 2012).

(12) $\overset{\text{re}}{\overline{[\text{TODAY MAN}_3\text{ PIE BRING PE}_3]}}$ YESTERDAY (INDEX$_3$) DANCE

'The man that brought the pie today danced yesterday'

The same way raised eyebrows are found to mark relative clauses and polar questions, they mark **conditionals** (Cecchetto, 2012). This non-manual feature tend to be associated with a raised chin, notably in ASL, but also in DGS, LIS and in NGT. Pfau (2016) observes that in these languages, the lexical marker of conditionality is optional. In French Sign Language (LSF), Millet (2020, chap.7) expressed that IF can be signed with dactylology, though many signers contest this and she explains that conditionality is more often formulated through the use of a back body lean and a facial expression.

Additionally, raised eyebrows act as **topic** markers in sign languages (Pfau and Quer, 2010). Like is the case for sentence type marking or even negation, non-manual components are often layering (appearing concomitantly) to mark sentence topics. As there exists different kinds of topics, ASL distinguishes them by varying the non-manual components it uses. Widened eyes, mouth open and diverse head movements discern the many topics. For LSF, Millet (2020, chap.10) explains that a topic is marked by a specific signing rythm alongside a facial expression, a head nod as well as a shoulder movement.

Finally, **agreement** may also be marked with non-manual features. According to in Neidle and Nash (2012), non-manual components such as lip-pointing, eye gaze pointing and head tilt are used as agreement markers. These non-manual markers may serve to mark subject or object. When their manual counterpart are omitted, eye gaze takes on a pronominal analysis.

The range of possible syntactic roles taken on by the non-manual markers in all of these sign languages is as large if not larger than that of the morphological non-manuals.

In LSF, we have seen that non-manual markers are notably described to have a role in the marking of conditionals and topics. In the next section, we dive into the semantic roles that non-manual markers may carry in sign languages.

2.1.1.4 Semantic-pragmatic non-manuals

Non-manual markers can carry a semantic-pragmatic function. They notably do so when marking focus. Focus relates to the part of an utterance that is new, informative or important and that is marked linguistically by a specific means. In spoken languages, focus can be marked with a pitch accent appearing on the relevant element of the utterance or with focus particles such as *only*, *even* or *also* which have different focus meaning, namely restrictive, scalar and additive. In sign languages, such particles are also used and non-manual markers may accompany them or may occasionally appear on their own.

Body leans (bl) often accompany focus particles (Baker and van den Bogaerde, 2016).In ASL, a backward body lean or a shrug may appear with restrictive focus particles. Restrictive focus meaning can be formulated in many ways in ASL. The particles JUST and ONLY, which has no less than three various forms in the language, namely ONLY, ONLY-ONE and THAT'S-ALL, are examples of the expression of restrictive focus (Herrmann, 2013). In example 13 from Herrmann (2013) we see that ONLY-ONE is accompanied by a backward body lean. '

(13) INDEX$_1$ RECENTLY FOUND-OUT WHAT $\overset{\text{bl-backward}}{\overline{\text{ONLY-ONE}}}$ $\overset{\text{hn}}{\overline{\text{KIM}}}$ GET-A

'I just found out that Kim got an A.'

JUST and THAT'S-ALL are non-manually marked by the shrug which carries functions analogous to the backward body lean. In contrast to restrictive focus particles, a forward lean of the body accompanies additive particles such as SAME/ALSO when it carries the scalar meaning 'even'. We give an example of SAME exhibiting the meaning 'even' in example 14 from Herrmann (2013).

(14) (...) ALL KNOW-THAT Bill $\overset{\text{bl-forward}}{\overline{\text{SAME}}}$ PT TEST PT GET-A

'(What an easy test!) Everyone knows that even Bill got an A.'

However, these non-manuals can be used without their manual counterparts and still carry the focus particle meaning. In these situations, they accompany the focus constituent.

Herrmann (2013) investigates the expression of focus particles in DGS, NGT and Irish Sign Language (ISL). She reports that except for an occurrence of fingerspelled ‘even’ in ISL and a single appearance of the Signed German ‘even’ in DGS, signers use a combination of an additive particle with specific non-manuals to express the meaning of ‘even’. Herrmann (2013) adds that in rare occasions, the non-manual components appeared by themselves carrying the scalar additive meaning of ‘even’.

Contrastive focus may also be expressed by means of a body lean. Baker and van den Bogaerde (2016) explain that in NGT and in RSL, left and right body leans mark contrast. Aternatively, Pfau and Quer (2010) add that ASL signers mark contrastive focus with forward-backward body leans. Likewise, these body leans are used to mark positive and negatives responses. Baker and van den Bogaerde (2016) note that while positive reactions may involve a forward body lean, negative responses trigger a backward body lean.

Another way non-manual markers carry this semantic-pragmatic function is by accompanying role shifts. Baker and van den Bogaerde (2016) distinguish two types of role shifts, namely the body shift used in quotational situations and the perspective shift or constructed action. Body shifts are compared to direct quotations in spoken languages and are achieved by a subtle change in body position and facial expression. The signers use these body shifts to express the point of view of a character of the story they are telling. Role shifts or perspective shifts are used in a wider context than body shifts and allow not only to express the thoughts of the character embodied but also illustrate the actions that the character is doing. The characters are distinguished by using different non-manual elements: postures and facial expressions.

We have taken a look at the various grammatical and lexical functions that non-manual markers have in sign languages. We will now observe the prosodic functions that these non-manual markers may carry and will see that these NMMs may have overlapping functions.

2.1.2 Prosodic functions of non-manual markers

Sandler (2012) defines prosody as a means to separate or relate elements of an utterance through the employment of “timing, prominence, and intonation”. She explains that through her work, Liddell (1978) observed that specific non-manual elements consistently occurred along with manual signs in specific sentence types in ASL. These non-manual markers are said to be prosodic markers. However, non-manual components are not the only means used to express prosody. Indeed, intonation is often expressed on the face but timing tends to be carried through the use of the hands, which is also true for prominence (Sandler, 2012).

A clarification needs to be provided before moving further: prosodic constituents and syntactic ones tend to be isomorphic (Sandler, 2012; Brentari et al., 2018), that is their boundaries are aligned. Although this is not always the case as one sees in example 15 from Pfau and Quer (2010), it is important to note that because it often happens, non-manual markers previously introduced as syntactic markers are likely to carry a prosodic function as well. Pfau and Quer (2010) give the example of topicalized constituents which constitute intonational phrases in themselves (that is a chunk of elements with its own intonation pattern) and for which it cannot assuredly be said that the non-manual markers fulfill a specific function, whether it is a prosodic or a syntactic one.

(15) syntax: TOMORROW [MAN (IDEX$_{3a}$) $\overset{\text{re(}\qquad\text{)}}{\overline{\text{RPRO}_{3a}\text{ TIE BUY}}}$ CONFERENCE$_{3b}$ GO-TO$_{3b}$

‘Tomorrow the man who is buying a tie will go to a conference’

(16) prosody: [TOMORROW MAN (IDEX$_{3a}$)] $\overset{\text{re(}\qquad\text{)}}{\overline{\text{[RPRO}_{3a}\text{ TIE BUY]}}}$ [CONFERENCE$_{3b}$ GO-TO$_{3b}$]

‘Tomorrow the man who is buying a tie will go to a conference’

Prosodic chunks are organised hierarchically, i.e. larger chunks are built up from elements at the preceding level (Selkirk, 1996).

syllable > foot > prosodic word > phonological phrase > intonational phrase
> phonological utterance

We will focus here on the prosodic word and the intonational phrase.

Before discussing these prosodic non-manuals, it is important to define the notions of domain and boundary markers. A non-manual domain marker is one that spreads over the whole clause. In Israeli SL conditionals, the signs of the first clause are marked by a squint of the eyes and raised eyebrows as illustrated in figure 2.3 from Dachkovsky and Sandler (2009). In contrast to domain markers, boundary markers do not extend to the whole phrase but instead identify the edge of a prosodic domain (van der Kooij and

Crasborn, 2016). If we take the example of the Israeli SL conditionals, at the boundaries, a forward lean of the head marks the last sign.

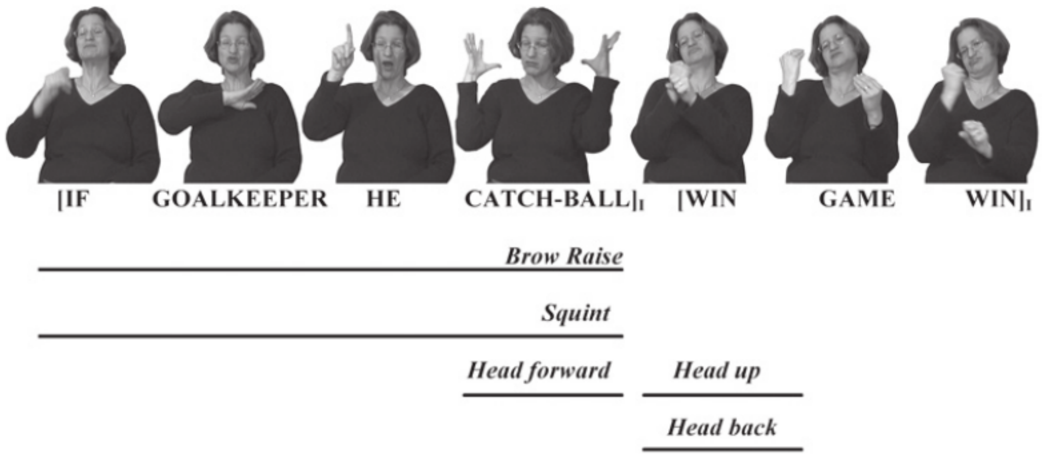


Figure 2.3: Illustration of domain and boundary markers, from (Dachkovsky and Sandler, 2009)

In English, examples of prosodic words different from a simple syntactic word would be ‘don’t’ (that is ‘do not’) or ‘she’s’ (that is ‘she is’) where two morphosyntactic words cliticize and make a single prosodic word (Sandler, 2012). Pfau and Quer (2010) give an example of a type of cliticization from Israeli Sign Language taken from Sandler (1999), notably coalescence. In sign languages, syllables are connected to movements, one movement gives one syllable. In coalescence, the two morphosyntactic words, specifically a host followed by a pronoun, combine and from initially being one syllable each become one single syllable together. Coalescence arises through the articulation of the host sign by the non-dominant hand while the dominant hand, halfway through the production of the host sign, produces the clitic, combining the normally two syllables into one (Sandler, 2012).

Aside from coalescence, cliticization can be marked non-manually, namely by means of mouthing Pfau and Quer (2010). The authors explain that the non-manual marker occurs on the lexical sign and spreads onto the following functional sign. These findings are based on research on NGT, BSL and SSL. They note that the spread of mouthing is used in other context than cliticization, for instance from one lexical sign onto the following lexical sign.

Sandler (2012) explains that an intonational phrase in spoken language is defined as the domain in which we find the most “salient pitch excursions”. As written in Sandler

(2012), intonational phrases are:

> marked by a salient break that delineates certain syntactically coherent elements, such as (fronted) topics, extraposed elements, non-restrictive relative clauses, the two clauses of conditional sentences, and parentheticals

The intonational phrase is the principal domain in which intonation is expressed. In sign languages, intonation is said to be expressed by the upper face articulations (Dachkovsky and Sandler, 2009).

The authors explain that seeing facial expression as working similarly to intonation in spoken languages is motivated by these facial expressions achieving the same pragmatic functions as that of intonation, notably in marking the difference between dissimilar sentence types. The facial non-manuals also communicate an attitude toward a proposition made prior in the discourse, for example incredulity or highlight of shared information. They also explain that the non-manual markers are aligned with the intonational phrase in that at their boundaries, facial articulators switch their position no matter what their initial position was.

In Israeli Sign Language, Sandler (2012) explain that a change in head or body position reveals the boundaries of an intonational phrase. This change is accompanied by a switch of all facial expressions. Dachkovsky and Sandler (2009) give the example of a conditional where the first phrase is marked with two domain markers namely raised eyebrows and a squint, and at the boundary, a forward lean of the head marks the last sign. When the second phrase starts, the eyes and eyebrows switch to their neutral position and the head position changes from forward to up and back. Dachkovsky and Sandler (2009) also note that non-manual markers are multifunctional. For instance, raised eyebrows not only mark polar questions but also arise with other types of sentences, notably conditionals. They also find that the squint tends to appear on shared information but does surface with topics and relative clauses as well.

We saw that in Israeli SL the squint and raised brows were domain markers, on the other hand, head nods and eye blinks appear at boundaries and can therefore be considered to be edge markers (Pfau and Quer, 2010; Brentari et al., 2018). We have already seen such an example in figure 2.3, with the Israeli SL conditional in which a forward lean of the head appeared on the last sign of the phrase, marking its boundary. This head thrust acts similarly in ASL's conditionals (Sandler, 2012). In addition, the author claims the forward lean of the head materializes as well on the last sign of when-clauses. Eye blinks are different in that they materialize in between phrases. These are present in ASL and Israeli SL and mark the breaking point between two intonational phrases (Sandler, 2012). Eye blinks represent a complex matter as there are various kinds of blinks as we will see in the following section 2.2, however, we can already note that some appear at the junction between two intonational phrases. This type of blink arises in DGS (Pfau and Quer, 2010) but also in Hong Kong Sign Language (HKSL) (Sze, 2004). Pfau and Quer (2010) additionally claims that the eye blink sometimes behaves similarly to the head nod thus appearing not, or not only in between the two intonational phrases but also with the last word of the first phrase.

We have now seen that non-manual markers have various roles in sign languages ranging from marking specific lexical signs to having morphological and syntactic functions or bearing a semantic-pragmatic meanings. We also show in this section that NMMs can have prosodic functions as well which are sometimes even difficult to distinguish from the previously introduced syntactic functions. In the next sections we explore eye blinks in all their complexity. We will look at their physiological role as well as their function in sign language communication. Eventually, we will see how eye blinks can automatically be detected.

2.2 Eye blinks

2.2.1 Physiological aspects of eye blinks

To be able to explain the use of eye blinks in sign language communication, it is necessary to understand the physiological aspect of eye blinks. In this section, we try to understand how humans blink and then briefly trace blink research history.

2.2.1.1 On blink physiology

Researchers have agreed that three main types of blinks distinctly arise. In their article, Bour et al. (2000) make the distinction between these three kinds of blinks, specifically reflex, voluntary and spontaneous blinks and Kaneko and Sakamoto (1999) have studied them to determine the characteristics that discriminate them from one another. Two muscles are mainly responsible for the eyelid movements. These are the levator palpebrae superioris (LP) and the orbicularis oculi (OO) muscles illustrated in figure 2.4. To be more specific, LP controls the elevation of the upper lid while OO is the muscle involved in the closing of the eyelid.

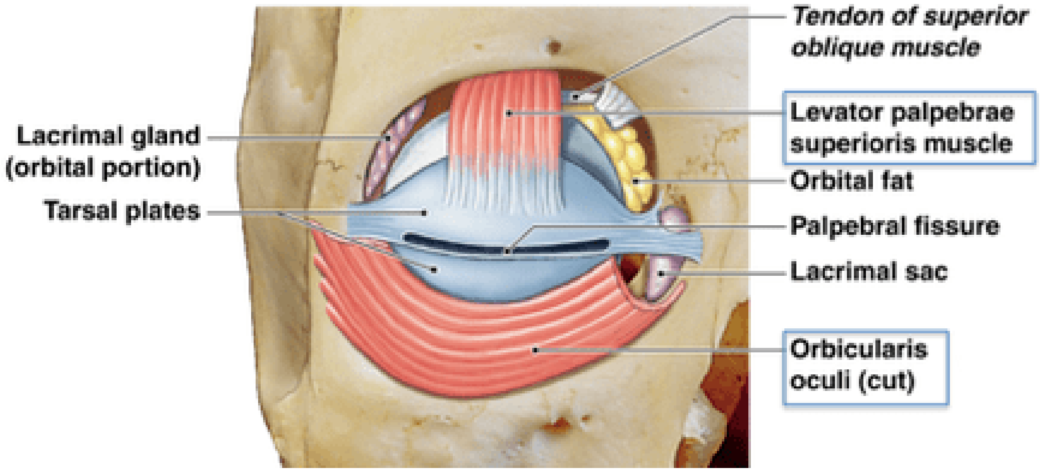

Tendon of superior oblique muscle
Levator palpebrae superioris muscle
Lacrimal gland (orbital portion)
Orbital fat
Tarsal plates
Palpebral fissure
Lacrimal sac
Orbicularis oculi (cut)

Figure 2.4: Illustration of the eye muscles, notably the Levator Palpebrae Superioris and the Orbicularis Oculi involved in blinking

Bour et al. (2000) give a detailed description of the physiological mechanisms of blinking from which we note that blinking is accompanied by a movement of the eye, specifically a “disconjugate oblique eye movement” which accompanies the three kinds of blinks mentioned. This movement of the eye starts before the closing of the eyelids and after the blink the eye goes back to its initial position. As was first pointed out by Evinger et al. (1984) and confirmed in Bour et al. (2000), during blinking, the eyeball is retracted into its orbit, this retraction movement is explained by the contraction of the inferior and superior rectus muscles, responsible for the upward and downward movements of the eyeball. Kaneko and Sakamoto (1999) used EMG (eletromyogram[1]) at the OO muscle and EOG (electro-oculogram[2]) to measure the amplitude and duration of the various blinks. They report that with the EMG, “spontaneous blinks had the smallest mean amplitude and duration for the EMG”. Kaneko and Sakamoto (1999) explain that:

> the triggers of the generation of the reflex, the voluntary, and the spontaneous blinks were the external stimulus, a motor command from the cerebral cortex, and neurogenic change in the cortical or subcortical level, respectively.

We have learnt about the muscles involved in blinking and we now know there exists

[1] “Electromyography measures muscle response or electrical activity in response to a nerve’s stimulation of the muscle” Hopkins (n.d.).

[2] “Electro-oculography is a technique for measuring the coreno-retinal standing potential that exists between the front and the back of the human eye” Macula-Retina-Institute (n.d.).

three types of blink, let us now talk about blink distribution and blink rates by reviewing studies conducted on the matter in section 2.2.1.2

2.2.1.2 On blink distributions and blink rates

The first researchers studying eye blinks were Ponder and Kennedy (1927) with their famous article on spontaneous blinks *On the Act of Blinking.* In their study, 50 participants were observed reading what the author called light literature in periods ranging from 30 minutes to 2 hours and the researchers disregarded the first 5 to 10 minutes as these were considered to be an adaptation period to the task. Ponder and Kennedy (1927) differentiate four types of distributions of intervals between blinks, namely the J-shape blink distribution, the Irregular Plateau distribution, the Bimodal distribution, and the Symmetrical distribution. They found that the most common distribution is the J-shape blink distribution, out of their 50 participants, 31 displayed this type of distribution. This pattern consists of numerous intervals of short duration between the blinks. The researchers however explain that these intervals decrease in number as their duration last longer. The second most common type of blink distribution is the Irregular Plateau; 11 participants exhibited this distribution type which consists of long and short inter-blink periods. Ponder and Kennedy (1927) further proclaim that the longer intervals are most common and may last up to over 60 seconds. People who display a bimodal eye blink distribution exhibit two distinctive blink patterns, namely one with a great many inter-blink periods that last about half a second and one in which the inter-blink duration is longer, specifically, these intervals may range from 4 to 6 seconds. Out of the 50 participants, 6 showed this pattern. Finally, the last type of distribution, the symmetrical eye blink distribution, was only exhibited by 2 participants among the 50. It corresponds to many inter-blink periods that last for 6 to 10 seconds along with both shorter and longer interval periods occurring in equal numbers.

Hall (1936) distinguishes two types of blinks: reflex blinks and blinks of uncertain origin. The former type is further divided in three sub-types, namely, the corneal reflex blink (refers to blinks as a response to contact on or close to the eye), the dazzle reflex blink (refers to blinks as a response to a sudden change in luminosity) and the menace reflex blink (refers to blinks as response to the sudden appearance of a threatening object). The latter type is one that does not arise because of an external stimulus and that tends to appear at regular intervals. This type of uncertain origin is the kind investigated by Ponder and Kennedy (1927) and that will later be called ‘spontaneous’ (Kaneko and

Sakamoto, 1999).

Hall (1945) gives additional information on the types of blinks exhibited by men. He differentiates against three types of reflex or automatic blinks. The group B blinks are concerned with the protection of the eye, they bring together the corneal and the dazzle reflex blinks. The group C is concerned with "the preservation of the organism as a whole", Hall (1945) adds that they are the type most commonly exhibited by humans. Finally, he describes Group D as blinks of technique observed while participants are reading. He further explains that these blinks are "controlled" by the reader and usually appear at punctuation mark. In this same article, Hall (1945) measured blink rates in normal subjects[3]. He measured the blink rate of 57 participants (37 men and 20 women) during conversation. The average blink rate in conversation, according to him, is 25.4 blinks per minute[4]. Hall (1945) also investigated the blink rate of 29 subjects (17 men and 12 women) as they were reading aloud. He reported an average blink rate of 3.29 blinks per minute[5] for his participants.

Karson et al. (1981) studied spontaneous blinks in varying conditions as well. They observed 41 participants (14 female participants and 27 male participants) in 8 different settings: during conversation, while being silent and chewing gum, while interpreting proverbs, while memorizing text, while having explicitly asked them to suppress their blinking for as long as they could, while having explicitly asked them to speed up their blinking[6], and once again in conversation with the examiner. They explain that they let the subjects get used to the environment for 5 minutes before starting to measure the blinks. A minute break was given to the participants between each task and the entire experiment lasted 35 minutes (Karson et al., 1981). They report an average rate of 19 ± 12.6 blinks per minute while being silent[7]. The mean blink rate while speaking is reported to be at 24.7 ± 12.6 which is indicated to be significantly higher than during quiet time. Karson et al. (1981) report a mean blink rate of 12.3 ± 7.4 for reading, that is significantly different from the blink rate reported while staying quiet. The mean blink rate while reading aloud was also significantly higher than that of the quiet time (25.2 ± 13.3).

[3] To serve as a baseline for research on chronic encephalitics.

[4] A mean of 29.3 blinks per minute for the male participants against 18.3 blinks per minute for the female participants.

[5] A mean of 3.57 blinks per minute for male participants against 2.58 blinks per minute for female participants.

[6] Only the last 12 subjects, lasted 1 minute.

[7] 18.5 ± 9.0 blinks per minute while being silent and chewing gum.

Karson (1983) reports new blink rates in similar conditions as he had done previously with his colleagues. He recorded the blink rates of 49 participants (29 male participants and 20 female participants) in 4 different settings: during conversation, while being silent, while reading cards and while memorizing text. He reports a mean reading blink rate of 14 ± 8 blinks per minute that is significantly smaller compared to that of the silent period for which the mean rate was noted to be of 19 ± 14 blinks per minute. He reports an average speaking blink rate of 25 ± 14 blinks per minute, significantly higher than the average blink rate recorded during the silent period, and he also mention that the mean recitation (after memorization) blink rate of 29 ± 15 blinks per minute, also significantly different from the average rate of the silent period. Overall, we see that the reported average blink rates are similar between the two studies. However, if the average blink rate during conversation is similar to that reported in Hall (1945), the mean reading rates reported by Karson et al. (1981) and Karson (1983) differ from that of Hall (1945).

In a later study, Bentivoglio et al. (1997) also analyze blink rates in normal subjects. They indicate inconsistent results across studies on the reported average spontaneous blink rates. As was already mentioned in the previous studies, Bentivoglio et al. (1997) acknowledge that cognitive and emotional states of a person might influence his or her blinking rate. As noted earlier by Karson et al. (1981), Bentivoglio et al. (1997) note that activities involving speech or memory increase a person's blinking rate while activities such as reading or daydreaming, requiring visual fixation, slow down the blinking rate. They further explain that during speech, blinking marks phrases and blinks tend to appear at the end of sentences. In their studies, they observe and measure the blink rate of 150 normal subjects in three different tasks, specifically while at rest, during conversation and while reading. They report the average blink rate while resting at 17 blinks per minute, while in conversation, the mean blink rate rises to 26 blinks per minute and the average reading blink rate drops at 4.5 blinks per minute. Oddly, in contrast with previous studies, the mean blink rates recorded for female participants are higher than that of men though only significantly higher while reading[8].

Doughty (2002) investigates spontaneous eye blink rate as he himself also found results reported across studies to be inconsistent. He indicates the lack of agreement regarding

[8] male participants while resting: 15.6, while reading: 3, while conversing: 24, female participants while resting: 18; while reading: 6.2, while conversing: 26.7.

the methods, notably in terms of environment for the measurements as well as the length of the time period during which participants' blinking is recorded. He investigated what he calls primary gaze-SEBR which refers to spontaneous eye blink rate while the participants are silently sitting and looking straight-ahead of themselves. Doughty (2002) studies the primary gaze-SEBR of 61 participants, specifically 30 men and 31 women. A 5 minute period is allocated to all participants to get acquainted with the environment of the study. He records their primary gaze-SEBR for a 5 minute period. The participants have an average SEBR of 10.3 ± 3.1 blinks per minute. Doughty (2002) further reports that no significant gender difference is found in men and women blinking frequency. The eye blink patterns of the participants are studied based on the work of Ponder and Kennedy (1927) and Carney and Hill (1982). He finds that the participants could be separated in three groups. A third of the participants exhibits the irregular eye blink pattern, the next third shows the J-type inter eye blink distribution and the last third exhibits a symmetrical blink pattern, the distribution that is the less displayed by the participants of Ponder and Kennedy (1927). It is shown that difference in blink rate arise between the three groups of participants. Participants showing the irregular eye blink pattern have a lower blink rate with 7.5 blinks per minute. Participants with the J-type distribution have a blink rate of 10.7 blinks per minute. Finally, participants showing the symmetrical eye blink pattern have an average blink rate of 12.3 blinks per minute. Doughty (2002) reports that the comparison of the blink rate across the three different blink patterns showed that all groups are statistically different from each other.

In the context of the development of a non-invasive method for the detection of movements of the eyelids, Sforza et al. (2008) have investigated spontaneous blinking in normal people. They study SEBR in 44 normal people, 23 men and 21 women. They group subjects according to their age. Participants younger than 30 years old (25 participants, 13 men and 12 women) are considered young, the older participants are older than 50 years old and up to 77 (19 participants, 10 men and 9 women). They use video cameras[9] to record the participants blinks. All participants go through 2 recording sessions lasting 90 seconds each and with a 5 minute break in between sessions. They also collected participants staying static for reference, one with eyes shut and one with eyes open. They measure the two eyes of each participant separately. They briefly mention incomplete blinks and the fact that they are recorded as well. Blinks in which the two eyelids do not

[9] "Optoelectronic three-dimensional motion analyzer", "six high-resolution infrared sensitive charge-coupled device video cameras coupled with a video processor".

meet are still considered blinks (Sze, 2004). They report that women blink significantly more frequently than men. They also report that women moved their eyelid faster than men but the maximum velocity is significantly decreased with age. Sforza et al. (2008) note that young men moved their eyelids 70% faster than older men and the same is true for older women who move their eyelids 80% slower than younger women. They also add that blinking is symmetrical, the same number of blinks are recorded for participants' left and right eyes. As is pointed out in Collins et al. (2006), eyelids do not always meet while blinking. In their study, participants are recorded for about 3 minutes, the objective of the researchers is to record approximately 40 blinks per participant and they consider that the mean primary gaze-SEBR was of 14.5 blinks per minute. Incomplete closure of the eyelid is observed in 22% ± 23% of the recorded blinks. This is also demonstrated by Sforza et al. (2008), where blinks are considered incomplete when the eyelid is closed to less than 50% of the way. They also add that differences are observed depending on the age and gender of the participants. They observe that young male participants are the group that most often closed their eyes completely, 44% of the time. They additionally note that older men exhibited a larger number of incomplete blinks. Finally, Sforza et al. (2008) report an average spontaneous blink rate of 15.1 blinks per minute for all participants and more specifically an average blink rate of 10.6 blinks per minute for men[10] and of 19.65 blinks per minute for women[11].

We have now learnt more about blinks, from mean frequency to blink physiology, we have reviewed numerous studies that explain which muscles are involved in blinking, that have measured blink rate and taken an interest in blink distribution. What is apparent is that blink rate seems to vary across individuals. We saw that blink frequency varies depending on the task performed and is consequently dependent on the cognitive workload. What interests us most in the context of this thesis is that spontaneous blinks in conversation which are reported to be most frequent than that measured while staying silent or during reading (conversation > rest > reading). We can consider that the mean eye blink rate while resting should be 15 ± 5 blinks per minute while the average conversation blink rate should approximate and may even exceed 20 blinks per minute. In section 2.2.2, we take a closer look at how blinks are used in sign language communication.

[10] An average of 10.1 blinks per minute for young men against 11.1 blinks per minute for older men.

[11] An average of 16 blinks per minute for young women against 23.3 blinks per minute for older women.

2.2.2 Eye blinks in sign language communication

Blinks, in the context of sign language communication, have been studied in a few sign languages among which we find ASL (Wilbur, 1994) and HKSL (Sze, 2004) but also LSF (Chételat-Pelé, 2010). In their work, Baker and Padden (1978) note that blinks do not simply serve the purpose of watering the eyes and they point out that signers of ASL blink at constituent boundaries, predicting prosodic phrases. Wilbur (1994) goes further and investigate blinking thoroughly. She aims at defining the linguistic use of blinks in ASL and she observes signers in various elicited situations. The data is manually annotated, glosses were indicated along with pauses, head nods and blinks which are all thought to play a role in sign language prosody. She notes that two types of blinks emerged. The first one is named boundary blinks and is claimed to be involuntary or periodic blinks[12] (which are defined as the ones whose function is to water the eyes but which rate depends on the cognitive load). Boundary blinks can have different functions, namely marking syntactic phrases, prosodic phrases, discourse units, and narrative units. The other category forms the lexical blinks: these are said to belong in the voluntary blink category which she defines by their length: they have a longer duration than periodic blinks. Wilbur (1994) adopts the "75% rule" to account for blinks which category was unclear: the duration of the blink would have to overlap with the duration of the sign for at least 75% of its duration to be considered a lexical one, in other cases, the blink was classified as a boundary blink. It is explained that the duration of lexical blinks is longer than that of boundary blinks, therefore she attributes them a semantic or prosodic function of marking assertion, stress or emphasis. This would be in line with the findings of Brentari and Crossley (2002). She observes that boundary blinks tend to occur at Intonational Phrase boundaries:

> We observed that in ASL, eyeblinks occur more regularly at these locations than do any other nonmanuals: pauses, head nods, or eye gaze changes. (Wilbur, 1994)

Brentari and Crossley (2002) note that blinks co-occur with body leans in the marking of contrastive focus in ASL.

Sze (2004) investigates the relation between blinks and intonational phrases in Hong Kong Sign Language. She rejects the two categories defined by Wilbur (1994) explaining

[12] What Wilbur (1994) names involuntary or periodic blinks was defined as spontaneous blinks in our previous sections.

that these cannot account for all of the blinks exhibited by HKSL signers. She notes:

> In particular, [...] blinks induced by physiological factors are unlikely to serve linguistic functions and [...] blinks occurring toward the end of or after a sign possibly co-occur with syntactic boundaries of constituents equivalent to or smaller than a clause. (Sze, 2004)

Sze (2004) also adds that a large rate of blinks appear alongside head movements and gaze direction change; these observations would need to be taken into account to avoid the overestimation of blinks carrying a linguistic function. Sze (2004) addresses some of the weaknesses of Wilbur's (1994) work. Wilbur (1994) omits to define the duration of the blinks. Since blinks can be divided into three phases: closing, closed and reopening phases, it is important to explicitly define the duration. The same way, the measurement criteria relative to the manual signs are not signaled, therefore the rule Wilbur (1994) put into place cannot be reproduced in other studies with the assurance that it is done the same way the original researcher did it. Sze (2004) further adds that the two categories described by Wilbur (1994) fail to account for blinks that may not be purely linguistic blinks and whose occurrence could be attributed to other factors. In Sze's 2004 study, the annotation of the blinks is more detailed: the closing of the lid, the closure itself and the reopening of the lids are coded separately and blink duration is calculated only using the closing and closure of the lid: the reopening of the lid is not taken into account. Eventually, Sze (2004) divides blinks into five categories:

- Blinks induced by physiology
- Blinks occurring at boundaries
- Blinks co-occuring with head movement and gaze change
- Lexical blinks
- Blinks accompanying hesitation or self-correction

In her data, boundary blinks are the most common type of blinks, they cover almost 70% of all the blinks, while hesitation/self-correction blinks are the next most common category, only 10% of the blinks are of this type. Blinks induced physiologically represent about 7% of the blinks in the data, the head movement/eye gaze change blinks 6% of the data and lexical blinks appear to be the minority category with only about 5% of the blinks falling into this category (Sze, 2004).

In a follow-up article, Sze (2011) claims that the appearance of blinks at different types of grammatical boundaries prevents her from assuming that blinks may function as topic markers in HKSL even though they are the most frequent non-manual marker surfacing after fronted object topics, as they arise less than 50% of the times. The same way, blinks do not seem to occur frequently enough to mark scene-setting topics in HKSL (Sze, 2011).

