A Survey on Mouth Modeling and Analysis for Sign Language Recognition

Epameinondas Antonakos*, Anastasios Roussos* and Stefanos Zafeiriou†
Department of Computing, Imperial College London, U.K.

Abstract— Around 70 million Deaf worldwide use Sign Languages (SLs) as their native languages. At the same time, they have limited reading/writing skills in the spoken language. This puts them at a severe disadvantage in many contexts, including education, work, usage of computers and the Internet. Automatic Sign Language Recognition (ASLR) can support the Deaf in many ways, e.g. by enabling the development of systems for Human-Computer Interaction in SL and translation between sign and spoken language. Research in ASLR usually revolves around automatic understanding of manual signs. Recently, ASLR research community has started to appreciate the importance of non-manuals, since they are related to the lexical meaning of a sign, the syntax and the prosody. Non-manuals include body and head pose, movement of the eyebrows and the eyes, as well as blinks and squints. Arguably, the mouth is one of the most involved parts of the face in non-manuals. Mouth actions related to ASLR can be either mouthings, i.e. visual syllables with the mouth while signing, or non-verbal mouth gestures. Both are very important in ASLR. In this paper, we present the first survey on mouth non-manuals in ASLR. We start by showing why mouth motion is important in SL and the relevant techniques that exist within ASLR. Since limited research has been conducted regarding automatic analysis of mouth motion in the context of ALSR, we proceed by surveying relevant techniques from the areas of automatic mouth expression and visual speech recognition which can be applied to the task. Finally, we conclude by presenting the challenges and potentials of automatic analysis of mouth motion in the context of ASLR.

I. INTRODUCTION

Sign languages (SLs) commonly serve as an alternative or complementary mode of human communication. Recent statistics suggest that around 5% of the worldwide population suffers from hearing loss to some degree [44]. Furthermore, it is estimated that about 1% of the worldwide population (around 70 million Deaf people) use SLs as their native languages, even though it is hard to measure the exact number. Note that SL is not only used by the Deaf community, but also by people who cannot physically speak. The limited usage and popularity of SLs among people that use spoken languages has lead to the dominance of several misconceptions about them. For example, many people believe that there is a unique international SL or that SL is simply a pantomime. However, each country has each own SL (in some cases more than one) and there are hundreds of different SLs worldwide.

* The authors contributed equally and have joint first authorship. Their names appear in alphabetical order.
† The work of E. Antonakos and S. Zafeiriou was partially funded by the EPSRC projects EP/J017787/1 (4DFAB) and EP/L026813/1 (ADAManT).
There are many difficulties that Deaf people encounter related to facial features and head movements. There are many open questions regarding the role of the mouth during signing. Some researchers argue that mouthings are not really part of the lexical scope of a SL and are not linguistically significant [57], [30]. However, recent research has shown that mouthings and mouth gestures contribute significantly to the semantic analysis of various SLs [37], [39], [9]. The frequency of mouthings during signing is different for each SL [47], [16] and is dependent on both context and the grammatical category of the manual sign they occur with [39], [9]. The mouth actions can contribute to the signing in various ways [56], [50]. Most mouthings have a prosodic interpretation while others have lexical meaning. An example of this is shown in Fig. 1. In ASL, the signs “not yet” and “late” have the exact same manual gesture. The only lexical distinction difference between them is that in order to articulate “not yet” (Fig. 1b) the signer needs to touch the lower lip with the tongue and make a slight rotation of the head from side to side that declares negation [37]. Another example is shown in Fig. 2 for the words “brother” and “sister” in German SL. Finally, there are cases in which the mouth may articulate physical events, emotions or sensations, such as types of sounds, noise, disturbances, heaviness, types of textures etc. These are usually referred to as non-linguistic mouth gestures [55].

