Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT), volume 2, number 1, October 2013

Enhancing Gloss-Based Corpora with Facial Features Using Active Appearance Models

Christoph Schmidt, Oscar Koller, Hermann Ney
{surname}@cs.rwth-aachen.de
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University

Thomas Hoyoux, Justus Piater
{firstname}.{surname}@uibk.ac.at
Intelligent and Interactive Systems
University of Innsbruck, Austria

ABSTRACT

In text-to-avatar translation systems, facial expressions and mouth patterns are a vital part of a natural sign language avatar animation. However, gloss-based corpora often lack detail with respect to such non-manual features. To create a translation system which can produce facial expressions and mouthings, a more fine-grained annotation is necessary. In this work, we apply a clustering algorithm to automatically distinguish between different facial patterns using an active appearance model. The resulting translation system is then able to produce such expressions based on the written language text. In our experiments, the system produced suitable expressions with an accuracy of 78.4%.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language generation

General Terms

Algorithms, Experimentation

Keywords

Sign Language, Facial Expressions, Mouthing, Active Appearance Models, Gloss Notation, Clustering

1. INTRODUCTION

Sign languages are multi-modal in the sense that they express meaning simultaneously via different communication channels. Besides the manual information such as hand shape, orientation and movements, non-manual aspects such as body posture and facial expressions play a vital role in expressing meaning. These aspects are also important in the field of text-to-avatar translation. User studies such as [7] and [9] have shown that facial expressions are necessary both for intelligibility and perceived fluency of sign language avatars.

One non-manual aspect of sign languages is mouthing. In some countries which have a strong oralist tradition, sign languages frequently use mouthings, that is, mouth patterns derived from the spoken language. In these languages, mouthings are also used to derive new signs by using the manual components of a similar sign and changing only the mouthing.

One open question in the sign language research community is how to capture this multi-modal nature of sign languages in a comprehensive annotation system. Existing sign language corpora vary greatly in the notation used and in the way multi-modality is treated. Often, the annotation system chosen for the corpus is highly influenced by the task which the scientists want to tackle.

A simple annotation method is gloss annotation, where a sign is annotated by one or several words which roughly correspond to its meaning, usually written in the stem form in upper case. Since the same sign can have several meanings in different contexts, it can be transcribed differently depending on its context. In contrast to this, the term ID-gloss [8] is used if one sign is always annotated with the same gloss, independent of its meaning in a particular context. In our corpus and experiments, we use ID-glosses.

Gloss annotation has several advantages, but also drawbacks when compared to more fine-grained notation systems such as the Hamburg Notation System [14]. One advantage is that signing variants which slightly differ in a component are transcribed differently on the "phonetic" level but can be denoted as a variant in the gloss system, e.g. by adding a variant suffix to a common stem ("lemma"). Moreover, annotation work using glosses can be done faster and requires less training for the annotators, and thus larger corpora can be annotated with the same amount of resources available.

However, gloss annotation also has disadvantages and limits. Since glosses encode a whole sign by a single token, the multi-modal nature of sign languages which use not only the hands but also facial expression, eye-gaze, torso or shoulder movements and head movements such as head shakes is not

adequately captured. For a critical discussion of gloss notation see [13]. Gloss notation can only approximate multimodal features by introducing gloss variants. For example, a prefix *NEG-* can be added to a gloss to indicate that a sign is negated by a head shake. Similarly, mouthings can be added to a gloss as an additional information, leading to an enriched gloss system. This leads to the question of granularity. How fine-grained does the gloss system have to be to be used in an automatic machine translation framework?

In this work, we present a solution providing mouthings and facial expressions for text-to-avatar translation systems that are trained based on simple gloss corpora lacking any non-manual annotation. Active appearance models allow for extracting specific facial features from the video corpus. A clustering approach then helps to automatically identify the underlying variations and finally spots representative expressions to be used in the avatar animation.