In her analysis of blinks in French Sign Language, Chételat-Pelé (2010)[13] defined a blink as having two phases: a closing phase and an opening one. She annotated a total of 200 blinks and defined seven different types of blinks:

- Segmentation blinks (31% of the annotated blinks)
- Iconic blinks (22% of the annotated blinks)
- Highlight blinks (18% of the annotated blinks)
- Repetition blinks (10% of the annotated blinks)
- First person blinks (9% of the annotated blinks)
- False question blinks (6% of the annotated blinks)
- Diverse (5% of the annotated blinks)

Segmentation blinks were further divided into four subtypes; phrase segmentation, topic segmentation, pause segmentation, and start/end segmentation (Chételat-Pelé, 2010). Iconic blinks appear in depictive structures. Highlight blinks are concerned with highlight parts of the discourse, namely to highlight a focus, emphasize a pointing gesture, and to frame a correction made to what has been uttered. Repetition blinks, as they suggest, are repeated several times, the duration of the blinks is brief and Chételat-Pelé (2010) indicates that these mark repetition of an action or abundance. First person blinks arise to put emphasis on the fact that the signer is talking about himself. Finally, false question blinks appear at the end of the wh-question followed by its answer. Chételat-Pelé (2010) does not define what is meant by "Diverse" but she mentions in her thesis that the different types of blinks may be accompanied by other non-manuals.

[13] See Braffort and Chételat-Pelé (2011) for detail in English.

Blinks have been investigated in a few sign languages and their description has become richer as the years went on. If Baker and Padden (1978), Wilbur (1994) and Brentari and Crossley (2002) all agree that blinking plays a role in the prosody of sign languages, at least in ASL, Wilbur (1994) distinguishes two types of blinks, namely lexical ones and boundary blinks. Wilbur (1994) and Brentari and Crossley (2002) agree that lexical blinks last longer than boundary blinks. Divergence appears as Sze (2004) investigates blinks in HKSL. She notes that Wilbur's (1994) description of blink types does not consider the fact that not all blinks carry a linguistic functions and that some blinks are purely physiological. Sze (2004) proposes five different blink types which not only take into account non-linguistic blinks but also consider blinks occurring with gaze direction change and head movement, as well as blinks marking hesitation or revision of one's discourse. Later, for LSF, Chételat-Pelé (2010) describes seven types of linguistic blinks, of which only the segmentation blinks were described by Sze (2004) in terms of boundary blinks. Aside from the types of blinks that exists, the definition of blinks given by Sze (2004) and Chételat-Pelé (2010) are also divergent: Sze (2004) considers that the duration of a blink starts as the eye starts shutting and ends as it starts reopening, the reopening phase in itself is not considered while Chételat-Pelé (2010) explains that blinks have two phases, namely the closing phase and the reopening phase, neglecting the phase where the eye is shut.

In this thesis, we consider blinks to start as the eyes start shutting and ends once it finished reopening. In the next section, we will discover methods that have been established to handle the issue of blink automatic detection.

2.2.3 On blinks automatic detection

Automatic detection of blinks is an issue that has been widely studied for the past twenty years (Moriyama et al., 2002; Wang et al., 2009) and that is still widely researched today (Ibrahim et al., 2021; Nousias et al., 2022; Phuong et al., 2022; Dewi et al., 2022). Ibrahim et al. (2021) introduce their research topic by indicating that eye tracking and eye blink detection are popular subjects in today's computer vision research world. Not only is eye blink detection interesting in the context of sign language research but it has been developed in contexts such as for human-computer interactions, or for driver fatigue analysis systems. Before Ibrahim et al. (2021) move on and present their method, they state that to these days, challenges in solving the issue of eye blinks detection remain. In the following paragraphs, we present these challenges along with the methods that have been proposed to detect eye blinks reliably.

Moriyama et al. (2002) are one of the firsts to create an algorithm that is able to detect blinks by tracking motion and appearance information. They list some of the challenges that arise in the development of methods that reliably detect eye blinks. They note that difficulties such as: “rigid head motion, non-frontal pose, occlusion from head motion, glasses, and gestures, talking, low intensity action units, and rapid facial motion” are aspects that make the issue a complex one and these all need to be taken into account when developing a blink detector. Wang et al. (2009) point out the same challenges, adding that technical instability such as variation in the quality or luminosity of images are also difficulties that emerge when looking to automatically detect blinks.

Moriyama et al. (2002) base their method on Ekman and Friesen’s 1978 FACS, namely Facial Action Coding System that categorizes the different movements of the face a human can make by looking at the muscles involved in these motions. The system defines Action Units (AU) as the action of a muscle in question, that is its contraction or its relaxation. In 2007, Lalonde et al. also mention the FACS and its relevance in the context of their study. Lalonde et al. (2007) detected and tracked the eyes on low contrast images to propose a method of eye blink detection that they claim is more stable. They first locate the participant’s eyes using regions of interest, they make sure to keep tracking the eyes and estimate affine transformation[14] between the participant’s head position in the first frame where the participant is supposed to face the camera frontally and the following frames. To detect the blinks, they use an algorithm that detects motion by looking at differences inside the regions of interest between one frame and the subsequent one.

Wang et al. (2009) distinguish two main types of methods used in the detection of eye blinks, namely contour template-based methods and appearance-based methods. The former creates a model of the eye that is based on its shape. Once the model of the eye is made, they perform template-matching to look for the image of the eyes. They note that downsides of these methods include the need for the templates of both open and closed eyes to be processed separately along with a sensitivity to luminosity variation. In the latter type of methods, a binary classification is performed that distinguishes whether the eye is open or closed. These appearance-based methods require a large amount of data. Another downside of these methods is the inability of the classifier to extract the eye contours. In their work, Wang et al. (2009) choose to combine both types of methods.

[14] “An affine transformation is any transformation that preserves collinearity (i.e., all points lying on a line initially still lie on a line after transformation) and ratios of distances (e.g., the midpoint of a line segment remains the midpoint after transformation).” (Weisstein, 2004)

They start by locating the position of the eyes and they define 16 landmarks around the eyes to outline their contour. These landmarks give a representation of an open eye and a closed eye. Afterwards, they use a classifier trained to recognize the landmarks. They use images of participants with open and closed eyes facing the camera. They create an eye blink estimation procedure that involves the calculation of the degree of the eyes open and closed illustrated in figure 2.2.3.

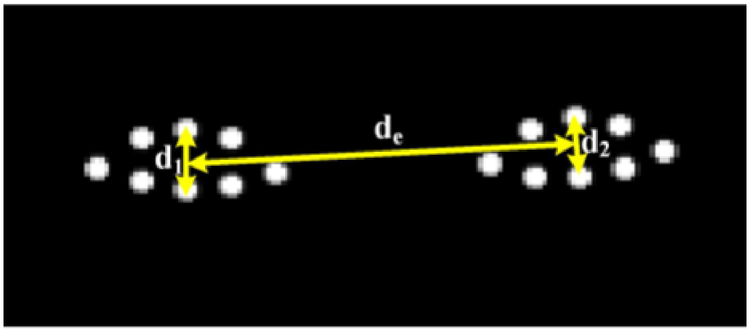


Figure 2.5: Illustration of the degree of eye openness calculation points from Wang et al. (2009).

It calculates the distance between the upper and lower lids of both eyes (d_1 and d_2) as well as the distance between the left eye and the right eye (d_e) as is shown in equation 2.1 where Dl denotes the degree of openness of the left eye while Dr denotes the degree of openness of the right eye.

$$Dl = d_1/d_e, Dr = d_2/d_e \tag{2.1}$$

The calculation outputs a degree included between 0 and 1. In figure 2.6 is an example taken from Wang et al. (2009) where they show instances of an open eye and a closed one for which they give the degree of the eye open which is equal to 0.158 and the degree of the closed eye is 0.016. Eye blink detection relies on the results of this degree calculation. A blink is detected when the degree is lower than 0.02.

Choi et al. (2011) also started by locating the face on the images before detecting the eyes and eventually refining the results. They used the AdaBoost face detector for the detection of a participant's face and the AdaBoost eye detector to detect the eyes. The two AdaBoost detectors process whole images and use cascade classifiers, that is a classifier that is built on the concatenation of many classifiers. The information outputed by one classifier is given to the next. The first cascade classifier, the Haar Cascade Classifier

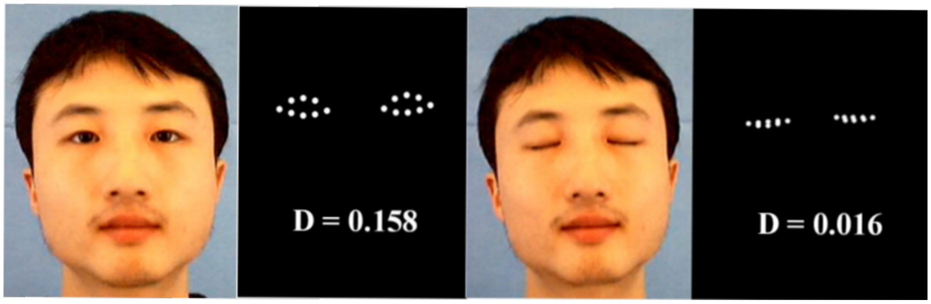


Figure 2.6: Example of the degrees of openness when the eye is open and when the eye is closed taken from Wang et al. (2009).

was created by Viola and Jones (2001) and was mentioned in many works interested in the question of eye blink detection (Wang et al., 2009; Ayudhya and Srinark, 2009; Choi et al., 2011; Ibrahim et al., 2021). In their work, Choi et al. (2011) also use what they call Modified Census Transform (MCT) features[15] as input to their AdaBoost face detector and AdaBoost eye detector. The MCT features of the latter detector use multiple image pyramids[16]. Later, a refinement of the various scanning of the eye images is performed. Eventually, they propose an eye blink detection algorithm based on AdaBoost learning with MCT features that consists of a binary classification, where the eye can be classified as open or closed. They add that their method is suitable for use on smartphones, and other devices with low computing power.

Soukupová and Čech (2016) propose a new way to evaluate eye openness and introduce the Eye Aspect Ratio (EAR). They explain that the EAR is an estimation of the degree of openness of an eye. The computation of the EAR consists of the calculation of the distances between the lower lid and the upper lid (two calculations are made for this distance for each eye) and the distance between the left and right corners of each eye. The exact calculation is shown in 2.2 and the placement of the points P is shown in equation 2.2.3, where 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.

$$EAR = \frac{|| P_2 - P_6 || + || P_3 - P_5 ||}{2 \, || P_1 - P_4 ||} \tag{2.2}$$

[15] MCT features consist of a Census Transform which attibute the pixels in the image in grayscale, a binary encoding giving an information on its pixel's intensity compared to the pixel's neighbors. (Choi et al., 2011)

[16] Pyramids consists of the same image represented in different sizes.(Choi et al., 2011)

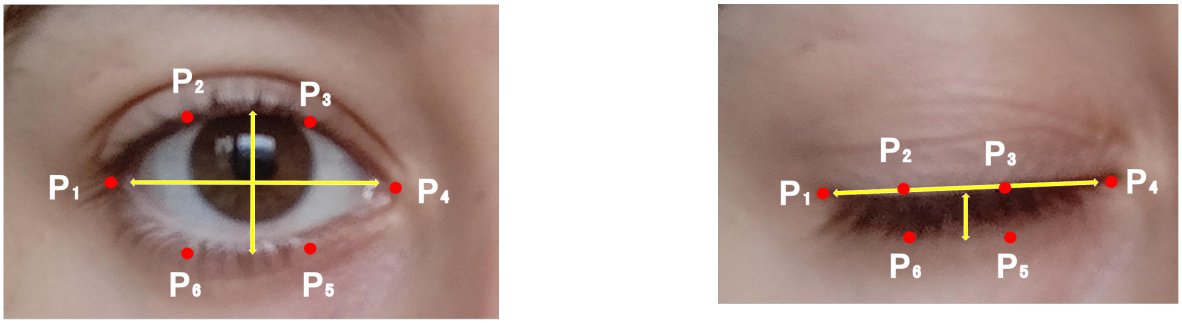


Figure 2.7: Eye landmarks position with open eye and with closed eye.

Soukupová and Čech (2016) note as people blinks happen on both eyes at the same time, they average the EAR of both eyes as one can see in equation 2.3.

$$AVG\ \ EAR = \frac{1}{2}(EAR_{Left} + EAR_{Right}) \tag{2.3}$$

Soukupová and Čech (2016) note that looking at the EAR measure frame by frame may not allow them to detect the blinks dependably, so instead, they train a classifier which takes several frames into account. The EAR is computed for all frames. An issue that was expressed by Lalonde et al. (2007) notably was that head movements were causing errors in facial detection and therefore in the detection of eyes and blinks. Soukupová and Čech (2016) precise that the EAR has the advantage of not being subject to variation when the face rotates horizontally. They also note that a threshold set at 0.2 works well.

Ibrahim et al. (2021) aim to use Raspberry Pi camera and Raspberry Pi 3 to solve an eye blink detection task. They explain that:

> [t]he raspberry pi 3 is a credit card-sized computer that just needs a mouse, keyboard, power supply, display, and micro-SD with an installed Linux framework. After attaching this hardware and software to this small and cheap platform, [they] can run all applications like a traditional computer.

Much like their predecessors, Ibrahim et al.'s Ibrahim et al. (2021) first aim is to detect the faces of the participants and the same way their predecessor did, they use Viola and Jones's 2001 Haar Cascade algorithm to solve the issue. Once this first step is over, they want to retrieve facial landmarks and therefore use an algorithm that detects them. Eventually, they detect the regions of interest where the eyes lie and calculate

Soukupová and Čech's (2016) Eye Aspect Ratio (EAR). They dispute Soukupová and Čech (2016)'s claim that the EAR is not showing variation across participants but they still set their threshold at 0.2, which appear to compromise well the variation in the eye size of the three participants. They consider that when the EAR goes lower than 0.2, the participant is blinking.

Dewi et al. (2022) create a method which solves some of the challenges mentioned at the beginning of this section, specifically, their method show reliability when luminosity changes and when the participants move their head or when the participants have smaller eyes and wear glasses. In their paper, they develop a classifier that categorizes blinks automatically, relying on the primarily detected facial landmarks. To detect the facial landmarks, Dewi et al. (2022) use the Dlib library and their pre-trained implementation of the facial landmarks detector based on the work by Kazemi and Sullivan (2014). The Dlib detector features a total of 68 facial landmarks. Dewi et al.'s 2022 method is based on the creation of a new EAR that they name Modified EAR. Modified EAR is created to counter eye size variation in participants. Instead of having one measurement independent of the state of openness of the eye, they present two measurements: one for the closed eye and one for the open eye. They apply their method in the detection of what they call 'strong' eye blink detection, testing it on different datasets. Their EAR threshold varies from 0.2 to 0.3 and is different from dataset to dataset and even from one video to the other.

To sum up, the development of methods tackling the issues of face recognition and automatic blink detection has been ongoing for over twenty years (Viola and Jones, 2001; Moriyama et al., 2002) and interest in these topics remains as important as ever today (Nousias et al., 2022; Phuong et al., 2022). Over the years, new methods have been designed (Soukupová and Čech, 2016) but some challenges related to the study's environment or to participants' specific features are persistant and we have seen that latest works have started to address these questions (Soukupová and Čech, 2016; Ibrahim et al., 2021; Dewi et al., 2022).

Now that we know more about non-manual markers in sign languages from lexical and grammatical non-manuals to non-manuals carrying a prosodic function, that we have taken a closer look at blinks, their physiology as well as their function in sign language communication, and that we have learnt about research on eye blink automatic detection, it is time to describe the methods used to meet the aims of this thesis. No research so far has been devoted to the automatic annotation of blinks' functions in signed languages. This is what this thesis aims to do. In the following section, we sketch the goal of this work and specifically, we describe the methods used to investigate blinks in French Sign Language in order to create an automatic blink detector, the first step we take toward blink automatic annotation. Section 3 is organized into four subsections in which we mention the data and the participants of this study in section 3.1.1, the annotation process is described in section 3.1.2, the qualitative analysis is explained in section 3.1.4 and finally, the automatic blink detection method is presented in section 3.2.

Chapter 3

Methods

This thesis aims at investigating the types of blinks which are used by signers of LSF. The second aim of this thesis is to automatically detect blinks. These two aims put together will lead us toward the achievement of a more global goal, that is the automatic annotation of blinks types.

To achieve our aim, we need to ask ourselves a few questions: what types of blinks are found in LSF data? Do all of these blinks have a linguistic function? What method or what kind of algorithm would allow us to detect blinks reliably? Do rule-based models offer more dependable results than machine learning-based models? If the two types of approach are combined, would the results be better? To provide answers to these questions, we take the following steps: we use the LSF part of the Dicta-Sign Corpus introduced in section 3.1.1.1. The participants appearing in the corpus videos are presented in section 3.1.1. The corpus was partly annotated and indications about the coding of the data are presented in section 3.1.2.1. The blinks were not annotated by the researchers who created the corpus and we therefore had to code them ourselves. Using the Elan Software (see section 3.1.3), we manually annotated the blinks of 26 videos, regardless of their function. The results of this first round of coding are available in section 3.1.3.1. After paying close attention to the various kinds of blinks observed in two of the annotated videos (see section 3.1.4.1), we coded the types of the blinks using a subpart of the original 26 videos, specifically 8 videos. The results are presented in section 3.1.4.2. Eventually, we moved on to the detection of the blinks and we describe our hybrid method combining both approaches in section 3.2.

3.1 Annotations and blink types definition

3.1.1 Data and participants

3.1.1.1 The Dicta-Sign Corpus

The French Sign Language part of the Dicta-Sign corpus, Dicta-Sign–LSF–v2, is used for our study (LIMSI, 2022). Dicta-Sign–LSF–v2 was created in the context of a European project in 2010 and is available in other sign languages, namely British Sign Language, German Sign Language and Greek Sign Language. The corpus gathers videos recordings of loosely elicited content on the topic of European travel (LIMSI, 2022). Nine dyads of signers are having conversations, amounting to a total of 18 signers participating in the project for LSF. Between 3 and 9 tasks are performed by each dyad and the videos were made available together with partial annotation of the data. The annotations consist of glosses for the right hand, the left hand and signs articulated using both hands. Some videos also contain written translation of the signed utterances. In the context of our study, only part of the corpus is used, specifically a subset of the videos that had glosses.

The five participants whose recordings are used in this study belonged to pair 2, pair 4, pair 5 and pair 9. Signers $A11$ and $B15$ are conversing together while signers $B14$, $A9$ and $B5$ belong to the other pairs mentioned but their partners are not included in this study as the detection of their blinks is more arduous, compromised by the wearing of glasses or a head held down for longer periods of time. The age of each signer along with their gender, information on how they came to learn LSF, and whether or not they have Deaf relatives were provided with the corpus and are listed in Table3.1.

Signer	Dyad	Gender	Age	LSF learning	Other Deaf in family
A11	2	F	28	bilingual school	no
B15	2	M	38	primary school	yes
B14	4	F	28	kindergarten	yes
A9	5	F	28	birth	yes
B5	9	F	28	birth	yes

Table 3.1: Participants metadata

3.1.2 Annotations

3.1.2.1 Original corpus annotations

The dataset was made available with partial annotations. Belissen et al. (2020) explain that signs performed on the hands were annotated following the guidelines given in Johnston and De Beuzeville (2016, 15-17) for Australian Sign Language. Three types of signs were identified:

- Fully-lexical signs
- Partially-lexical signs
- Non-lexical signs

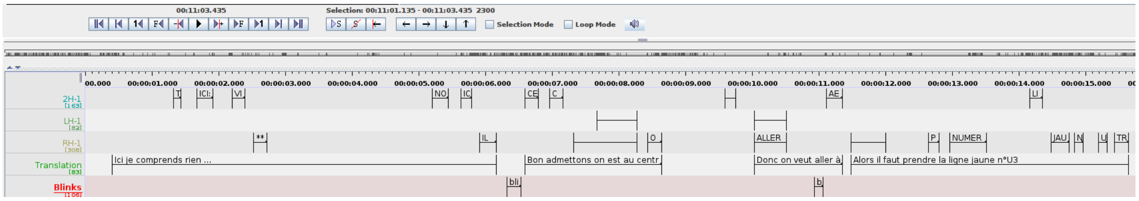
Fully-lexical signs are defined as those that can be registered in a dictionary. Johnston and De Beuzeville (2016) explain that Partially-lexical signs combine elements that are both conventional or dependent on the context of occurrence. Signs belonging in this category are classifier predicates (or polymorphemic signs)[1], and pointing signs. Finally, Non-lexical signs are described as gestures which do not carry a specific meaning or form and which are not specific to a language. Belissen et al. (2020) explain that the Partially-lexical signs annotated in their data could be divided into three categories, namely Pointing Signs, Depicting Signs (classifiers), and Fragment buoys which are defined as “hand shapes held in the signing space”. They also note that fingerspelling and numerals were annotated as Non-lexical signs.

Left-handed, right-handed as well as both-handed signs were glossed. Belissen et al. (2020) indicate that whenever a sign bore several meanings, these meanings were denoted in the glosses (e.g. AIR/FRAIS/PRINTEMPS/MENTHE, “air/fresh/Spring/mint”). If multiple signs represented one unique meaning in French, the French gloss would bear a number (e.g.: GRIS1, GRIS2, “grey”). Finally, Belissen et al. (2020) explain “If the sign if a variant of the usual sign, add: VAR after the gloss (e.g. OUI:VAR)”.

3.1.3 ELAN

In this study, the EUDICO Linguistic Annotator (ELAN) 6.2 software program (Sloetjes and Wittenburg, 2008) is used to further the annotations of the corpus. ELAN is a tool

[1] Classifiers serve to describe the size or the shape of an object or give information on motion or location.



Ici je comprends rien ...

Bon admettons on est au centr

Donc on veut aller à

Alors il faut prendre la ligne jaune n°U3

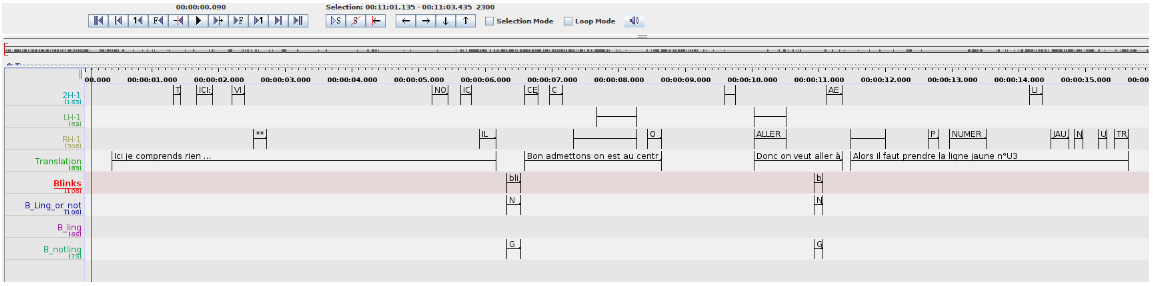
Figure 3.1: Elan partition example: example from the first annotation phase, where only one tier is added, namely 'Blinks'.

used to annotate video and audio recordings. It was initially created for linguistic data and can be used to annotate but also visualize and search the annotations of video and audio files to analyze them.

The annotation process consists of three main steps. First, *tiers* are defined. Tiers are the rows on which the annotations that share similar attributes will be annotated. For reference, *tiers* in the originally annotated corpus, consist of the *tier* for left-handed signs, the *tier* for right-handed signs, the *tier* for both-handed signs, and in some cases, the translation *tier*. Once a tier is defined, the time interval of the excerpt needing annotation is selected. Finally, the selected extract is labeled. Possible labels for specific tiers can be defined beforehand by creating a Controlled Vocabulary that holds a definite number of possible values that can be attributed to elements on a specific tier. Controlled Vocabularies avoids the attribution of an incorrect label to an element (Sloetjes and Wittenburg, 2008).

Our annotation process is divided into two phases illustrated with examples of an Elan partition in figure 3.1 and figure 3.2. In the first instance, blinks are annotated regardless of their types. One tier is created to indicate their presence in the video. In the second phase of the annotation process, three tiers are added. The first consists of a clarification as to whether a blink is linguistic or non-linguistic, a third possible annotation is 'overlap'[2]. The second tier is the linguistic blink tier which is used in order to specify the blink's subtype, it is only used when the blink in question is a linguistic one or again, when there is an overlap between two categories. The third tier consists of the non-linguistic blink tier which is completed to specify the subcategory of the blink in question, only when this blink is annotated as a non-linguistic blink or when there is an overlap. Blink categories and overlaps are described in section 3.1.4.1.

[2] explained in section 3.1.4.1.



Ici je comprends rien ...

Bon admettons on est au centr

Donc on veut aller à

Alors il faut prendre la ligne jaune n°U3

Figure 3.2: Elan partition example: example from the second annotation phase, where three more tiers are added, namely the ‘B_ling_or_not’ tier, ‘B_ling’ tier and ‘B_notling’ tier.

3.1.3.1 Annotations of blinks

Videos were captured at 25 fps and are ranging from about 5500 frames for the shortest video to a bit over 16600 frames for the longest video. The *.csv* documents containing the annotations of the Dicta-Sign–LSF–v2 corpus (LIMSI, 2022) are transformed such that frames are converted into time intervals using a Python script. We write a second Python script to link the annotation ID to the gloss of the sign associated with it in the annotation ID file provided with the corpus in order to see the gloss while annotating the blinks. The annotation ID consists of a 5 digit number referring to the gloss of a sign. The files containing the annotations are then linked to the videos to which they refer and are saved as *.eaf* documents, using ELAN. Signers’ blinks are annotated using the software. As described in section 3.1.3, a single tier is created in the first instance and the only possible annotation is “blink” whenever one occurs. A total of 26 videos are annotated (see table 3.2), representing a total of 2 hours and 59 minutes and 4342 blinks, giving an average of 24 blinks per minute. The average time recorded per signer is of 35 minutes and 36 seconds but it varies from 26 minutes to 44 minutes (see table 3.2).

In her work, Sze (2004) divides the annotation of a blink into three phases, and Chételat-Pelé (2010) divides the blinks into two phases. In our study, we decide to have one annotation per blink. What is considered a blink is the three phases described by Sze (2004), namely the action of shutting the lid, the closure of the lid and the reopening. Chételat-Pelé (2010, 127-129) explains that what is considered closure, is the closure of the lid for over 40 milliseconds, she notes that only the closing phase and the reopening phase are considered blinking. She further observes that in her data which consists of 200 blinks, most commonly, that is for 70% of the blinks, the closing of the lid takes 80 milliseconds, less commonly, specifically for 24% of the blinks, the closing motion takes 120 milliseconds and in rare occasions, namely less than 6% of the blinks, the closing phase takes 40 milliseconds. The majority of blinks, that is 46%, have a reopening phase lasting 120 milliseconds, but Chételat-Pelé (2010) finds that about 37% of the blinks have a reopening phase lasting 80 milliseconds, 14% a reopening phase lasting 160 milliseconds and finally 3% of the blinks have a reopening phase lasting for 200 milliseconds or longer. Considering Chételat-Pelé's (2010) findings, the shortest blink lasts around 160 milliseconds, while the longest do not exceed 380 milliseconds. Even though the method differs and our results are not as detailed, the overall results obtained in our study are similar. The mean blink duration across all signers is of 230 milliseconds (see table 3.2).

Video	Video duration	Total number of blinks	Average blink duration	Median blink duration	Shortest blink duration	Longest blink duration
Signer A11 - Dyad 2						
T1	00 : 11 : 05.000	100	0,23 s.	0,21 s.	0,9 s.	0,5 s.
T2	00 : 03 : 51.000	66	0,256 s.	0,22 s.	0,13 s.	0,65 s.
T3	00 : 04 : 28.000	63	0,2595 s.	0,22 s.	0,13 s.	0,73 s.
T4	00 : 06 : 46.000	116	0,2556 s.	0,25 s.	0,11 s.	0,51 s.
T5	00 : 09 : 15.291	104	0,2339 s.	0,2 s.	0,11 s.	1,95 s.
All videos	00 : 35 : 25	449	0,24 s.	0,22 s.	-	-
Signer B15 - Dyad 2						
T1	00 : 11 : 05.000	229	0,2159 s.	0,21 s.	0,12 s.	0,6 s.
T2	00 : 03 : 51.000	100	0,24	0,22 s.	0,13	0,68
T3	00 : 04 : 28.000	124	0,2044	0,2	0,11	0,4
T4	00 : 06 : 46.000	172	0,2237 s.	0,2 s.	0,11 s.	0,63 s.
T5	00 : 09 : 15.291	156	0,227 s.	0,23 s.	0,1 s.	0,38 s.
All videos	00 : 35 : 25	781	0,2222 s.	0,21 s.	-	-

Signer B14 - Dyad 4						
T2	00 : 05 : 05.360	175	0, 226 s.	0, 21 s.	0, 09 s.	0, 73 s.
T3	00 : 05 : 30.520	87	0, 24 s.	0, 21 s.	0, 11 s.	0, 73 s.
T4	00 : 06 : 14.560	204	0, 25 s.	0, 23 s.	0, 09 s.	0, 63 s.
T5	00 : 05 : 52.520	68	0, 228 s.	0, 22 s.	0, 14 s.	0, 48 s.
T6	00 : 07 : 36.960	270	0, 243 s.	0, 2 s.	0, 06 s.	0, 94 s.
T7	00 : 09 : 41.040	166	0, 219 s.	0, 21 s.	0, 07 s.	0, 55 s.
T9	00 : 04 : 40.840	81	0, 307 s.	0, 29 s.	0, 2 s.	0, 78 s.
All videos	00 : 44 : 38	1051	0, 244 s.	0, 21 s.	-	-
Signer A9 - Dyad 5						
T1	00 : 10 : 12.400	249	0, 239 s.	0, 23 s.	0, 15 s.	0, 94 s.
T3	00 : 05 : 47.680	159	0, 239 s.	0, 23 s.	0, 09 s.	0, 5 s.
T4	00 : 05 : 18.040	170	0, 223 s.	0, 22 s.	0, 09 s.	0, 46 s.
T9	00 : 05 : 21.823	206	0, 236 s.	0, 21 s.	0, 11 s.	0, 39 s.
All videos	00 : 26 : 38	784	0, 23425 s.	0, 225 s.	-	-
Signer B5 - Dyad 9						
T2	00 : 04 : 07.520	156	0, 243 s.	0, 22 s.	0, 11 s.	0, 61 s.
T3	00 : 03 : 46.240	98	0, 238 s.	0, 21 s.	0, 11 s.	0, 66 s.
T4	00 : 10 : 44.240	443	0, 224 s.	0, 22 s.	0, 09 s.	0, 79 s.
T6	00 : 08 : 22.160	280	0, 243 s.	0, 23 s.	0, 06 s.	0, 84 s.
T1	00 : 10 : 35.240	282	0, 204 s.	0, 2 s.	0, 08 s.	0, 48 s.
All videos	00 : 37 : 34	1259	0, 2304	0, 22	-	-

Table 3.2: Blink annotation statistics.

Blinks lasting from 100 to 199 milliseconds and from 200 to 299 milliseconds are most numerous in our data as can be seen on table 3.3. The first category contains a total of 1391 blinks across all videos representing 32% of all blinks, while the second amounts to 2354 blinks across all videos, that is 54% of all blinks in our data. These two categories therefore cover 86% of all blinks annotated in our data.

Video	Total n of blinks	Between $0-99$ ms.	Between $100-199$ ms.	Between $200-299$ ms.	Between $300-399$ ms.	Between $400-499$ ms.	lasting over 500 ms.
Signer A11 - Dyad 2							
T1	106	2	37	52	12	2	1
T2	71	0	21	33	10	5	2
T3	63	0	23	25	6	5	4
T4	116	0	20	73	16	6	1
T5	104	0	48	44	5	4	3
Total A11	460	2	149	227	49	22	11
Signer B15 - Dyad 2							
T1	234	0	88	127	15	2	2
T2	100	0	27	57	13	1	2
T3	126	0	51	72	2	1	0
T4	172	0	74	74	17	5	2
T5	156	0	40	104	12	0	0
Total B15	788	0	280	434	49	9	6
Signer B14 - Dyad 4							
T2	175	1	74	78	15	1	6
T3	87	0	32	45	4	1	5
T4	204	5	63	84	36	9	7
T5	68	0	19	44	3	2	0
T6	270	1	122	95	25	11	16
T7	166	2	66	81	13	2	2
T9	81	0	3	42	27	3	6
Total B14	1051	9	379	469	123	29	42
Signer A9 - Dyad 5							
T1	249	1	47	170	30	0	1
T3	159	1	27	111	17	2	1
T4	170	1	45	113	10	1	0
T9	206	0	68	132	5	0	1

Total A9	784	3	187	526	62	3	3
Signer B5 - Dyad 9							
T1	282	1	115	157	7	2	0
T2	156	0	47	79	18	8	4
T3	98	0	34	47	11	3	3
T4	443	2	131	248	42	16	4
T6	280	1	69	167	32	5	6
Total B5	1259	4	396	698	110	34	17
All	4342	18	1391	2354	403	97	79

Table 3.3: Distribution of blink duration.

