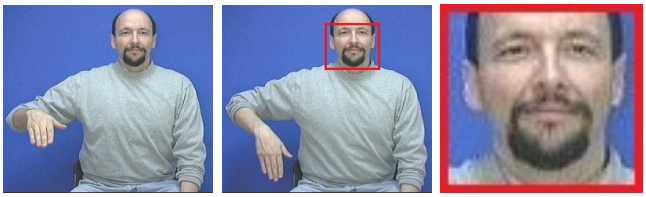
A Survey on Mouth Modeling and Analysis for Sign Language Recognition

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

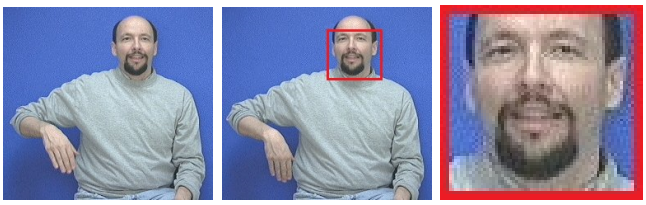
***Abstract*— Around 70 million Deaf worldwide use Sign Languages (SLs) as their native languages. At the same time, they have limited reading/writing skills in the spoken language. This puts them at a severe disadvantage in many contexts, including education, work, usage of computers and the Internet. 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.



(a) LATE



(b) NOT YET

Fig. 1: In American SL, the signs "late" and "not yet" are performed with the same manual gesture. Their lexical distinction is only based on the mouth action of the tongue touching the lower lip. (Images copyright ASL University)

Opposite to another common misconception, SLs are as rich and grammatically complex as spoken languages and even though they usually have different grammatical structure, they exhibit very similar properties [32], [57]. For example, similar to spoken languages, SLs consist of basic semantic components, referred to as phonemes [57]. The phonemes are mainly expressed through combinations of manual features, such as shape, posture (orientation), location and motion of the hands. However, SLs convey much of their prosody through non-manual signs [45], [67], [47], [68], [56], [41], such as the pose of head and torso, facial expressions (combinations of eyes, eyebrows, cheeks and lips) and mouth movements. These non-manual articulators play an important role in lexical distinction, grammatical structure and adjectival or adverbial content. For example, yes/no questions in American SL (ASL) are associated with raised eyebrows, head tilted forward and widely-opened eyes, and wh-questions with furrowed eyebrows and head forward [36], [56]. Topics are described by raised eyebrows and head slightly back and negations are expressed with a head shake [36], [56]. The head pose and eye gaze describe turn taking during a story narration [6].

In this paper we are interested in the mouth actions within SL. As explained in [8], [47], the mouth lexical articulators can be separated in two categories: mouth gestures and

* 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).

(a) BRUDER (translated as “brother” in English)

(b) SCHWESTER (translated as “sister” in English)

Fig. 2: In German SL, the signs “bruder” and “schwester” are performed with the same manual gesture and only differentiate on the lips patterns. (The figure is used with permission from [66].)

mouthings. Mouth gestures (also referred to as oral components), which include deformations of the mouth’s shape, movements of the tongue and visibility of the teeth, are not related to the spoken language [47]. The term mouthing refers to the silent articulators of the mouth that correspond to a pronounced word or part of it. If part of the word is articulated then, in most SLs, it is its first syllable. Note that there is a debate in literature regarding the importance and 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 (please see [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 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. 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-

ploys 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

Method	Year	Face modeling & tracking	Facial features	Recognition approach	SL phenomena modeled	Problem(s) tackled	SL studied	Size of dataset used
Parashar [46]	2003	Elliptical masking, PCA	Global face appearance	Bottom-up, pruning based on facial info	Negation	Continuous sign recognition	American	1 signer, 25 sentences, 39 signs
v. Agris et al. [66], [67]	2008	AAMs	Geometric measures	HMMs	(no specific phenomenon)	Isolated & continuous sign recognition	German	25 signers, 450 signs
Nguyen et al. [43]	2011	KLT, PPCA mixture of feature points	Geometric measures	HMMs, SVMs	Questions, topic, negation, assertion, conditional and relative clause	Grammatical markers recognition	American	7 signers, ∼300 sequences
Schmidt et al. [61]	2013	AAMs, 3D PDM	Geometric measures	HMMs	Mouthings	SL recognition translation	German	7 signers, ∼15K glosses
Schmidt et al. [60]	2013	AAMs, 3D PDM	Geometric measures	HMMs	Mouthings	Mouthings clustering for SL synthesis	German	7 signers, ∼15K glosses
Koller et al. [33]	2014	AAMs, 3D PDM	Geometric measures, appearance	HMMs	Mouthings	Lip reading in signing	German	7 signers, ∼15K glosses
Koller et al. [34]	2014	AAMs, 3D PDM	Geometric measures, appearance	HMMs	Mouthings	Automatic mouthing transcription	German	7 signers, ∼15K glosses
Pfister et al. [48]	2013	KLT, local appearance descriptor	Appearance-based	Multiple-Instance Learning, SVM	Mouthings	SL videos clustering for automatic sign learning	British	17 signers, 30 hours of continuous signing, 1000 words
Benitez-Quiroz et al. [7]	2014	Manual annotations	Qualitative relative temporal features	ATL, RLDA	Conditionals, Questions, Assertions, Positive & negative polarities	Analysis of discriminant non-manual features	American	15 signers, 129 sentences per signer
Antonakos et al. [2]	2014	Global and local AAMs	Shape, appearance, geometric measures	Hierarchical clustering	Sign boundaries, alternative construction	Extreme facial events detection, Analysis of their links to linguistics	Greek, American	2 signers, continuous signing