This paper is structured as follows: related work is discussed in Section 2. We describe the basic approach of this work in Section 3. Section 4 gives an overview of the RWTH-Phoenix-Weather corpus which we use for our experiments. In Section 5, we review the technique of active appearance models for the extraction of facial features and mouthings. Section 6 presents our experiments to cluster glosses and choose a suitable mouthing according to its translation in the spoken language. A conclusion and outlook on future work is given in Section 7.

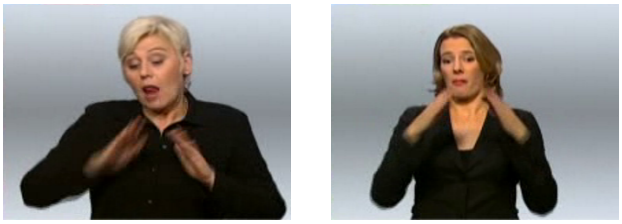
2. RELATED WORK

In sign language and avatar animation research fields, deformable model methods such as active appearance models and constrained local models are particularly suited for detecting facial expressions and mouth patterns and have naturally found many applications. These model-based marker-free techniques are powerful tools to extract meaningful facial features in video corpora.

Many publications use active shape models or active appearance models for tracking facial point features and then inferring facial cues from them in the context of video-based automatic sign language analysis. [2] and [15] focus on recognizing a predefined set of facial expressions and use training corpora with annotated facial expressions. [12] and [19] provide facial features for the use in a sign language recognition framework, i.e. they integrate low-level facial features into their system to improve the recognition of the glosses.

More closely related to our work, active appearance models can also be used to synthesize facial expressions in a picture or video. In [1], a discrete set of emotional expressions is synthesized and projected into a picture. A dynamic transfer of facial expressions onto an avatar is presented in [17], which describes a real-time puppetry system based on a constrained local model and which focuses on the realism of the transferred expressions.

In our work, we extract facial features and mouthings from a gloss-annotated corpus to provide a more natural avatar animation. In contrast to previous work, no annotation of facial features is necessary. Our paper focuses on the selection of appropriate video samples for the animation of facial expressions based on the gloss label and its translation in the spoken language text.



ALPS (mouthing "Alpen") MOUNTAIN (mouthing "Berg")

Figure 1: Two signs with the same manual component, differing only in the mouthing. At the time of the snapshots, the underlined letters are pronounced.

3. ENHANCING GLOSS-BASED CORPORA BY CLUSTERING

In the following, we will describe the clustering approach developed in this paper.

The annotation of a sign language video corpus highly depends on the task at hand. For example, if a linguist wants to study certain linguistic patterns, the annotation should be detailed with respect to these patterns. In the same way, an annotation suitable for an automatic sign language recognition system should be tailored according to the features which the system can actually recognize. Since the RWTH-Phoenix-Weather corpus, which will be described in the next section, was originally developed for the recognition of hand-based features, both the time boundaries of the ID-glosses and their label were mainly based on the signing hands. This means that signs which are identical in the hand components but differ in their mouthing received the same label. For example, the sign of a specific mountain is formed by mouthing its name and performing the general sign for mountain with the hands (see Figure 1). In the corpus, this sign is glossed as "MOUNTAIN". Using the corpus in a text-to-avatar scenario implies that the system can only reliably produce hand patterns but not other features such as facial expressions or mouthings.

In order to create a translation system which distinguishes between signing variants featuring different mouthings and facial expressions, the gloss annotation has to be more fine grained. Since a manual revision of all glosses is very time-consuming and thus expensive, we opt for an automatic clustering technique which distinguishes glosses with different facial patterns based on their translation in the spoken language.

The baseline approach to a gloss-based text-to-avatar system is to translate the spoken language text into a sequence of glosses and then to create an avatar animation based on the glosses. The issue is that a gloss is always animated in the same way, because the context information is lost, and gloss variants are not considered.

In our approach, when translating the spoken language text into glosses, the translation system output also contains the

original spoken word which corresponds to the gloss, thus consisting of (gloss,translation) pairs. E.g., if the system translates the spoken word "Alps" into the gloss MOUNTAIN, it produces the pair (MOUNTAIN,Alps). The clustering algorithm then determines a suitable sample video to animate each specific gloss variant.