Unlike Sze (2004), who considered the duration of the closing of the lid and the complete closure as being part of the blink and who did not include the reopening of the eye, in this study, the duration of a blink covers not only the two first phases but also the reopening of the lid. This is motivated by the definition of a blinks given by physiologists, though it seems that there does not exist a consensus as to which part of a blink event are actually considered part of the blink as we have seen with Sze (2004) and Chételat-Pelé (2010) diverging definitions. As it sometimes happens that the lids only reopen in squinted eye for example, the annotation of the blink stops when the lid excursion is back to what is was prior to the closing or as the lid stops opening further, therefore taking the context into account, that is the frames following the frames in which the blink occurs.

3.1.4 Qualitative analysis

3.1.4.1 Defining the types of blinks

After looking through two videos from both signers of dyad 2 (signers $A11$ and $B15$), specifically $T1$ and $T2$, a list of blink types was drawn up.

The blinks are divided into two main categories (see table 3.4). Within the first category, *non-linguistic blinks*, three subcategories are defined. Within the *linguistic blinks* category, seven subcategories are defined, making a total of ten types of blinks across the different categories.

The *non-linguistic* category is divided into a *change of gaze direction* blinks category, a *signed on the face* or *reflex* blinks category and finally an *involuntary* blinks category. The linguistic blinks are divided into a *addressee feedback* blinks category, a *prosodic* blinks category, a *turn taking* blinks category, a *punctuation* blinks category, a *contrast* blinks category, a *lexical association* blinks category, and eventually an *emphasis* blinks category.

Blink type	**Blink subtype**	**Specifications**
Non-linguistic	Gaze direction (GD)	Change in gaze direction
Non-linguistic	Reflex blinks (SFACE)	Blinks in response to a sign signed on or near the face
Non-linguistic	Involuntary blinks (INVB)	Periodic blinks considered to be physiological
Linguistic	Addressee Feedback (ADFE)	Show that one is listening, show agreement (may be accompanied by a head nod or the sign 'OUI'.), show disagreement (may be accompanied by a headshake), add complementary information after receiving feedback.
Linguistic	Prosodic blinks (PROS)	Act as prosodic boundary markers
Linguistic	Turn taking blink (TURNTC)	Appear to mark the beginning of a turn or its end and mark the beginning and end of short pauses
Linguistic	Repetition blinks (REP)	Mark the repetition of an action or punctuate enumerations
Linguistic	Contrast blinks (CONT)	Occur with conjunctions 'OU2 (OU BIEN)' and in comparisons
Linguistic	Lexical association blinks (LEXASS)	Appear with specific signs
Linguistic	Emphasis blinks (EMPH)	Accompany pointing signs

Table 3.4: Types of blinks

Among the non-linguistic blinks, we find three subtypes, that are *gaze change* blinks, *reflex* blinks and *involuntary* blinks. The blinks called *gaze direction* blinks are the ones which happen as a signer moves his gaze in a different direction, they happen with and without an accompanying head movement. These were described in Doughty (2002) where he writes that "it seems likely that far greater changes in SEBR can occur if the experimental paradigm changes, e.g., the subject changes their direction of gaze [...]". In our data, *gaze direction* blinks occur quite often as we explain in section 3.1.4.2. This is primarily due to the location of the screen supporting the elicitation data, which is placed in front of the signer but below their eye gaze as their addressee was sitting directly in front of them. The contribution of a third person monitoring the exchange would also trigger these head movements in turn activating a blink response.

Reflex blinks in general are triggered by contact on or close to the eye as explained in section 2.2.1.2, thus, it seems evident that signs produced on the face and close to the eyes may not carry a linguistic function but may instead arise as a physiological response. In LSF, such blinks appear quite consistently with signs such as MUSÉE ('museum') or PENSER ('to think').

Finally, *involuntary* blinks are devoid of any linguistic meaning. Those blinks were described as spontaneous in previous sections and that Wilbur (1994) also calls involuntary, although her involuntary blinks cover more types than the ones described here and carry a linguistic function which is not the case in this present work. These have a physiological function and arise when the signer has not blinked for a certain period of time. Sforza et al. (2008) have written that they occur around 15 times per minute but in our context, they are not really quantifiable. These are sometimes difficult to differentiate from linguistic *addressee feedback* blinks as they tend to arise when the signer is not signing, although the contraction of the muscles involved in the production of *addressee feedback* blinks seems stronger. *Involuntary* blinks may also be triggered by a previous blink that may not contain a full closure of the eyelids.

Among the linguistic blinks, *addressee feedback* blinks are those which we can describe as facilitating communication in a way. They arise as a response to something that the other person has signed, or more generally, are a sign of the addressee's attention.

Prosodic blinks are various and appear at the edge of prosodic domains (or syntactic ones for the two are sometimes difficult to differentiate), marking boundaries. We find that such blinks are used in the expression of consequence, following the introduction of the cause. *Prosodic* blinks are also found when signer introduce a modification or propose a correction to their speech (an instance of this blink is available in example 21 with the blink occurring on the second iteration of 'POUVOIR'), we consider these blinks to mark the beginning of a new information unit. *Prosodic* blinks also seem to occur to place the topic in focus and thus may arise between the topic and the rest of the utterance. Finally, and more generally, *prosodic* blinks punctuate the boundaries of utterances as we show in example 17.

(17) ENVIE QUESTION*[3] **blink** pres1* mer1 pouvoir* nager* pouvoir*

WISH QUESTION **blink** CLOSE SEA CAN SWIM CAN

'I want to ask a question... It's close to the sea so we can go bathe?'[4]

The next linguistic blink category is *turn-taking* blinks. They appear as the signer starts signing and as the signer concludes his/her turn. As the signer begins to sign, the blink may overlap with the first sign. *turn-taking* blinks occurring at the end of a turn sometimes look similar to addressee feedback blinks. Another kind of blink categorized as *turn-taking* in our data were blinks indicating that the signer was taking a quick break to think, example 18 shows an instance of this blink subtype. 'DEUX:NUM:VOT_DEUX1:VAR' is one sign, in example 18, the signer blinks at the start of the sign, the sign is held while the signer thinks and the signer blinks one more time as he starts signing again.

(18) [$\overset{\text{blink}}{\overline{\text{DEUX2:NUM:}}}$ VOT_DEUX1 $\overset{\text{blink}}{\overline{\text{:VAR}}}$]

The following subtype concerns blinks appearing as the signer repeats a sign or enumerates things, these are called *repetition* blinks. In our data, we note that signer *A*11 repeats 'CHANGER' three times, and each of these iterations is accompanied by a blink. *Repetition* blinks also arise as the signer lists steps to take to achieve an action, lists numbers and lists items. In these occasions, the blinks seem to act as commas.

(19) **blink**[5] $\overset{\text{blink}}{\overline{\text{UN:NUM:VAR}}}$ IL-FAUT* $\overset{\text{blink}}{\overline{\text{APPO}}}$RTER PHOTOGRAPHIE* TON* $\overset{\text{blink}}{\overline{\text{DEUX2:NUM:*}}}$ AVEC SEUL1 FAMILLE/EQUIPE

blink $\overset{\text{blink}}{\overline{\text{ONE}}}$ ONE-NEEDS $\overset{\text{blink}}{\overline{\text{BRING}}}$ PHOTOGRAPH YOUR $\overset{\text{blink}}{\overline{\text{TWO}}}$ WITH ONLY FAMILY/TEAM

[3] All signs followed by '*' appear in their full form in Appendix 1.

[4] Original annotator's translation: 'J'ai envie de poser une question... C'est proche de la mer donc on peut aller se baigner ?'

[5] **turntc**

‘You have to bring a photo of yourself. Are you going by yourself or with your family?’[6]

Contrast blinks may occur with conjunctions, specifically with the sign ‘OU2 (OU BIEN)’ (or (or else)) as is demonstrated in example 20.

(20) CHOISIR DEUX2:NUM:* PIED-À *#indecipherable#* $\overset{\text{blink}}{\overline{\text{OU2/(OU BIEN)}}}$ DEUX2:NUM:* TRANQUILLE1* DEUX2:NUM:*

CHOOSE TWO FOOT-ON *#indecipherable#* $\overset{\text{blink}}{\overline{\text{OR/(OR ELSE)}}}$ TWO COOL TWO

‘So we can choose two solutions: by foot or else really play it cool.’[7]

Contrast blinks may also occur as the signer compares two things. In such circumstances, two blinks are produced, one with the first part of the comparison, the second with the second half of the comparison as is shown in example 21 where the signer can choose to get their ticket physically or online.

(21) BILLET:VAR $\overset{\text{blink}}{\overline{\text{SUR1}}}$ PLACE1 $\overset{\text{blink}}{\overline{\text{INTERNET}}}$ PAGE3 PAPIER* POUVOIR* $\overset{\text{blink}}{\overline{\text{POUVOIR*}}}$ INTERNET $\overset{\text{blink}}{\overline{\text{COMMANDER}}}$ /ORDRE* $\overset{\text{blink}}{\overline{\text{PARFOIS}}}$*

TICKET $\overset{\text{blink}}{\overline{\text{ON}}}$ PLACE $\overset{\text{blink}}{\overline{\text{INTERNET}}}$ PAGE PAPER CAN $\overset{\text{blink}}{\overline{\text{CAN}}}$ INTERNET $\overset{\text{blink}}{\overline{\text{ORDER}}}$ /COMMAND $\overset{\text{blink}}{\overline{\text{SOMETIMES}}}$

‘Ah yes yes, you can also you can also go get your ticket on site or in an agency. Or online indeed, so you order.’[8]

Lexical association blinks arise with specific lexical items. Examples of signs accompanied by these blinks include: TOTAL (‘total’), BERLIN (‘Berlin’), JUSTE/PRÉCIS (‘accurate/precise’), PERMUTER (‘to permute’), OUBLIER (‘to forget’), PENSER (‘to think’), OU2 (OU BIEN) (‘or (or else)’), MAIS (‘but’), SUPER/BIEN (‘super/good’), or again IL-N’Y-A-PAS/PERSONNE (‘there is not/no one’). As you may notice, PENSER made it into the list but in truth, it is complicated to confirm or deny whether the blinks which occurs with PENSER is a *reflex* blink or a lexical association one.

[6] Original annotator’s translation: ‘Il faut apporter une photo de vous. Vous partez seul ou en famille ?’

[7] Original annotator’s translation: ‘Alors on peut choisir deux solutions: à pied ou alors vraiment se la jouer cool.’

[8] Original annotator’s translation: ‘Ah oui oui tu peux aussi tu peux aussi aller chercher ton billet sur place ou dans une agence. Ou par internet effectivement donc tu commandes.’

The last subtype of the linguistic blink category is the *emphasis* blink category which arise with the index pointing sign to emphasize the reference the signer is making.

Looking at the work from Chételat-Pelé (2010), the blinks categorized as *first person* blinks are classified as *reflex* blinks (or *signed on the face* blinks) in our study and are not considered to be linguistic blinks: as mentioned by Chételat-Pelé (2010), they often occur with signs like THINK, KNOW or HEAR which are signed on or near the face and are considered to be the result of a physiological response.

A type of blinks defined in Chételat-Pelé (2010) are segmentation blinks which she subdivide into four smaller categories, namely *phrase segmentation*, *topic segmentation*, *pause segmentation*, and *start/end segmentation*. In this present work, the main *segmentation* blinks category as described by Chételat-Pelé (2010) is divided into two main categories, namely *prosodic* blinks and *turn taking* blinks. The former encapsulates *phrase segmentation* blinks. The latter, *turn taking* blinks, is the equivalent of Chételat-Pelé's (2010) *start/end segmentation* blinks and *pause segmentation* blinks. Chételat-Pelé (2010) also use a *highlight* blinks category which gathers blinks occurring with a pointing gesture, in our data, these were annotated as *emphasis* blinks. This *highlight* category also contains elements put in focus in discourse, these are categorized as *prosodic* blinks in our case. Corrections in discourse as well as added information are surrounded by blinks, these are categorized as *highlight* blinks by Chételat-Pelé (2010), while in our study, these blinks belong in the *prosodic* blinks category. Our *prosodic* blinks category also contains blinks serving the expression of consequence.

Chételat-Pelé (2010) define a *repetition* blinks category which description is similar to that of our *punctuation* blinks category, although they do not mention enumerations in their work.

Finally Chételat-Pelé (2010) describe *Wh-questions with answer* blinks category. In our study, these blinks are annotated as *prosodic* blinks.

We define three other linguistic categories: the *addressee feedback* blinks category, the *contrast* blinks category, and the *lexical association* blinks category. The *addressee feedback* blinks are most often found as the person appearing in the video watches and reacts to the addresser's discourse. They are used to show attention but may also be used to display agreement or disagreement in which case they co-occur with a head nod or a headshake. They can potentially be used as the signer is signing, to confirm that the signer has received/understood his addressee's feedback. The *contrast* blinks category contains instances of blinks occurring with conjunctions or in comparisons. Finally, the *lexical association* blinks category occur with specific lexical signs in a 'systematic' manner.

Signs making it into the list are signs which co-occur with a blink in most occurrences. These blinks may depend on specific signers or may also belong to the *reflex* blinks category as they are cometimes signed on the face.

3.1.4.2 Annotation of the data according to their blink type

After defining the types of blinks, a subset of the originally annotated videos is annotated once more, this time more precisely. To annotate this data, three tiers are created as described in section 3.1.3. The detail of these annotations according to a blink's type is given in table 3.5.

Category	A11 T1	A11 T2	A11 T3	B15 T1	B15 T2	B15 T3	B14 T2	B14 T4	**TOTAL**
GD	36	6	14	45	2	4	3	20	130
SFACE	0	1	0	2	0	0	0	1	4
INVB	4	3	1	18	7	2	8	0	43
ADFE	10	13	20	46	40	22	85	13	249
PROS	13	15	6	29	21	42	32	67	225
TURNTC	2	7	3	18	10	12	15	6	73
REP	5	5	2	6	7	6	2	2	35
CONT	1	6	0	0	0	1	0	0	8
LEXASS	0	0	0	0	2	1	2	2	7
EMPH	0	0	0	0	1	0	0	1	2
Overlaps	35	15	15	66	9	36	27	89	292
TOTAL	106	71	61	230	99	126	175	201	1068

Table 3.5: Blinks annotation according to their types.

- For Signer $A11$, video $T3$, the linguistic count does not add up to 31 but only to 29: one of the blink was annotated as an overlap between the subtypes ADFE and LEXASS. The other was annotated as an overlap between the subtypes PROS and EMPH.
- For Signer $B15$, video $T1$, the non-linguistic count does not add up to 67 but only to 65: one of the blink was annotated as an overlap between the subtypes SFACE and GD and the other one was annotated as an overlap between the subtypes GD and INVB. The linguistic count does not add up to 101 but only to 99: one of the blink was annotated as an overlap between the subtypes TURNTC and PROS while the other was annotated as an overlap between the subtypes PROS and EMPH.
- For Signer $B14$, video $T2$, the non-linguistic count does not add up to 12 but only to

11: one of the blink was annotated as an overlap between the subtypes GD and INVB.
- For Signer B14, video T4, the non-linguistic count does not add up to 22 but only to 21: one of the blink was annotated as an overlap between the subtypes SFACE and GD. The linguistic count does not add up to 93 but only to 91: one of the blink was annotated as an overlap between the subtypes PROS and ADFE and the other was annotated as an overlap between the subtypes TURNTC and ADFE.

A total of 8 videos from three signers are selected and 1068 blinks are annotated according to their type in this second round. Three videos from signer A11 are annotated. They gather 240 blinks of which 110 are linguistic blinks, 65 blinks are non-linguistic blinks and the remaining 65 blinks are overlaps between linguistic and non-linguistic categories. Three videos from signer B15 are annotated. These videos gather 460 blinks of which 267 are annotated as linguistic blinks, 81 are marked as non-linguistic blinks and the last 112 blinks consist of an overlap between linguistic and non-linguistic blink types. Finally, two videos from signer B14 are annotated. They concentrate 379 blinks, 228 of which are categorized as linguistic blinks, 34 as non-linguistic and the 117 that are left are annotated as overlaps both of linguistic and non-linguistic types.

The systematic annotation of the blinks according to their types (described in section 3.1.4.1) in the videos show that the division of our blinks may have been too detailed. Some of the blinks, namely *lexical association*, *contrast*, and *emphasis* blinks do not appear abundantly in our data. We observe that *contrast* and *turn taking* blinks could be gathered into one category with *prosodic* blinks as both could be seen as marking the edges of clauses and utterances. We also note that it is sometimes difficult to attribute a unique category to a blink, hence the presence of so many overlaps.

We occasionally find overlaps between two subcategories, but occasionally we see them between three subcategories; these are presented in table 3.6.

Category	A11	A11	A11	B15	B15	B15	B14	B14	**TOTAL**
	T1	T2	T3	T1	T2	T3	T2	T4	

OVADGD	16	1	9	11	0	3	10	3	43
OVPRGD	9	5	2	38	1	19	2	69	145
OVTUGD	9	3	2	3	1	1	2	2	23
OVPUGD	0	2	0	1	0	0	0	3	6
OVCONGD	1	1	0	0	0	0	0	0	2
OVLEGD	0	0	0	0	0	1	0	3	4
OVADINV	0	2	2	9	6	6	12	0	37
OVPRINV	0	0	0	1	0	2	0	4	7
OVLESF	0	1	0	0	0	4	1	2	8
OVPRSF	0	0	0	2	0	0	0	1	3
OVTUSF	0	0	0	1	1	0	0	0	2
OVPRSFGD	0	0	0	0	0	0	0	1	1
OVLEPRGD	0	0	0	0	0	0	0	1	1
Linguistic OV	0	0	2	2	0	0	0	2	6
Non-linguistic OV	0	0	0	2	0	0	1	1	4
TOTAL	35	15	17	71	9	36	28	90	302

Table 3.6: Detail of the occurrences of overlapping categories in the annotation of the blinks according to their types.

We notably observe that the *lexical association* blinks category is most often than not used in overlaps with another subtype of blinks (13 occurrences of *lexical association* blinks overlapping with another subcategory against 4 *lexical association* blinks captured on their own.), be it a non-linguistic subtype or a linguistic one. Most of the time, two types of blinks may be attributed to the blink in question and usually, it concerns a linguistic type of blink along with a non-linguistic one. In two occasions, the hesitation concerns three subcategories. In the first case, one linguistic category, namely the *prosodic* subcategory and two non-linguistic ones, specifically the *signed on the face* and the *change in gaze direction* subtypes are found to overlap. In the second case, two linguistic categories, specifically, *prosodic* and *lexical association* and one non-linguistic category, namely the *change in gaze direction* subtype are found to overlap. In a few occasions, the indecision involves two subtypes of the same main category (these are reported in detail at the end of table 3.5). The most common overlap is that of a *prosodic* blink and a *gaze direction* blink. This kind of overlap occurs 145 times across the 8 videos, so that 48% of overlaps are co-occurrences of these two categories.

Now that we have annotated the blinks carefully and that we know more about the different types of blinks that exist, let us move on to the automatic detection of blinks.

3.2 Automated blink detection.

Researchers who have worked on the issue of eye blink detection have usually created methods that do not respond to whether a blink is happening or not but instead which tackle the problem of whether the eye is open or closed as we observed in section 2.2.3. Dewi et al. (2022) notably described blinks as follows:

> 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.

This definition is different from that given by physiologists and linguists. Specifically, 'blinks' as defined by physiologists can exhibit an incomplete closure of the lid. As described by linguists, 'blinks' are to be distinguished from 'closures' which last longer than blinks and which do not carry the functions and meanings blinks have in communication.

Researchers who have binarized the problem and considered it as an *open* or *closed* classification task have disregarded the *in-between* frames which show eyes that are neither *open* nor *closed* but rather which exhibit a lid that covers half of the eyeball, because from one frame to the other the state of openness of the eye lid changes. A blink extends on several frames as exhibited in figure 3.3 in which we can see that the only frame where the eyes are actually *closed* is the second frame.

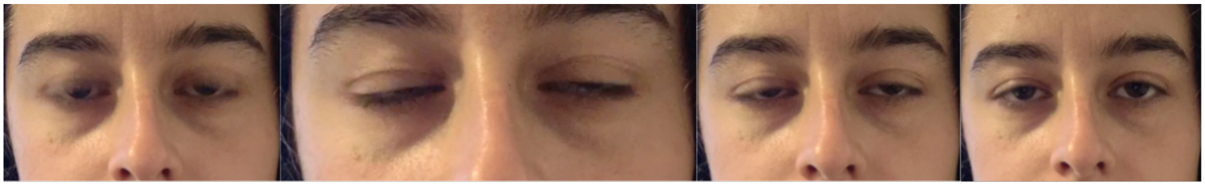


Figure 3.3: Illustration of a blink frame by frame -
Picture used for illustration purposes, not taken from the corpus.

Blinks last between 199 and 300 milliseconds in most occurrences and hence extend over 5 to 8 frames (respectively representing 200 and 320 milliseconds), therefore making a decision as to whether a blink is happening or not on a per-frame basis is an ill-posed problem.

Instead of returning a prediction for one frame at a time, we propose to classify blinks on windows of frames, creating a method based on a 'temporal analysis' of the frames.

This task is difficult as two aspects need to be taken into account, namely, the state of the eyes or their degree of openness needs to be detected before we agglomerate these degrees of openness together to make a decision as to whether a blink is happening over time.

As we have two subproblems, we divide the method into two components, phase one being the detection of the eyes' openness state and component two being the agglomeration of the degrees of openness over time. The method handling the first matter is presented in section 3.2.1, while the second part of the method, handling the agglomeration of the frames and the decision as to whether a blink is occurring is presented in section 3.2.2.

3.2.1 Phase one: detecting the state of openness of the eye using machine learning

To detect the state of openness of the eye, we implement three methods. The first is based on the Eye Aspect Ratio (EAR) calculation, it is presented in section 3.2.1.1. The second and third method are machine learning-based methods, one is based on the implementation of a classifier and is introduced in section 3.2.1.2.3 while the other concerns the implementation of a regressor detailed in section 3.2.1.2.4.

3.2.1.1 Eye Aspect Ratio - EAR

In this thesis, we use the Eye Aspect Ratio (EAR) calculation as a baseline to evaluate the machine learning method we aim at developing.

To calculate the EAR, one needs to extract eyes coordinates. We use MediaPipe Holistic (Grishchenko and Bazarevsky, 2020) which compiles models for the detection of body pose, hand tracking and facial landmarks detection. MediaPipe Holistic is therefore more complete than Mediapipe Face Detection solution or Face Mesh solution which respectively detect faces and estimates close to 500 facial landmarks. Using Holistic solution allows us to add additional features if necessary.

We bring down the features taken into account by only selecting a subset of 6 eye coordinates for each eye, making it a total of 12 coordinates for both eyes. These coordinates correspond to the 6 points used to calculate the EAR (Soukupová and Čech, 2016) which gives a value between 0 and 1 giving an estimation of the eye's aperture. The specific points are shown again in figure 3.4, marked by the red circles while the yellow arrows represent the eye height and width used to calculate the EAR.

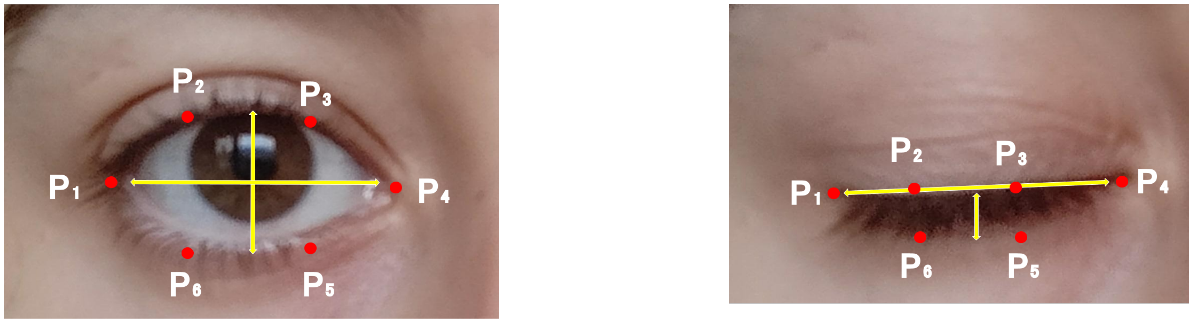


Figure 3.4: Eye landmarks position with open eye and with closed eye.

The EAR is computed for each frame using the detected landmarks, the calculation is shown again in equation 3.1. As the EAR typically ranges between 0 and 1, a high EAR value means that the eye is open while when the EAR value is low, that is closer to 0, we can expect the eye to be shut. Soukupová and Čech (2016) use the Eye Aspect Ratio to detect blinks in their work. They note that the EAR value is stable across frames when the eye is open, they also explain that the EAR does not exhibit variance across individuals though this last statement has been contested by other researchers (Ibrahim et al., 2021; Dewi et al., 2022).

$$EAR = \frac{|| p_2 - p_6 || + || p_3 - p_5 ||}{2 \, || p_1 - p_4 ||} \tag{3.1}$$

As mentioned in section 2.2.3, researchers who have used the EAR measure in the implementation of a blink detector have set thresholds of eye open vs. eye closed between 0.2 and 0.3. In our case, we test a variety of thresholds to see which best fits the signer whose blinks are being analyzed. These will be described in more detail in sections 3.2.2.1 and 3.2.2.2.

3.2.1.2 Machine learning-based method

In the following sections, we mention modifications brought to the dataset for the data to be used in the implementation of machine learning models to complete our first phase. We also present these machine learning models in detail.

3.2.1.2.1 The dataset

In order to use machine learning methods, one needs to prepare a dataset. As we have chosen to use the images of the signers, we need to save the videos as images. We specifically create two different crops of each frame, one focusing on the face of the signer, the other one zooming on the eyes.

We use the Mediapipe Face Landmarks (Grishchenko and Bazarevsky, 2020) to determine the area that needs to be cropped. Depending on the frames, the crops do not have the same dimensions. The head resembles a square or can easily be encapsulated in a square shaped crop while looking at the eyes, using the shape of a rectangle is more obvious. The area where the eyes are located is also smaller than that of the face which by definition includes the eyes. For this reason the resizing of the eyes is different from that of the face and smaller.

We save the images into two different folders, namely the *open* and *closed* folders. Frames making it in the *closed* folder are for the most part the frames on which the blinks are happening in the original round of annotations.

The files containing these original annotations that we will now call groundtruths were exported to a *.csv* file format from ELAN. They are organized such that we have one file per video and each file is a tabular dataset consisting of 7 important columns. The number of row varies depending on the number of blinks annotated for the video, that is, the data from one row covers the needed information for one blink. The first column is the name of the tier. The second column contains the starting time of the blink in hour format[9]. The third column contains the same information as column two but in second format[10]. The fourth column contains the end time of the blink in hour format while the fifth provides the same information in second format. The sixth column contains the duration of the blink in hour format and the seventh column contains the equivalent in second format.

We have previously attributed one annotation per blink. One annotation covers the three phases of the blinks. The annotations were created on a temporal basis, that is not at the level of individual frames. For this reason, frames that are now annotated as *closed* do not systematically display an entirely closed eye, that is the lid does not fully cover the eyeball. The same way, images contained in the *open* category sometimes display part of the lid covering the eyeball. Hence, this reinforces the idea that the task might not be binary (*open* vs. *closed*) and we decide to introduce a third class, namely the *in-between* category which gathers all of the instances that we consider ambiguous, instances in which

[9] hour format: 00:00:00.000

[10] second format: 00.000

the eye is neither fully open nor fully closed specifically when the lids are in motion or when the openness of the eye is difficult to infer in instances in which the signer has the eyes somewhat open but keep his head down. Finally, the *in-between* folder also gathers instances where the eyes are not visible at all. As it would take too long to annotate the whole dataset again, we choose to focus on one signer and to annotate two videos, namely video $S2T1B15$ and video $S2T2B15$. The former will be used for the training of Convolutional Neural Networks models, namely classifiers and regressors introduced in section 3.2.1.2.3 and 3.2.1.2.4 respectively. The latter video, $S2T2B15$, will be used to evaluate our machine learning algorithms on unseen data. For this re-annotation process, we look at the frames of the video one by one and make a decision based on the degrees of openness of the eye for each frame of the video. As we have two different types of crops, namely one face crop and one eye crop for each frame, we write down the frame numbers of all the face crops that require moving to the *in-between* category to make sure that the equivalent eyes crops are moved to that class as well. Proceeding that way allows us to perform a verification round.

Two things require attention here: the first is that it was sometimes difficult to decide whether the eye should be considered as belonging in the *open* or the *in-between* folder. We decide that when the iris is visible, it is considered open if the signer is looking directly at their addressee, in instances in which the signer is looking down, we consider the frames as belonging to the *in-between* folder as part of the lid obstructs part of the iris. The second observation concerns the decisions which might differ depending on the type of crops one is looking at. What sometimes seems almost closed in the face crops data appears more ambiguous in the eyes data. In such cases, we consider the decision made on the face crops to be more important and we keep both crops in the *in-between* folder.

3.2.1.2.2 Convolutional Neural Networks

We decide to use a Convolutional Neural Network model as we are working with images and CNNs were made to treat image data.

We create our CNN model inspired by the classic LeNet-5 architecture (LeCun et al., 1998) and consisting of several blocks, each block consisting of a convolutional layer followed by a pooling layer to capture spatial correlation at different scales in the image. The CNN ends with linear (or 'fully connected') layers. Convolutional Neural Networks or CNNs are a type of neural networks that were initially created to handle visual data. CNNs are based on convolution operations, that is, they apply filters, sometimes called kernels, moving on the image[11] represented as a matrix of values and for each kernel, a convolution operation is made. Convolutional layers are intended to capture local spatial correlation within the image.

3.2.1.2.3 Convolutional Neural Network classifier

Our CNN model varies depending on the type of the data, that is depending on whether we use the face crops or the eyes crops. We present a diagram of the CNN for the eyes crops in figure 3.5, while the model for the face crops is shown in figure 3.6.

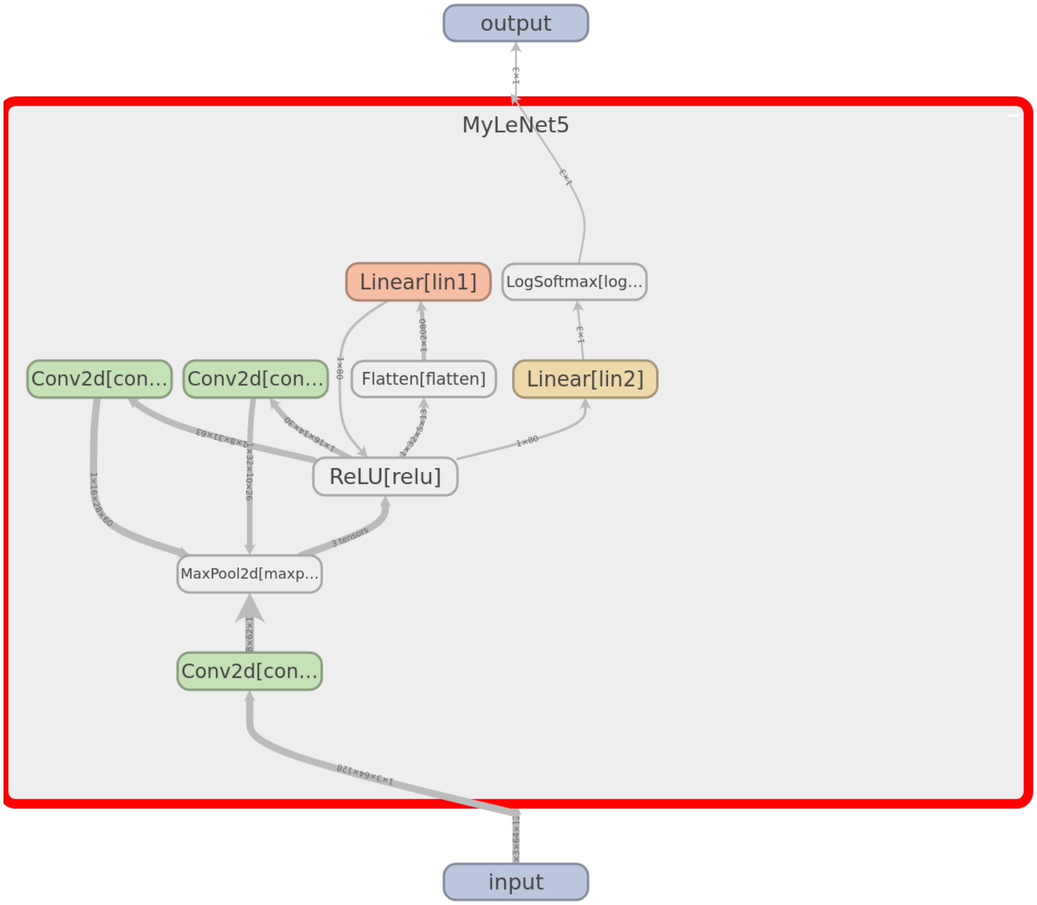


Figure 3.5: Diagram of the CNN classification model for the eyes crops.