TABLE I: List of ASLR methods that give emphasis to mouth modeling. KLT refers to KanadeLucasTomasi feature tracker. ATL refers to Allens Temporal Logic. RLDA refers to Regularized Linear Discriminant Analysis.

(a) Occlusion (b) Mouthing (c) Tongue (d) Pose

Fig. 3: Characteristic challenges of automatic analysis of mouth non-manuals. The frames are extracted from the DictaSign Greek SL Corpus [20].

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

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].

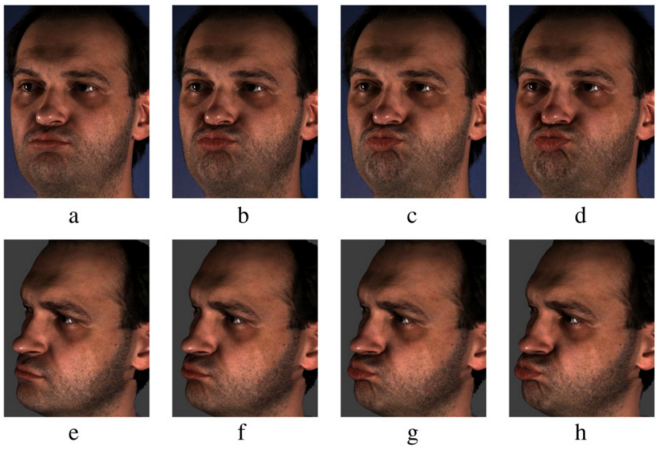


Fig. 4: AU18 (Lip Pucker) captured in both 2D and 3D. (a)(d) 2D nearly frontal view. (e)(h) 3D reconstructed data. (The figure is used with permission from [53].)

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 [1], 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,

[1]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 with the corresponding SL videos is solved. However, such an alignment is also a challenging problem, since the subtitles correspond only loosely to what is being signed. In the recent work of [48], the authors show that mouth non-manuals constitute an especially informative cue for tackling this problem. More precisely, the authors incorporate mouth motion information and develop a multiple instance learning strategy that leads to improvements in the computational performance, compared to previous related works. This is an especially promising line of research, which to the best of our knowledge, has not been further explored in the literature.

V. Conclusions

Development of Automatic Sign Language Recognition (ASLR) systems has the potential to support millions of Deaf people, as well as help linguists understand better sign languages. The past two decades research on ASLR has mainly concentrated on automatic understanding of manuals. Recently, it was argued that non-manual gestures play an important role, as well. In this paper, we surveyed the role of mouth non-manuals in ASLR, as well as methodologies that can be used for automatic understanding of mouth related motion in the context of an ASLR scenario. Important challenges include (a) the frequent occlusion of the mouth region by the hand while signing, (b) the low-resolution of the videos in the available datasets and (c) the lack of annotated corpora. A promising direction to mitigate for the lack of annotated data is to use unsupervised or semi-supervised techniques for training.

References

[1] Z. Ambadar, J. F. Cohn, and L. I. Reed. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of nonverbal behavior*, 33(1):17–34, 2009.

[2] E. Antonakos, V. Pitsikalis, and P. Maragos. Classification of extreme facial events in sign language videos. *EURASIP Journal on Image and Video Processing*, 2014(14), 2014.

[3] E. Antonakos, V. Pitsikalis, I. Rodomagoulakis, and P. Maragos. Unsupervised classification of extreme facial events using active appearance models tracking for sign language videos. In *IEEE Proc. Int'l Conf. on Image Processing*, 2012.

[4] O. Aran, T. Burger, A. Caplier, and L. Akarun. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition*, 42(5):812–822, 2009.

[5] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. The American Sign Language lexicon video dataset. In *CVPR-4HB*, 2008.

[6] C. Baker. Regulators and turn-taking in american sign language discourse. *On the other hand*, 1977.