Here is an outline of the procedure:

Input: a gloss-annotated video training corpus with corresponding sentences in a spoken language

1. Each gloss is aligned to a corresponding word in the spoken language using the open-source toolkit GIZA++[11]. With this alignment, each gloss is provided with its translation in the spoken language as additional context information, leading to (gloss,translation) pairs. This information is then used to guide the clustering process.
2. For each (gloss,translation) pair, all videos labelled with this pair are clustered:
   (a) Facial features of all videos are extracted using active appearance models.
   (b) The similarity of the facial features and mouthing between pairs of glosses is calculated based on a hidden Markov model, resulting in a distance matrix.
   (c) The videos are clustered according to the distance of the facial features.
   (d) The central element of the biggest cluster is selected to obtain an appropriate facial expression and mouthing variant when translating a spoken language word into a gloss.

Output: for each (gloss,translation) pair, a representative video is provided from which facial features can be extracted for avatar animation.

The clusters obtained by the algorithm are variants of a sign exhibiting different facial patterns. For signs with no or only one mouthing, the output contains only one cluster. If a cluster obtained by this procedure consists of only one or a few glosses, it can be considered an outlier or a seldom variant. Thus, on the one hand the algorithm distinguishes different facial features and mouthings, and on the other hand it helps avoiding less standardized variants not suitable for animation.

For clustering, we apply a medoid-shift clustering algorithm [3]. The algorithm clusters the elements around medoids, i.e. representative samples for each cluster, which can later be used to extract facial features for avatar animation.

When translating a new text from the spoken language and animating the (gloss,translation) pairs using an avatar, the algorithm has to select one video from the training data to model the facial features. The heuristic of the algorithm is to select the biggest cluster, because it was the variant which was seen most often in the context of the original spoken word. Within the cluster, the medoid is selected as the representative video. The avatar is then animated using the facial features of this video.

Table 1: Statistics of the RWTH-Phoenix-Weather corpus for DGS and announcements in spoken German

	DGS	German
# signers	7	
# editions	190	
duration[h]	3.25	
# frames	293,077	
# sentences	2,711	
# running glosses	17,744	33,190
vocabulary size	463	1,494
# singletons	537	536

In the next section, we will describe the RWTH-Phoenix-Weather corpus on which we will perform the clustering experiments.

4. THE RWTH-PHOENIX-WEATHER CORPUS

The RWTH-Phoenix-Weather corpus is a video-based, large vocabulary corpus of German Sign Language recorded and annotated for the use in statistical pattern recognition and statistical machine translation. The public TV broadcasting station Phoenix regularly broadcasts the major public news programs with an additional interpretation into German Sign Language using an overlay window which shows the interpreter.

The RWTH-Phoenix-Weather corpus contains the weather forecast portions of these news programs, which were manually annotated by a deaf expert and revised by a hard-of-hearing expert. The weather forecasts were chosen because weather forecasting forms a rather compact domain with a limited vocabulary. A complex domain such as news programs would require a much larger corpus to reliably estimate statistical models, but annotating such a corpus was infeasible due to time and budget constraints. The annotation of the RWTH-Phoenix-Weather corpus consists of glosses, and time boundaries have been marked on the sentence as well as the gloss level. The spoken German weather forecast has been transcribed semi-automatically using a state-of-the-art automatic speech recognition system. To train active appearance models on this corpus, facial landmarks have been manually labelled on a small set of images.

In the following, we will briefly describe the corpus setup and statistics. For a more thorough description see [5]. Baseline translation results both from German to German Sign Language and in the opposite direction can be found in [18].

Note that for this work, we only use sentences for which individual gloss boundaries have been annotated, so the statistics differ from the above reference. The time boundaries are necessary to extract facial features. The corpus statistics for the RWTH-Phoenix-Weather corpus can be found in Table 1. The database features a total of seven interpreters, consists of 2711 sentences and a total of 17,744 running glosses.