[11] This process is also called 'sliding window' procedure

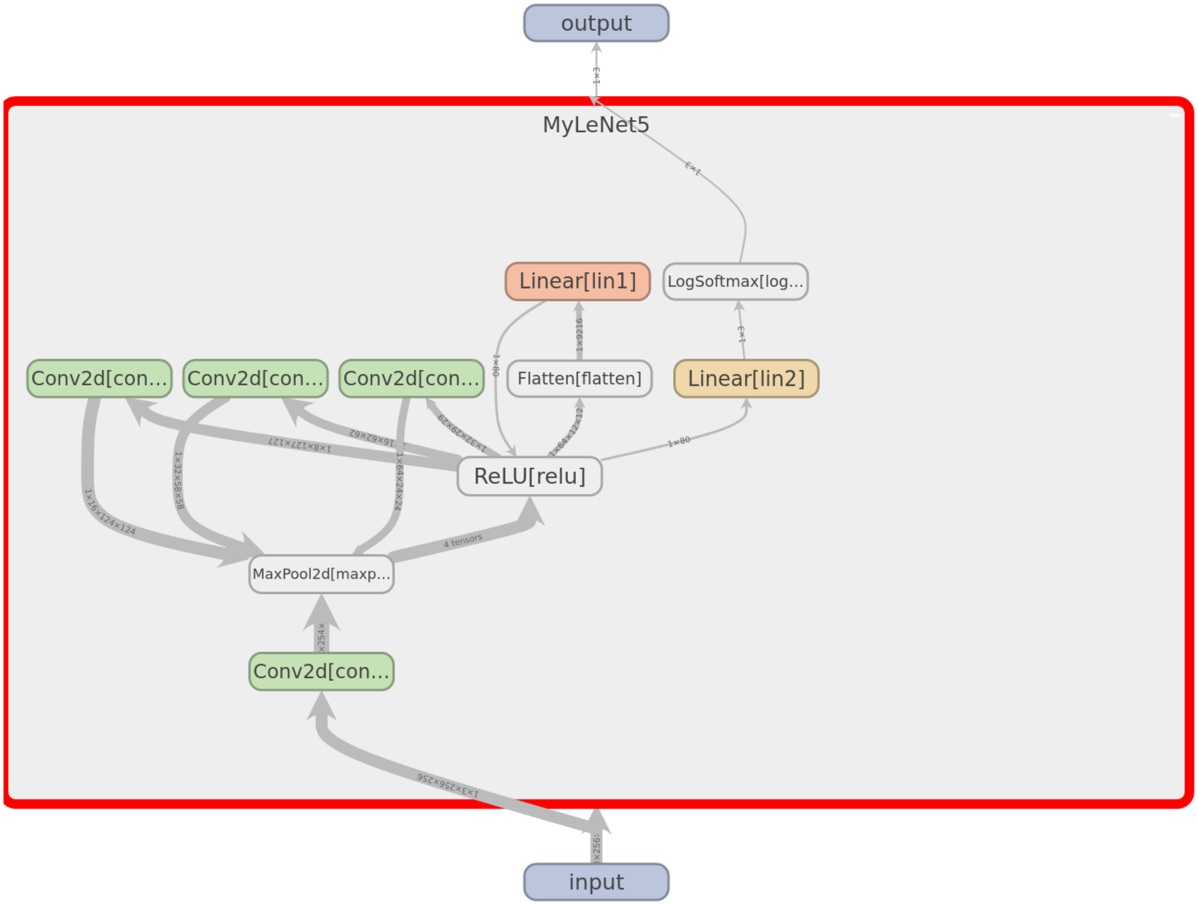


Figure 3.6: Diagram of the CNN classification model for the face crops.

When working with the face crops, the model is a bit more complex and contains one extra convolutional layer than when the classification is performed using the eyes crops to account for the larger spatial dimensions of input images ($256 * 256$ instead of $64 * 128$). The size of the first linear layers also varies depending on the data type, we have 2080 input features of the eyes crops while we have 9216 input features for the face crops. In the face crops model, we have a total of four convolutional layers, each combined with a maxpooling layer, followed by a flattening layer, and two linear (also called 'fully connected') layers. All layers but the last one are followed by the ReLu activation functions to account for non-linearity. For this model, the first linear layer takes 9216 input features and has an output of 80 nodes. Finally, the last linear layer takes 80 input features and has three output features, one per class in the *open* - *in-*

between - *closed* problem. In the eyes crops model, we have three convolutional layers, also combined with maxpooling layers, followed by a flattening layer and two linear layers. All layers, similarly to the face crops model, have a ReLu activation function. The first linear layer of this model takes 2080 as input features and outputs 80 nodes. Because the number of output features of the first linear layer for both the face crops model and the eyes crops model are identical, the second linear layer is identical for both models, taking 80 nodes as input and the three output features representing the three labels.

We use a softmax layer as the last activation layer of our network. The softmax activation function normalizes the output of our model into a vector of probabilities, that is all values in the output vector are in $[0, 1]$ and they sum to 1. We use the cross entropy loss to calculate the distance between the probabilities outputed by the model and the one-hot encoded groundtruths. Finally we use the Adam optimizer, a widely used optimizer, implemented by Kingma and Ba (2014) and which aims at minimizing complex non-linear functions.

The whole experimental pipeline is described in section 4.1.1, where we also report on the preprocessing of the data. We give our training and evaluation results in sections 4.1.2 for the face crops and 4.1.3 for the eyes crops.

3.2.1.2.4 Convolutional Neural Network regressor

We make the observation that the *open* - *in-between* - *closed* eye decision problem can be formulated as a regression problem rather than a classification problem, that is instead of predicting a probability vector for the different classes, we output instead a single value included between 0 and 1, where being close to 0 means we are confident that the eyes are closed and the opposite when the value is close to 1, that is similar to a normalized EAR value.

We also investigate this option and implement a regression model in complement of the classification model. We show the models for the eyes crops in figure 3.7, while the model for the face crops is exhibited in 3.8. In practice the classification and regression models differ in a few ways: in the classification task, labels are treated as categorical; the number of output in the model corresponds to the number of classes: the model outputs a vector of probabilities where one probability is associated to each class. The loss combines a softmax final activation and a cross-entropy function which compares the vector of probability to the one-hot encoded[12] labels which are the groundtruths. In the

[12] A one-hot encoding is a way of representing the categorical labels as vectors of probabilities. It contains as many values as there are classes, in our case, three values per classes but only on

regression model, classes are instead converted into a value in $[0, 1]$ instead indicating the level of confidence that the eye is *open* (1) or *closed* (0). For a matter of simplicity, we convert the *open*, *in-between*, and *closed* classes into 0, 0.5, and 1 groundtruth regression values; regression models output a single value and the loss now combines a sigmoid activation (which ensures that the value is in $[0, 1]$) and a mean squared error function. The mean squared error function is calculated by squaring the difference between the true y and the predicted $\hat{y}$ returned by the model. The rest is unchanged, notably the Adam optimizer is also used in the regression model.

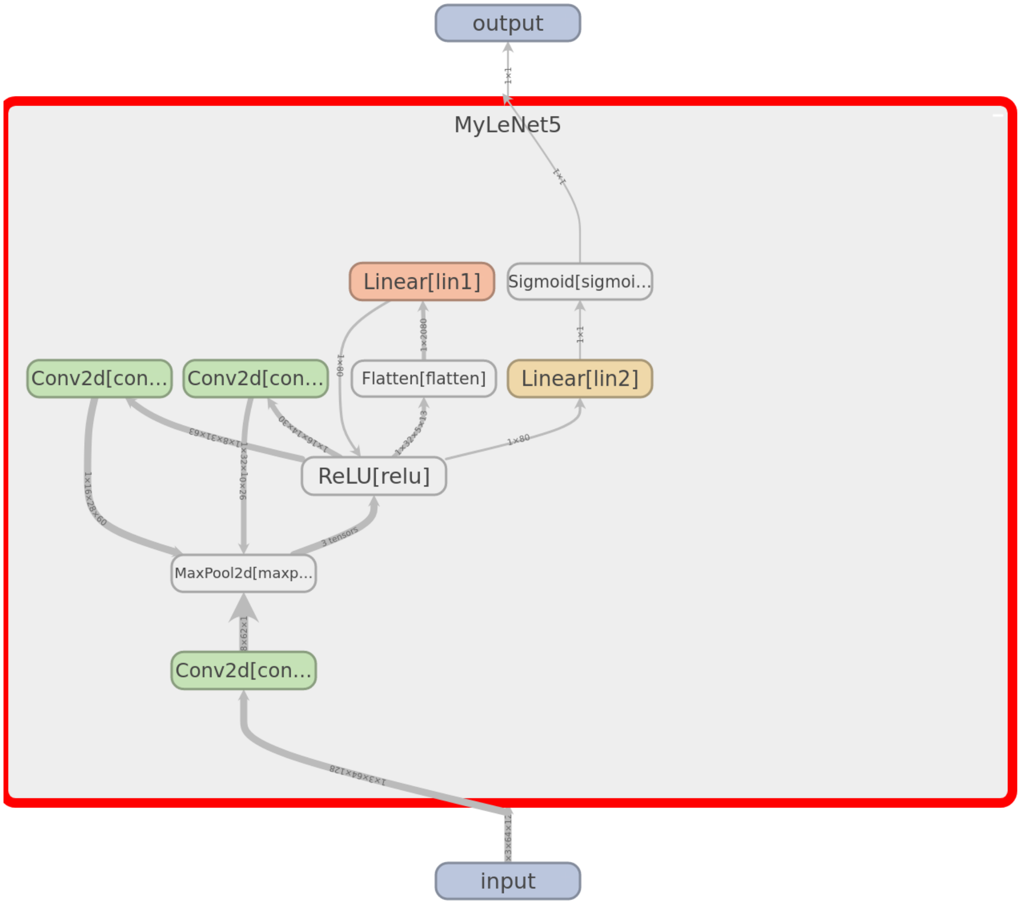


Figure 3.7: Diagram of the CNN regression model for the eyes crops.

The experimental pipeline is described in section 4.1.1, where we go over the data preprocessing. We report on the training and evaluation performances in section 4.1.4 for the model applied to the face crops and in section 4.1.5 for the regression task applied to

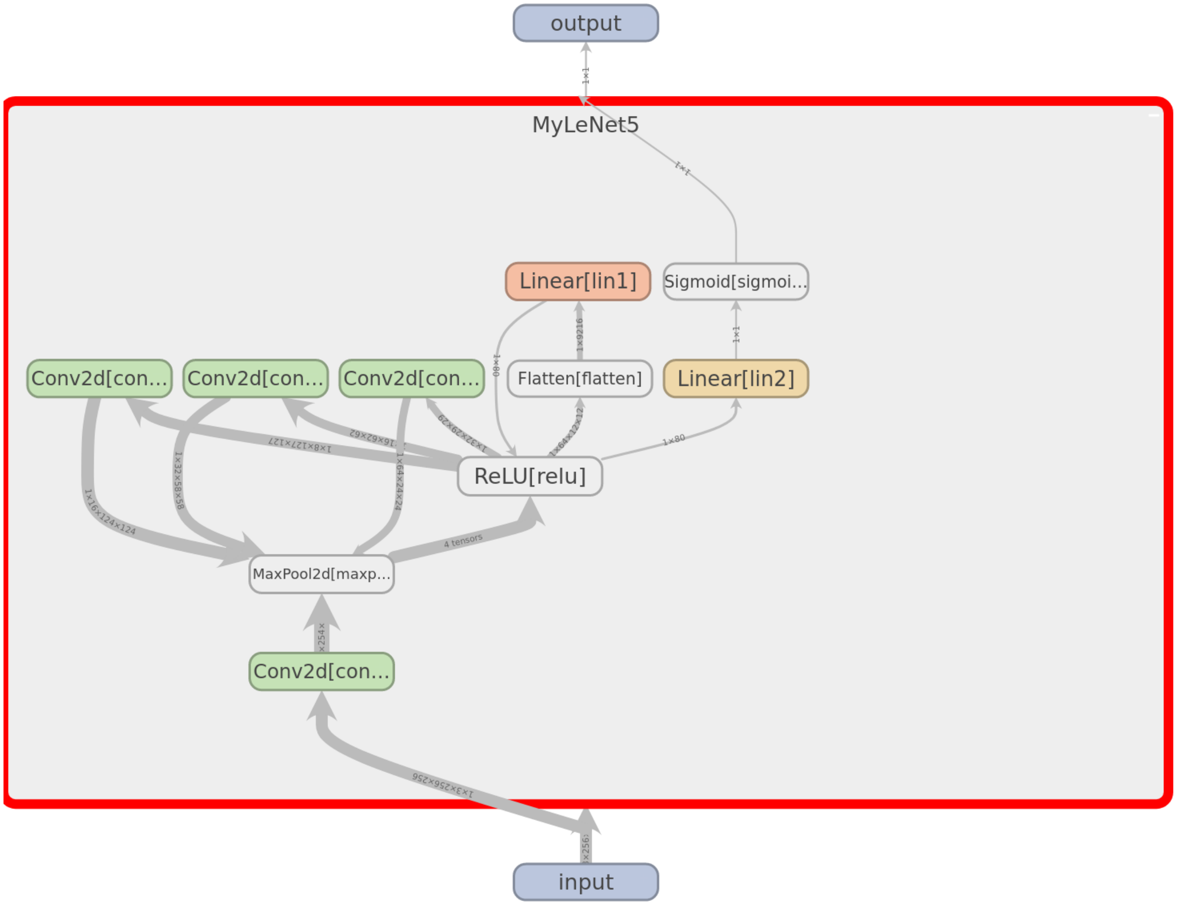


Figure 3.8: Diagram of the CNN regression model for the face crops.

the eyes crops.

3.2.2 Phase two: Agglomeration over time using logic-based rules

Once we have the classification and the regression results (see section 4.1) and we have empirical evidence that they work decently for the task of deciding whether eyes are *open*, *in-between*, or *closed*, we write the script that will take into account both the CNNs' outputs as well as logic rules we need to define. These will help us make a decision as to whether a blink is happening or not.

We use the original groundtruths, that is the annotated data from the first round of annotations for which we labeled the blinks of 26 videos. As explained in section 3.2.1.2.1, the groundtruths were exported from Elan and saved as *.csv* files, organized as one file per video.

As we mentioned earlier that we did not want to have one prediction per frame (as blinks last longer than 40 milliseconds[13]) but rather one prediction for a set of frames representing a time interval, we split videos into non-overlapping windows of frames (the exact process is discussed in section 4.2.1).

We then implement two different logic rules to try and detect whether a blink is occurring, namely the high-low-value-difference rule introduced in section 3.2.2.1 and the curve rule which is presented in section 3.2.2.2.

These logic rules will be based on the analysis of values in a given window of frames, where values can be EAR values or regression values of our CNNs; we note that for classification CNNs, one can simply convert outputed vectors of probabilities into a regression value: given p_c, p_b, p_o denoting respectively the probability of being *closed*, *in-between*, *open*, the simple operation $0 \times p_c + 0.5 \times p_b + 1 \times p_o$ gives a regression value in $[0, 1]$, where 0 corresponds to $p_c = 1$, 0.5 corresponds to $p_b = 1$, and 1 corresponds to $p_o = 1$.

3.2.2.1 The high-low-value-difference rule

We implement a logic rule that looks at the maximal amplitude between the values within the defined window. We call it the high-low-value-difference rule. When a blink is occurring, as we have annotated the data in such a way that the blink starts as the eye starts closing, the eye is still open at the beginning of the annotation and normally, has reopened at the end of the blink, therefore both high values and low values should be displayed in the window where a blink is occurring. We take the lowest value of the window and we subtract it from the highest value of the window. The difference should be higher than the defined threshold. If the value is higher, we consider that a blink is happening, otherwise, no blink is occurring.

3.2.2.2 The curve rule

The curve rule is the second rule we implement. We expect that if a blink is occurring the CNN and EAR values should be lower in the middle of the window than on the outskirts of the window where the values should be greater: we want to see whether the values in the window form a U-shape curve. This can be achieved by fitting a second-degree

[13] see section 3.1.3.1 and table 3.3 for more information on blinks and their length.

polynomial, for instance with a polynomial regression model from Scikit-Learn. A blink is occurring when the curve goes down and up again in a sufficiently steep manner (how steep depends on the set threshold).

Chapter 4

Results and discussion

In this section, we go over the experimental pipeline of both phases of our method, we present the results of both components and discuss them.

We build our CNN models as defined in section 3.2.1.2. We implement both a CNN classifier and a CNN regressor. We use the crops of the frames of video $S2T2B15$ to train the CNN models that make a decision as to whether the state of openness of the eye. The experimental pipeline for component 1 of our method is presented in section 4.1.1. The results of the training and the evaluation of the two CNN models are reported respectively in section 4.1.2, 4.1.4 for the face crops, and in sections 4.1.3 and 4.1.5 for the eyes crops. They are discussed in section 4.1.6.

Afterwards, we start building the hybrid model. To build the hybrid model, we import our CNN models. The EAR calculation as described in section 3.2.1.1 are implemented. We load the videos corresponding to our groundtruths, specifically the groundtruth from the initial annotations as described in section 3.1.2. We create the window sizes and the thresholds, and we define the rules that will distinguish closures from blink occurrences. The detailed experimental pipeline is available in section 4.2.1. Finally, results obtained on phase 2 of our method are show in sections 4.2.2 and 4.2.3, and discussed in section 4.2.4.

4.1 Eye *open* - *in-between* - *closed* problem

As we have seen in previous sections, methods developed in order to undertake the issue of automatic blink detection are numerous. Unfortunately, as we have argued in section 3.2, the definition of blinks according to the researchers who worked on the elaboration of these methods is simpler than the ones described in physiology or linguistics. For this reason, we decided to create our own blink detector divided into two components, namely phase one for which we create a machine learning model and phase two for which we create an approach based on logic rules, forming a hybrid model. For evaluation purposes, specifically to evaluate the performances of our CNN models, we also implement and use as baseline the popular EAR calculation combined with our rules.

In section 4.1.1, we report on the experimental pipeline used for the classification and regression of eyes openness states. Following, we report on the results obtained with our method in sections 4.1.2, 4.1.3, 4.1.4, and 4.1.5. Eventually, we discuss these results in section 4.1.6.

4.1.1 Experimental pipeline

We relate the creation and organization of our dataset in section 3.2.1.2.1. Once the dataset is ready, that is, once all the frames of video $S2T1B15$ are ordered properly in one of the three folders, *open* - *in-between* - *closed*, we proceed and load the images. We use the Pytorch library (Paszke et al., 2019) to develop our methods. We first start by splitting our data into train, validation, and test sets.

The images in the training set have a different preprocessing than the ones in the validation and test sets. What is common to all three sets are the resizing of the frames and the conversion of the images into numerical values. As mentioned above, all face crops do not have the same dimensions and this can also be said of the eyes crops which vary from frame to frame. The resizing of the images is therefore used to ensure that all frames have the same measurements. Face crops are resized to $256 * 256$ while eyes crops are resized to $64 * 128$.

The frames used in the training set are also subjected to data augmentation using the Trivial Augmentation Wide transform implemented in PyTorch and initially developed by Müller and Hutter (2021). Before introducing the method in more detail, let us explain why we do it.

We use such data augmentation because the distribution of the frames across classes is considerably unbalanced. For the video $S2-T1-B15$, a total of 16298 frames are distributed throughout the three folders, the *closed* folder only contains 690 frames while the *in-between* and the *open* folders share the remaining 15608 frames and respectively contain 7667 and 7941 frames.

We superficially restore balance across the categories by fixing the number of training images on a percentage of the minority class, that is the *closed* label. 70% of the 690 frames from the *closed* folder are used in the training set, giving 482 images. The training set also contains 482 frames from the *in-between* and the same amount from the *open* categories. The remaining 30% of the *closed* folder is divided in two, half of the frames are placed in the validation set while the other half is put in the test set. The large number of frames remaining in the *in-between* and *open* folders are also separated in half and placed in the validation and test sets.

Because the training dataset is so small following this undersampling, we apply a virtual data augmentation method that randomly modifies the images within the training set. From one batch to another, the 482 images do not appear the same way. As mentioned above, we use the TrivialAugment augmentation method which is a state-of-art automatic augmentation method. The degree to which TrivialAugment transforms an image varies randomly. Müller and Hutter (2021) note that only one augmentation method is applied to an image at a time. Examples of possible augmentation techniques applied to the images include variation in color, brightness, contrast, sharpness, blurring but also image rotation, or image flipping which are all examples of augmentation approaches that are considered interesting considering the nature of our data. Unlike other available implementations of these techniques in PyTorch, the Trivial Augment method does not require any argument, any specific manual parameter configuration. Examples of Trivial Augment output are exhibited in figures 4.1 and 4.2 below.

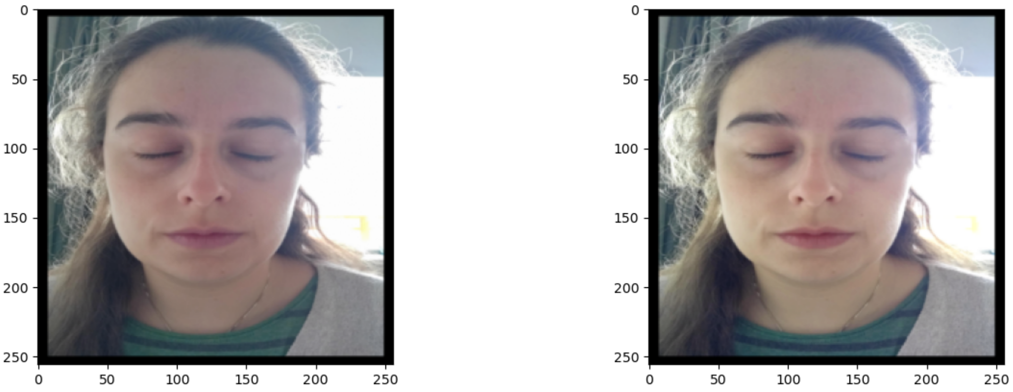


Figure 4.1: TrivialAugmentWide variation examples for the virtual augmentation showing the change in color.

We train both our CNN classification and regression models for 200 epochs on the eyes crops and 100 epochs on the face crops of the video $S2T1B15$. The batch size is set to 64 for the training set. We use the accuracy and $f1$-score metrics to evaluate our

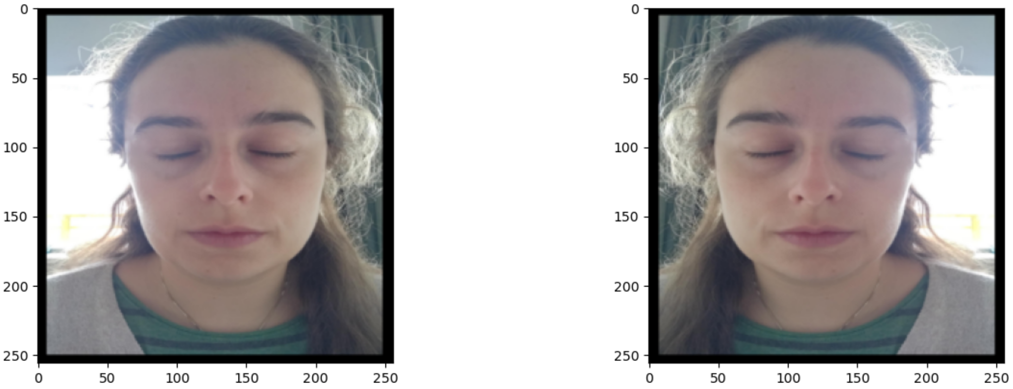


Figure 4.2: TrivialAugmentWide variation examples for the virtual augmentation exhibiting an horizontal flip of the image.

classifiers, whereas we use the $r2$-score for our regressors.

The evaluation of the models is performed on a newly annotated video, specifically $S2T2B15$, to assess the ability of our models to generalize to unseen data but on the same signer.

4.1.2 Results of the classification model used on face crops

We present, in the sections 4.1.2.1 and 4.1.2.2 the results obtained with our classification model during training and evaluation respectively. The following figures exhibit results for the CNN classifier applied to the face crops. Results are discussed in section 4.1.6.

4.1.2.1 The results from the training

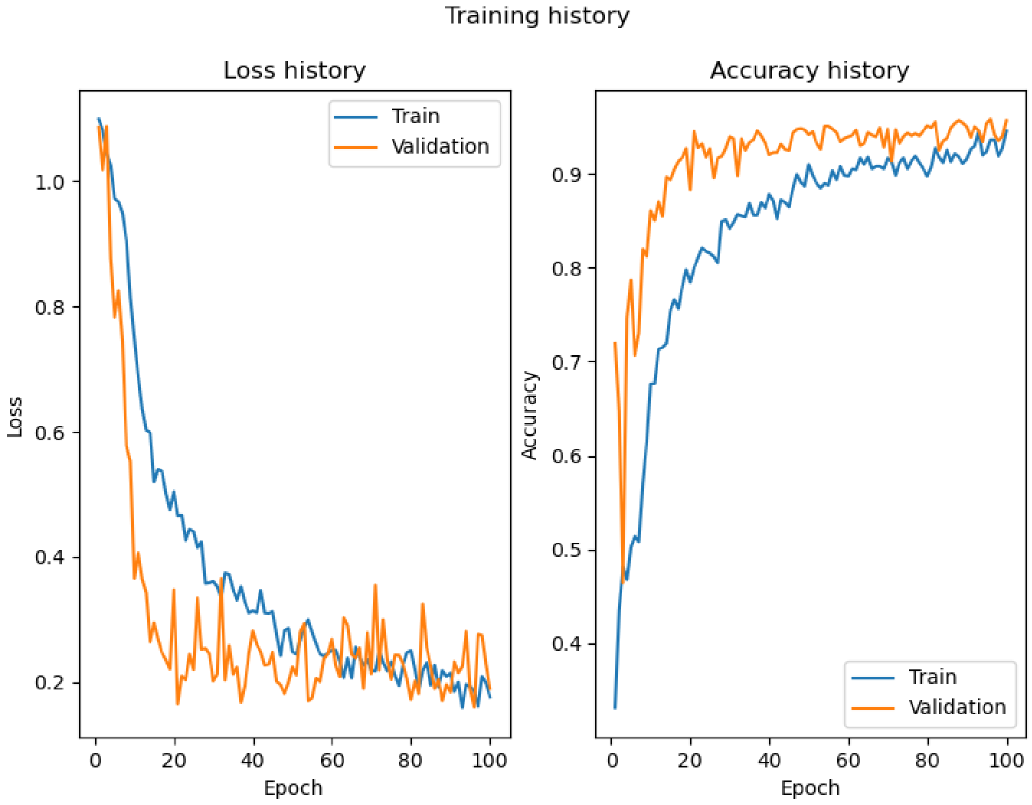


Figure 4.3: Loss and accuracy curves for the classification task using the face crops.

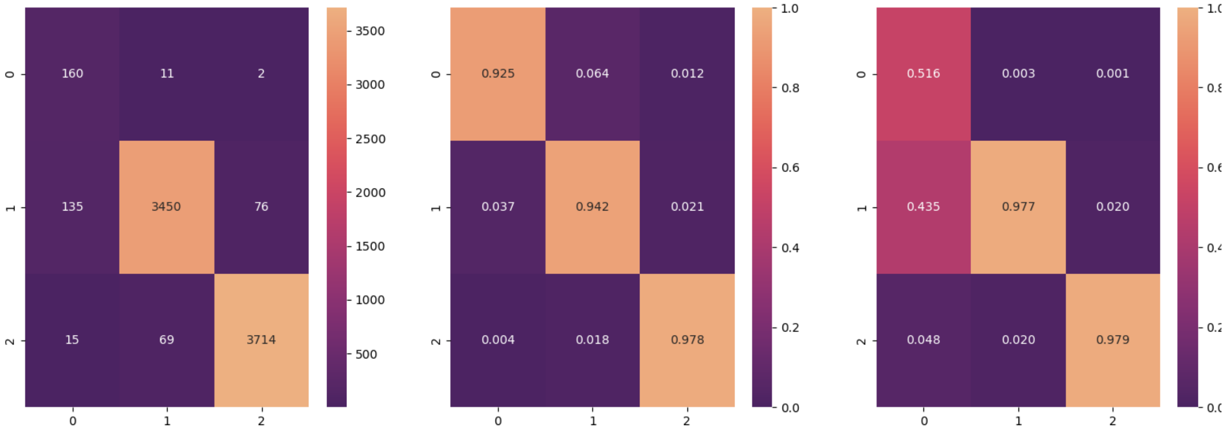


Figure 4.4: Confusion matrices depicting the predicted distribution of data across classes and both types of normalization for the classification task using the face crops on the validation data. (video $S2T1B15$, test set)

4.1.2.2 The results from the evaluation

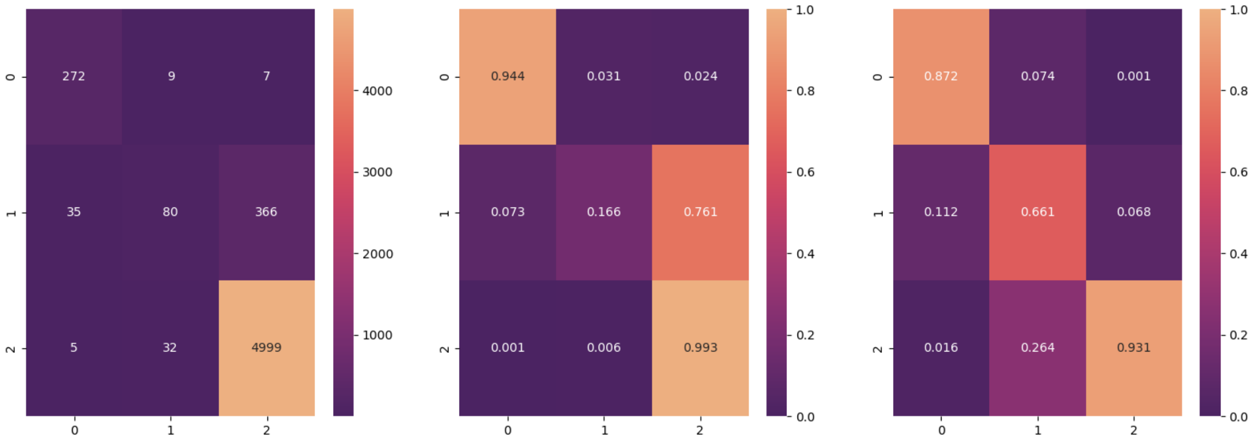


Figure 4.5: Confusion matrices for the evaluation of the CNN classification model on face crops. (video $S2T2B15$)

4.1.3 Results of the classification model used on eyes crops

In sections 4.1.3.1 and 4.1.3.2, we report the figures showing the results gotten with our classification model during training and in evaluation respectively. These figures exhibit results for the CNN classifier applied to the eyes crops. Results are discussed in section 4.1.6.

4.1.3.1 The results from the training

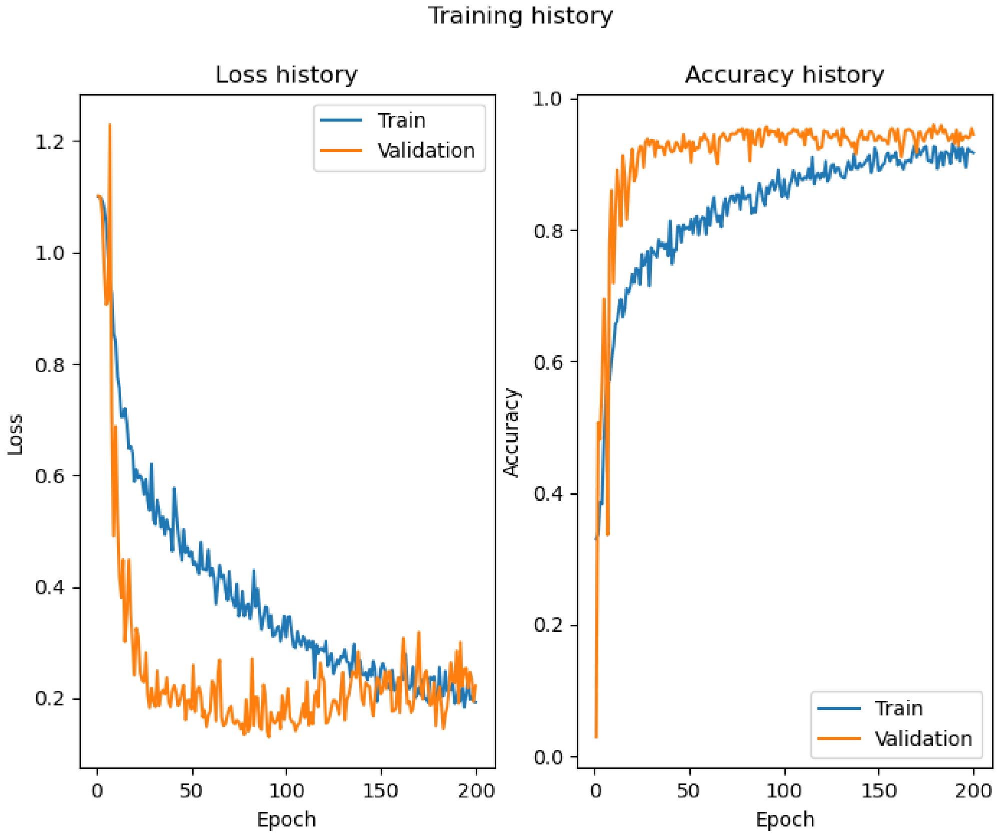


Figure 4.6: Loss and accuracy curves for the classification task using the eyes crops.

