Read My Lips: Continuous Signer Independent Weakly Supervised Viseme Recognition

Oscar Koller[1,2], Hermann Ney[1], and Richard Bowden[2]

[1] Human Language Technology and Pattern Recognition, RWTH Aachen, Germany
[2] Centre for Vision Speech and Signal Processing, University of Surrey, UK
{koller,ney}@cs.rwth-aachen.de, r.bowden@surrey.ac.uk

Abstract. This work presents a framework to recognise signer independent mouthings in continuous sign language, with no manual annotations needed. Mouthings represent lip-movements that correspond to pronunciations of words or parts of them during signing. Research on sign language recognition has focused extensively on the hands as features. But sign language is multi-modal and a full understanding particularly with respect to its lexical variety, language idioms and grammatical structures is not possible without further exploring the remaining information channels. To our knowledge no previous work has explored dedicated viseme recognition in the context of sign language recognition. The approach is trained on over 180.000 unlabelled frames and reaches 47.1% precision on the frame level. Generalisation across individuals and the influence of context-dependent visemes are analysed.

Keywords: Sign Language Recognition, Viseme Recognition, Mouthing, Lip Reading.

1 Introduction

Sign Languages, the natural languages of the Deaf, are known to be as grammatically complete and rich as their spoken language counterparts. However, their grammar is different to spoken language. They are not international and convey meaning by more than just the movements of hands. Sign languages make use of both 'manual features' (hand shape, position, orientation and movement) and linguistically termed 'non-manual features' consisting of the face (eye gaze, mouthing/mouth gestures and facial expression) and of the upper body posture (head nods/shakes and shoulder orientation). All of these parameters are used in parallel to complement each other, but depending on the context a specific component may or may not be required to interpret the sign, sometimes playing an integral role within the sign, sometimes modifying the meaning and sometimes providing context. Furthermore, the different information channels don't share a fixed temporal alignment, but are rather loosely tied together. For example, the mouthing 'ALPS' may span over the two manual signs 'MOUNTAIN' and 'REGION'. Historically, research on automatic recognition of sign language has

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 281–296, 2014.

focused extensively on the manual components [1–3]. These manual parameters are widely considered to cover an important part of the information conveyed by sign language. However, it is clear that a full understanding of sign language, particularly with respect to its lexical variety, language idioms and grammatical structures is not possible without further exploring the remaining information channels [4]. Computer vision methods exist to extract features for these non-manual channels. However, sign language constitutes an extremely challenging test bed as it incorporates huge variations inherent to natural languages. Further, ambiguity is inherent to sign languages, as each movement, each change in eye gaze or each appearance of the tongue may or may not have a grammatical or semantic function depending on the context. Thus, learning features and training classifiers that can be applied to sign language recognition must cope with a natural variation seldom present in other tasks.

The unsolved challenges in sign language recognition are to increase the number of signs to distinguish, recognise signs in a continuous fashion and generalise across different signers. Annotating the parallel information streams is cumbersome and time consuming, also due to the fact that sign languages don't have a standardised annotation system. Thus, possible annotation sources are noisy. This paper explores automatic identification and classification of mouthings in German Sign Language (DGS), as such it directly addresses each of these key challenges and our results are shown to generalise well across signers. They also scale well with increasing vocabulary (due to viseme sub-units) and the approach requires only weak supervision and no manual annotation. To our knowledge no previous work has modelled mouthings explicitly by sequences of visemes in the context of sign language recognition.

In Section 2 we specify the term 'mouthings' in the context of sign language and discuss difficulties when used for recognition. In Section 3 related work in viseme and facial recognition is shown. Further, the employed data sets and features are presented in Sections 4 and 5, respectively. In Section 6 the overall approach is explained. Results are given in Section 7 and finally the paper closes with a conclusion and future work in Section 8.

2 Mouthings in Sign Language, Challenging?

During signing the mouth of a signer performs notable and valuable actions. In sign language, two different types of actions are distinguished: mouth gestures and mouthings. Mouthings originate from speech contact [5] and represent at least part of a pronounced word, while mouth gestures are patterns that are unrelated to spoken language. Some signs are often accompanied by mouthings, others by mouth gestures and sometimes no mouth movement is present at all. Mouthing can be observed in many European sign languages, where it occurs more with nouns than with verbs. The latter are often accompanied by mouth gestures [6]. Nevertheless, the exact linguistic function of mouthings is still debated [7], but signing people state that it is evident they help to discriminate signs which are identical with respect to the manual components of the sign.

In audio-visual speech, recognising visemes, referring to visual patterns of the mouth while speaking, has been shown to be very challenging (even to humans) with error rates usually around 50% [8]. In sign language, and for this paper, additional challenges need to be tackled: 1. Mouthings may or may not occur with specific signs; 2. they can be stretched across several manual signs;

3. viseme sequences of a specific sign are not consistent (sign 'ALPS' sometimes is accompanied by the full mouthing 'A L P', but sometimes only an 'A' or an 'L' suffices); 3. phonemes and visemes don't share a one-to-one correspondence, rather a many-to-many [9]; 4. no standard viseme inventory for sign language exists; 5. huge variability in practises are observed, depending on context (see Fig. 1) and individuals; 6. sign language and spoken language sentence structure differs; 7. the video often has a low spatial resolution (mouth is small in videos); 8. there is an inherent lack of annotation, annotation is difficult; and time consuming especially due to ambiguity; 9. speech recognition cannot be used to bootstrap a viseme mapping. Our approach faces all these problems and suggests ways to solve them.