Figure 2: Visualization of facial annotations

To train active appearance models on this corpus, 38 facial landmarks for all seven interpreters have been labelled in a total of 369 images (that is, about 50 images per interpreter). Care was taken in selecting a set of images which contain many different expressions, including extreme ones, such that the trained models can approximately represent a large span of expressions for each interpreter. Two examples of the facial annotations are shown in Figure 2.

To evaluate the clustering algorithms developed in this study, we additionally labelled the mouthing for a subset of glosses. Starting with the most frequent glosses, we selected a subset of 23 glosses (2.5% of the vocabulary) for which more than one mouthing exists in the corpus. For these glosses, we select the pairs of glosses and their aligned spoken language word which were seen more than five times. This led to 64 (gloss,translation) pairs. For each pair, we labelled up to 25 instances in the corpus, though for many pairs the number of instances in the corpus was smaller. In total, we labelled 640 running glosses. These labels are not used in training but solely for the purpose of evaluating the quality of the resulting clusters.

5. ACTIVE APPEARANCE MODELS

The facial features which are used in the clustering approach described in Section 3 consist of the continuous measurements of some quantities related to mouthing and other facial cues for sign language. As shown in Table 2, these measurements are based on lower-level facial features which are defined as a set of consistent, salient point locations on the interpreter's face. As illustrated in Figure 2, these fiducial points – also called landmarks – correspond to key locations on the cheeks and chin outlines, the nose ridge and nose base, the eyelids and eye corners, the eyebrow outlines and the lip and mouth corners. We wish to track those point features accurately in the sign language videos in order to extract the higher-level facial features which will in turn be used in the context of the proposed clustering approach for enhancing the gloss-based RWTH-Phoenix-Weather corpus. 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.

Active appearance models (AAMs), first proposed in [4] and notably reformulated in [10], are a popular instance of the family of deformable model methods for image interpretation. Such model-based methods attempt to recover an object's structure as it appears in an image by registering a deformable shape model of the object to the image data. Mathematically, the shape $\mathbf{s}$ of an object is defined as the vector of stacked coordinates of its v landmark points:

$$\mathbf{s} = (x_1, y_1, x_2, y_2, \ldots, x_v, y_v)^\intercal$$

assuming here that each landmark is a 2-dimensional point representing a semantically meaningful part of the object, such as an eye corner in the human face.

Table 2: High-level facial features used in the proposed clustering approach and the related lower-level point features (Figure 2)

Semantic description	Related point features #
mouth vertical openness	{18, 21, 24, 25, 26, 27}
mouth horizontal openness	{18, 21}
lower lip to chin distance	{26, 27, 32, 33}
upper lip to nose distance	{15, 16, 17, 18, 21, 24, 25}
left eyebrow state	{0, 1, 2, 6, 8}
right eyebrow state	{3, 4, 5, 10, 12}
gap between eyebrows	{2, 3}

AAMs model shape deformation using a so-called point density model (PDM), which is a parametric linear subspace model learned statistically by principal component analysis (PCA) on a set of training shape examples. These examples are given as expert annotations of images of the object of interest, such as shown in Figure 2 for the human face. In such a representation, any shape $\mathbf{s}$ of the deformable object can be expressed by the generative model as a base shape $\mathbf{s}_0$ plus a linear combination of n shape vectors $\mathbf{s}_i$:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i$$

Registering a PDM to the image data then reduces to finding the optimal coefficient values p_i of this linear combination, i.e. the optimal PDM's parameters. AAMs propose to model the coupling between the PDM and the image data, i.e. the predictions on the PDM's landmarks locations given a target image, via a holistic appearance model of the pixel intensity values of the object's image. This appearance model is again a parametric linear subspace model, obtained by applying PCA to shape-normalized training example images of the object of interest. This shape normalization involves the warping of every example image to a reference frame, which is typically done by piecewise affine warping functions defined between each example shape and the base shape $\mathbf{s}_0$ of the PDM. The generative appearance model is then used to express any object's appearance $A(\mathbf{x})$ as a base appearance $A_0(\mathbf{x})$ plus a linear combination of m appearance images $A_i(\mathbf{x})$:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathcal{R}(\mathbf{s}_0)$$

where $\mathcal{R}(\mathbf{s}_0)$ denotes the set of pixel locations within the region defined by the base shape $\mathbf{s}_0$, i.e. the reference frame for the object's appearance.