[7] C.F. Benitez-Quiroz, K. Gökgöz, R.B. Wilbur, and A.M. Martinez. Discriminant features and temporal structure of nonmanuals in american sign language. *PloS one*, 9(2):e86268, 2014.

[8] P. Boyes-Braem. Functions of the mouthing component in the signing of deaf early and late learners of swiss german sign language. *Foreign vocabulary in sign languages: A cross-linguistic investigation of word formation*, pages 1–47, 2001.

[9] P. Boyes-Braem and R. Sutton-Spence. The hands are the head of the mouth. *Signum-Verlag, Hamburg*, 2001.

[10] D. Brentari and L. Crossley. Prosody on the hands and face. *Sign Language Studies & Linguistics*, 5(2):105–130, 2002.

[11] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *CVPR*, 2009.

[12] A. Chiţu[1] and L. Rothkrantz[1]. Automatic visual speech recognition. *Speech enhancement, modeling and recognition–algorithms and applications*, page 95, 2012.

[13] J. F. Cohn and P. Ekman. Measuring facial action. *The new handbook of methods in nonverbal behavior research*, pages 9–64, 2005.

[14] H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2568–2574. IEEE, 2009.

[15] H. Cooper, B. Holt, and R. Bowden. Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer, 2011.

[16] O. Crasborn, E. Van Der Kooij, D. Waters, B. Woll, and J. Mesch. Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*, 11(1):45–67, 2008.

[17] C. Darwin. 1965. the expression of the emotions in man and animals. *London, UK: John Marry*, 1872.

[18] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE T-PAMI*, 21(10):974–989, 1999.

[19] P. Dreuw and H. Ney. Towards automatic sign language annotation for the elan tool. In *Proc. Language Resources & Evaluation*, 2008.

[20] E. Efthimiou, S. E. Fontinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and F. Goudenove. Dicta-sign–sign language recognition, generation and modelling: a research effort with applications in deaf communication. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 80–83, 2010.

[21] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[22] P. Ekman and W. V. Friesen. Facial action coding system: A technique for the measurement of facial movement. palo alto, 1978.

[23] P. Ekman and E. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[24] U.M. Erdem and S. Sclaroff. Automatic detection of relevant head gestures in American Sign Language communication. In *IEEE Proc. of Int'l Conf. on Pattern Recognition*, 2002.

[25] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of the 2002 ACM SIGGRAPH*. ACM, 2002.

[26] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Language Resources and Evaluation*, pages 3785–3789, Istanbul, Turkey, 2012.

[27] M. G. Frank and P. Ekman. The ability to detect deceit generalizes across different types of high-stake lies. *Journal of personality and social psychology*, 72(6):1429, 1997.

[28] J. C. Hager, P. Ekman, and W. V. Friesen. Facial action coding system. *Salt Lake City, UT: A Human Face*, 2002.

[29] M. Heller and V. Haynal. Les visages de la dépression et du suicide. *Cahiers psychiatriques genevois*, (16):107–117, 1994.

[30] A. Hohenberger and D. Happ. The linguistic primacy of signs and mouth gestures over mouthing: Evidence from language production in german sign language (dgs). In *The hands are the head of the mouth: The mouth as articulator in sign language*, pages 153–189. Signum, 2001.

[31] M. Hrúz, Z. Krňoul, P. Campr, and L. Müller. Towards automatic annotation of sign language dictionary corpora. In *Text, speech and dialogue*, 2011.

[32] E. S. Klima. *The signs of language*. Harvard University Press, 1979.

[33] O. Koller, H. Ney, and R. Bowden. Read my lips: Continuous signer independent weakly supervised viseme recognition. In *Computer Vision–ECCV 2014*, pages 281–296. Springer, 2014.

[34] O. Koller, H. Ney, and R. Bowden. Weakly supervised automatic transcription of mouthings for gloss-based sign language corpora. *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, 2014.

[35] Y. Lan, R. Harvey, and B. J. Theobald. Insights into machine lip reading. In *ICASSP*, pages 4825–4828. IEEE, 2012.

[36] S. K Liddell. *American sign language syntax*. Mouton The Hague, 1980.

[37] S. K Liddell. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003.

[38] D. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, and C. Neidle. Recognition of nonmanual markers in ASL using non-parametric adaptive 2D-3D face tracking. In *Proc. Language Resources & Evaluation*, 2012.

[39] M. A. Nadolske and R. Rosenstock. Occurrence of mouthings in american sign language: A preliminary study. *Trends in linguistics studies and monographs*, 188:35, 2007.

[40] C. Neidle, N. Michael, J. Nash, D. Metaxas, I. E.B Bahan, L. Cook, Q. Duffy, and R.G. Lee. A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus. In *Proc. of ESSLLI Wrks. Formal Applications to SLs*, 2009.