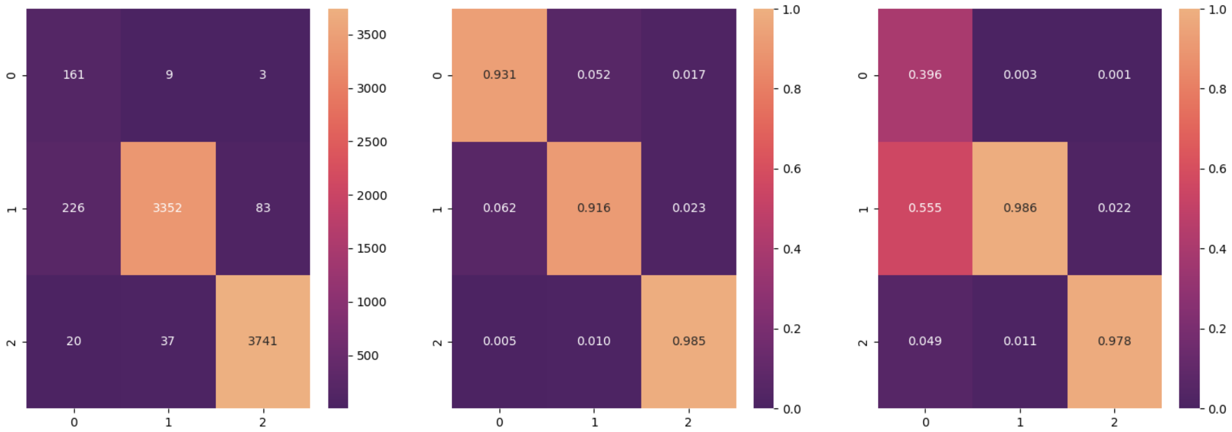


Figure 4.7: Confusion matrices depicting the predicted distribution of data across classes and both types of normalization for the classification task using the eyes crops on the validation data. (video $S2T1B15$, test set)

4.1.3.2 The results from the evaluation

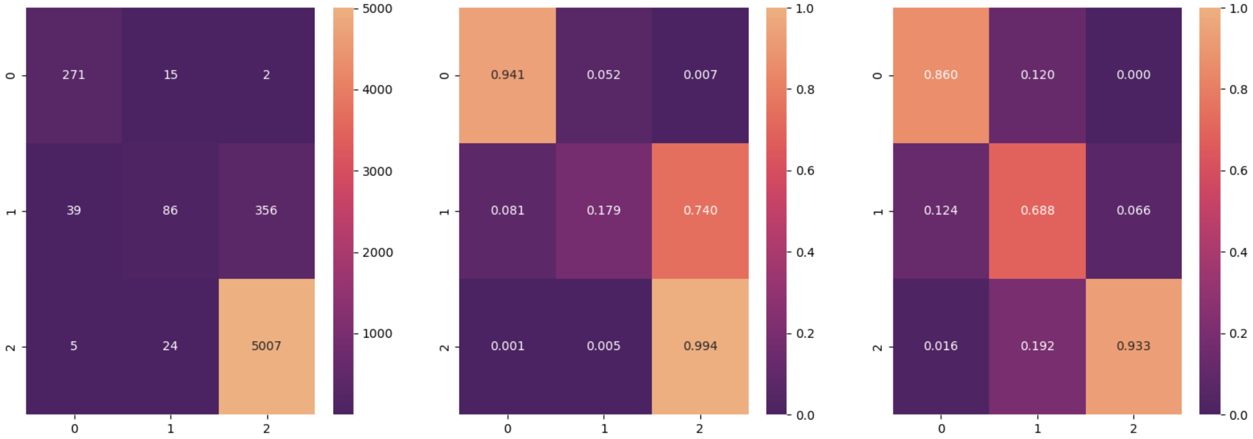


Figure 4.8: Confusion matrices for the evaluation of the CNN classification model on eyes crops. (video $S2T2B15$)

4.1.4 Results of the regression model used on face crops

The figure represented in section 4.1.4.1 shows the results obtained with our regression model during training. The following figures displays the results for the CNN regressor applied to the face crops. Discussion of these results are available in section 4.1.6.

4.1.4.1 The results from the training

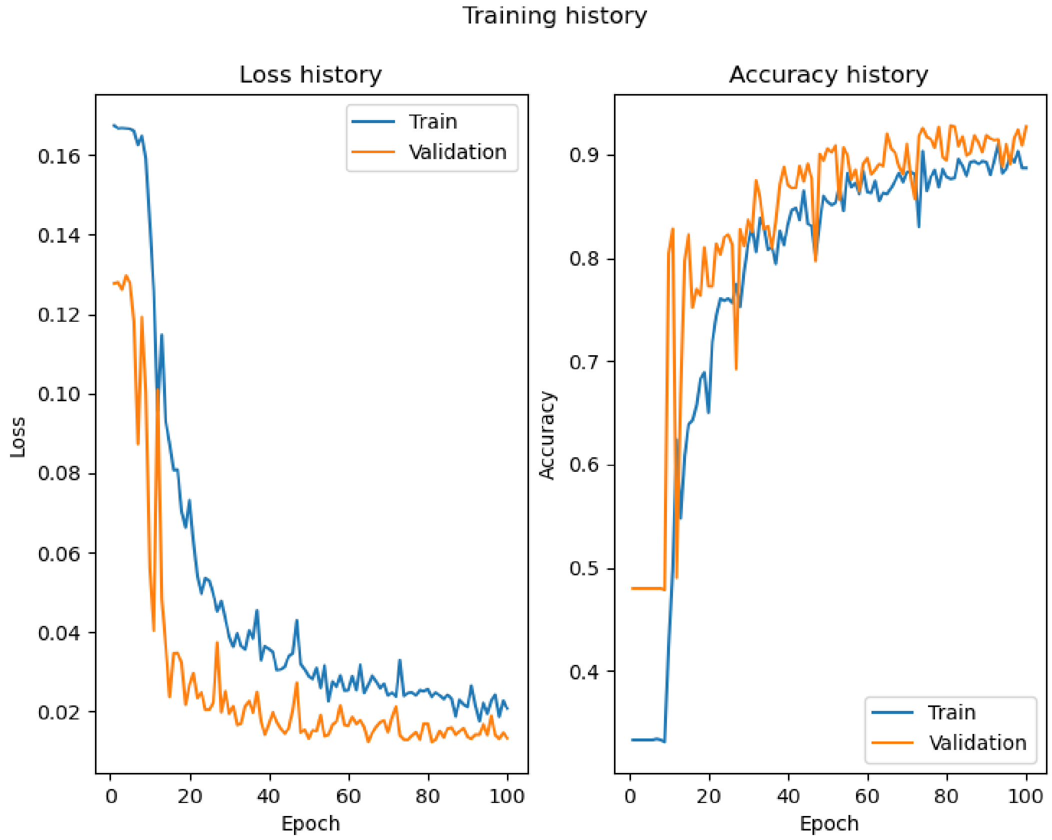


Figure 4.9: Loss and accuracy curves for the regression task using the face crops.

4.1.5 Results of the regression model used on eyes crops

Section 4.1.5.1 displays the results achieved with our regression model after training. The figure below reveals the results of the CNN regressor applied to the eyes crops. We discuss these results in section 4.1.6.

4.1.5.1 The results from the training

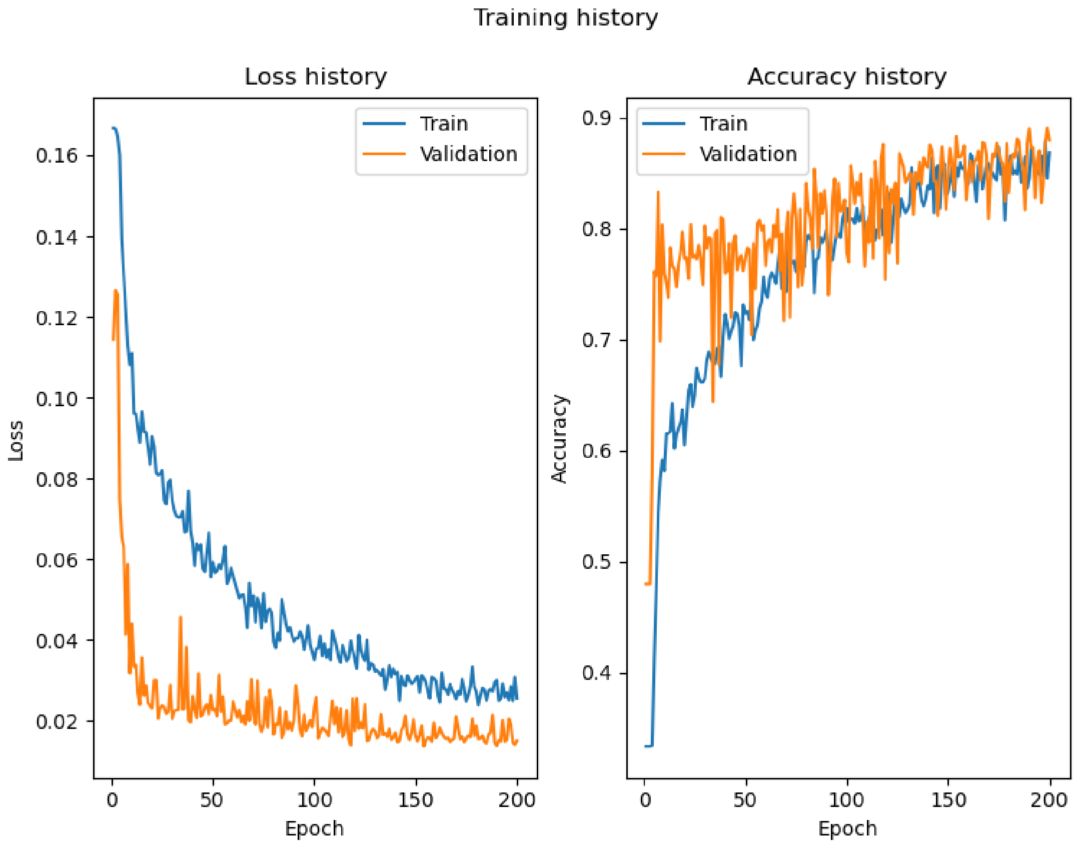


Figure 4.10: Loss and accuracy curves for the regression task using the face crops.

4.1.6 Discussion

After creating the new dataset (consisting of the unique video: $S2T1B15$), we realize that the problem might not be categorical in nature but rather, the lids opening and closing exhibited all degrees of openness imaginable. In the real world, openness of the eyes can be described as a continuum.

In this section, we compare the training, validation and evaluation results of the classification and the regression tasks whose corresponding figures are displayed in the previous sections. Results from the training were garnered on a subset of the frames from video $S2T1B15$, namely the test set. We evaluated the models on a video of the same signer, namely $S2T2B15$.

Let us start by taking a look at the classification training and validation results for the face crops. We train our CNN model for 100 epochs. The loss history and accuracy history of the model are presented in figure 4.3. The confusion matrix is shown in figure 4.4, where 0 represents the *closed* category, 1 represents the *in-between* class and 2 represents the *open* category. The first confusion matrix represents the raw distribution of the data across the classes. The two other confusion matrices are normalized versions of the confusion matrix on the left. The middle confusion matrix has its rows normalized, and is a visualization of the predictions of the model given the groundtruths, the columns of the right confusion matrix are normalized, giving the probability of the groundtruths knowing the predictions.

We train the classification model for the eyes for 200 epochs. Results of the training of the classification model on the eyes crops are displayed in figures 4.6 and figure 4.7. We choose to have a longer training for the eyes crops because the eye crops sizes show more variation while containing less information per image, and therefore take longer to train to achieve a similar level of performance. Because of this difference, the crops sizes variation and the training period, we cannot truly compare the classification results for the eyes and for the face but we do have an idea of what the accuracy of the eye classification model looks like at 100 epochs for the training and we see that had we stopped there, the face classification model would perform better. As visible on figure 4.6, at around 100 epochs, the eyes crops model has an accuracy of about 87% on the training set and 94% on the validation set while after the same amount of epochs, the face crops model has an accuracy approaching 95% on the training set and 96% on the validation set.

After training, the eyes crops and face crops models are evaluated on a test set. They get similar accuracy results with respectively 95% and 96%. Their macro $f1$-score is at 83% for the eyes crops classification model and 87% for the face crops model. The face crops model is a little stronger than the eye crops model.

We can see in figures 4.7 and 4.4 as we had noted in section 4.1.1 that data is pretty unbalanced, specifically, the *closed* class only has 173 instances, representing 2.2% of all of the data (7632 instances). The models make similar errors, specifically, the model seem to predict crops showing closed eyes (0) with a high accuracy as we note that for the face crops, 160 images are correctly classified out of 173 frames. On the eyes crops, 161 closed eyes images are correctly classified. This means that for the *closed* category, 7.5% for the face crops and 7% for the eyes crops are misclassified. It is important to give these percentages as it may appear on the confusion matrices that the *in-between* images are hard to label. Indeed, for the *in-between* face crops, a total of 211 images are misclassified, the majority of which are miscategorized as *closed*. For the eyes crops, this misclassification of the *in-between* category images shows even more as 309 images are not labeled correctly, mostly put in the *closed* category. Considering the size of the *in-between* class, these misclassification only represent respectively about 6% and 8% of the data. The class that shows a higher accuracy is that of the *open* eyes instances where for the face crops, 2.2% of the images are mislabeled, while only 1.5% of eyes crops frames are misclassified as *closed* or *in-between.* Considering all categories, the face crops model has a total of 4% of misclassified images against almost 5% for the eyes crops, making it a better model, more easily trained as well. Now that we have taken a look at the classification results, let us see whether the results obtained with the regression task are similar or not.

Results of the regression model applied to the face crops are shown in figure 4.9 with the loss and accuracy curves. We obtain an accuracy of about 93% on the validation set which is similar even though a little lower than what we get with the classification model. We also get a score of 93% accuracy on the test set. The $r2$-score is at 83%.

Results for the regression on the eyes crops can be seen in figures 4.10 which show the loss and accuracy curve. We note that the accuracy of the model is slightly below 90% after 200 epochs which is close to the results we obtained with the classification task where we got an accuracy neighboring 90% as well. The $r2$-score is at 79.5%. The $r2$-score shows us that there is a three points difference between the face crops model and the eyes crops model.

We evaluated the classification models and the regression ones on the video $S2T2B15$, which we have previously saved as images and which frames we have reorganized into the three categories, namely *closed*, *in-between*, and *open*, the same way the video $S2T1B15$ was organized. We reshaped the images.

The confusion matrices for the evaluation of the classification task on the face crops are shown in figure 4.5. We obtain a macro $f1$-score of 71.1% on the evaluation data.

In figure 4.8 the confusion matrices for the evaluation of the classification model on the eyes crops are displayed. The $f1$-score obtained is at 71.5%, which means that the classification models on both the face crops and the eyes crops perform similarly.

What we observe looking at the confusion matrices of these classification tasks is that the categories are this time again highly unbalanced. Not many images belong in the *closed* and the *in-between* categories and by itself, the *open* category gathers 86% of the data, against 8.2% for the *in-between* category and almost 5% for the *closed* category. Once again, we note that the *closed* eyes instances are well classified, with 5.5% of all instances misclassified for the face crops and 5.9% of misclassified instances for the eyes crops. *In-between* instances are interestingly misclassified as *open* rather than *closed* as was observed during the training. For the face crops, 83% of the *in-between* instances are misclassified which is much higher than was has been observed on the original test set but there were not as many data instances in this class this time. For the eyes crops, it's 82% of the data points that have been mislabeled.

We evaluated the regression model on video $S2T2B15$. On the face crops, we obtained an $r2$-score or 74.4% while on the eyes crops, our $r2$-score is 74.9%.

Now that we have taken a look at the results of the CNN models, we report in section 4.2 the results on our hybrid models.

4.2 Blink problem

Now that we have taken a look at the results of the CNN models for the eye *open - in-between - closed* problem, both for the classification model and for the regression model, let us have a look at the results for the blink problem.

In the next section, we report on the experimental pipeline of the combination of the CNN and the rules. In section 4.2.2, we show the results obtained with the high-low-value-difference rule while in section 4.2.3, we exhibit the results obtained using the curve rule. Finally, in section 4.2.4, we discuss the limits and the possible improvements that can be implemented in order to achieve stronger results and better reliability.

4.2.1 Experimental pipeline

As discussed in section 3.2.2, we create windows of frames as we want to make decision on the occurrence or non-occurence of a blink based on a time interval longer than the duration of a single frame.

These windows may vary in size: the window size in instances in which a blink is occurring depends on the length of the original blink in the dataset. That is if the blink annotated in the groundtruths lasts for 3 frames then the window size for which the model will need to return a prediction will be set at 3, if the blink lasts 7 frames, then the window size will be set at 7, and so on. When no blink is occurring, the window size is set at 5 by default and we make sure that windows do not overlap.

As we have a classification task (blink or not blink) and a large imbalance, that is more non-blink intervals than blink intervals, we use the $f1$-score as a metric.

We compare 2 rules, namely the high-low-value-difference rule and the curve rule described in section 3.2.2, within 5 methods, that is the classification CNN model on face crops, on eyes crops, the regression CNN model on face crops, on eyes crops, and eventually, the EAR calculation model.

For each rule, we test different thresholds both for the CNNs and the EAR to see which will come out as the best threshold.

For rule 1, we select sixteen threshold values to be tested on the EAR, namely: 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.10, 0.21, 0.22, 0.23, 0.24, 0.25. We also set different thresholds for the CNN outputs as we are not sure how large the difference between the lowest and highest values need to be to best represent the occurrence of a blink. The different thresholds are set at 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85.

The thresholds for rule 2 are different than that defined for the first rule, we define the following set of threshold for the CNN models: 0.25, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, and the following ones for the EAR method: 0.25, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0.

4.2.2 Results of the high-low-value-difference rule

We display in this section the figures relating the results obtained on the combination of our different CNN models or the EAR calculation implementation and our first rule, specifically, the high-low-value-difference rule.

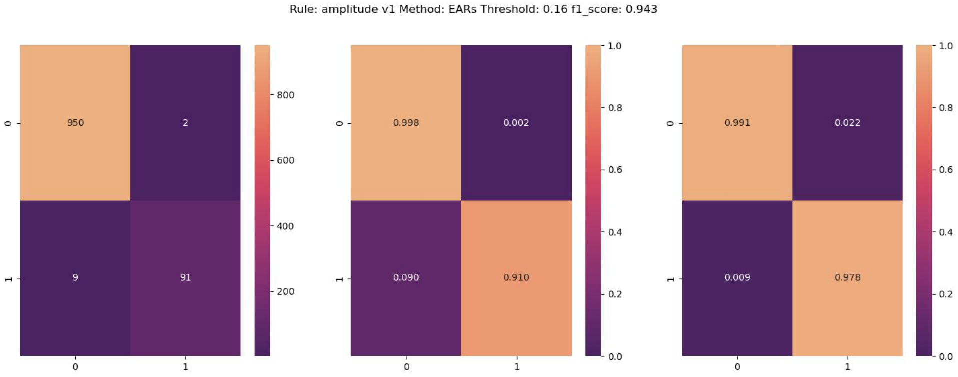


Figure 4.11: Confusion matrices for the best threshold of the hybrid model using the EAR calculation with the high-low-value-difference rule. (video $S2T2B15$)

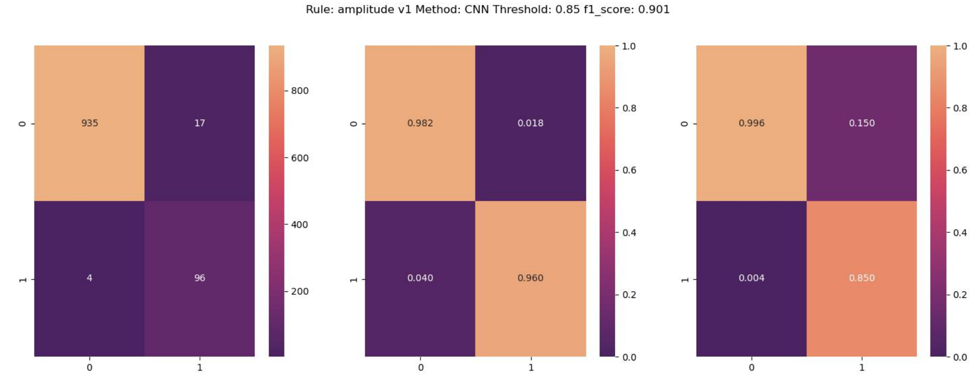


Figure 4.12: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the high-low-value-difference rule. (video $S2T2B15$).

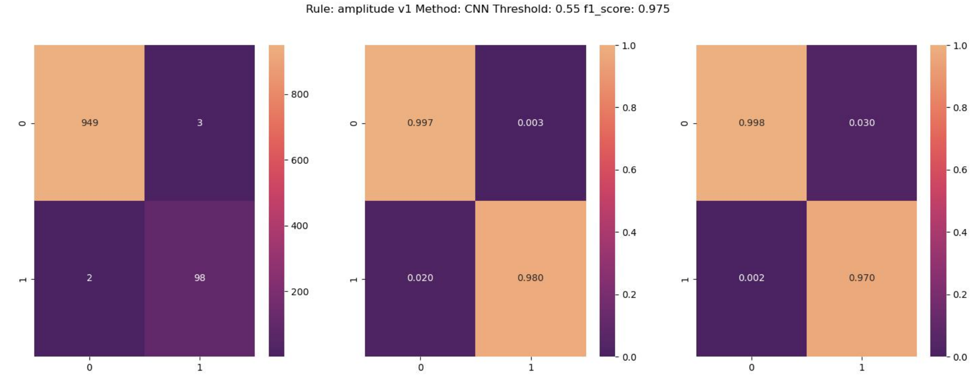


Figure 4.13: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the high-low-value-difference rule. (video $S2T2B15$).

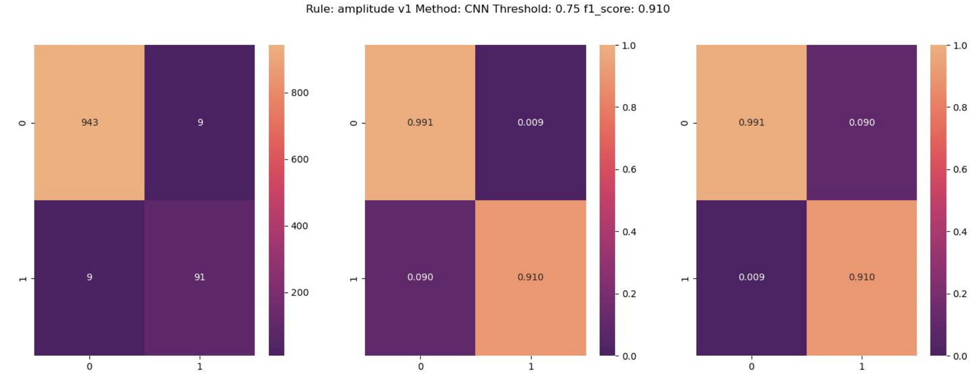


Figure 4.14: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the high-low-value-difference rule. (video $S2T2B15$).

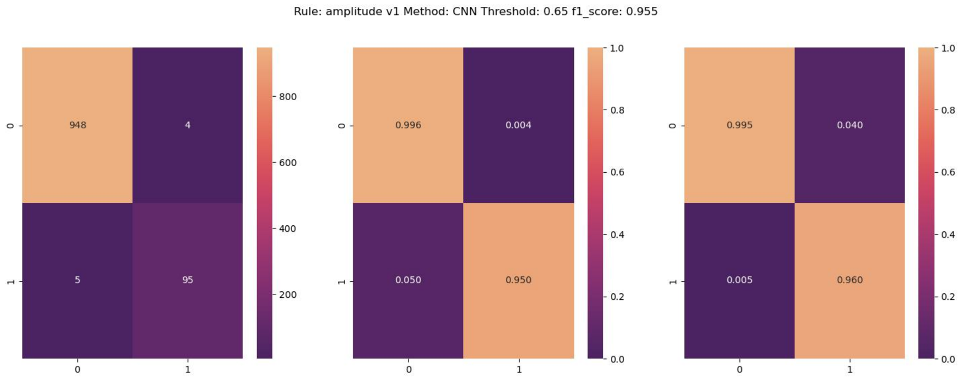


Figure 4.15: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the high-low-value-difference rule. (video $S2T2B15$).

4.2.3 Results of the curve rule

Results yielded using the combination of the EAR calculation implementation or one of our CNN model are reported in this section in the form of figures. The results displayed on the following confusion matrices are discussed in section 4.2.4.

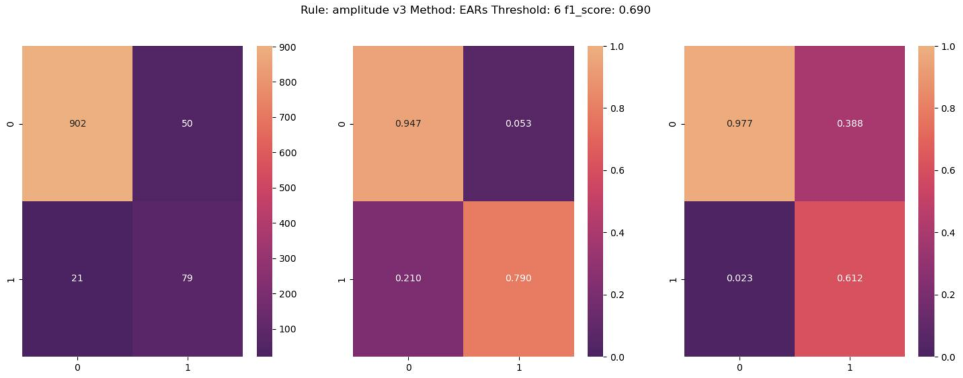


Figure 4.16: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the curve rule. (video $S2T2B15$).

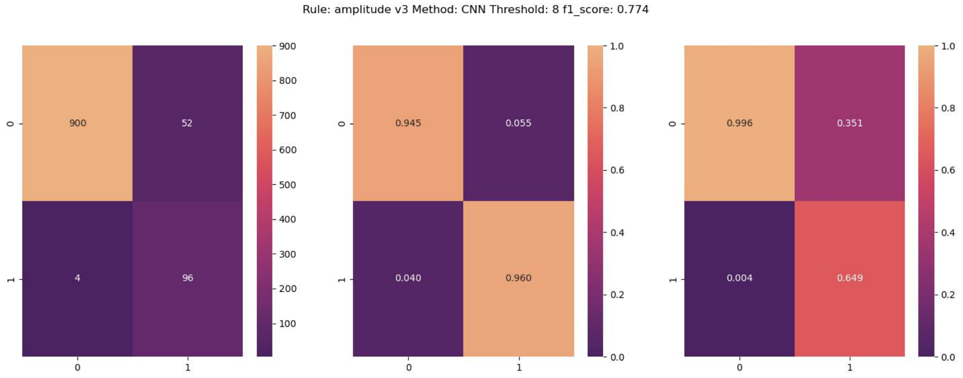


Figure 4.17: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the curve rule. (video $S2T2B15$).

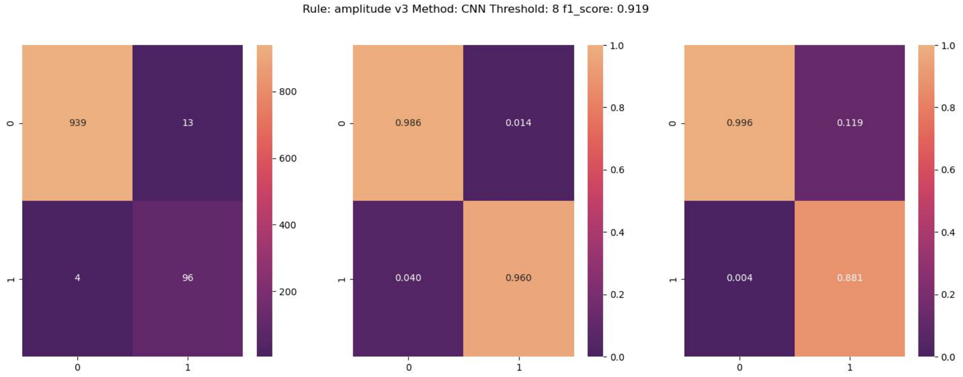


Figure 4.18: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the curve rule. (video $S2T2B15$).

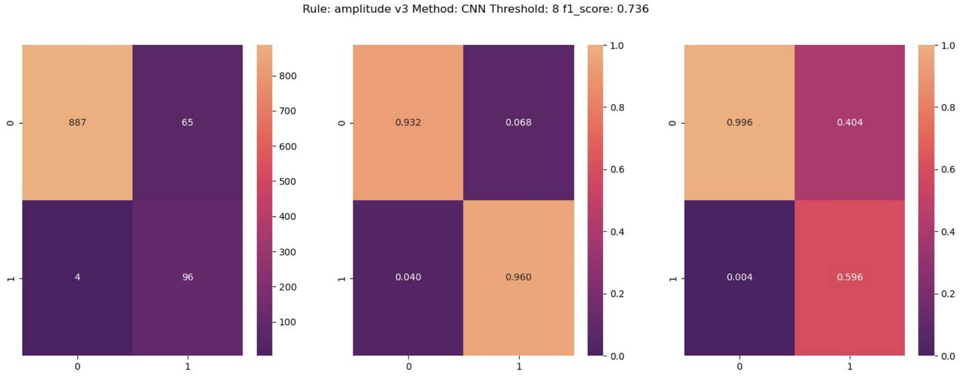


Figure 4.19: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the curve rule. (video $S2T2B15$).

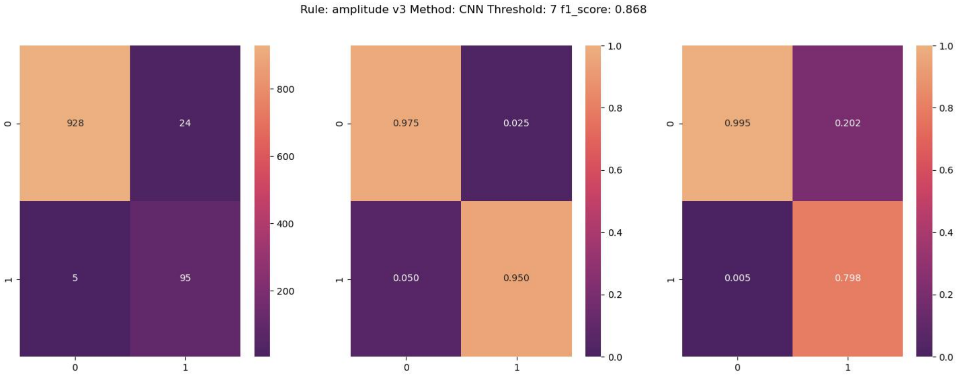


Figure 4.20: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the curve rule. (video $S2T2B15$).

4.2.4 Discussion

As explained in section 3.2, our blink detection algorithm consists of the combination of a CNN model [1], for which we have reported and discussed the results in section 4.1, along with rules detailed in section 3.2.2.

We create a first rule that looks at the difference between the highest and lowest values in the defined window to determine whether a blink is happening or not.

The second rule proposed looks at whether a U-shaped bell curve sufficiently steep is occurring within the window size, if so we consider that a blink is occurring. The figures displaying the results that we obtain using the first rule are shown in section 4.2.2 while the results concerning the second rule are displayed in section 4.2.3.

We tested our method on the other videos of signer $B15$, that is, videos $S2T2B15$, $S2T3B15$, and $S2T4B15$. We also applied the method on one video of each signer for which we have annotated data. We applied it to the following videos: $S2T2A11$, $S4T6B14$, $S5T1A9$, and $S9T6B5$. The figures showing the results for these videos will be available in Appendix 2.

In figure 4.12, we see the results of the best threshold obtained on the face crops for video $S2T2B15$ using the classification as our CNN model. Figure 4.11 show the results of the best threshold obtained on the face crops of video $S2T2B15$, this time using the EAR calculation.

Both methods get similar results even though the EAR method shows a higher $f1$-score which attains 94% while the $f1$-score yielded with the CNN classifier on the face

[1] see section 3.2.1 for more detail

crops is 90%. Though, if we look at the detail of the first confusion matrix, we see that if the EAR method is better at predicting the non-blinks, the blink occurrences in which we are truly interested are better detected using our combination of the CNN model and the rule. While we have an error percentage of 4% on the detection of the blinks with our method, the percentage of errors made by the EAR model is 9%. The percentage of errors for the non-blinks is 1.7 for the CNN model while only being 0.2% for the EAR model.

In figure 4.13 we see the results for the CNN classification model applied to the eyes crops, which we compare to the EAR results from figure 4.11.

First, what is interesting to note is that the threshold for the CNN classifier is different depending on whether we apply the model on the face crops or on the eyes crops. When looking at the results, it is interesting to see that this time, the CNN hybrid method gets a higher $f1$-score, specifically a value of 97.5%, exhibiting an error rate for the blink occurrences of 2% for frames where a blink truly occurs.