Given these two generative models and following the so-called "independent" AAMs formulation proposed in [10],

registration can be seen as an image matching problem between the synthetic model image and the shape-normalized target image; the fitting goal can therefore be expressed as finding the parameters $\mathbf{p} = (p_1, p_2, \ldots, p_n)^\intercal$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^\intercal$ that minimize the following sum of squared differences:

$$\sum_{\mathbf{x} \in \mathcal{R}(\mathbf{s}_0)} \left[A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) - I(\mathbf{W}(\mathbf{x}; \mathbf{p})) \right]^2$$

where I is the target image and $\mathbf{W}(\mathbf{x}; \mathbf{p})$ is a (piecewise affine) warping function which projects a pixel location $\mathbf{x}$ from the reference frame to the target image frame, depending on the PDM's parameters $\mathbf{p}$. The minimization of this quantity is non-linear in the parameters $\mathbf{p}$ and must be solved iteratively by linear approximation, typically using the Gauss-Newton algorithm.

Variants met in the AAM-related literature mostly differ in the way they parameterize this linear approximation to derive the parameters update equation. In this work, we chose to use the efficient version of the simultaneous inverse-compositional AAM (SICAAM) proposed in [6]. This variant is more robust than others to large variations in shape and appearance, which typically occur when dealing with facial expressions in the context of sign language. Moreover, in order to cope with large off-plane head rotations, which are also common in sign language and can lead a 2D AAM to failure, we used the refinement proposed in [20]. In this work, a 3D PDM is estimated using a non-rigid structure-from-motion algorithm on the training shapes, and is then involved in the optimization process which incorporates a regularization term encouraging the 2D shape controlled by the 2D PDM to be a valid projection of the 3D PDM. Similar to the 2D PDM, the 3D PDM expresses any 3D shape $\mathbf{S}$ as a 3D base shape $\mathbf{S}_0$ plus a linear combination of $\bar{n}$ 3D shape vectors $\mathbf{S}_i$:

$$\mathbf{S} = \mathbf{S}_0 + \sum_{i=1}^{\bar{n}} \bar{p}_i \mathbf{S}_i$$

Notice that the 3D PDM is also involved in the calculation of the high-level facial features described below.

The procedure for the production of the high-level facial features includes a training stage:

1. Extrude the set of 2D training shape examples to 3D by means of the 3D PDM.
2. Remove global translations and rotations by aligning every extruded shape to the base shape $\mathbf{S}_0$ of the 3D PDM.
3. Project the aligned extruded shapes to 2D and, for each, estimate local area-based measurements corresponding to the point features subsets given in Table 2.
4. For each point features subset, store as the training output the minimum and maximum values of the corresponding local area-based measurements.

Extracting high-level facial features from the tracked lower-level point features is then done in the following way:

1. Extrude the registered shape and remove its global translation and rotation by means of the 3D PDM
2. Project the aligned extruded shape to 2D and, for each point features subset given in Table 2, estimate the corresponding local area-based measurement.
3. Normalize each local area-based measurement between 0 and 1 according to the minimum and maximum values obtained during training for the corresponding point features subset.
4. Each registered shape is then associated with a vector of D (in our work $D = 7$) continuous values in the range $[0, 1]$, corresponding to our high-level facial features.