3 State of the Art

In 1968 Fisher [10] was the first to mention differences between spoken phonemes and corresponding visemes in the mouth area. Nowadays lipreading and viseme recognition is a well established, yet challenging research field in the context of audio-visual speech recognition. The first system was reported in 1984 by Petajan [11] who distinguished letters from the alphabet and numbers from zero to nine and achieved 80% accuracy on that task. Since then the field has advanced in terms of recognition vocabulary, features and modelling approaches. In 2011 Zhou et al. [12] achieve a Frame Recognition Accuracy (FRA) of 56% on the speaker independent OuluVS database [13] proposing a method to project visual mouthing features to a low dimensional graph representation. Lan et al. [8] achieve an accuracy of 45% on their challenging 12 speakers audio-visual corpus. A good overview of the field is given in [14] and [15]. Neti et al. [16] present audio-visual but also visual only recognition results. In their report they briefly evaluate phonetic decision trees and context-dependent modelling of visemes. Not much work has been done training viseme models in an unsupervised or weakly supervised fashion. Most deals with the problem of clustering visemes in order to find an optimal phoneme to viseme mapping [17].

In facial expression recognition mouth features and classifiers can also be found [18], e.g. [19] recognizes action units (and models the mouth with only three different states: open, closed, very closed).

With respect to sign language, several works exist that exploit weak supervision to learn hand-based sign models [20–24]. Facial features have also been used before. Michael et al. [25] employs spatial pyramids of Histogram of Oriented Graphs (HOG) and SIFT features together with 3D head pose and its first order derivative to distinguish three grammatical functions trained on isolated American Sign Language (ASL) data of three signers. Vogler and Goldstein [26] present a facial tracker specifically for ASL.

Pfister et al. [27] employ mouth openness as feature to distinguish signing from silence. This is used to reduce the candidate sequences in multiple instance learning (which besides manual features employs a sift descriptor of the mouth region). However, to our knowledge no previous work has explicitly modelled dedicated visemes in the context of sign language recognition.

4 Corpora

The proposed approach uses the publicly available RWTH-PHOENIX-Weather corpus, which contains 7 hearing interpreter's performing continuous signing in DGS. The corpus consists of a total of 190 TV broadcasts of weather forecast recorded on German public TV. It provides a total of 2137 manual sentence segmentations and 14717 gloss annotations, totalling to 189.363 frames. Glosses constitute a less labour intense way of annotating sign language corpora. They can be seen as an approximate semantic description of a sign, usually annotated w.r.t. the manual components (i.e. the hand shape, orientation, movement and position), neglecting many details. For instance, the same gloss 'MOUNTAIN' denotes the sign alps but also any other mountain, as they share the same hand configuration and differ only in mouthing. Moreover, the RWTH-PHOENIX-Weather corpus contains 22604 automatically transcribed and manually corrected German speech word transcriptions. The boundaries of the signing sentences are matched to the speech sentences. It is worth noting that the sentence structures for spoken German and DGS do not correlate. This is a translation rather than a transcript.

For the purpose of evaluating this work, we annotated 5 sentences per signer on the frame level with viseme labels totalling 3687 labelled frames. The annotation was performed three times by a learning non-native signer with profound knowledge of sign language. While annotating, the annotator had access to the video sequence of signing interpreters showing their whole body (not just the mouth), the gloss annotations and the German speech transcriptions. In each of the three annotation iterations the frame labels varied slightly due to the complexity and ambiguity of labelling visemes (see [8] for a human evaluation of viseme annotations). We consider each annotation to be valid, yielding more than a single label per frame for parts of the data. Refer to Tab. 1 for details.

5 Mouthing Features

The features extracted from the mouth region consist of ten continuous distance measurements around the signers mouth and the average colour intensity of three areas inside the mouth (to capture tongue and teeth presence), as shown in Fig. 2. First and second order derivatives and an additional temporal window of 1 frame are added to the feature vector. In a later stage of the proposed algorithm Linear Discriminant Analysis (LDA) is used to reduce the dimensionality to 15.

The mouth-distance measurements are based on lower-level facial features, which are defined as a set of consistent, salient point locations on the interpreter's

Fig. 1. Illustration of context-dependency of visemes in the annotated data. All frames share the same annotation, but occur in different context. They stem from the phoneme /s/ which is mapped to 'T'. The first two frames originate from the pronounced sequence 'Island' (engl: Iceland), while the second two occurred within 'Küste' (engl: coast).

face. Since the structure of the human face as described by a set of such point features exhibits a lot of variability due to changes in pose and expression, we chose to base our tracking strategy on the deformable model registration method known as Active-Appearance-Models (AAMs).