There are many difficulties that Deaf people encounter in the every day life. Many of them have limited skills in reading/writing in the spoken language, which for them is a foreign language with a fundamentally different grammatical structure. At the same time, the vast majority of the rest of the population does not understand and is unable to use SL. Additionally, the current technological advances within the domain of Human-Computer Interaction (HCI), ranging from text-based interaction to speech recognition and synthesis, are almost entirely oriented towards hearing people. The aforementioned issues put the Deaf community in a disadvantaged position within many contexts, including education, work, usage of computers and the Internet. Automatic SL Recognition (ASLR) can support the Deaf community in overcoming these disadvantages by enabling the development of reliable systems for HCI in SL and translation between sign and spoken language. In contrast to speech recognition, which is now ubiquitous in real-world HCI and other applications, ASLR is still far from being a mature technology. Nevertheless, during the last decades, there have been some significant developments in ASLR [45], [67], [15] for some general surveys). However, the most of the research effort is oriented towards the SL recognition using hands-based features. The research efforts that employ non-manual features, including mouth actions, are very limited.

In this paper we present an overview of the literature around the employment and interpretation of mouth actions in ASLR systems. Since, limited research has been conducted regarding automatic analysis of mouth motion in the context of ASLR, we proceed by surveying relevant techniques from the areas of automatic mouth expression and visual speech recognition which can be applied to the task. As mentioned above, there are many open questions regarding the role of the mouth in SL, both on a linguistic and computing level. Herein, we aim to provide a clear and comprehensive overview of the ongoing research on this problem in order to highlight the achievements that have already been accomplished, but most importantly emphasize the challenges that are still open and need to be addressed.

II. NON-MANUAL FEATURES IN ASLR

The vast majority of ASLR methods use solely hand features. However, the research in ASLR has recently started appreciating the importance of non-manual parameters. This relatively new research direction is especially promising and is yet to be explored in depth. The non-manual parameters play an essential role in SL communication because they are related to the meaning of a sign, the syntax or the prosody [54], [10], [69], [70]. There are methods related to the direct recognition of non-manual linguistic markers [43], [42], [38], as applied to negations, conditional clauses, syntactic boundaries, topic/focus and wh-, yes/no questions. Moreover there are methods for the detection of important facial events such as head gestures [40], [24], [42], eyebrows movement and eyes blinking/squint [42], along with facial expression recognition [40], [64], [65] within the context of SL. The authors in [42] employ a 2-layer Conditional Random Field for recognizing continuously signed grammatical markers related to facial features and head movements. [38] em-
employs geometric and Local Binary Pattern (LBP) features on a combined 2D and 3D face tracking framework to automatically recognize linguistically significant non-manual expressions in continuous ASL videos. The challenging task of fusion of manuals and non-manuals for ASLR has also received attention [58], [67], [66], [4]. Due to the timewise cost and the lack of annotations, recently there is a more explicit trend by works towards preliminary tools for semi-automatic annotation via a recognition and a translation component [19] at the sign level concerning manuals, by categorizing manual/non-manual components [31], providing information on lexical signs and assisting sign searching. Early enough, [65], [64] have contributed in this direction.

III. MOUTH NON-MANUALS IN EXISTING ASLR SYSTEMS

As explained in Sec. I, among all non-manual features, the shape and motion of mouth, in particular, define crucial cues of information for ASLR systems. For example, in ASL, tongue through front teeth might indicate that something is done carelessly, without paying attention [45]. As another example, in British SL (BSL), some signs are disambiguated solely by the lips shapes that co-occur with them [15].

There are a few existing ASLR methods that give emphasis to mouth modeling; see Table I for a list containing the main characteristics of each method. Parashar [46] proposed one of the first approaches of combining manual and non-manual information for ASLR. This approach uses the facial information to prune the word hypotheses generated by manual information. After masking the face with an elliptical structure, a signer-specific Principal Component Analysis (PCA) - based model of image appearance variation is employed. Applying this training procedure in various SL sentences reveals that some of the most dominant dimensions of facial expressions describe the movement of the lips. In the conducted ASLR experiments, the incorporation of the facial motion cue was found to increase the continuous words recognition accuracy from 88% to 92%.