Seven SICAAMs specific to the seven interpreters of RWTH-Phoenix-Weather have been trained for the end purpose of extracting high-level facial features from the gloss-annotated videos as shown in Figure 3. Training and tracking with one single SICAAM for all seven interpreters would have been a viable choice as well because of the enhanced robustness of this AAM variant to variability in identity. However, we wanted to obtain the best possible accuracy in the tracking of the low-level point features. On the other hand, the calculation of our high-level features is rather sensitive to identity changes and as such had to be designed in an identity-dependent fashion. The extraction of reliable identity-*independent* facial features similar to those used in this work is part of the advanced computer vision research topic known as "expression transfer" and is beyond the scope of this paper, where our primary goal is to give a proof of concept that gloss-based corpora can be enhanced by automatic face analysis methods.

Regardless of whether or not the extraction of high-level facial features is identity-independent, driving animation of an avatar's face can be done using our method's output. The grid of fitted AAM shape points shown in the top-left part of Figure 3 have known positions in 3-space, as one can see illustrated in the top-right part of the figure where the grid has been normalized in 3D to get a frontal pose. These accurate point positions along with the high-level features extracted from them (shown at the bottom of the figure) convey all the necessary information for modeling and transferring continuously facial expressions to an articulated avatar's face, using mapping techniques such as the ones proposed in [17] where a geometrical transfer matrix (from the deformable shape model to the avatar's control nodes) is combined with a higher-level, semantical transfer map.

6. EXPERIMENTS

As mentioned in Section 4, the RWTH-Phoenix-Weather corpus was mainly annotated for the use in statistical sign language recognition and translation. Since the sign language recognition research at the time focused on hand-based features, both the time boundaries of the glosses and their identity were labelled with regard to the hand parameters. For example, if two signs only differ in the mouthing, they share the same gloss label. In a text-to-avatar translation scenario, which would consist of a text-to-gloss translation and consecutive avatar animation based on the glosses,

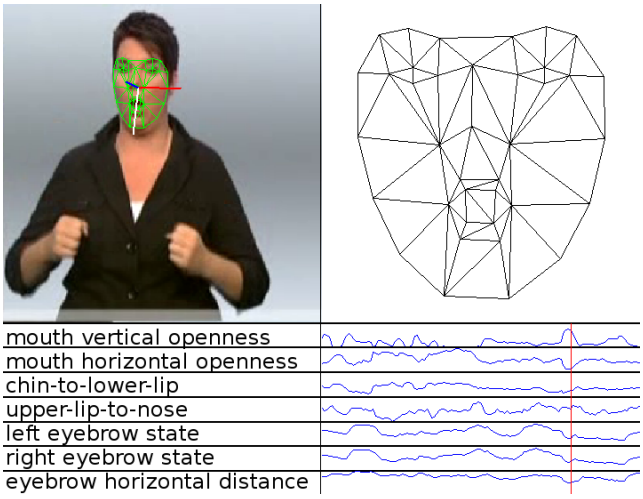


Figure 3: High-level feature extraction
Top left: the grid of fitted AAM points
Top right: rotated and normalized AAM points
Bottom: high-level feature values over time

this lack of detail with regard to facial features and mouthing implies that the resulting avatar animation could not produce these features, because the information is not contained in the annotation. A gloss would be animated with the same facial expression and mouthing irrespective of its context.

An optimal solution to this granularity problem would be the manual refinement of the annotation, but this process would be both time-consuming and expensive. Moreover, as one focuses on more and more aspects of sign language, the annotators would need to refine the annotation again and again.

In this work, we want to automate the refining process by providing the computer with facial features and performing an automatic clustering of the glosses based on these features. Moreover, for the task of text-to-avatar translation, the facial expressions and mouthing which accompany the signing are based on the source sentence of the spoken language, and consequently we use the source text information to select suitable facial features for the avatar animation.