As we have seen in the previous section, the regression model was performing slightly worse than the classification model for the *open/closed* problem. Let us see whether it remains true when combining the CNN regressor with the rule for the *blink/not-blink* problem. In figure 4.14, we show the results of the CNN regressor.

Using the regressor, we see that the $f1$-score is a little better even though only slightly and when we take a look at the confusion matrices, we see that if the regressor does a better job at giving a correct prediction on the windows in which no blink is occurring, its error rate when it comes to the blink occurrences is higher, specifically, at the same level of that of the EAR method.

In figure 4.15, we consider the results obtained on the eyes crops with the CNN regression model.

This time, if the $f1$-score stays higher than that of the EAR method, it is lower than the ones gotten while using the classification model. Not only does the model make more mistakes predicting the non-blink periods, it is also worse at predicting the blinks occurrences.

As we have seen, testing the method on this video with our high-low-value-difference rule, we obtain stronger results than the EAR method only when using the eyes crops. We apply this method to other videos as mentioned earlier and what we have observed is that if it works well on the participant on which the CNN has been trained, when using data from other signers, the results are not always as good as the EAR which does not rely on training and can be applied to any participant. However, sometimes, the results yielded using the combination of a CNN model with the high-low-value-difference rule are very positively surprising, notably if we look at results from video $S5T1A9$ and video $S9T6B5$ where we obtain $f1$-scores close to 70% on the former and 73% on the latter which are lower than that obtain with the EAR calculation.

Using video $S2T2B15$, we also note that the difference in performance between the two methods, the first combining one of the CNN model with the rule, the second combining the EAR calculation with the rule, is only subtle. On other videos of this signer, we note that our method obtains better results than the EAR method, this is notably true when taking a look at figures showing results for video $S2T4B15$. We conclude that with more training and training on other people, we are hopeful that our method could also be used on more people more reliably.

The curve rule is a bit more sophisticated than the high-low-value-difference rule. Let us see whether the results we get with this rule exceed the ones obtained with the first rule or with the EAR method.

In figure 4.17, we see the confusion matrices of the best threshold using the CNN classifier on the face crops. In figure 4.16, we have the confusion matrices of the best threshold this time using the EAR calculations.

Using the curve rule, we obtain a $f1$-score of 77.4% using the classification CNN model. This method performs better than the one using the EAR calculations for which we obtain a $f1$-score of 69%. The hybrid EAR model has an error rate of 21 per 100 blinks while the hybrid CNN method is slightly worse at correctly predicting the instances where no blinks are happening but only have a percentage of error of 4 on the blink occurrences predictions. Let us see in figure 4.18 how the eye model does using the CNN classifier on the eyes crops with the curve rule.

Considering the hybrid model with the CNN classifier on the eyes crops, we see that the results are sensibly better than those we get working with the face crops as we get a $f1$-score of 91.9%. However, we note that the error rate for the blink detection is not lower, only the error rate for the non-blink instances has changed and is improved. Even if the CNN classifier combined with the curve rule does well on the eyes crops frames, the results are not as good as those reported for the first rule for which we obtained a $f1$-score of 97.5% and had an error rate on the blink category of 2%.

Now that we have looked at the results from the hybrid CNN classifier combined with the curve rule, let us turn to the results of the hybrid CNN regression method associated with the curve rule. Results of the application of the model on the face crops are presented in figure 4.19.

The results of the regression CNN model associated with the curve rule applied to the face crops are worse than those obtained using the classification CNN, as the $f1$-score has decreased of almost 4 points, though it is interesting to note that the error rate for the blink detection has not gone up, only the one for the non-blink instances predictions has changed, going from an error rate of 5.4% to 6.8%. In figure 4.20, we note that the same observation can be made for the hybrid model using the CNN regression model used on the eyes crops.

The $f1$-score dropped even more as the model is applied to the eyes crops compared to when it is used on the face crops as we lost 5 points. In this case, not only has the error rate of the non-blink instances gone up, the error rate of the blink prediction has risen as well.

Now that we have taken a look at the results using the curve rule, we note that the less sophisticated rule, namely the high-low-value-difference rule has proven to perform better. Overall, we can also say that the classification model is stronger than the regression model regardless of the rule we used. The only exception we notice is with the first rule where using the face crops, the model using the CNN regressor performed slightly better than the model using the CNN classifier.

We also find it worthy to note that the EAR method is robust on all signers but is often outperformed by the method using a CNN model on the signer on which the CNN models were trained. Using the first rule, both the CNN classifier and regressor on the eyes crops proved to exceed the EAR calculations. With the second rule, namely the curve rule, a large gap is visible between the EAR results which $f1$-score is at 69% and the results from both CNN models regardless on the crops on which they were applied. We report $f1$-scores of 77% on face crops and 91% on eyes crops with the CNN classifier. The $f1$-scores for the CNN regressor on the face crops using the curve rule is 73% while we obtain a score of 86% on the eyes crops.

The results obtained with the association of the CNN models and the first rule on the face crops are outperformed by the EAR results, but importantly, we note that when using the eyes crops, the method using the CNN is more solid than the EAR method.

The second rule is not as powerful as the first but we observe that we are able to get sturdier results with the CNN method than with the EAR method when applying it.

Finally, another issue that arise with our method as it stands today is the lack of generalization. In Appendix 2, one can see that even if the EAR method also varies from one person to another, from one video to another, the results obtained with this method vary less than the ones we get using the CNN models. That is because the quality of the EARs results rely on the quality of MediaPipe eyes landmarks. When a signer puts his head down for a while, the EAR calculation cannot be computed. On the other hand, the quality of the CNN relies on the similarity of the facial characteristics between the signer on which the CNNs have been trained and that on which it is applied. This lack of generalization is predominantly due to a lack of training, be it training on this participant but also on other participants, be they younger, older, female, or with another ethnic background. The CNN method has room for improvement. We think that a way to enhance our results would be focus on training on more data instances but we could also look at the iris landmarks and add this information to the CNN values used in the windows. We also think it is important to try to detect incomplete blinks and for this, one would need to annotate the CNN dataset in a way such that several degrees of eye openness are considered.

Chapter 5

Conclusion

The initial aim of this thesis was to implement a classification task to be able to distinguish correctly the various types of blinks. To achieve this goal, it was necessary for us to start by learning more about non-manual markers in sign languages, the way blinks work physiologically, what their patterns are, how they are used to communicate. Ultimately, we also needed to know what these many types of blinks were and what methods were out there to detect the blinks automatically because before being able to classify them according to their types, these types needed to be defined and we needed to be able to detect the eye blinks reliably.

We started by annotating the LSF part of the Dicta-Sign corpus. We annotated blink occurrences in a total of 26 videos after which we looked at two videos of two different signers to try to determine the types of the blinks happening in these videos. We found 10 types of blinks, namely 3 non-linguistic types and 7 linguistic types. Following, we annotated the blink types for a subset of the initial 26 videos, specifically 8 videos from three signers.

Once this step was over, we turned to the automatic blink detection and we noted that the definition of 'blink' for researchers working on their automatic detection was different from the one presented by both physiologists and linguists; what computer science researchers called blinks were considered as eye closures to us and instances of incomplete eye closures during blinkind were not addressed. Most papers published in the field combined a machine learning algorithm, specifically a cascade classifier, and the use of the EAR calculation.

We propose a hybrid method that combines a CNN model and rules. The CNN models are used in phase 1 or our method, to get an idea as to whether the eye is in the *open*, *closed* or what we call *in-between* state. Component 2 of our method consists in resolving the blink problem, that is making a decision as to an actual blink is occurring. In this part, we create rules which we associate with the trained CNN models presented previously.

To evaluate our method, we also implement the EAR calculation and compare the results obtained with these measurements combined with the rules with the results yielded in the same conditions with our CNN models.

We implement two rules, one that we name the high-low-value-difference rule and the other one that is slightly more sophisticated and that we call the curve rule.

We have seen that our method does not systematically display better results than that obtained with the combination of the EAR calculation but even if not always, our CNN models exceeds in many instances the scores obtained with the EAR measurement implementation on the signer on which the CNN models have been trained. As we applied this method to other videos of the corpus, we were able to observe that without prior training of the machine learning algorithms on the other participants, the CNN models combined with the rules were sometimes getting very promising results. We also find it worthy to note that the CNN model, applied to video $S2T2B15$, is always better at predicting the blink occurrences than the EAR method.

It is interesting to add that the CNN model required a different type of annotating, not relying on the blinks and their different phases but rather taking into account the degree of openness of the eyes, therefore, we only were able to categorize the images of two videos, one on which we trained the CNN algorithm and the other one on which we evaluated our model. We believe that with more training data, the CNN model could outperform the EAR method in most instances. As we have noted, there is still a way to go, both for the further elaboration of this method which we present here as a proof of concept but also in the creation of the classification algorithm to categorize the various types of blinks.

Appendix 1: Annotations from the dataset used in examples in section 3.1.4.1

Annotation in example	Full annotation	English translation
QUESTION	QUESTION - POSER UNE / POINT D'INTERROGATION	question - ask a/interrogation point
PRES1	PRES1-TOUT / PROCHE	close by / nearby
POUVOIR	POUVOIR / CAPABLE	to be able to, can / capable
NAGER	NAGER / PLAGE / PISCINE	swim / beach / swimming pool
IL-FAUT	IL-FAUT / DEVOIR	one has to / to have to, must
PHOTOGRAPHIE	PHOTOGRAPHIE / APPAREIL PHOTO	photograph / camera
TON	TON/SON/APPARTENIR-À	your / his, her, its / to belong to
DEUX2:NUM:	DEUX2:NUM:VOT$_d$EUX1:VAR (PAUME VERS SOI)	number 2 (palms toward oneself)
TRANQUILLE	TRANQUILLE1 / PEINARD / AUTOMATIQUE	tranquil, calm / chilled / automatic

PAPIER	PAPIER - FEUILLE DE	paper - sheet of
COMMANDER / ORDRE	COMMANDER / ORDRE / DIRECTIVE	to order / an order / a command
PARFOIS	PARFOIS / ÇA DÉPEND	sometimes / it depends

Appendix 2: Results of the hybrid model applied on videos of other signers.

Figures for video $S2T3B15$. In order, the figures 5.1, 5.2, 5.3, 5.4, and 5.5 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation which do not vary from face to eyes crops. We use the first rule, namely the high-low-value-difference one.

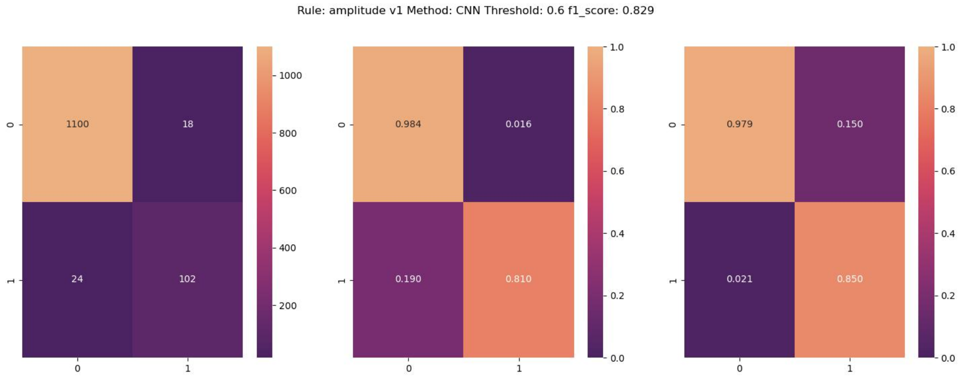


Figure 5.1: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the high-low-value-difference rule. (video $S2T3B15$).

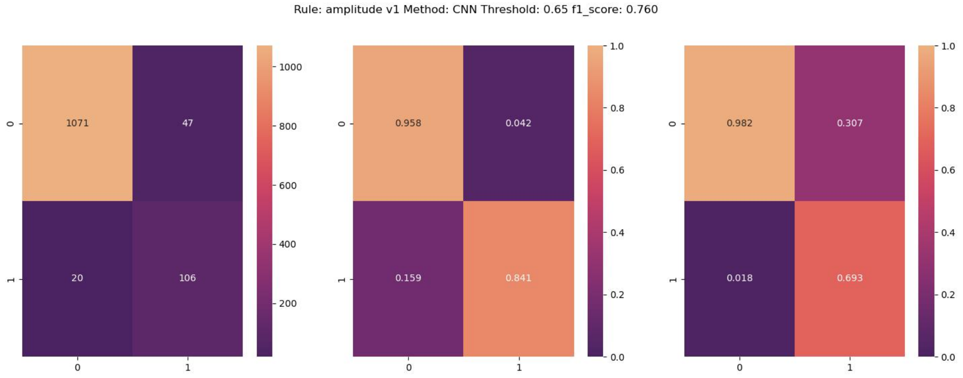


Figure 5.2: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the high-low-value-difference rule. (video $S2T3B15$).

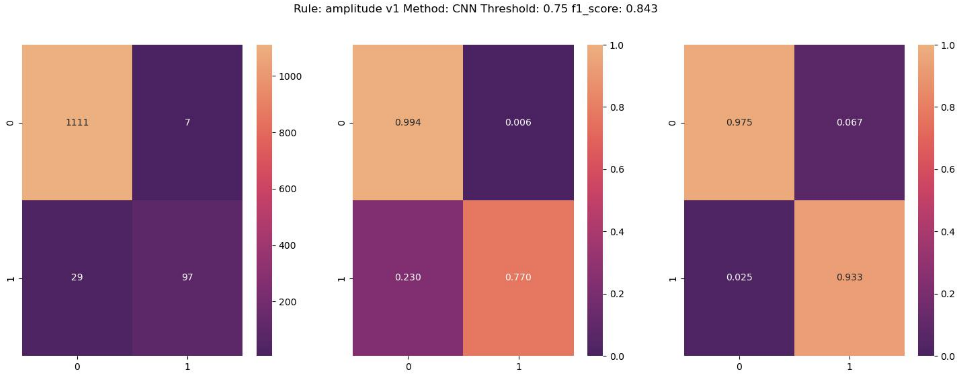


Figure 5.3: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the high-low-value-difference rule. (video $S2T3B15$).

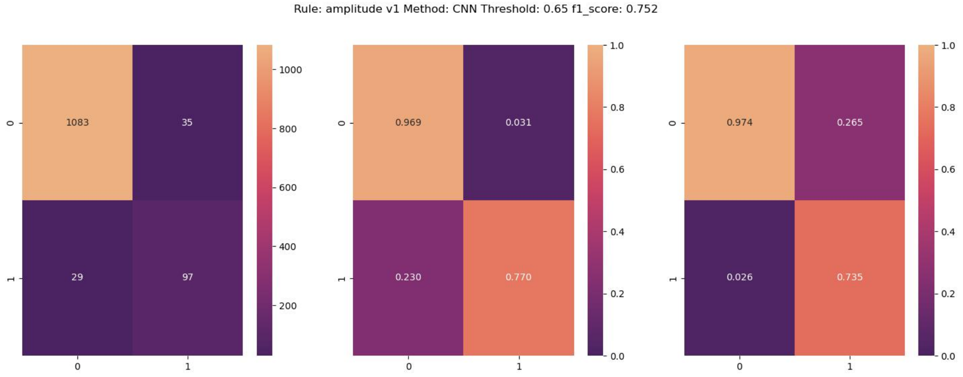


Figure 5.4: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the high-low-value-difference rule. (video $S2T3B15$).

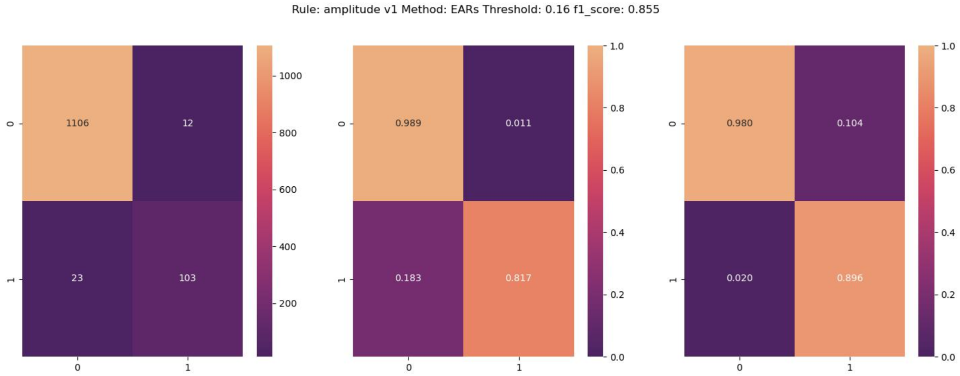


Figure 5.5: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the high-low-value-difference rule. (video $S2T3B15$).

Figures for video $S2T3B15$. In order, the figures 5.6, 5.7, 5.8, 5.9, and 5.10 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation, all while using the second rule, namely the curve rule.

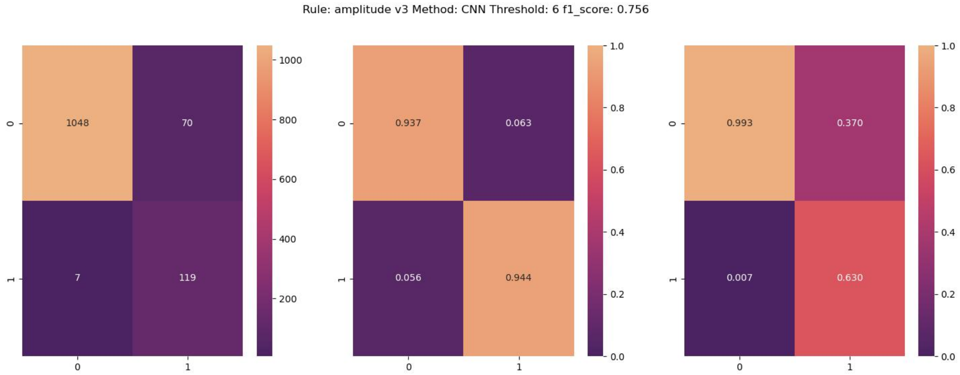


Figure 5.6: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the curve rule. (video $S2T3B15$).

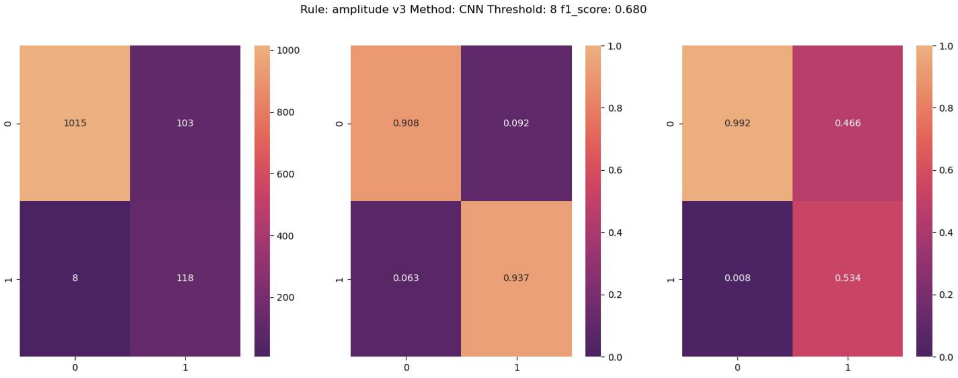


Figure 5.7: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the curve rule. (video $S2T3B15$).

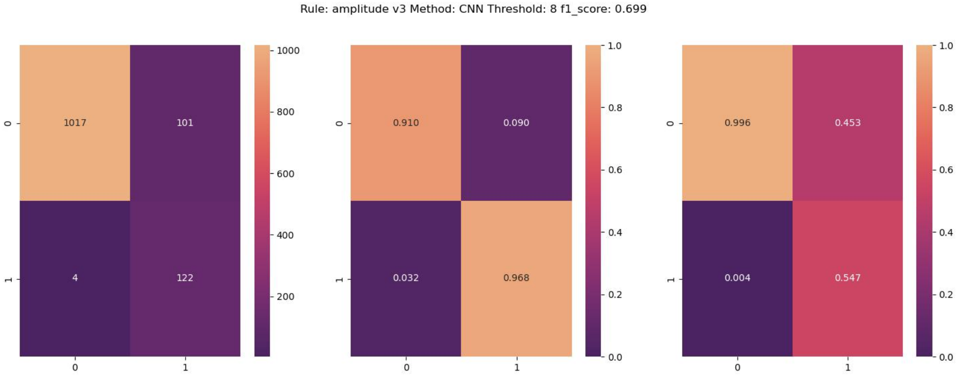


Figure 5.8: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the curve rule. (video $S2T3B15$).

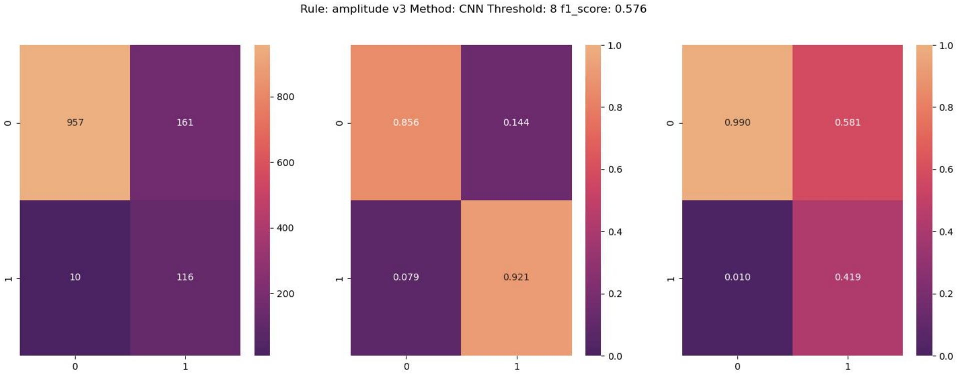


Figure 5.9: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the curve rule. (video $S2T3B15$).

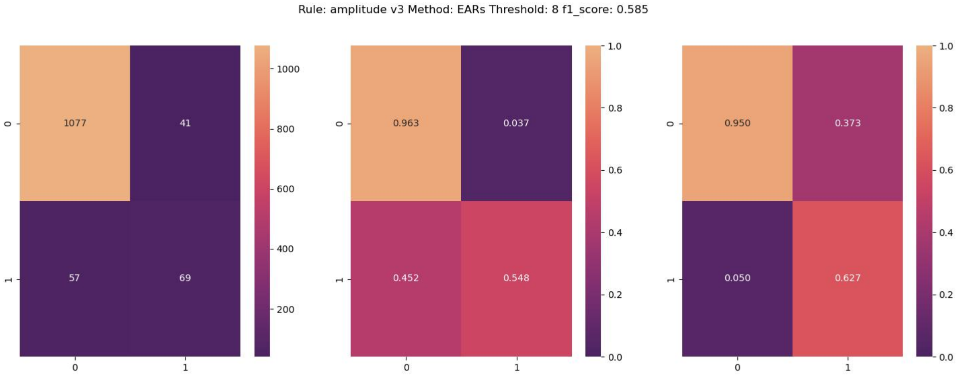


Figure 5.10: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the curve rule. (video $S2T3B15$).

Figures for video $S2T4B15$. In order, the figures 5.11, 5.12, 5.13, 5.14, and 5.15 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation which do not vary from face to eyes crops. We use the first rule, namely the high-low-value-difference one.

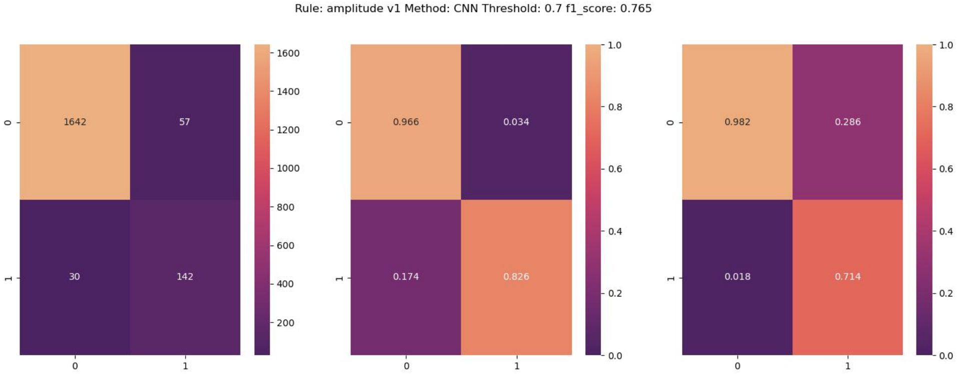


Figure 5.11: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the high-low-value-difference rule. (video $S2T4B15$)..

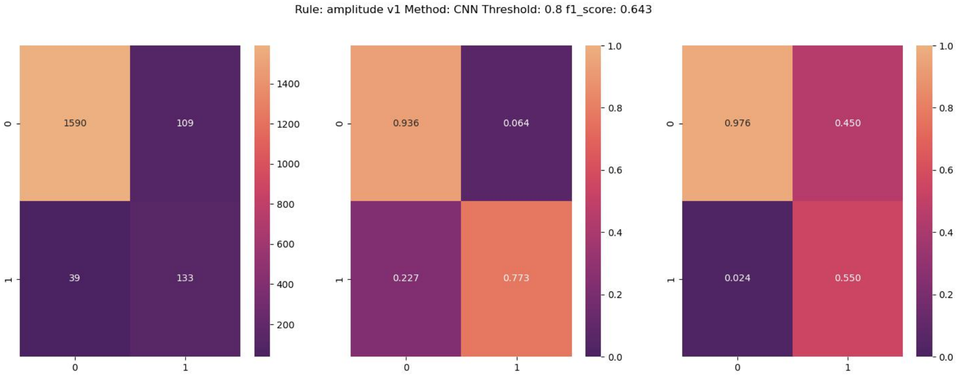


Figure 5.12: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the high-low-value-difference rule. (video $S2T4B15$)..

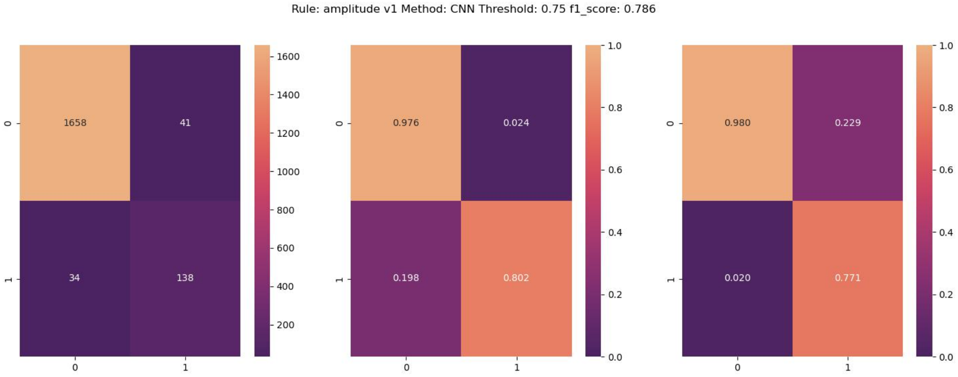


Figure 5.13: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the high-low-value-difference rule. (video $S2T4B15$)..

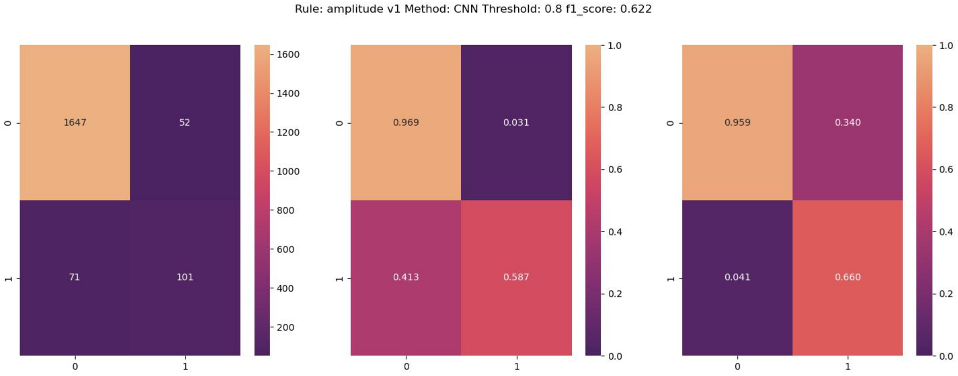


Figure 5.14: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the high-low-value-difference rule. (video $S2T4B15$)..

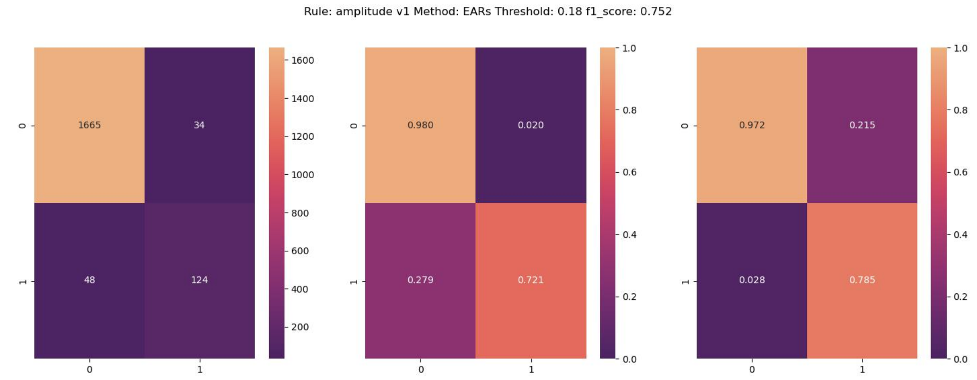


Figure 5.15: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the high-low-value-difference rule. (video $S2T4B15$).

Figures for video $S2T4B15$. In order, the figures 5.16, 5.17, 5.18, 5.19, and 5.20 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation, all while using the second rule, namely the curve rule.

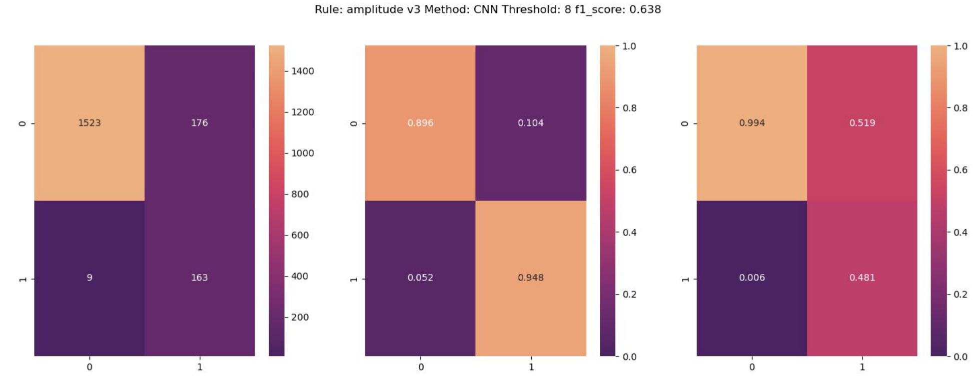


Figure 5.16: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the curve rule. (video $S2T4B15$).

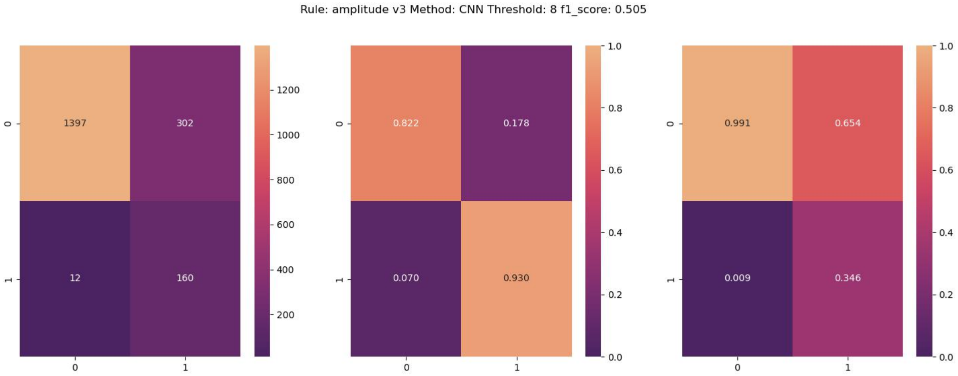


Figure 5.17: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the curve rule. (video $S2T4B15$).

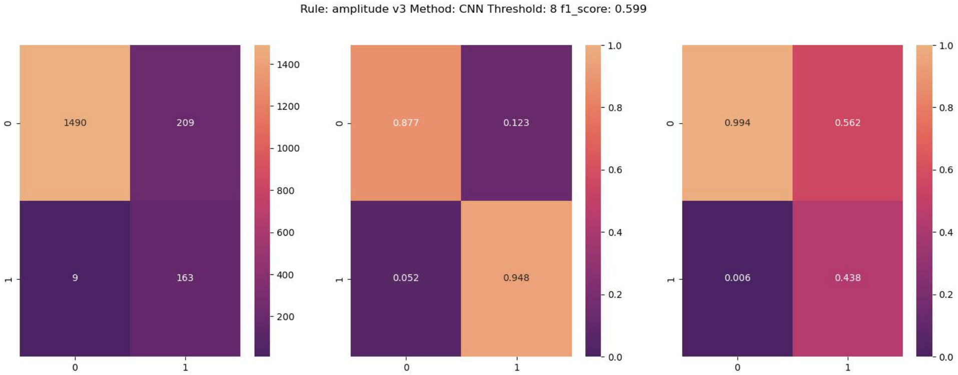


Figure 5.18: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the curve rule. (video $S2T4B15$).