[41] C. J. Neidle. *The syntax of American Sign Language: Functional categories and hierarchical structure*. MIT Press, 2000.

[42] T. Nguyen and S. Ranganath. Recognizing continuous grammatical marker facial gestures in sign language video. *ACCV*, pages 665–676, 2011.

[43] T.D. Nguyen and S. Ranganath. Facial expressions in american sign language: Tracking and recognition. *Pattern Recognition*, 2011.

[44] N. Oishi and J. Schacht. Emerging treatments for noise-induced hearing loss. *Expert opinion on emerging drugs*, 16(2):235–245, 2011.

[45] S. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE T-PAMI*, 27(6):873–891, 2005.

[46] A.S. Parashar. *Representation and interpretation of manual and non-manual information for automated american sign language recognition*. PhD thesis, Univ. of South Florida, 2003.

[47] R. Pfau and J. Quer. *Nonmanuals: their grammatical and prosodic roles*. Cambridge University Press, 2010.

[48] T. Pfister, J. Charles, and A. Zisserman. Large-scale learning of sign language by watching tv (using co-occurrences). In *Proceedings of the British machine vision conference*, 2013.

[49] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[50] J. S. Reilly, M. McIntire, and U. Bellugi. The acquisition of conditionals in american sign language: Grammaticized facial expressions. *Applied Psycholinguistics*, 11(04):369–392, 1990.

[51] G. Sandbach, S. Zafeiriou, and M. Pantic. Binary pattern analysis for 3d facial action unit detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, Guildford, UK, September 2012.

[52] G. Sandbach, S. Zafeiriou, and M. Pantic. Local normal binary patterns for 3d facial action unit detection. In *ICIP*, pages 1813–1816. IEEE, 2012.

[53] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[54] W. Sandler. The medium and the message: Prosodic interpretation of linguistic content in Israeli Sign Language. *Sign Language Studies & Linguistics*, 2(2):187–215, 1999.

[55] W. Sandler. Symbiotic symbolization by hand and mouth in sign language. *Semiotica*, 2009(174):241–275, 2009.

[56] W. Sandler. Prosody and syntax in sign languages. *Transactions of the Philological Society*, 108(3):298–328, 2010.

[57] W. Sandler and D. Lillo-Martin. *Sign language and linguistic universals*. Cambridge University Press, 2006.

[58] S. Sarkar, B. Loeding, and A.S. Parashar. Fusion of manual and non-manual information in american sign language recognition. *Handbook of Pattern Recognition & Computer Vision*, 2010.

[59] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer, 2008.

[60] C. Schmidt, O. Koller, H. Ney, T. Hoyoux, and J. Piater. Enhancing gloss-based corpora with facial features using active appearance models. *International Symposium on Sign Language Translation and Avatar Technology*, 2013.

[61] C. Schmidt, O. Koller, H. Ney, T. Hoyoux, and J. Piater. Using viseme recognition to improve a sign language translation system. In *International Workshop on Spoken Language Translation*, pages 197–203, 2013.

[62] K.C. Scott, D.S. Kagels, S.H. Watson, H. Rom, J.R. Wright, M. Lee, and K.J. Hussey. Synthesis of speaker facial movement to match selected speech sequences. *NASA technical report*, 1994.

[63] M. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE T-SMCB*, 42(4):966–979, 2012.

[64] C. Vogler and S. Goldenstein. Analysis of facial expressions in american sign language. In *Proc. of Int'l Conference on Universal Access in HCI*, 2005.

[65] C. Vogler and S. Goldenstein. Facial movement analysis in ASL. *Universal Access in the Information Society*, 6(4):363–374, 2008.

[66] U. von Agris, M. Knorr, and K. F. Kraiss. The significance of facial features for automatic sign language recognition. In *Automatic Face & Gesture Recognition, 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[67] U. Von Agris, J. Zieren, U. Canzler, B. Bauer, and K.F. Kraiss. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6(4):323–362, 2008.

[68] R. B. Wilbur. Effects of varying rate of signing on asl manual signs and nonmanual markers. *Language and speech*, 52(2-3):245–285, 2009.

[69] R.B. Wilbur. Eyeblinks & ASL phrase structure. *Sign Language Studies*, 1994.

[70] R.B. Wilbur and C. Patschke. Syntactic correlates of brow raise in asl. *Sign Language Studies & Linguistics*, 2(1):3–41, 1999.

[71] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T-PAMI*, 31(1):39–58, 2009.

[72] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[73] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE T-Multimedia*, 11(7):1254–1265, 2009.

[74] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen. A compact representation of visual speech data using latent variables. *IEEE T-PAMI*, 36(1):181, 2014.