Von Agris et al. [66], [67] propose an ASLR method that incorporates both manual and facial feature extraction. The latter is based on facial Active Appearance Model (AAM) fitting, followed by estimating geometric measures of facial expressions, with particular emphasis on measures of the mouth region. After systematic evaluation on an SL corpus with 25 signers in both signer-dependent and signer-independent scenarios, the experimental results verify that the recognition performance is significantly improved when facial features are incorporated. Nguyen et al. [43] tackle the problem of recognizing facial expressions that are used in SL as grammatical markers. Wh-questions, yes/no questions, rhetorical questions, topic, negation, assertion, conditional clause and relative clause are considered. A facial shape subspace is learned by a mixture of Probabilistic PCA (PPCA) model applied on a set of robustly tracked facial feature points. The proposed recognition framework adopts Hidden Markov Models (HMM) accompanied with Support Vector Machines (SVM) modeling. Among the most prominent facial features that are used in this framework are the mouth movements and lips shapes.

Schmidt et al. [61], [60] employ a signer-specific facial AAM fitting followed by 3D Point Distribution Model (PDM) estimation, to extract high-level facial features, which include mouth vertical and horizontal openness, chin-to-lower-lip and upper-lip-to-nose normalized geometrical measurements. This facial feature extraction is applied on the German SL RWTH-Phoenix-Weather [26] corpus, which is formed by SL-interpreted TV weather forecasts and contains experts-driven manual annotations as well as semi-automatic transcriptions based on speech recognition. In [61], the authors employ this facial feature extraction to build a viseme recognizer that implements automatic lip reading. This is integrated in a combined sign language recognition and translation system. In [60], the facial feature extraction is used to perform clustering of different mouthings within the RWTH-Phoenix-Weather corpus. This clustering is motivated by the vision of using this method for facial expression animation in avatar-based SL synthesis. Koller et al. [33], [34], utilize the same corpus (RWTH-Phoenix-Weather) as well as a similar approach of extracting high-level facial features to model mouthings in SL. In [33], the authors develop a novel viseme recognition method that is specifically designed for SL, does not require any manual annotation and is signer-independent. In [34], they propose an algorithm that automatically annotates mouthings in SL videos. This algorithm is based on the semi-automatic transcriptions as a source of weak supervision and the only manual annotation required is a gloss-level annotation.

Pfister et al. [48] employ mouth patterns as features that prove to be highly informative for isolating words in SL videos. This enables the automatic learning of signs from TV footage of signers by exploiting the subtitles that are broadcast simultaneously (see also Sec. IV-C). Benitez-Quiroz et al. [7] propose a framework that combines linguistic and computational modeling to analyze the discriminant non-manual features of SL. By applying this framework on five types of sentences of ASL, the experiments reveal that the mouth and teeth features are among the most discriminant non-manual features, in terms of separating conditionals from non-conditionals. They also discover a complex interaction between head position and mouth shape. Antonakos et al. [2], [3] propose a novel semi-supervised approach for Extreme States Classification (ESC) on feature spaces of facial cues in SL videos. Their method is built upon AAM face tracking and feature extraction of global and local AAMs and applied for detection of sign boundaries and alternative constructions.

