Centroid-Based Exemplar Selection of ASL Non-Manual Expressions using Multidimensional Dynamic Time Warping and MPEG4 Features

Hernisa Kacorri[1], Ali Raza Syed[1], Matt Huenerfauth[2], Carol Neidle[3]
[1]The Graduate Center, CUNY, Computer Science Program, New York, NY, USA
[2]Rochester Institute of Technology, Golisano College of Computing & Information Sciences, Rochester, NY, USA
[3]Boston University, Linguistics Program, Boston, MA, USA
{hkacorri,asyed2}@gradcenter.cuny.edu, matt.huenerfauth@rit.edu, carol@bu.edu

Abstract

We investigate a method for selecting recordings of human face and head movements from a sign language corpus to serve as a basis for generating animations of novel sentences of American Sign Language (ASL). Drawing from a collection of recordings that have been categorized into various types of non-manual expressions (NMEs), we define a method for selecting an exemplar recording of a given type using a centroid-based selection procedure, using multivariate dynamic time warping (DTW) as the distance function. Through intra- and inter-signer methods of evaluation, we demonstrate the efficacy of this technique, and we note useful potential for the DTW visualizations generated in this study for linguistic researchers collecting and analyzing sign language corpora.

Keywords: American Sign Language, Non-Manual Expressions, Dynamic Time Warping, Exemplar Selection, Animation Synthesis

1. Introduction

Technology to partially automate the process of producing animations of a virtual human character producing American Sign Language (ASL) could make it easier and more cost-effective for organizations to provide sign language content on websites. As compared to providing videos of ASL, animations that are automatically synthesized from a symbolic specification of the message would be easier to update and maintain, as discussed in (Huenerfauth, 2004; 2008). In this study, we examine whether multidimensional dynamic time warping (DTW) is suitable for evaluating the similarity of recordings of face and head movements of ASL non-manual expressions (NMEs).

For the purposes of generating animations of ASL, given a set of recordings of human face movements for ASL NMEs, it is valuable to identify an exemplar recording that could be used as the basis for generating the movements of virtual human character, to produce an understandable ASL animation containing an NME. The goal of the current study is to evaluate the potential of centroid-based exemplar selection for ASL NMEs. Specifically, we are investigating whether, given a set of recordings of humans producing some category of ASL NME, a multidimensional DTW metric can serve as the basis for selecting a "centroid," a member of the set with the minimum cumulative pairwise distance from the other members, such that this centroid serves as a representative exemplar of the set. Given idiosyncratic differences among individual productions, which are naturally found in any collection of recordings of ASL NMEs, it is reasonable to expect that some recordings may be outliers, and others may be more similar to the other items in the set, where "outlier" is defined as a member of the set with the maximum cumulative pairwise distance from the other members.

In this paper, we present an algorithm for selecting a centroid exemplar of a sign language NME from a set of human recordings (and for comparison sake, we also define a method for determining the maximal outlier in such a set of recordings). Through various forms of intra-signer and inter-signer comparison, we evaluate the effectiveness of this technique for identifying an NME performance that is typical of a specific linguistic type. In addition, we present several forms of data visualization based on this algorithm, some of which may be useful for linguistic researchers who are collecting or analyzing sign language corpora.

2. Background and Related Work

An ASL production consists of movements of the eyes, face, head, torso, arms, and hands; in fact, the movements of the face and head, during non-manual expressions (NMEs), are essential to the meaning of utterances. For example, these expressions can convey grammatical information about individual words or entire phrases or clauses during the utterance. The upper face and head movements of these NMEs occur in parallel with phrases containing manual signs (Neidle et al., 2000).

In current collaborative work involving RIT, Boston University, and Rutgers University, we are videorecording and annotating a set of human ASL productions, including several categories of syntactic NMEs. These recordings include markers of wh-questions, yes/no-questions, rhetorical questions, negation, and topics. This dataset has been essential for research on automatic ASL recognition technologies (Neidle et al., 2014), and RIT researchers are using this dataset for research on ASL animation synthesis. These recordings serve as the source of human movement data for the methods presented in section 3.

Several groups of researchers internationally have investigated how to generate animations of sign language that include linguistically meaningful facial expressions; we have compared and surveyed their methods and contributions in (Kacorri, 2015). In the work most closely related to the methods discussed in section 3, researchers have investigated the potential of clustering and centroid selection for identifying variants of German Sign Language (DGS) lexical items, based on the co-occurring lexical NMEs, primarily given differences in mouthing (Schmidt et al., 2013). However, the potential for centroid-based exemplar selection for syntactic NMEs (consisting of head and upper face movements and spanning over one or more glosses) has not been previously investigated.

2.1 Dynamic Time Warping (DTW)

DTW is utilized as the distance metric for the methods in section 3. DTW is a methodology commonly used to evaluate similarity among time-series data, e.g. (Sakoe and Chiba, 1978), and has been previously adopted to evaluate the similarity of animated characters' facial expressions (Mana and Pianesi, 2006; Ouhyoung et al., 2012).