Since the mouthings of a sign often mimic the words of the spoken language, providing the spoken word as a context can help to select a sign with a specific mouthing. We therefore align the glosses to the spoken language text in order to obtain the meaning of a gloss in a given context. We use the open-source toolkit GIZA++ to align each gloss to at most one word. This process leads to a set of (gloss,translation) pairs. For each instance of such a pair, we also have the corresponding video of the persons signing. To extract the facial features and mouthing for such a pair, we need to select a representative video from this set of videos. This leads to two problems. First, one (gloss,translation) pair might have several variants with respect to facial expression and mouthing. The variants might be caused by regional dialects or personal preferences. Second, some videos might be of a poor quality and not suitable for extracting features, e.g. if the mouth is occluded by the signing hands. To solve both problems, we cluster the videos with respect to to their AAM-based facial features.

Facial expressions and mouthings are seen as changing descriptors in a time series of images. We use the publicly available open source speech recognition system RASR [16] to model these sequences. This approach allows us to automatically calculate the degree of similarity between all gloss instances present in the data and store it in a global distance matrix.

We model each facial feature by a separate Hidden Markov model (HMM), which constitutes a stochastic finite state automaton. The number of states is chosen based on the actual frame length of the original feature sequence. Coarticulation effects are accounted for by a single state garbage model which can optionally be inserted at the beginning or end of a sequence. Single Gaussian densities, a globally pooled covariance matrix, global state transition penalties and the EM-algorithm with Viterbi approximation and maximum likelihood criterion are employed for training the models in a nearest neighbor fashion. The free HMM parameters, such as the time distortion penalties, are optimized in an unsupervised manner using the German translation as weak labels.

A trained set of HMMs is then used to calculate the distance between all pairs of gloss instances. By using an adaptive medoid-shift algorithm, we find several modes based on the distances. These modes are calculated for a given German context. By selecting the biggest cluster, we avoid outliers which are separated into smaller clusters. Moreover, we select the medoid of the biggest cluster to obtain a video which is representative of the whole cluster. The facial features of this video can then be used to drive the animation of an avatar system.

As described in Section 4, we labelled a subset of the glosses to evaluate the quality of the clustering algorithm. Moreover, we also want to evaluate the quality of the medoid by checking whether the medoid, i.e. the representative video, has the same mouthing as the glosses in the same cluster.

The external evaluation results of the clustering algorithm can be seen in Figure 4. The plots show the distribution of precision, recall and f-measure between the clusters provided by the algorithm and the hand-labelled mouthings for each (gloss,translation) pair. On the average, about two thirds of the (gloss,translation) pairs are correctly classified.

Besides the quality of the clustering, we are mainly interested in whether the adaptive medoid-shift algorithm selects a good representative video. For this, we also labelled the medoids resulting from the above clustering. The accuracy of the selected medoids is the fraction of the labelled data which has the same label as the medoid of the cluster they are in. The distribution of the accuracy is presented in Figure 5. On average, the algorithm has an accuracy of 78.4%, which means that in about four of five cases, the algorithm selects a good representative facial expression or mouthing.

Figure 6 shows two image sequences extracted from the corpus. The upper sequence shows the sign “Allgäu” (a hilly

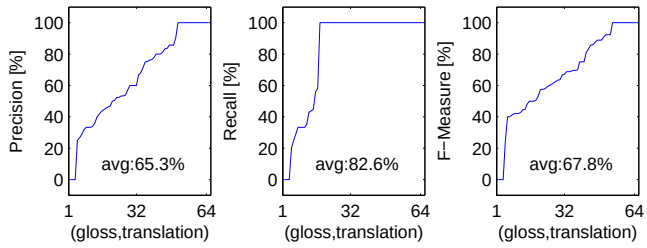


Figure 4: External evaluation of the clustering with respect to the labelled data

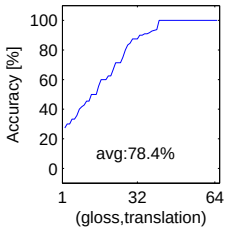


Figure 5: Accuracy of the selected medoid with respect to the labelled data

region in southern Germany) in which the hands perform the sign for mountain and the word "Allgäu" is mouthed. The lower image sequence in the same figure shows the medoid of the cluster the upper sign was placed into. The example shows that the algorithm is able to recognize similar mouthings between different signers even if they sign and mouth at different speeds. In the overall text-to-avatar pipeline, the word "Allgäu" would be translated into the gloss MOUNTAIN, but the suitable mouthing "Allgäu" would be selected for avatar animation.