IV. AUTOMATIC ANALYSIS OF MOUTH NON-MANUALS: CHALLENGES AND POTENTIALS

The automatic analysis of mouth non-manuals is a challenging problem. As shown in Fig. 3, some of the most common difficulties are occlusion by hands, intense mouthings, expressions and pose, tongue visibility and low resolution of the mouth region. In general, automatic analysis of mouth
non-manuals can be separated to (a) automatic understanding of mouth-related expressions and (b) automatic understanding of mouthings. Even though limited research has been conducted regarding automatic analysis and understanding of expressions in the context of ASLR, a lot of relevant research has been conducted in the general framework of automatic analysis of facial expressions. Similarly, despite the fact that limited work has been conducted towards the understanding of mouthings, this problem bears a lot of similarities with the more extensively explored fields of automatic visual speech recognition and machine lip reading. Both facial expressions analysis and visual speech recognition constitute distinct research fields, hence a thorough review on these would fall outside the scope of this survey paper. Nevertheless, in the following Secs. IV-A, IV-B, we briefly mention techniques from these fields, which are relevant to the problem of automatic analysis of mouth non-manuals.

Furthermore, the training of ASLR systems using large-scale data can benefit from weakly-supervised techniques for analysis of mouth non-manuals. Such techniques are discussed in Sec. IV-C.

A. Automatic analysis of facial expressions

Analysis of facial expressions is a popular research study in many scientific disciplines spanning from Psychology to Human Computer Interaction (HCI) and Robotics. In the context of HCI applications facial expressions are directly linked to human emotions, while in the context of ASLR system facial expressions can be crucial for the meaning of the particular manual and, in general, are not directly linked to a particular emotion.

Research on how facial expression measurement is performed revolves around two main lines: message judgement and sign judgement [13]. Message judgment aims to immediately recognize the meaning conveyed by a facial expression. The meaning is usually related to a particular emotion such as being happy, angry or sad. On the other hand, sign judgement studies the physiological manifestation of facial expressions into its fundamental and, arguably, irreducible

---

**Table I:** List of ASLR methods that give emphasis to mouth modeling. KLT refers to Kanade-Lucas-Tomasi feature tracker. ATL refers to Allen’s Temporal Logic. RLDA refers to Regularized Linear Discriminant Analysis.

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>Face modeling &amp; tracking</th>
<th>Facial features</th>
<th>Recognition approach</th>
<th>SL. phenomena modeled</th>
<th>Problem(s) tackled</th>
<th>SL. studied</th>
<th>Size of dataset used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parashar [46]</td>
<td>2003</td>
<td>Elliptical masking, PCA</td>
<td>Global face appearance</td>
<td>Bottom-up, pruning based on facial info</td>
<td>Negation</td>
<td>Continuous sign recognition</td>
<td>American</td>
<td>1 signer, 25 sentences, 39 signs</td>
</tr>
<tr>
<td>v. Agris et al. [66], [67]</td>
<td>2008</td>
<td>AAMs</td>
<td>Geometric measures</td>
<td>HMMs</td>
<td>(no specific phenomenon)</td>
<td>Isolated &amp; continuous sign recognition</td>
<td>German</td>
<td>25 signers, 430 signers</td>
</tr>
<tr>
<td>Nguyen et al. [43]</td>
<td>2011</td>
<td>KLT, PPCA mixture of feature points</td>
<td>Geometric measures</td>
<td>HMMs, SVMs</td>
<td>Questions, topic, negation, assertion, conditional and relative clause</td>
<td>Grammatical markers recognition</td>
<td>American</td>
<td>7 signers, ~300 sequences</td>
</tr>
<tr>
<td>Schmidt et al. [60]</td>
<td>2013</td>
<td>AAMs, 3D PDM</td>
<td>Geometric measures</td>
<td>HMMs</td>
<td>Mouthings</td>
<td>SL recognition transcription</td>
<td>German</td>
<td>7 signers, ~15K glosses</td>
</tr>
<tr>
<td>Schmidt et al. [61]</td>
<td>2013</td>
<td>AAMs, 3D PDM</td>
<td>Geometric measures</td>
<td>HMMs</td>
<td>Mouthings</td>
<td>Mouthings clustering for SL synthesis</td>
<td>German</td>
<td>7 signers, ~15K glosses</td>
</tr>
<tr>
<td>Koller et al. [33]</td>
<td>2014</td>
<td>AAMs, 3D PDM</td>
<td>Geometric measures, appearance</td>
<td>HMMs</td>
<td>Mouthings</td>
<td>Lip reading in signing</td>
<td>German</td>
<td>7 signers, ~15K glosses</td>
</tr>
<tr>
<td>Koller et al. [34]</td>
<td>2014</td>
<td>AAMs, 3D PDM</td>
<td>Geometric measures, appearance</td>
<td>HMMs</td>
<td>Mouthings</td>
<td>Automatic mouthing transcription</td>
<td>German</td>
<td>7 signers, ~15K glosses</td>
</tr>
<tr>
<td>Pfister et al. [48]</td>
<td>2013</td>
<td>KLT, local appearance descriptor</td>
<td>Appearance-based</td>
<td>Multiple-Instance Learning, SVM</td>
<td>Mouthings</td>
<td>SL videos clustering for automatic sign learning</td>
<td>British</td>
<td>17 signers, 30 hours of continuous signing, 1000 words</td>
</tr>
<tr>
<td>Antonakos et al. [2]</td>
<td>2014</td>
<td>Global and local AAMs</td>
<td>Shape, appearance, geometric measures</td>
<td>Hierarchical clustering</td>
<td>Sign boundaries, alternative construction</td>
<td>Extreme facial events detection, Analysis of their links to linguistics</td>
<td>Greek, American</td>
<td>2 signers, continuous signing</td>
</tr>
</tbody>
</table>