In this work, we chose to use the efficient version of the simultaneous inverse-compositional AAM (SICAAM) proposed in [28]. The implementation is more robust to large variations in shape and appearance, which typically occur when dealing with facial expressions in the context of sign language. Moreover, it copes well with large out-of-plane head rotations, also commonly present in sign language, which can lead a 2D AAM to fail. We also use the refinement proposed in [29]. Following the work in [30] a 3D Point Density Model (PDM) is estimated using a non-rigid structure-from-motion algorithm on the training shapes, and is then involved in the optimisation process which incorporates a regularisation term encouraging the 2D shape controlled by the 2D PDM to be a valid projection of the 3D PDM. To estimate the high-level mouth distances

Table 1. Frame annotation statistics for 11 visemes on the RWTH-PHOENIX-Weather corpus. The penultimate line shows relative annotation per viseme in [%]. 'gb' denotes frames labelled as non-mouthings/garbage. 'ratio' refers to the average labels per frame (last row) or per viseme (last line), which reflects the uncertainty of the annotator.

	frames	A	E	F	I	L	O	Q	P	S	U	T	gb	ratio
Signer 1	489	45	34	42	48	12	73	55	62	19	36	112	240	1.6
Signer 2	484	66	46	38	30	28	47	59	36	31	44	94	298	1.7
Signer 3	556	69	27	26	57	20	65	105	65	21	29	127	326	1.7
Signer 4	517	92	62	47	42	21	58	70	40	26	45	116	161	1.5
Signer 5	596	62	62	64	97	44	53	57	50	36	54	121	268	1.6
Signer 6	522	76	42	68	29	13	73	77	36	16	42	136	241	1.6
Signer 7	523	46	29	40	87	23	71	57	57	9	36	127	256	1.6
$\sum$	3687	12.4	8.2	8.8	10.6	4.4	11.9	13.0	9.4	4.3	7.8	22.6	48.6	1.6
ratio	1.6	1.8	1.9	1.9	2.0	2.0	1.8	2.2	1.8	1.9	1.9	2.0	1.9	

we project the registered shape and remove its global translation and rotation by means of the 3D PDM. Then, for each point features subset given in Fig. 3, we estimate the corresponding local area-based measurements and normalise it between 0 and 1 according to the minimum and maximum values obtained during training. To capture the mouth cavity, we extract the pixels in the quadrilateral defined by its four mouth corners and project it to a fixed-sized square. The pixel intensities are averaged over three regions: patch top, centre and bottom, yielding 3 features.

6 Weakly Supervised Mouthing Recognition

6.1 Overview

The approach exploits the fact that mouthings are related to the corresponding spoken words, for which automatic spoken language transcripts are part of the RWTH-PHOENIX-Weather corpus. However, there is a loose relation between speech and mouthings, which holds for some signs only. An overview of the scheme is given in Fig. 4

Visual features of the mouth region are extracted and clustered using Gaussian clustering and Expectation Maximization (EM) while constraining the sequence of features to the sequence of automatically transcribed German words in a Hidden-Markov-Model (HMM) framework. For increased accuracy, the word sequence can be optionally reordered by using manual gloss annotations and techniques commonly used in statistical machine translation to align source and target language. Furthermore, a lexicon is built that includes a finite set of possible pronunciations for each German word. This lexicon consists of different phoneme sequences for each word and an entry for 'no-mouthing'. Finally, to account for the difference in articulatory phonemes and visual visemes, we need to map phonemes to visemes. Two different ways are explored to achieve this: either apply the mapping directly to the lexicon or to include it later in the pipeline in the estimation of context-dependent visemes. During the EM-iterations, the pronunciation probabilities in the lexicon are constantly updated based on the pronunciation counts in the current cluster. In the last step, context-dependent

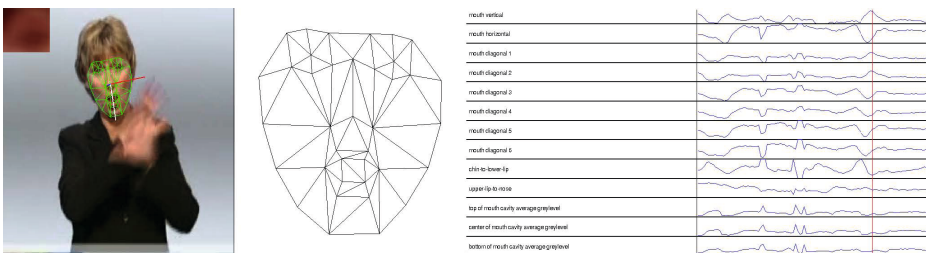
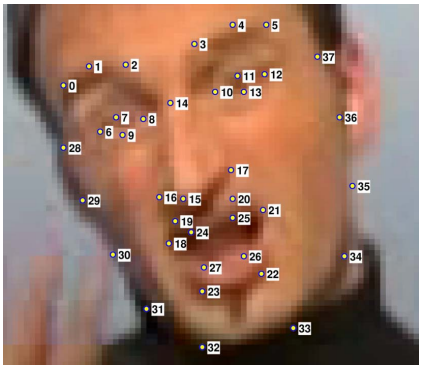


Fig. 2. Feature extraction, left: fitted AAM grid and inner mouth cavity patch, centre: rotated and normalised AAM grid, right: high-level feature values over time