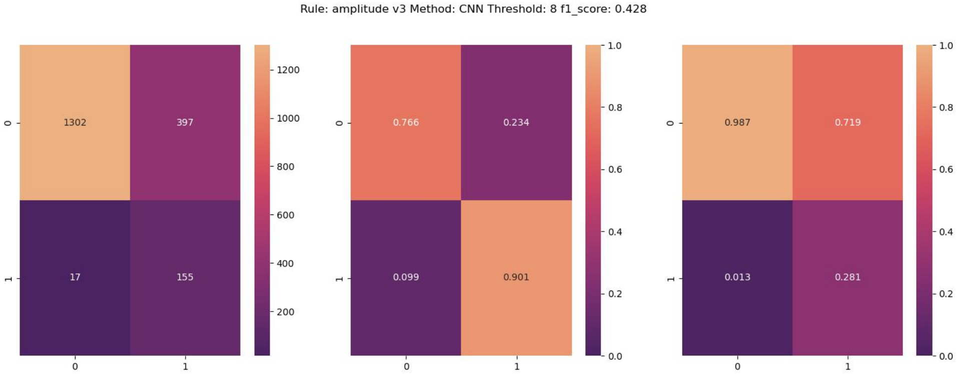


Figure 5.19: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the curve rule. (video $S2T4B15$).

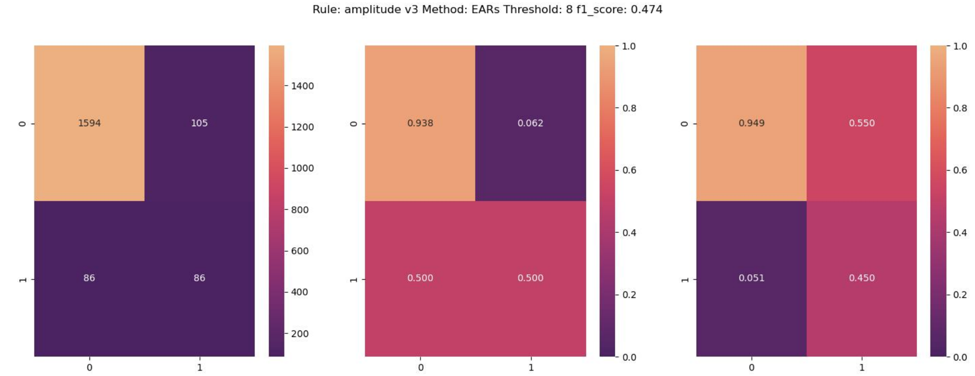


Figure 5.20: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the curve rule. (video $S2T4B15$).

Figures for video $S2T2A11$. In order, the figures 5.21, 5.22, 5.23, 5.24, and 5.25 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation which do not vary from face to eyes crops. We use the first rule, namely the high-low-value-difference one.

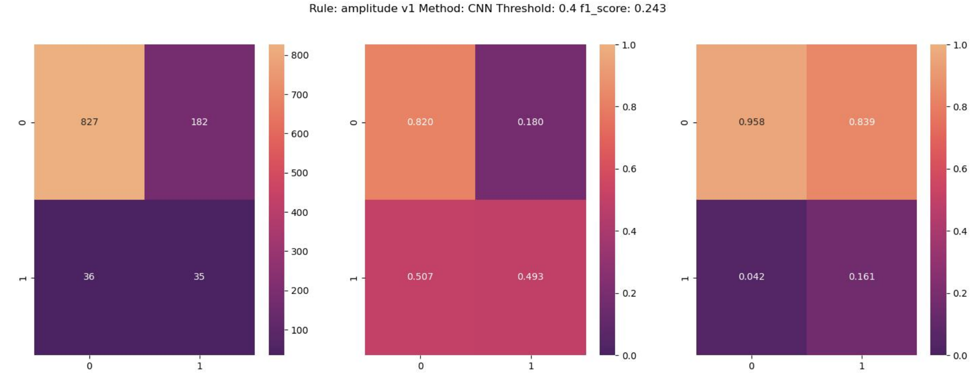


Figure 5.21: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the high-low-value-difference rule. (video $S2T2A11$).

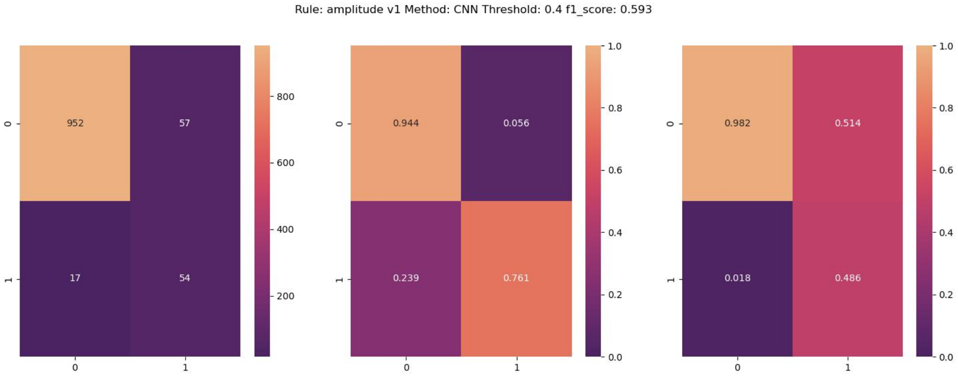


Figure 5.22: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the high-low-value-difference rule. (video $S2T2A11$).

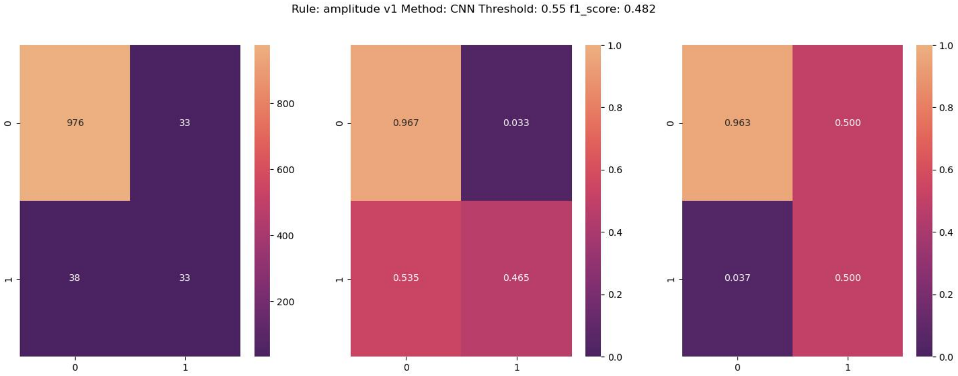


Figure 5.23: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the high-low-value-difference rule. (video $S2T2A11$).

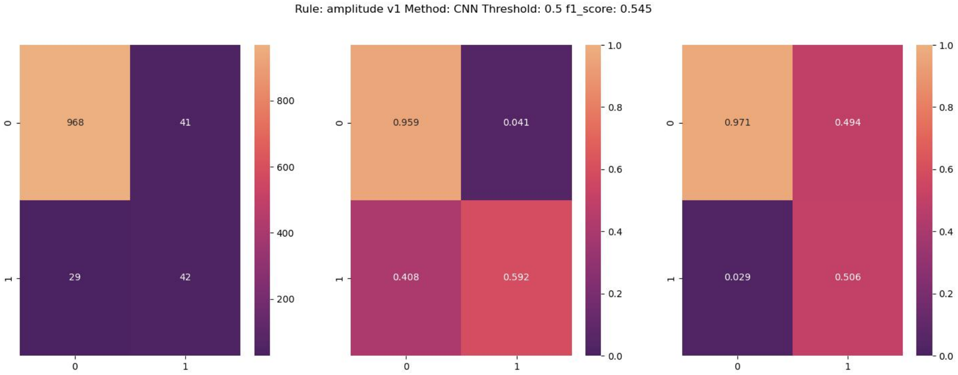


Figure 5.24: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the high-low-value-difference rule. (video $S2T2A11$).

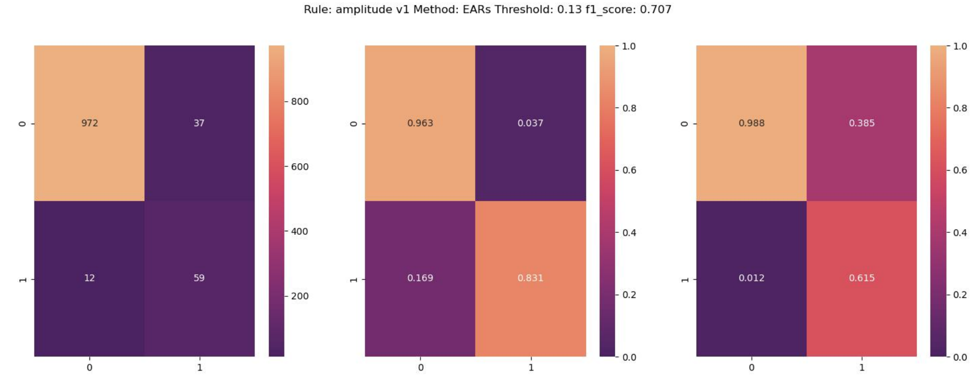


Figure 5.25: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the high-low-value-difference rule. (video $S2T2A11$).

The following figures represent the results obtained using the curve rule, on video $S2T2A11$. In order, the figures 5.26, 5.27, 5.28, 5.29, and 5.30 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation.

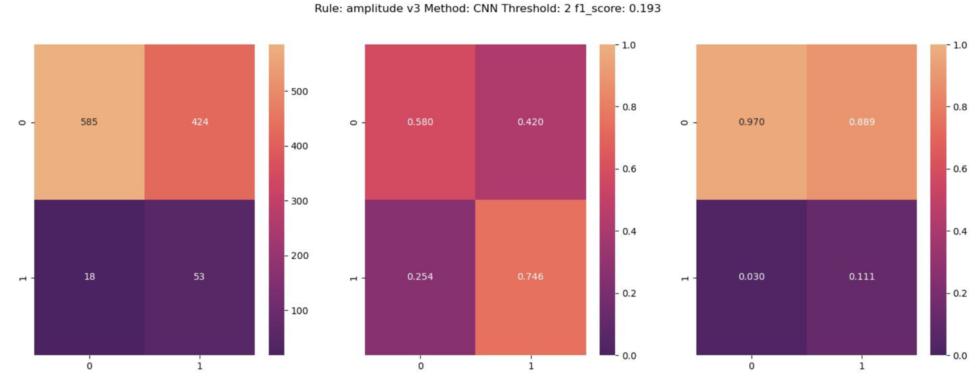


Figure 5.26: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the curve rule. (video $S2T2A11$).

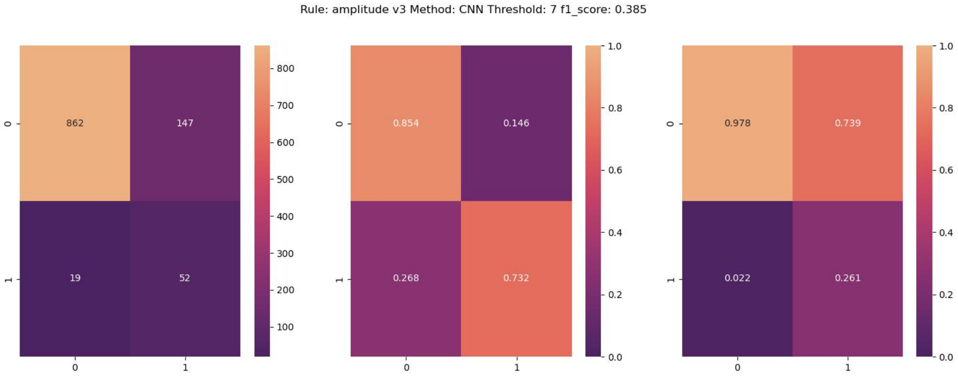


Figure 5.27: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the curve rule. (video $S2T2A11$).

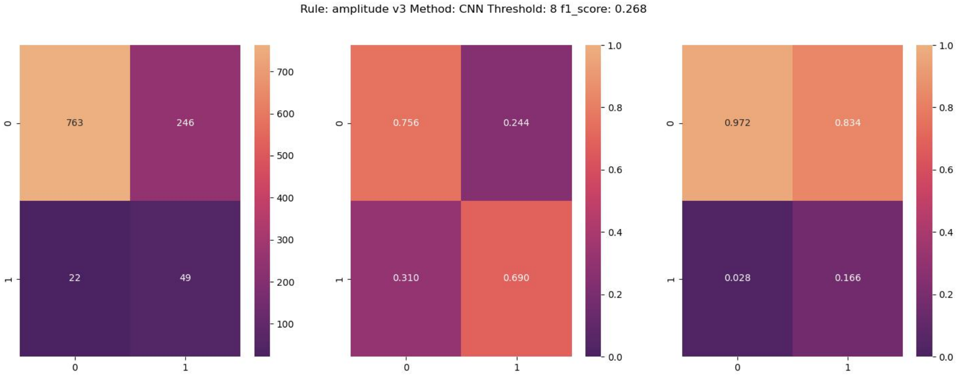


Figure 5.28: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the curve rule. (video $S2T2A11$).

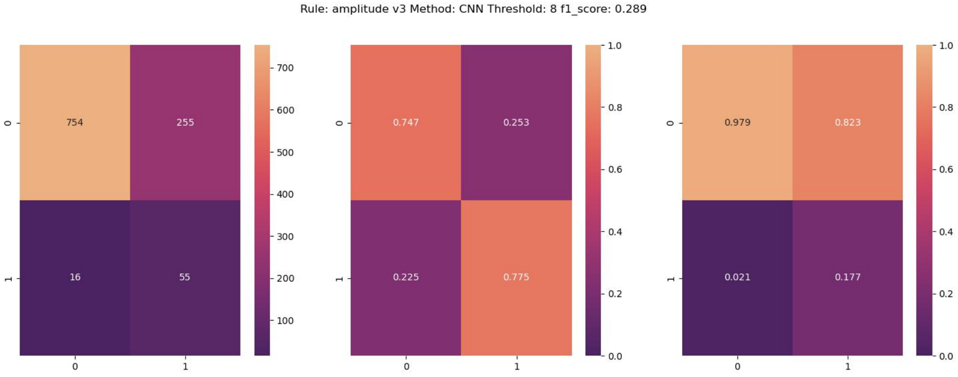


Figure 5.29: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the curve rule. (video $S2T2A11$).

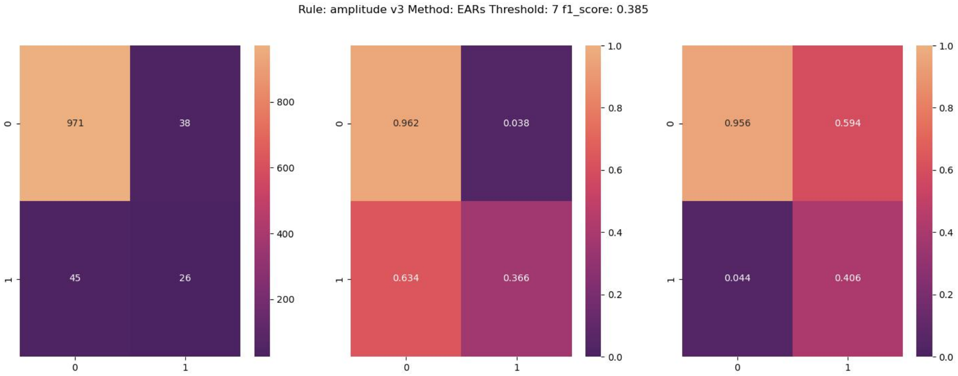


Figure 5.30: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the curve rule. (video $S2T2A11$).

Figures for video $S4T6B14$. In order, the figures 5.31, 5.32, 5.33, 5.34, and 5.35 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation which do not vary from face to eyes crops. We use the first rule, namely the high-low-value-difference one.

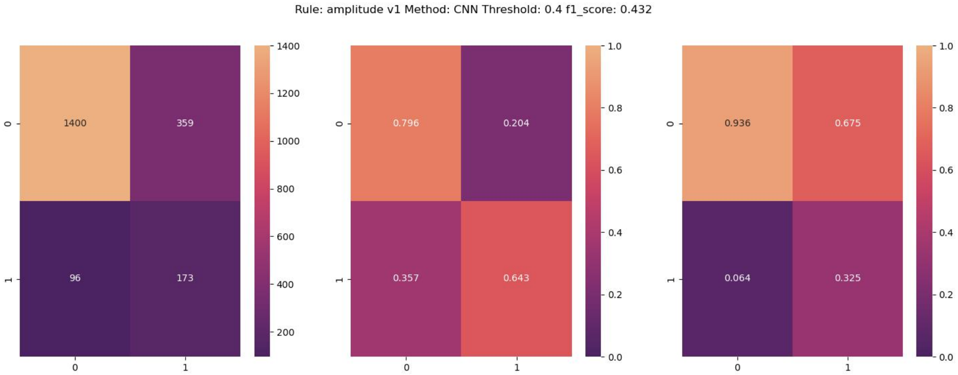


Figure 5.31: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the high-low-value-difference rule. (video $S4T6B14$).

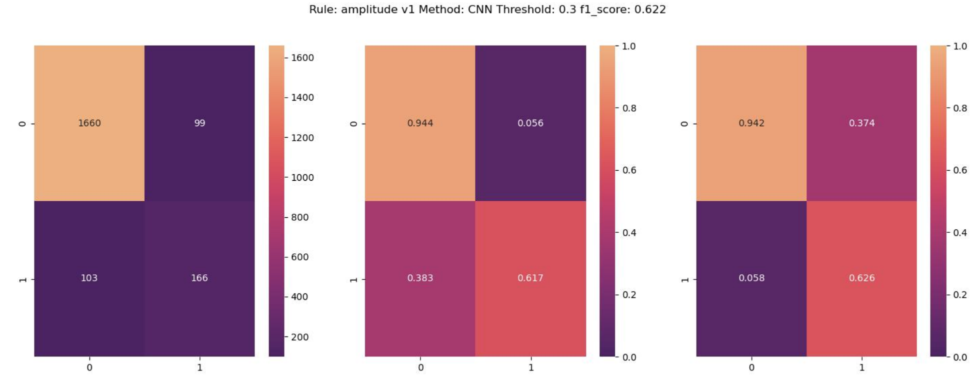


Figure 5.32: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the high-low-value-difference rule. (video $S4T6B14$).

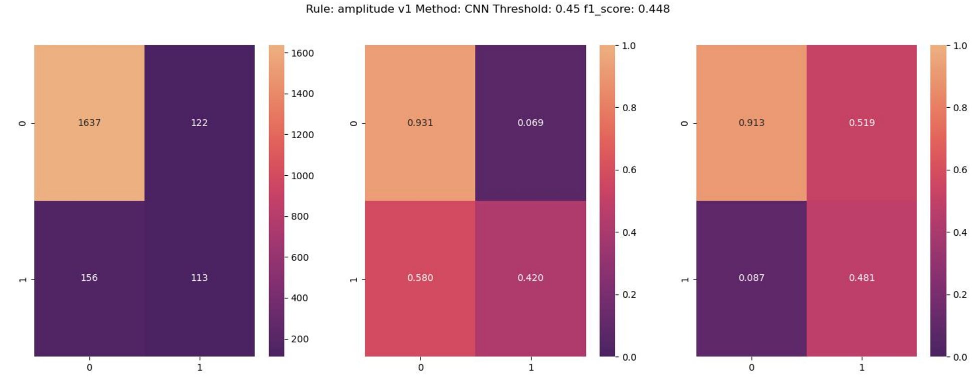


Figure 5.33: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the high-low-value-difference rule. (video $S4T6B14$).

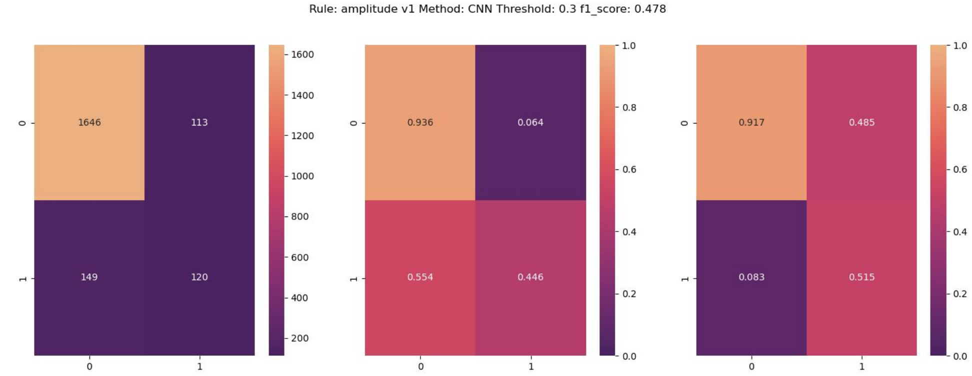


Figure 5.34: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the high-low-value-difference rule. (video $S4T6B14$).

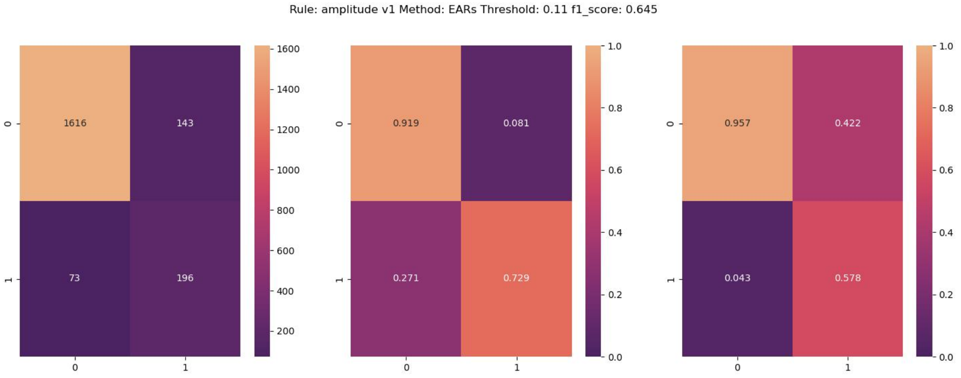


Figure 5.35: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the high-low-value-difference rule. (video $S4T6B14$).

The following figures represent the results obtained using the curve rule, on video $S4T6B14$. In order, the figures 5.36, 5.37, 5.38, 5.39, and 5.40 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation.

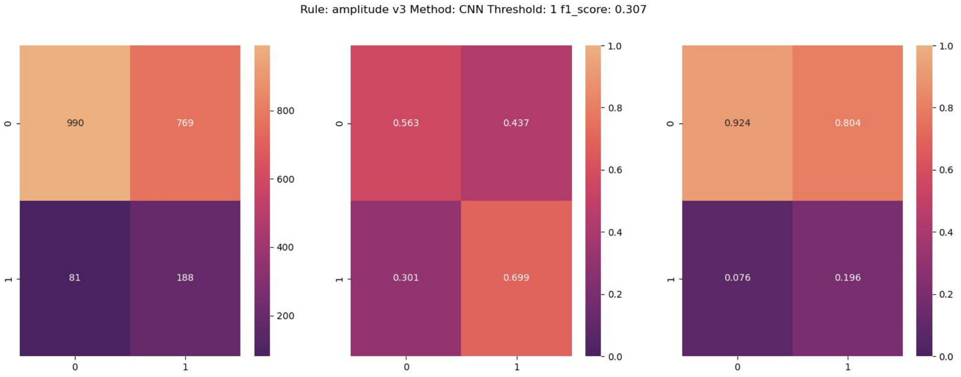


Figure 5.36: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the curve rule. (video $S4T6B14$).

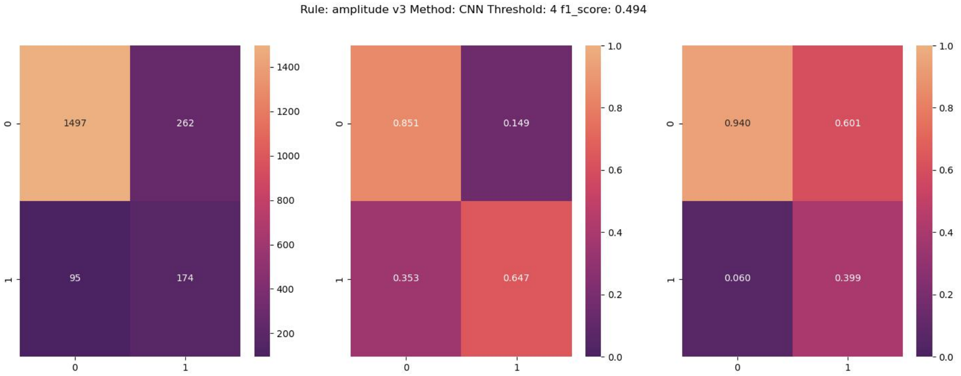


Figure 5.37: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the curve rule. (video $S4T6B14$).

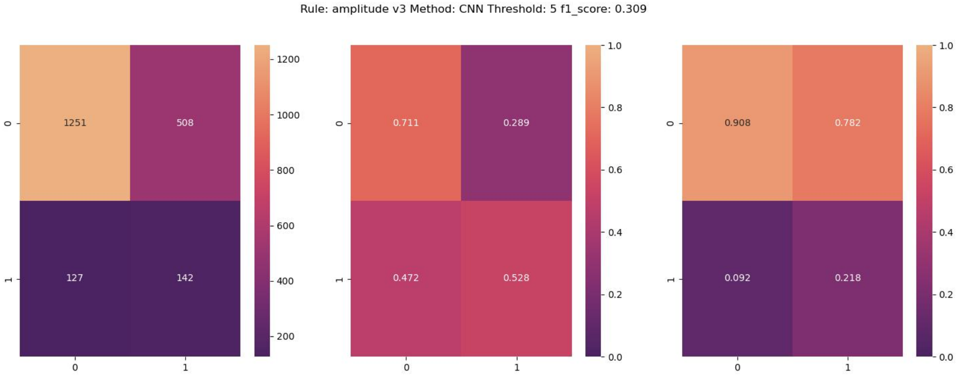


Figure 5.38: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the curve rule. (video $S4T6B14$).

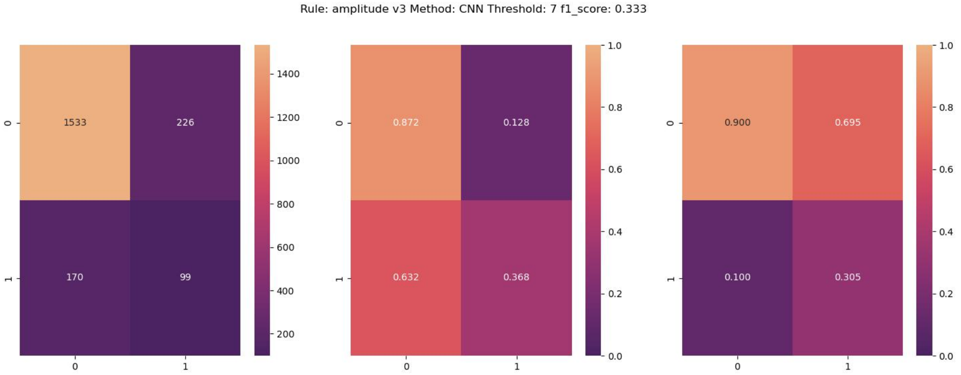


Figure 5.39: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the curve rule. (video $S4T6B14$).

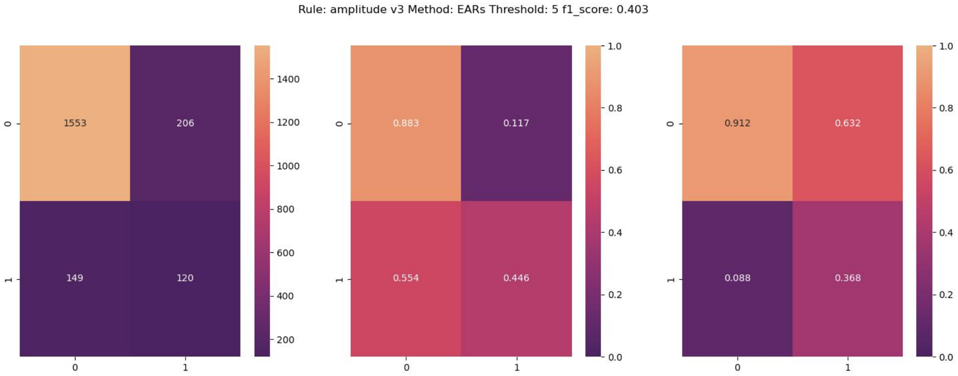


Figure 5.40: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the curve rule. (video $S4T6B14$).

Figures for video $S5T1A9$. In order, the figures 5.41, 5.42, 5.43, 5.44, and 5.45 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation which do not vary from face to eyes crops. We use the first rule, namely the high-low-value-difference one.

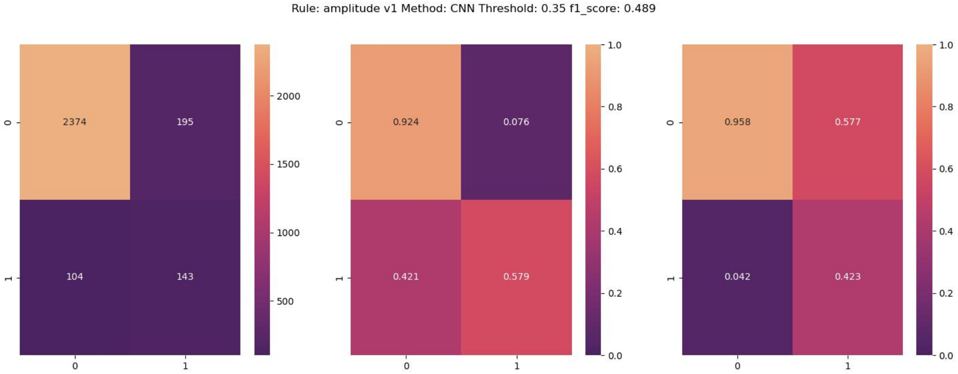


Figure 5.41: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the high-low-value-difference rule. (video $S5T1A9$).

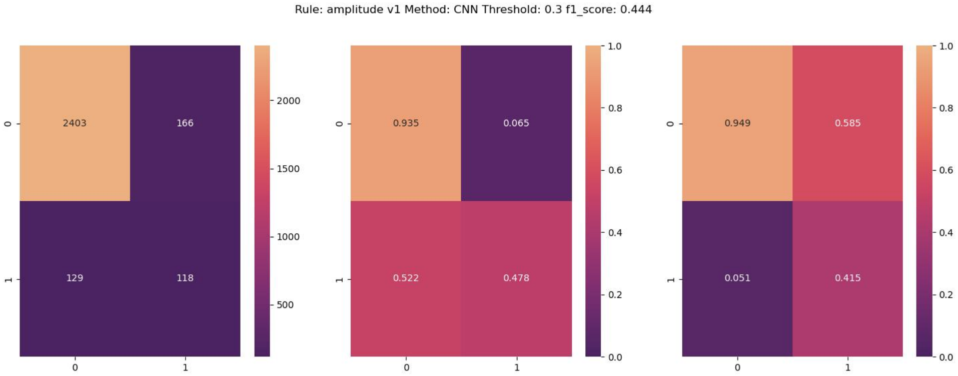


Figure 5.42: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the high-low-value-difference rule. (video $S5T1A9$).

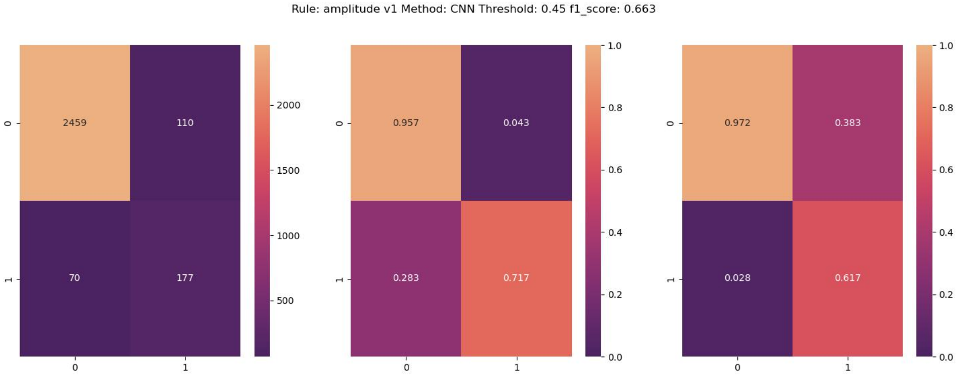


Figure 5.43: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the high-low-value-difference rule. (video $S5T1A9$).