**Fig. 3:** Characteristic challenges of automatic analysis of mouth non-manuals. The frames are extracted from the DictaSign Greek SL Corpus [20].
Atoms, such as the movement of individual facial muscles (e.g. raised cheeks or depressed lips).

Arguably, the pillar of the first line of research, i.e. on message judgement approaches, is the theory of the six basic expressions first suggested by Darwin [17] and later extended by Paul Ekman [21]. They argued and suggested that there is a set of six basic emotions, namely anger, fear, disgust, happiness, sadness and surprise, which are manifested through universally facial expressions. Due to the simplicity of the above discrete representation, as well as to the fact that is is feasible to record posed facial expressions of the six basic emotions by providing a simple set of instructions to people, the above message judgement approach became very popular and well-studied.

There are two major drawbacks of message judgement approaches which make their application to analysis of ASLR rather limited. Firstly, it cannot explain the full range of expressions, as the set of expressions that can be explained is restricted by the discrete set of predefined messages (i.e., anger etc.). Secondly, it is very difficult to define a predefined set of messages (expressions) that are universally used in the context of ASLR.

More relevant to the case of analysis of facial and mouth expressions for ASLR are sign-judgement approaches. The most commonly used set of descriptors in sign-judgement approaches is that specified by the Facial Action Coding System (FACS) [22], [28]. The FACS is a taxonomy of human facial expressions. It was originally developed by Ekman and Friesen in 1978, and revised in 2002. The revision specifies 32 atomic facial muscle actions, named Action Units (AUs), and 14 additional Action Descriptors (ADs) that account for miscellaneous actions, such as jaw thrust, blow and bite. The FACS is comprehensive and objective, as opposed to message-judgement approaches. Since any facial expression results from the activation of a set of facial muscles, every possible facial expression can be comprehensively described as a combination of AUs. And while it is objective in that it describes the physical appearance of any facial display, it can still be used in turn to infer the subjective emotional state of the subject, which cannot be directly observed and depends instead on personality traits, context and subjective interpretation [71].

Over the past 30 years, psychologists and neuroscientists have conducted extensive research using the FACS on various aspects of facial expression analysis [1], [27], [23], [29], [18].

A major impediment to the widespread use of FACS is the time required both to train human experts and to manually score a video. It takes over 100 hours of training to achieve minimal competency as a FACS coder, and each minute of a video takes approximately one hour to score [18]. Due to (a) the difficulty to collect and annotate databases with AUs, and (b) the complexity of the AU detection problem, for which there is a high number of classes, more subtle patterns, and small inter-class differences, machine analysis of AUs is still an open challenge [63].