Feature Description	Related Points #
vert. openness	[24-25 ; 26-27]
horiz. openness	[18 ; 21]
mouth diagonal 1	[25 ; 27]
mouth diagonal 2	[24 ; 26]
mouth diagonal 3	[20 ; 18]
mouth diagonal 4	[19 ; 21]
mouth diagonal 5	[18 ; 22]
mouth diagonal 6	[21 ; 23]
lip to chin	[26-27 ; 32-33]
lip to nose	[16-16-17 ; 18-24-25-21]
cavity	[24 ; 25 ; 26 ; 27]

Fig. 3. Visualisation of distance measures employed as features

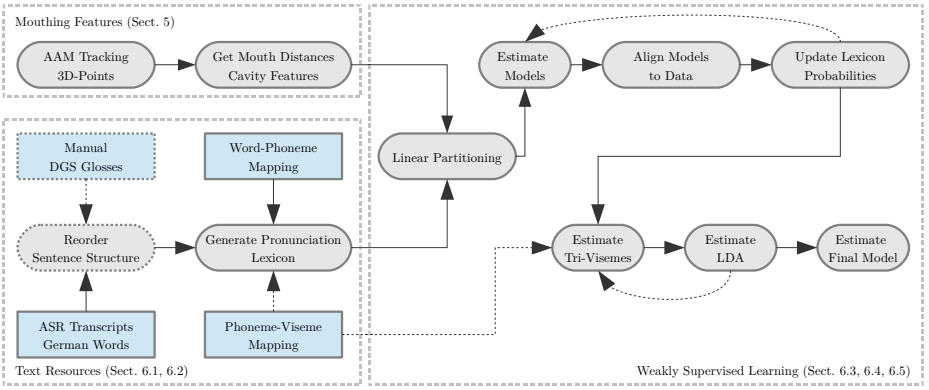


Fig. 4. Overview of the proposed approach. Dotted lines represent optional usage for better results. Round boxes represent procedures, while squared boxes are resources.

tri-visemes are estimated. In order to cope with limited data, a visemic Classification And Regression Tree (CART) is used to cluster those tri-visemes that share similar characteristics. The fine-grained tri-viseme alignments are used to perform a LDA on the input features, while adding more temporal context to the features.

6.2 Reordering Sentence Structure

Sign languages and their spoken counterparts do not share the same word order, nor does one word always translate to exactly one sign. Spoken German typically follows the 'subject (S), verb (V), object (O)' structure, while DGS prefers 'SOV'. Inspired by statistic machine translation, we employ a technique presented in [31], which maximises the alignment likelihood on a training corpus of sentence pairs each with a pair of sequences of German words $\mathbf{w} = w_1^J := w_1, \ldots, w_J$ and DGS

glosses $\mathbf{g} = g_1^I := g_1, \ldots, g_I$ $(\mathbf{w}, \mathbf{g})$. The approach uses an alignment variable $\mathbf{a} = a_1^J$, which describes the mapping from a source position j to a target position a_j for each sentence pair. We try to find the best Viterbi alignment by maximising the statistical alignment model p_θ, which depends on a set of unknown parameters θ that is learnt from the training data:

$$\hat{a}_1^J = \arg\max_{a_1^J} p_\theta(w_1^J, a_1^J | g_1^I) \tag{1}$$

The technique includes the so called IBM Models as alignment models, which account for lexical translation and reordering. For more details refer to [31]. However, the resulting alignment is very noisy, due to the limited amount of training data available and due to the fact that not every source word has a single target. We thus apply filtering to the generated (w, g) pairs constituting a mapping $M : \mathcal{G} \rightarrow \mathcal{P}(\mathcal{W})$, where $w \in \mathcal{W} = \{\text{all spoken words}\}$ and $g \in \mathcal{G} = \{\text{all sign glosses}\}$. We employ an absolute and relative filtering criterion, such that

$$M(g)' = \left\{ w \in M(g) \,\middle|\, c(w,g) > \vartheta_A \wedge \frac{c(w,g)}{\sum\limits_{w' \in M(g)} c(w',g)} > \vartheta_R \right\}, \tag{2}$$

where $c(w, g)$ counts the number of occurring pairs (w, g) and ϑ_A and ϑ_R are the thresholds.

6.3 Pronunciation Lexicon and Viseme Mapping

Based on German words, we can build a pronunciation lexicon, which defines the finite set of possible pronunciations that occur with a sign. We first need a phoneme representation of the German words. For this purpose we use a word-phoneme mapping which has been generated with the publicly available Sequitur Grapheme-to-Phoneme converter [32].

However, mouthings produced by signers often do not constitute fully pronounced words, but rather discriminating bits of words. Thus, for each full pronunciation we add multiple shortened versions to our lexicon ψ by truncating the word w which consists of a sequence of phonemes $s_1^N = s_1, \ldots, s_N$, such that

$$\psi = \left\{ w' : s_1^{N-\phi} | \phi \in \{0, \ldots, \phi_{trunc}\} \wedge N - \phi \geq \phi_{min} \right\} \tag{3}$$

Moreover a 'no-mouthing' is added to the lexicon for each word. We are aware of the fact that visemes have a different inventory than phonemes. In the literature there is some specific work on viseme sets for Deaf people. Elliott [33] suggests a phoneme to viseme mapping resulting in 11 visemes (A, E, F, I, L, O, P, Q, S, T, U).