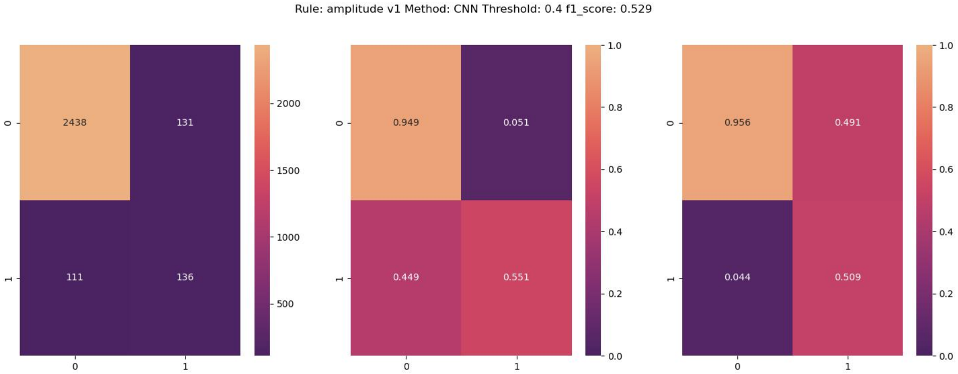


Figure 5.44: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the high-low-value-difference rule. (video $S5T1A9$).

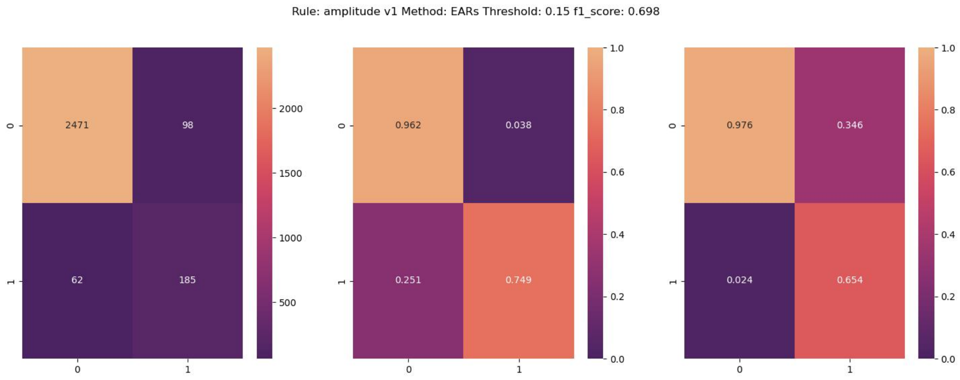


Figure 5.45: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the high-low-value-difference rule. (video $S5T1A9$).

The following figures represent the results obtained using the curve rule, on video $S5T1A9$. In order, the figures 5.46, 5.47, 5.48, 5.49, and 5.50 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation.

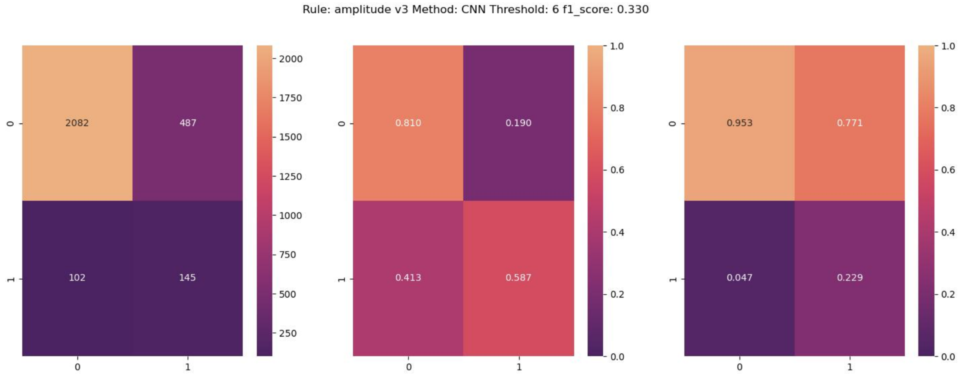


Figure 5.46: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the curve rule. (video $S5T1A9$).

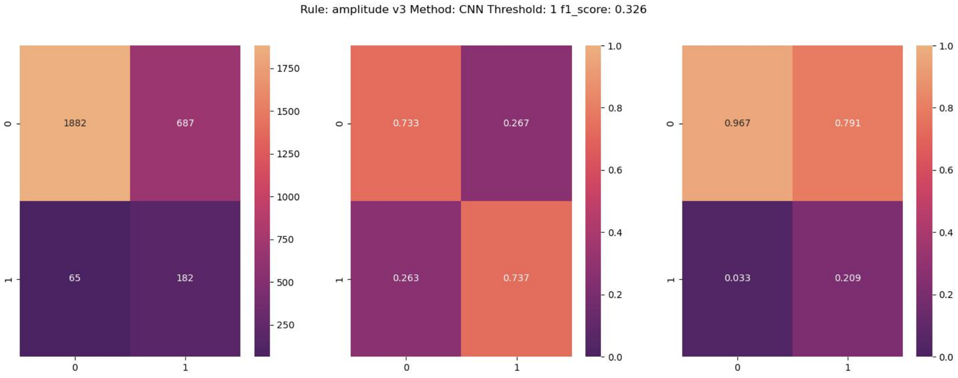


Figure 5.47: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the curve rule. (video $S5T1A9$).

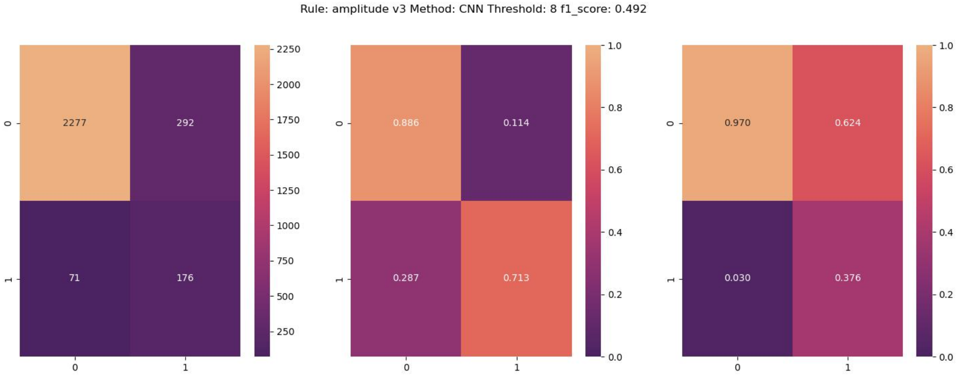


Figure 5.48: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the curve rule. (video $S5T1A9$).

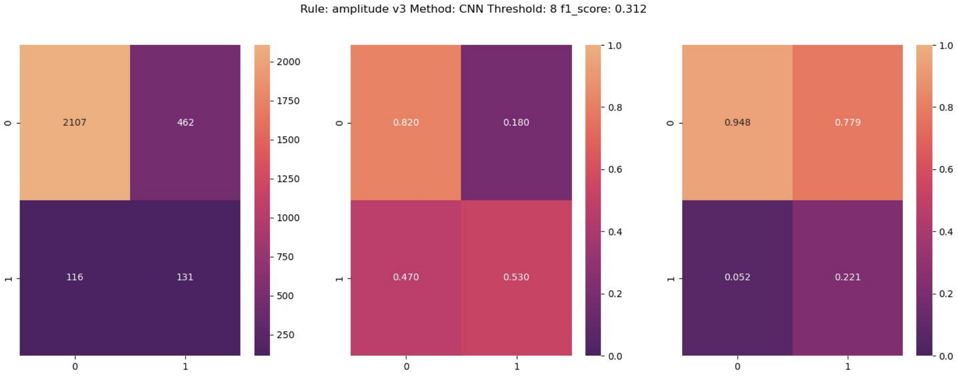


Figure 5.49: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the curve rule. (video $S5T1A9$).

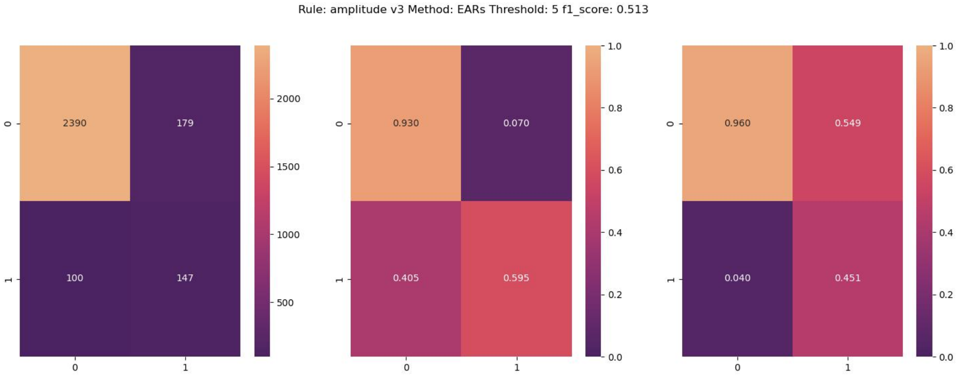


Figure 5.50: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the curve rule. (video $S5T1A9$).

Figures for video $S9T6B5$. In order, the figures 5.51, 5.52, 5.53, 5.54, and 5.55 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation which do not vary from face to eyes crops. We use the first rule, namely the high-low-value-difference one.

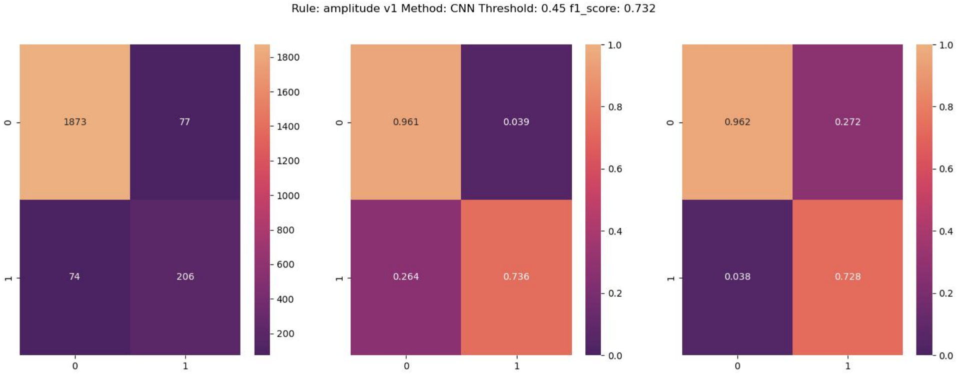


Figure 5.51: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the high-low-value-difference rule. (video $S9T6B5$).

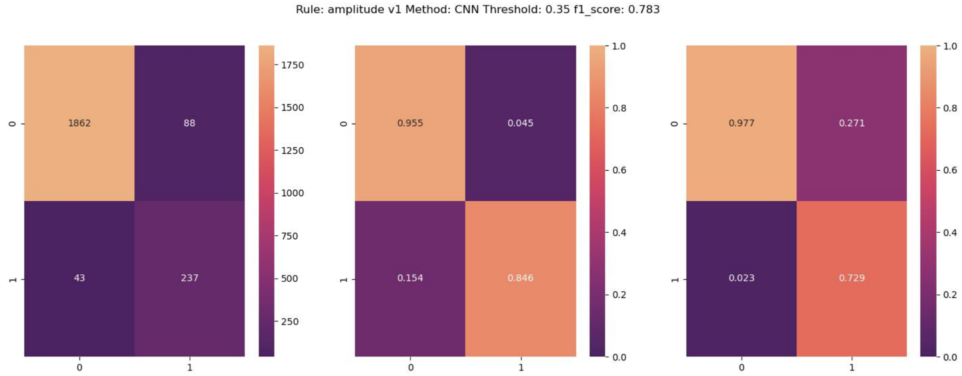


Figure 5.52: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the high-low-value-difference rule. (video $S9T6B5$).

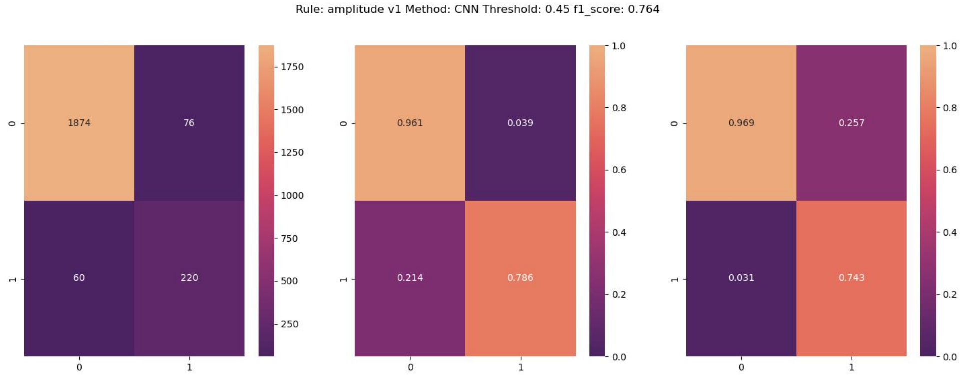


Figure 5.53: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the high-low-value-difference rule. (video $S9T6B5$).

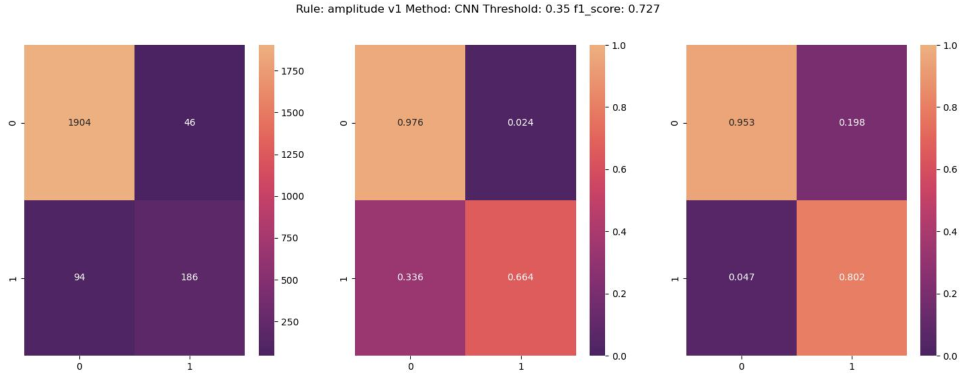


Figure 5.54: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the high-low-value-difference rule. (video $S9T6B5$).

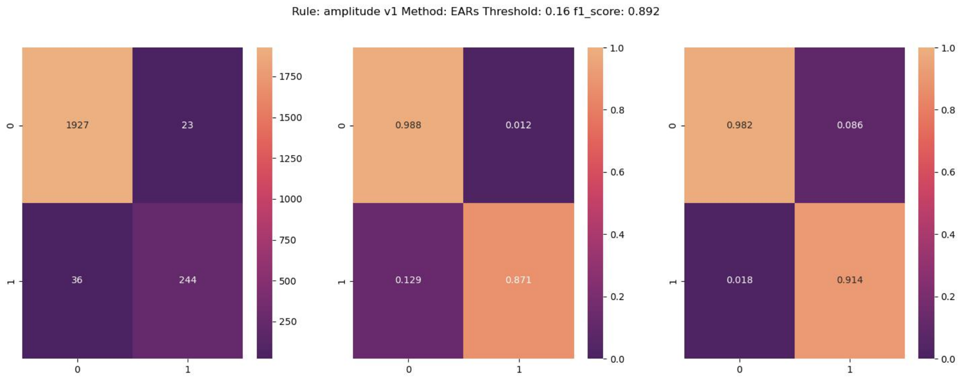


Figure 5.55: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the high-low-value-difference rule. (video $S9T6B5$).

The following figures represent the results obtained using the curve rule, on video $S9T6B5$. In order, the figures 5.56, 5.57, 5.58, 5.59, and 5.60 show the confusion matrices of the hybrid algorithm using the classification model on eyes, on faces, the regression model on eyes, on faces, and the EAR calculation.

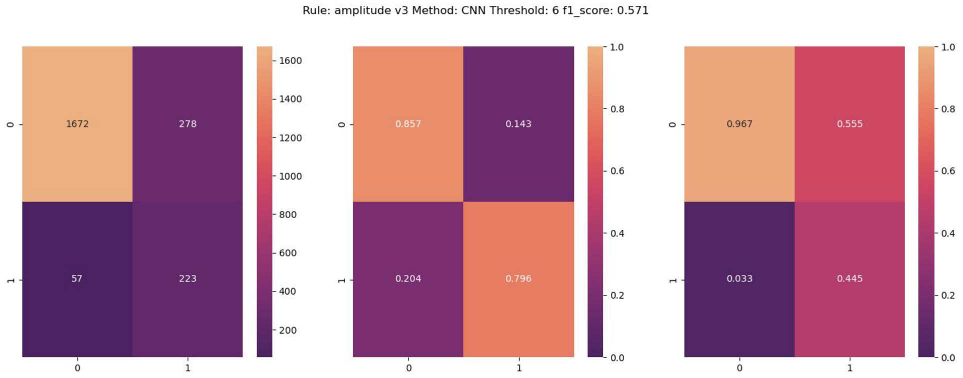


Figure 5.56: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN classifier with the curve rule. (video $S9T6B5$).

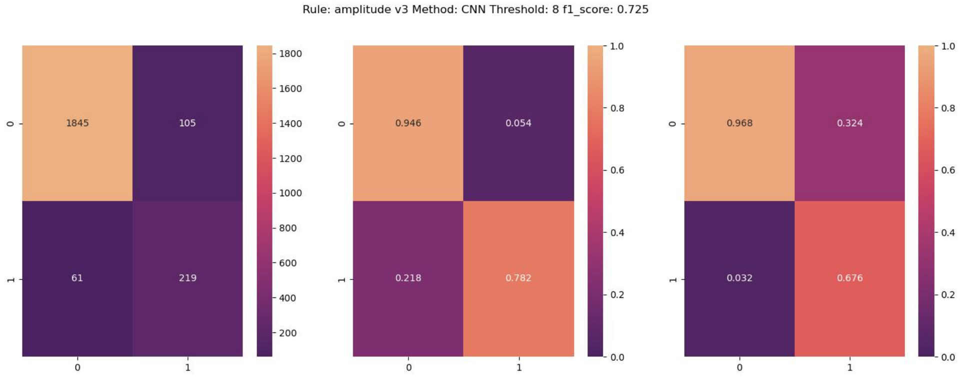


Figure 5.57: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN classifier with the curve rule. (video $S9T6B5$).

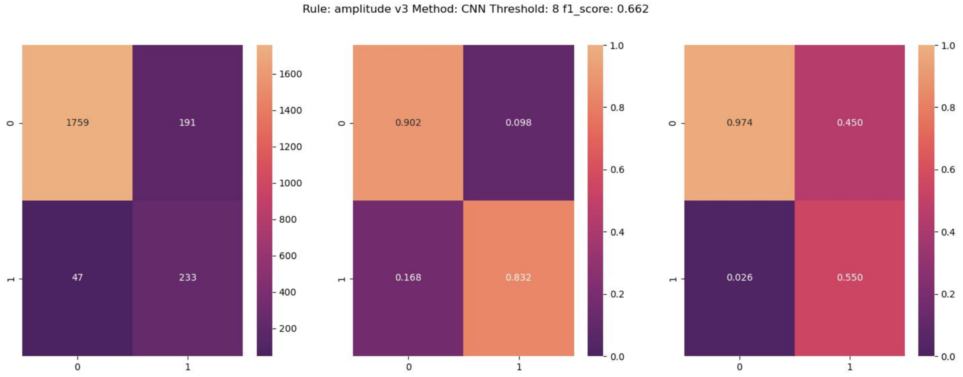


Figure 5.58: Confusion matrices for the best threshold of the hybrid model on eyes crops using the CNN regressor with the curve rule. (video $S9T6B5$).

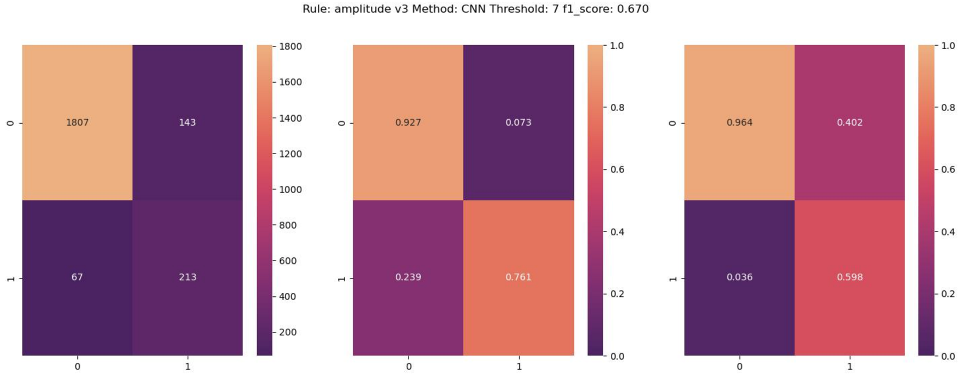


Figure 5.59: Confusion matrices for the best threshold of the hybrid model on face crops using the CNN regressor with the curve rule. (video $S9T6B5$).

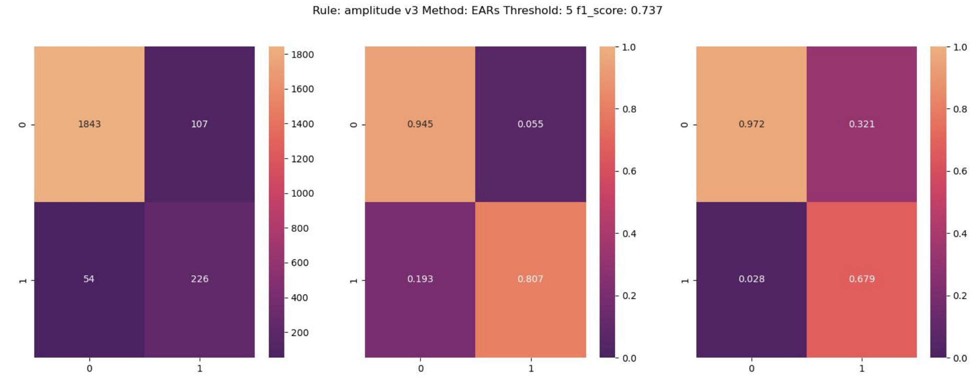


Figure 5.60: Confusion matrices for the best threshold of the hybrid model using the EAR measurements with the curve rule. (video $S9T6B5$).

Bibliography

C. D. N. Ayudhya and T. Srinark. A method for real-time eye blink detection and its application. In *The 6th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 25–30, 2009.

A. Baker. *The linguistics of sign languages: An introduction.* John Benjamins Publishing Company, 2016.

A. Baker and B. van den Bogaerde. *Interaction and discourse.* John Benjamins Publishing Company, 2016.

C. Baker and D. Cokely. *American Sign Language: A teacher's resource text on grammar and culture.* Silver Spring, MD: TJ Publishers, 1980.

C. Baker and C. Padden. Focusing on the nonmanual components of american sign language. understanding language through sign language research, ed. by p. siple, 27-57, 1978.

V. Belissen, A. Braffort, and M. Gouiffès. Dicta-sign-lsf-v2: remake of a continuous french sign language dialogue corpus and a first baseline for automatic sign language processing. In *LREC 2020, 12th Conference on Language Resources and Evaluation*, 2020.

A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali, and A. Albanese. Analysis of blink rate patterns in normal subjects. *Movement disorders*, 12(6):1028–1034, 1997.

L. J. Bour, M. Aramideh, and B. W. Ongerboer De Visser. Neurophysiological aspects of eye and eyelid movements during blinking in humans. *Journal of neurophysiology*, 83 (1):166–176, 2000.

A. Braffort and E. Chételat-Pelé. Analysis and description of blinking in french sign language for automatic generation. In *International Gesture Workshop*, pages 173–182. Springer, 2011.

C. Branchini and C. Donati. Relatively diflerent italian sign language relative clauses. *Correlatives cross-linguistically*, 1:157, 2009.

D. Brentari. Phonology. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

D. Brentari and L. Crossley. Prosody on the hands and face: Evidence from american sign language. *Sign Language & Linguistics*, 5(2):105–130, 2002.

D. Brentari, J. Falk, A. Giannakidou, A. Herrmann, E. Volk, and M. Steinbach. Production and comprehension of prosodic markers in sign language imperatives. *Frontiers in psychology*, 9:770, 2018.

L. G. Carney and R. M. Hill. The nature of normal blinking patterns. *Acta ophthalmologica*, 60(3):427–433, 1982.

C. Cecchetto. Sentence types. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

E. Chételat-Pelé. *Les Gestes Non Manuels en langue des signes française; Annotation, analyse et formalisation: Application aux mouvements des sourcils et aux clignements des yeux.* PhD thesis, Université de Provence-Aix-Marseille I, 2010.

I. Choi, S. Han, and D. Kim. Eye detection and eye blink detection using adaboost learning and grouping. In *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*, pages 1–4, 2011. doi: 10.1109/ICCCN.2011.6005896.

M. J. Collins, D. R. Iskander, A. Saunders, S. Hook, E. Anthony, and R. Gillon. Blinking patterns and corneal staining. *Eye & contact lens*, 32(6):287–293, 2006.

O. A. Crasborn. Nonmanual structures in sign language. In *Encyclopaedia of Languages and Linguistics, 2nd ed.* Amesterdam: Elsevier, 2006.

S. Dachkovsky and W. Sandler. Visual intonation in the prosody of a sign language. *Language and speech*, 52(2-3):287–314, 2009.

C. Dewi, X. Jiang, and H. Yu. Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks. *PeerJ Computer Science*, 8:e943, 2022. https://doi.org/10.7717/peerj-cs.943.

M. J. Doughty. Further assessment of gender-and blink pattern-related differences in the spontaneous eyeblink activity in primary gaze in young adult humans. *Optometry and Vision Science*, 79(7):439–447, 2002.

P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

C. Evinger, M. Shaw, C. Peck, K. Manning, and R. Baker. Blinking and associated eye movements in humans, guinea pigs, and rabbits. *Journal of Neurophysiology*, 52(2): 323–339, 1984.

S. D. Fischer. Questions and negation in american sign language. In *Interrogative and negative constructions in sign language.* Ishara Press, 2006.

I. Grishchenko and V. Bazarevsky. Mediapipe holistic—simultaneous face, hand and pose prediction, on device, 2020. Retrieved from: https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html.

K. Gökgöz. Negation in turkish sign language: The syntax of nonmanual markers. In *Nonmanuals in sign language (Vol. 53).* John Benjamins Publishing, 2013.

A. Hall. The origin and purposes of blinking. *The British journal of ophthalmology*, 29 (9):445, 1945.

A. J. Hall. Some observations on the acts of closing and opening the eyes. *The British Journal of Ophthalmology*, 20(5):257, 1936.

A. Herrmann. *Focus Particles in Sign Languages.* De Gruyter Mouton, 2013.

Hopkins. Electromyography, n.d. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electromyography-emg.

B. R. Ibrahim, F. M. Khalifa, S. R. M. Zeebaree, N. A. Othman, A. Alkhayyat, R. R. Zebari, and M. A. M. Sadeeq. 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, 2021. doi: 10.1109/BICITS51482.2021.9509908.

T. Johnston and L. De Beuzeville. *Auslan corpus annotation guidelines*. 2016.

T. Johnston and A. Schembri. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007.

K. Kaneko and K. Sakamoto. Evaluation of three types of blinks with the use of electro-oculogram and electromyogram. *Perceptual and motor skills*, 88(3):1037–1052, 1999.

C. N. Karson. Spontaneous eye-blink rates and dopaminergic systems. *Brain*, 106(3): 643–653, 1983.

C. N. Karson, K. F. Berman, E. F. Donnelly, W. B. Mendelson, J. E. Kleinman, and R. J. Wyatt. Speaking, thinking, and blinking. *Psychiatry research*, 5(3):243–246, 1981.

V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. doi: 10.1109/CVPR.2014.241.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau. Real-time eye blink detection with gpu-based sift tracking. In *Fourth Canadian Conference on Computer and Robot Vision (CRV'07)*, pages 481–487. IEEE, 2007.

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

D. Lewin and A. C. Schembri. Mouth gestures in british sign language: A case study of tongue protrusion in bsl narratives. *Sign Language & Linguistics*, 14(1):94–114, 2011.

S. K. Liddell. Nonmanual signals and relative clauses in american sign language. In *Proceedings of the First National Symposium on Sign Language Research and Teaching.*, page 193–228, 1978.

LIMSI. Dicta-sign-lsf-v2, 2022. URL `https://hdl.handle.net/11403/dicta-sign-lsf-v2/v1`. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Macula-Retina-Institute. Electro-oculogram, n.d. `https://www.maculaandretinainstitute.com/tests-treatments/electro-oculogram-eog/`.

I. Meir. Word classes and word formation. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

A. Millet. *Grammaire descriptive de la langue des signes française.* UGA Editions, 2020.

M. W. Morgan. Interrogatives and negatives in japanese sign language (jsl). In *Interrogative and negative constructions in sign language.* Ishara Press, 2006.

T. Moriyama, T. Kanade, J. F. Cohn, J. Xiao, Z. Ambadar, J. Gao, and H. Imamura. Automatic recognition of eye blinking in spontaneously occurring behavior. In *Object recognition supported by user interaction for service robots*, volume 4, pages 78–81. IEEE, 2002.

S. G. Müller and F. Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 774–782, 2021.

C. Neidle and J. Nash. The noun phrase. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

G. Nousias, E.-K. Panagiotopoulou, K. Delibasis, A.-M. Chaliasou, A.-M. Tzounakou, and G. Labiris. Video-based eye blink identification and classification. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3284–3293, 2022. doi: 10.1109/JBHI.2022.3153407.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

R. Pfau. *Syntax: complex sentences.* John Benjamins Publishing Company, 2016.

R. Pfau and H. Bos. *Syntax: simple sentences.* John Benjamins Publishing Company, 2016.

R. Pfau and J. Quer. *Sign Languages*, chapter Nonmanuals: their grammatical and prosodic roles, pages 381–402. Cambridge: Cambridge University Press, 2010.

R. Pfau, M. Steinbach, and B. Woll. Tense, aspect, and modality. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

T. T. Phuong, L. T. Hien, N. D. Vinh, et al. 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. IEEE, 2022.

E. Ponder and W. Kennedy. On the act of blinking. *Quarterly journal of experimental physiology: Translation and integration*, 18(2):89–110, 1927.

J. Quer. Negation. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

J. Reilly. How faces come to serve grammar: The development of nonmanual morphology in american sign language. In *Advances in the sign language development of deaf children*. Oxford University Press on Demand, 2005.

W. Sandler. The medium and the message: Prosodic interpretation of linguistic content in israeli sign language. *Sign Language & Linguistics*, 2(2):187–215, 1999.

W. Sandler. Visual prosody. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

E. Selkirk. The prosodic structure of function words. In *Earlier versions of the chapter were presented in the Phonology Proseminar at U Massachusetts, Fall 1992; at the conference" Signal to Syntax" held at Brown U, Feb 1993; at talks at the U Tübingen and the U Konstanz, Sum 1993; and at the 1st Rutgers Optimality Workshop, Oct 1993.* Lawrence Erlbaum Associates, Inc, 1996.

C. Sforza, M. Rango, D. Galante, N. Bresolin, and V. F. Ferrario. Spontaneous blinking in healthy persons: an optoelectronic study of eyelid motion. *Ophthalmic and Physiological Optics*, 28(4):345–353, 2008.

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

T. Soukupová and J. Čech. Eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop*, 2016.

M. Steinbach. Plurality. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

F. Sze. Blinks and intonational phrasing in hong kong sign language. In *Signs of the time*, pages 83–107, 2004.

F. Sze. Nonmanual markings for topic constructions in hong kong sign language. *Sign Language & Linguistics*, 14(1):115–147, 2011.

G. Tang and P. Lau. Coordination and subordination. In *Sign Language: An International Handbook*, Handbooks of Linguistics and Communication Science. De Gruyter Mouton, 2012.

R. Thompson, K. Emmorey, and R. Kluender. The relationship between eye gaze and verb agreement in american sign language: An eye-tracking study. *Natural Language & Linguistic Theory*, 24(2):571–604, 2006.

E. van der Kooij and O. Crasborn. *Phonology.* John Benjamins Publishing Company, 2016.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

L. Wang, X. Ding, C. Fang, C. Liu, and K. Wang. Eye blink detection based on eye contour extraction. In *Image Processing: Algorithms and Systems VII*, volume 7245, pages 222–228. SPIE, 2009.

E. W. Weisstein. Affine transformation, 2004. `https://mathworld.wolfram.com/AffineTransformation.html`.

R. Wilbur. Eyeblinks & asl phrase structure. *Sign Language Studies*, 84(1):221–240, 1994.

R. B. Wilbur. Phonological and prosodic layering of nonmanuals in american sign language. In *The signs of language revisited*, pages 196–220. Psychology Press, 2013.

S. Wilcox and B. Shaffer. Modality in american sign language. In *The expression of modality*, pages 207–238. De Gruyter Mouton, 2006.

U. Zeshan. Negative and interrogative constructions in sign languages: A case study in sign language typology. In *Interrogative and negative constructions in sign language.* Ishara Press, 2006a.

U. Zeshan. Negative and interrogative structures in turkish sign language. In *Interrogative and negative constructions in sign language.* Ishara Press, 2006b.

S. Zucchi. Along the time line. *Natural Language Semantics*, 17(2):99–139, 2009.