Many systems for AU detection and recognition in intensity video sequences have been proposed over the past twenty years (please see [63] and the references therein). However, most of these systems are still highly sensitive to variations in recording conditions such as illumination, occlusions and other changes in facial appearance such as makeup and facial hair. The problem of occlusion is more intense in signing videos where, often, the signer occludes part of the mouth region with her hands. Furthermore, in most cases when 2D facial intensity images are used, it is necessary to maintain a consistent facial pose (preferably a frontal one) in order to achieve a good recognition performance, as even small changes in the facial pose can reduce the system’s accuracy. In order to tackle this problem, 3D data can be acquired and analyzed [59], [52], [51], [53].

When it comes to recognition of many mouth related AUs, the subtle changes occurring in the depth of the facial surface are captured in detail when 3D data are used, which does not happen with 2D data. For example, AU18 (Lip Pucker, please see Fig. 4) is not easily distinguished from AU10+AU17+AU24 (Upper Lip and Chin Raising and Lip Presser) in a 2D frontal view video. On the other hand, in a 3D capture the action is easily identified [53]. Unfortunately, even though it has been argued that the problem of automatic analysis of AUs becomes easier in case that 3D data are available, there are very few databases that contain annotated data with respect to AUs [53], [72] and none of them recorded in conditions required by an ASLR application (i.e., the subject to move freely etc.).

B. Visual speech recognition

The automatic analysis of non-verbal facial/mouth behaviour could employ two alternative approaches, originating from visual speech recognition, which has received a lot of attention during the past twenty years. The first one corresponds to the recognition of the specific word or phrase,

The recent advances in structured light scanning, stereo photogrammetry and photometric stereo have made the acquisition of 3D facial structure and motion a feasible task.
while the second one performs lip reading by first recognizing a set of predefined mouth shapes (or appearances) or sequence of mouth dynamics that are required to generate the visual letters or words.

The state-of-the-art in the first line of research includes the method of [74], which proposes a low-dimensional dynamic manifold for representing visual words, achieving a Frame Recognition Accuracy (FRA) of 56% on the speaker independent OuluVS database [73]. Comprehensive overviews of the field can be found in [49] and [12].

Regarding the second line of research, to the best of our knowledge the only method for viseme recognition in the context of ASLR is [33]. Drawbacks of using visemes for description and recognition of mouthings include the fact that there is not a standardized set of visemes (e.g., 13 visemes were used in [35], 16 were used in [25] while 50 visemes were used in [62] to model the effects of co-articulation) and using a low number of visemes may cover a small subspace of the mouth motions represented in the visual domain. Furthermore, the task of viseme recognition is very challenging (even for humans) with reported error rates around 50% [35]. The interested reader may refer to [33] for additional technical challenges in viseme recognition in a setting when there is no manual annotation.

C. Mouth non-manuals for weakly supervised ASLR training

One of the biggest challenges in training ASLR systems is the creation of adequately labelled, realistic datasets, which is an exceptionally time-consuming and expensive procedure. There are very few publicly available datasets that are suitable for the training of such recognition systems, e.g. [20], [5].

In order to bypass the shortage of SL datasets and the difficulties in their creation, a promising solution is to exploit the TV footage of signers broadcast simultaneously with subtitles [11], [14], [48]. More precisely, weakly supervised training can be applied, once the problem of aligning the subtitles [11], [14], [48]. More precisely, weakly supervised training can be applied, once the problem of aligning the subtitles is solved. How- ever, such an alignment is also a challenging problem, since the subtitles correspond only loosely to what is being signed. The new handbook for automatic understanding of mouth movements in sign language videos. In IEEE Proc. Int’l Conf. on Image Processing, 2012.