We choose two different ways to include this viseme knowledge into our pipeline: 1. We map our phoneme pronunciations to viseme sequences. 2. We use phoneme classes as models and include the viseme mapping in a visemic clustering of tri-visemes, as described later in this paper (Section 6.5).

6.4 Training Viseme Models

We use EM with Gaussian clustering in an HMM-framework to train viseme models from our data. Thus, we consider the weakly supervised viseme training to be a search problem of finding the sequence of visemes $v_1^Z := v_1, \ldots, v_Z$ belonging to a sequence of mouthings (pronounced words) $m_1^N := m_1, \ldots, m_N$, where the sequence of features $x_1^T := x_1, \ldots, x_T$ best matches the viseme models. We maximise the posterior probability $p(v_1^N|x_1^T)$ over all possible viseme sequences for the given sequence of glosses.

$$x_1^T \rightarrow \hat{v}_1^Z(x_1^T) = \arg\max_{v_1^Z} \left\{ p(m_1^N) p(x_1^T|v_1^Z) \right\}, \tag{4}$$

where $p(m_1^N)$ denotes the pronunciation probability for a chosen mouthing. In a first step we model each viseme by a 3 state HMM and a no-mouthing model having a single state. The emission probability of an HMM state is represented by a single Gaussian density with a diagonal covariance matrix. The HMM states have a strict left to right structure. Global transition probabilities are used for the visemes. The no-mouthing model has independent transition probabilities. We initialise the viseme models by linearly partitioning the data. We then use the EM algorithm to iteratively 1. estimate the best alignment based on the current models and 2. to accumulate updated viseme models and 3. update pronunciation probabilities based on the alignments. To prevent abrupt changes in the pronunciation probabilities due to limited data, we average the probabilities over the last three alignments.

6.5 Context-Dependent Visemes with a Visemic Classification and Regression Tree

Visemes are known to be context dependent, e.g. the viseme /s/ in the words 'sue' and 'sea' is likely to have very different properties. Refer to Fig. 1 for a visual example. Co-articulation effects stem from the constraints enforced by the human muscular system, which does not allow immediate, ad-hoc execution or stops of motions, but rather blends one movement into another [34, 35].

We model the viseme context using both the previous and subsequent viseme (so-called tri-visemes). However, due to data limitations, not all tri-visemes can be observed during training. It is necessary to tie states of less frequent tri-visemes together and pool their model parameters. We follow the approach of phonetic decision trees presented in [36] for Automatic Speech Recognition (ASR). We cluster the tri-visemes with respect to visual properties of the visemes. The method uses a decision tree whose internal nodes are tagged with questions on these properties, as listed in Tab. 2. The leafs of the tree represent the actual tri-visemes. The

Table 2. Common visemic properties of the mouthings, used for decision tree based clustering

Common Property	Visemes
Consonant	F, P, T
Vowel	A, E, I, O, U, Q
Alveolar	T, Q
Labial	F, P
Round	U, O ,S
Not-round	I, E
Open	A, Q ,L
Semi-open	U, Q, L, E, T

observations within each node are modelled by a single Gaussian density with diagonal covariance. Starting at the root, the leafs of the tree are consecutively split by the questions regarding the visemic properties, where the order of questions is based on the maximum local gain in likelihood. Splitting is stopped, when there are less than 200 observations in a leaf or when the likelihood gain falls below a threshold. The tree can also be used to incorporate further linguistic knowledge such as the mapping from phonemes to visemes.

6.6 Linear Discriminant Analysis

LDA helps to find a linear transformation of our feature vectors to a lower dimensional space, while maximising class separability. Inspired from a quasi-standard in ASR [37], we apply LDA to the estimated tri-visemes. At this stage we also take into account the temporal context by concatenating the preceding $n = 3$ frames to the feature vector x_t^T, which yields a context feature vector X_t^T consisting of context frames plus the current frame. Finally, a reduced feature representation y_t is achieved by projecting X_t into a subspace of reduced dimensionality 15 with $y_t = V^T X_t$. The transformation matrix V^T is constructed by LDA such that it maximises interclass, while minimising intra-class variance.

7 Results

In this section, we present results that allow assessment of all training steps proposed in this framework. We evaluate four different setups in terms of their alignment performance during weakly supervised training and in classification performance on the frame level ground truth annotation (see Tab. 3). Due to the weakly supervised nature, the latter can be understood as a recognition constrained by the accompanying manual signs. If not otherwise stated, all results have been trained in a signer independent fashion, i.e. leaving one signer's data out of the training and averaging over all signers. Furthermore, we show how the visemes generalise across different signers comparing a multi-signer setup (no unseen signer in test) with a signer independent setup (see Tab. 4) and how the systems behave with a variation of precision and recall based on classifier

confidence thresholding (see Fig. 5). Finally, we also analyse the classification errors on the viseme level (see Fig. 6)

We perform classification based on the highest pooled posterior probability per frame $\hat{p}(v|x)$ of the viseme v given the feature vector x,

$$\hat{p}(v|x) = \max_{v \in \mathcal{V}} p(v|x) = \max_{v \in \mathcal{V}} \sum_{v_c \in \mathcal{C}_v} p(v_c|x), \tag{5}$$

where $\mathcal{C}_v = \{v_{c1}, \ldots, v_{cN}\}$ contains all context-dependent tri-visemes of v. The classification does not rely on any priors, such as a grammar. The standard classification task distinguishes 11 visemes and a 'no-mouthing' class, whereas in a second task ('excl. Garbage') we exclude all frames that have been manually labelled with 'garbage' and evaluate only the 11 viseme classes.