7. CONCLUSIONS / OUTLOOK

In this paper, we propose a method to automatically enhance a gloss-based corpus to provide facial expressions and mouthings for the use in avatar animation. We applied the method to the RWTH-Phoenix-Weather corpus and evaluated the clustering results using a set of hand-labelled data. The overall algorithm achieved a high accuracy of 78.4%. Since the clustering is an unsupervised method, the only additional data necessary to improve the corpus is a small set of labelled images to train the active appearance models. Thus, the method is a viable way to improve an existing corpus without the effort of additional manual annotations.

Both the features selected for this work and the evaluation mainly focused on mouthing variants present in the corpus. In the future, we want to extend this method to other features and aspects of sign languages. To improve the features for the recognition of mouthings, we want to extend the set of high level features to recognize for example different tongue positions. One issue we also want to address is the spreading of the mouthing which sometimes is not synchronous to the manual component of the signing.

8. REFERENCES

[1] B. Abboud, F. Davoine, and M. Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19(8):723–740, 2004.

[2] I. Ari, A. Uyar, and L. Akarun. Facial feature tracking and expression recognition for sign language. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*, pages 1–6, 2008.

[3] A. Asghar and N. I. Rao. Color image segmentation using multilevel clustering approach. In *Proceedings of the 2008 Digital Image Computing: Techniques and Applications*, DICTA '08, pages 519–524, Washington, DC, USA, 2008. IEEE Computer Society.

[4] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 300–305, June 1998.

[5] 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 *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May 2012.

[6] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093, 2005.

[7] M. Huenerfauth, P. Lu, and A. Rosenberg. Evaluating importance of facial expression in american sign language and pidgin signed english animations. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 99–106. ACM, 2011.

[8] T. Johnston. The lexical database of auslan (australian sign language). *Sign Language and Linguistics*, 4:145–169(25), 2001.

[9] M. Kipp, Q. Nguyen, A. Heloir, and S. Matthes. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 107–114. ACM, 2011.

[10] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[11] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, Mar. 2003.

[12] J. Piater, T. Hoyoux, and W. Du. Video analysis for continuous sign language recognition. In *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 22–23, 2010.

[13] E. Pizzuto, P. Rossini, and T. Russo. Representing signed languages in written form: questions that need to be posed. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages: "lexicographic matters and didactic scenarios", International Conference on Language Resources and Evaluation LREC 2006*, pages 1–6. Genoa, Italy - 28th May 2006, 2006.

[14] S. Prillwitz. *HamNoSys. Version 2.0; Hamburg Notation System for Sign Language. An Introductory Guide*. Signum Verlag, Hamburg, Germany, 1989.

Figure 6: Comparison of a sign labelled with the mouthing "Allgäu" and the medoid of the biggest cluster

[15] I. Rodomagoulakis, S. Theodorakis, V. Pitsikalis, and P. Maragos. Experiments on global and local active appearance models for analysis of sign language facial expressions. In *9th International Gesture Workshop on Gestures in Embodied Communication and Human-Computer Interaction*, pages 96–99, 2011.

[16] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. The RWTH aachen university open source speech recognition system. In *10th Annual Conference of the International Speech Communication Association*, pages 2111–2114, 2009.

[17] J. M. Saragih, S. Lucey, and J. F. Cohn. Real-time avatar animation from a single image. In *Proc. Automatic Face & Gesture Recognition and Workshops*, pages 117–124, 2011.

[18] D. Stein, C. Schmidt, and H. Ney. Analysis, preparation, and optimization of statistical sign language machine translation. *Machine Translation*, 26(4):325–357, Dec. 2012.

[19] 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.

[20] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+ 3d active appearance models. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages II–535, 2004.