As evaluation criterion we chose precision $= \frac{\text{tp}}{\text{tp+fp}}$ and recall $= \frac{\text{tp}}{\text{tp+fn}}$, where a classification is counted as true positive (tp) if it corresponds to any of the annotated ground truth labels (1.6 labels/frame, see Section 4). The reference labels count a false negative (fn) if no classified label matches them. If the chosen label was other than 'garbage' it counts additionally as false positive (fp).

In Tab. 3 we see four different experiments. The first experiment does not compensate for different word order (see Section 6.2) and applies the viseme mapping at an early stage straight to the lexicon (see Section 6.3), while experiment (3) and (4) incorporate the phoneme-viseme mapping into the clustering of tri-visemes (see Section 6.5). Precision and recall are given for each training step of all experiments: after the initial linear bootstrapping of the models, after 25/50 iterations of the EM-algorithm (see Section 6.4) and after successive tri-viseme clustering and incorporation of temporal context with a LDA (see Section 6.6). The results in Tab. 3 show the strength of our weakly supervised learning approach in detail. Furthermore, Fig. 6 shows the confusions on the viseme level achieved by the best system, split up by each signer, allowing to asses the quality of the approach in general and also qualify its signer independent capabilities. Following statements can be drawn from the results:

1. **Reordering is important.** The alignment precision during training improves in all cases (see right columns in Tab. 3, 34.1 $\rightarrow$ 41.3% and 34.3 $\rightarrow$ 41.3%). Reordering has in all cases a positive impact on the final classification performance (43.4 $\rightarrow$ 44.1% and 40.8 $\rightarrow$ 47.1%). Earlier EM-iteration steps in some cases show a slight degradation, which may be due to introduced noise by the reordering technique. In Fig. 5 we also see that systems (2) and (4) outperform the others.
2. **Integration of a viseme mapping through a visemic decision tree is advisable when reordering is applied.** The late integration outperforms the early viseme mapping with 44.1 $\rightarrow$ 47.1%.
3. **Visemes have signer independent properties.** Tab. 4 shows that the recognition precision only degrades by 3.2% (32.1% $\rightarrow$ 29.0%) on average from the multisigner to the unseen signer (signer independent) case. Signer specific models have a slightly better performance, but their data is very

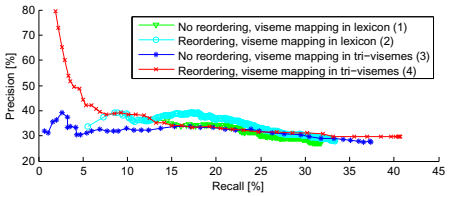


Fig. 5. Performance curves of the four competing systems. Precision and recall varied by applying a confidence threshold to the joint classifier

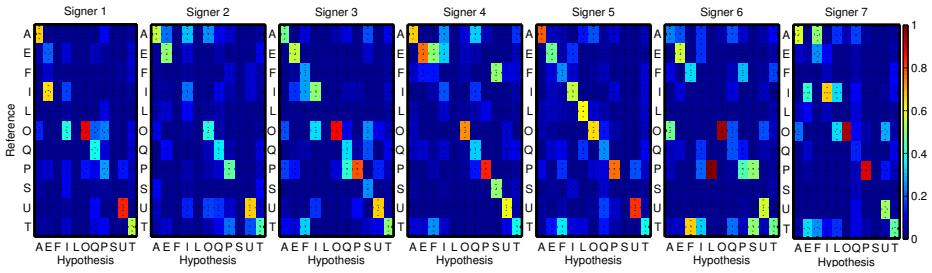


Fig. 6. Confusion matrices per signer of the best system (4) ('Reordering & viseme mapping in tri-visemes') excluding frames manually labelled as garbage on a signer independent task. Colours on the diagonal correspond to the precision of a certain viseme. This setup achieves 47.1% precision averaged over all signers.

limited. However, Fig. 6 shows that not all visemes are equally well recognized across all signers. Thus, improved adaptation methods are still required.

4. **Context-dependent modelling is very important.** Context-dependent outperform context-independent visemes heavily (e.g. 27.0 → 44.9%)
5. **Frames ground-truthed as 'garbage' are problematic.** Results excluding 'garbage' are constantly better than including it.
6. **LDA with added temporal context seems to require more and cleaner training alignments.** In cases without applied reordering , the LDA does not improve results. This may be due to low recall and limited precision achieved by the weakly supervised training (see 'Alignment during Training' in Tab. 3: 35.8 → 36.9% and 35.9 → 38.5%).
7. **Normalization of features w.r.t. the signer and to the out of plane rotation is important.** Comparative experiments have been done replacing the AAM distance features by a sift descriptor (128 dim., placed in the centre of the mouth, resized to match the mouth opening). This only yields 26.4% prec. and 26.0% recall in the 'no garbage' task and compares to 44.9% and 41.7% with the original features.

In terms of computational complexity, the algorithm requires around 50 minutes to train using all 189.363 frames on a single core of a AMD Opteron

Table 3. Precision and recall in [%] measured on the frame-level of continuous viseme recognition without any grammar constraints in a signer independent task averaged over all seven signers

	Testing				Alignment during Training			
	Standard		no Garb.		Standard		no Garb.	
	prec.	recall	prec.	recall	prec.	recall	prec.	recall
No Reordering & viseme mapping in lexicon (1)								
Partition linearly	11.9	10.6	17.9	11.8	23.7	23.5	33.8	25.8
25 EM-iterations	11.4	12.4	18.1	13.9	33.9	32.8	49.1	35.7
50 EM-iterations	11.5	12.5	18.3	14.1	34.0	32.9	49.2	35.8
1st Tri-visemes	29.3	34.6	43.5	38.3	34.1	33.0	49.4	35.9
LDA	29.0	38.0	43.4	42.3	”	”	”	”
Reordering & viseme mapping in lexicon (2)								
Partition linearly	10.2	9.6	16.0	10.7	30.0	26.4	40.3	28.3
25 EM-iterations	11.9	13.4	19.1	15.1	40.9	36.0	55.9	38.3
50 EM-iterations	11.9	13.4	19.1	15.1	41.2	36.2	56.2	38.4
1st Tri-visemes	29.0	35.5	43.5	39.6	41.3	36.3	56.3	38.5
LDA	29.5	39.0	44.1	43.7	”	”	”	”
No Reordering & viseme mapping in tri-visemes (3)								
Partition linearly	16.8	17.7	25.1	19.8	24.1	24.2	33.9	26.4
25 EM-iterations	16.7	20.9	26.1	23.9	33.9	32.9	47.7	35.5
50 EM-iterations	17.0	21.3	26.6	24.4	34.5	33.4	48.4	36.0
1st Tri-visemes	27.3	34.0	41.7	38.1	34.3	33.2	48.0	35.8
LDA	26.6	36.9	40.8	41.6	”	”	”	”
Reordering & viseme mapping in tri-visemes (4)								
Partition linearly	17.0	23.2	26.4	26.4	31.4	28.2	42.1	30.2
25 EM-iterations	17.4	22.3	27.2	25.5	41.0	34.6	54.5	36.4
50 EM-iterations	17.2	22.1	27.0	25.3	41.4	35.1	55.1	36.9
1st Tri-visemes	29.7	37.2	44.9	41.7	41.3	35.1	55.1	36.9
LDA	**31.3**	**43.2**	**47.1**	**48.2**	41.3	35.1	55.1	36.9
Chance	13.3	-	13.9	-	-	-	-	-

Table 4. Precision and recall in [%] on the frame-level of continuous viseme recognition without grammar constraints. Results are given for signer specific models (Single Signer), all signers trained jointly (Multi Signer) and for all signers trained jointly with exclusion of any data of the tested signer (Signer Independent).

	Single Signer		Multi Signer		Signer Indep.	
	prec.	recall	prec.	recall	prec.	recall
Average	33.5	36.1	32.1	38.1	29.0	35.5
Signer 1	41.9	45.1	31.5	38.2	24.1	28.8
Signer 2	29.9	37.2	22.3	33.3	25.0	37.4
Signer 3	27.7	30.4	22.9	28.7	17.9	23.7
Signer 4	39.5	34.8	49.6	41.2	38.1	37.2
Signer 5	34.3	42.0	37.9	46.9	36.4	45.2
Signer 6	30.8	31.7	31.5	37.4	30.2	36.4
Signer 7	29.8	30.2	31.4	37.5	30.4	37.2

Processor 6176 with 2300 Mhz. Each of the 25 EM iterations takes approximately 20 minutes. Frame recognition runs at around 9000 frames per second (fps), whereas feature extraction (matlab implementation) runs at only 0.07 fps.

8 Conclusions

This paper has proposed a framework to build a mouthing recogniser for continuous sign language. To our knowledge no previous work has achieved to apply a dedicated viseme recognition to the particularities of sign language recognition. We use no hand labelled training data, but just a pool of 189.363 frames. Our approach reaches 47.1% precision on the frame level on a challenging signer independent task, facing low quality 'real-life' data recorded from TV, with low spatial resolution. The approach requires only weak supervision and does not rely on any grammar priors.

The approach uses AAM-based distance features around the mouth to model 11 visemes and a 'no-mouthing class. The visemes are modelled as context-dependent tri-visemes which are clustered using a visemic decision tree. An extensive quantitative analysis in four different experimental settings allows to deduce new knowledge about recognition of mouthings in sign language.

We find that the modelling of visemes drastically improves with context dependent tri-visemes. Furthermore, accounting for differences in sentence structure between spoken and sign language improves the visual models. We further show that the visemes generalise well to unseen signers with a drop of only 3.2% precision.

Besides adding adaptation methods to enhance generalisation across signers, we identify the task of distinguishing between mouthings and mouth gestures in sign language as important future research. Moreover, work is needed to integrate the viseme recognition into a multimodal recognition pipeline. Finally, finding the actual number and properties of visemes best suited for sign language recognition also remains an open question.

Acknowledgements. The work presented has been supported by the EPSRC project "Learning to Recognise Dynamic Visual Content from Broadcast Footage" (EP/I011811/1). Special thanks to Thomas Hoyoux (University of Innsbruck) for continuous support related to the AAMs.

References

1. Starner, T., Weaver, J., Pentland, A.: Real-time American sign language recognition using desk and wearable computer based video. IEEE Pattern Analysis and Machine Intelligence 20(12), 1371–1375 (1998)
2. Vogler, C., Metaxas, D.: Handshapes and movements: Multiple-channel American sign language recognition. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 247–258. Springer, Heidelberg (2004)

3. Zaki, M.M., Shaheen, S.I.: Sign language recognition using a combination of new vision based features. Pattern Recognition Letters 32(4), 572–577 (2011)
4. Ong, S.C., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Pattern Analysis and Machine Intelligence 27(6), 873–891 (2005)
5. Lucas, C., Bayley, R., Valli, C.: What's your sign for pizza?: an introduction to variation in American Sign Language. Gallaudet University Press, Washington, D.C (2003)
6. Emmorey, K.: Language, Cognition, and the Brain: Insights From Sign Language Research. Psychology Press (November 2001)
7. Sandler, W.: Sign Language and Linguistic Universals. Cambridge University Press (February 2006)
8. Lan, Y., Harvey, R., Theobald, B.-J.: Insights into machine lip reading. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4825–4828 (March 2012)
9. Hilder, S., Theobald, B.J., Harvey, R.: In pursuit of visemes. In: Proceedings of the International Conference on Auditory-Visual Speech Processing, pp. 154–159 (2010)
10. Fisher, C.G.: Confusions among visually perceived consonants. Journal of Speech, Language and Hearing Research 11(4), 796 (1968)
11. Petajan, E.D.: Automatic Lipreading to Enhance Speech Recognition (Speech Reading). PhD thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA (1984)
12. Zhou, Z., Zhao, G., Pietikainen, M.: Towards a practical lipreading system. In: Computer Vision and Pattern Recognition, pp. 137–144 (2011)
13. Zhao, G., Barnard, M., Pietikainen, M.: Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia 11(7), 1254–1265 (2009)
14. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.: Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE 91(9), 1306–1326 (2003)
15. Chi†u, A., Rothkrantz, L.J.M.: Automatic visual speech recognition. In: Ramakrishnan, S. (ed.) Speech Enhancement, Modeling and Recognition- Algorithms and Applications. InTech (March 2012)
16. Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J.: Audio-visual speech recognition. In: Final Workshop 2000 Report, vol. 764 (2000)
17. Aghaahmadi, M., Dehshibi, M.M., Bastanfard, A., Fazlali, M.: Clustering persian viseme using phoneme subspace for developing visual speech application. Multimedia Tools and Applications, 1–21 (2013)
18. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing 27(6), 803–816 (2009)
19. Tian, Y.L., Kanade, T., Cohn, J.: Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(2), 97–115 (2001)
20. Buehler, P., Everingham, M., Zisserman, A.: Employing signed TV broadcasts for automated learning of British sign language. In: Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, pp. 22–23 (2010)
21. Cooper, H., Ong, E.J., Pugeault, N., Bowden, R.: Sign language recognition using sub-units. The Journal of Machine Learning Research 13(1), 2205–2231 (2012)

22. Kelly, D., McDonald, J., Markham, C.: Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41(2), 526–541 (2011)
23. Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) Visual Analysis of Humans, pp. 539–562. Springer, London (2011)
24. Koller, O., Ney, H., Bowden, R.: May the force be with you: Force-aligned SignWriting for automatic subunit annotation of corpora. In: IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai, PRC (April 2013)
25. Michael, N., Neidle, C., Metaxas, D.: Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation. In: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC, Malta (2010)
26. Vogler, C., Goldenstein, S.: Facial movement analysis in ASL. Universal Access in the Information Society 6(4), 363–374 (2008)
27. Pfister, T., Charles, J., Zisserman, A.: Large-scale learning of sign language by watching TV (using co-occurrences). In: Proceedings of the British Machine Vision Conference, U. K. Leeds (2013)
28. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing 23(12), 1080–1093 (2005)
29. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2D+ 3D active appearance models. In: CVPR (2), pp. 535–542 (2004)
30. Schmidt, C., Koller, O., Ney, H., Hoyoux, T., Piater, J.: Enhancing gloss-based corpora with facial features using active appearance models. In: International Symposium on Sign Language Translation and Avatar Technology, Chicago, IL, USA, vol. 2 (2013)
31. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. 29(1), 19–51 (2003)
32. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication 50(5), 434–451 (2008)
33. Elliott, E.A.: Phonological Functions of Facial Movements: Evidence from deaf users of German Sign Language. Thesis, Freie Universität, Berlin, Germany (2013)
34. Jiang, J., Alwan, A., Bernstein, L.E., Auer, E.T., Keating, P.A.: Similarity structure in perceptual and physical measures for visual consonants across talkers. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. I–441–I–444 (May 2002)
35. Turkmani, A.: Visual Analysis of Viseme Dynamics. Ph.d., University of Surrey (2008)
36. Beulen, K.: Phonetische Entscheidungsbäume für die automatische Spracherkennung mit großem Vokabular. Mainz (1999)
37. Haeb-Umbach, R., Ney, H.: Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 13–16 (1992)