The rate at which ASL NMEs change may vary by signer or context. For example, in Figure 1, the head roll movement in Recording 2 peaks first, then falls and rises faster than the movements in Recording 1. Comparing the two series requires alignment to match points by, for example, finding corresponding points between peaks and valleys as well as the rising and falling portions of the two series. The two curves can be aligned through a nonlinear stretching and compressing the time axis ("warping"). The distance between the two series is taken to be the sum of distances between the matching points in the warped time domain. DTW yields an optimal alignment by using dynamic programming to determine a warping function that minimizes the total distance between the series.

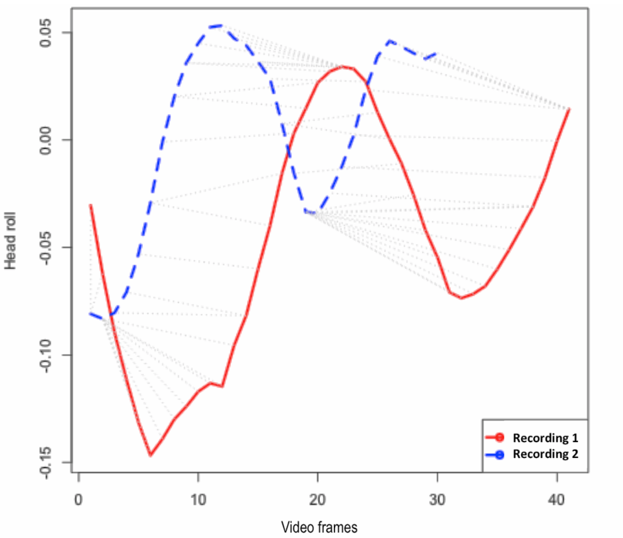

Figure 1: Example of DTW alignment between the "head roll" values detected during two Negative NMEs extracted from human ASL recordings in our dataset.

In a prior study (Kacorri and Huenerfauth, 2015), we demonstrated the potential of DTW for evaluating the similarity of face and head movements for ASL NMEs. We examined whether the judgments of native ASL signers about the quality of animations correlated with DTW-based similarity between that animation and a gold-standard human recording that had been analyzed to yield a stream of MPEG4 Facial Action Parameters (FAPs) (Pandzic and Forchheimer, 2003), representing the movements of the face. We found a significant correlation, indicating that DTW may be useful for automatically evaluating the similarity of ASL NMEs. In that work, we used the normalized distance from a multidimensional variant of DTW (Giorgino, 2009) on the parallel streams of MPEG4 values. This multidimensional DTW is also used in the centroid exemplar-selection methods in section 3.

3. Centroid-Based Exemplar Selection

This section describes our method for selecting an exemplar recording of a human ASL NME from a dataset containing multiple examples of each type of NME. As discussed in section 3.2, this method operates by identifying a centroid item within the set, using normalized multivariate DTW as the distance function.

3.1 NME Dataset

Our data for this study consists of the recordings of human ASL productions that were collected and analyzed at Boston University, as discussed in section 2. We analyzed 173 of these annotated video recordings of a female ASL native signer using an MPEG-4 face tracker (Visage Technologies, 2016), and we extracted the head pose and MPEG-4 facial features for each video frame. Since some recordings contained more than one NME, our dataset included a total of 199 multivariate time series of syntactic NMEs distributed in the following categories: wh-questions (14), yes/no-questions (21), rhetorical questions (13), negation (55), and topics (96). As shown in Table 1, there was a high variability across recordings in video length and number of manual glosses (individual signs performed on the hands) that occurred in parallel with each NME.

Given that the NME categories we are investigating mostly involve head and upper face movements, e.g. (Neidle et al., 2000), for this study we are interested only in a subset (a total of 9) of the extracted features by Visage face tracker, which includes:

- **Head orientation** (FAP48-FAP50): orientation parameters defined as pitch, yaw, and roll.
- **Eyebrow vertical displacement** (FAP31-FAP36): 6 parameters describing vertical movements of the inner, middle, and outer points of the left and right eyebrow.

NME Dataset	**Video Frames**	**Num. of Glosses**
(Num. of examples)	min – max (mean)	min – max (mean)
Topic (96)	5 – 54 (15.5)	1 – 4 (1.43)
Negation (55)	10 – 76 (38.1)	2 – 7 (3.56)
Y/N-question (21)	9 – 78 (34.6)	2 – 6 (3.6)
Wh-question (14)	15 – 69 (31.2)	1 – 5 (2.2)
Rhetorical (13)	11 – 46 (28.3)	1 – 4 (3.0)

Table 1: NME Dataset Characteristics.

3.2 Selecting Centroid and Outlier

We used multivariate DTW (Giorgino, 2009) to obtain the normalized distances between all pairs of recordings in the same NME category. Within each NME category, we labeled one recording as the "centroid" and one recording as the "outlier", which were defined in the following way:

$$\text{centroid} = \arg\min_{u \in S} \sum_{v \in S} \text{DTW}(u, v)$$

$$\text{outlier} = \arg\max_{u \in S} \sum_{v \in S} \text{DTW}(u, v)$$

S is the set of all recordings within an NME category. Thus "centroid" and "outlier" are the recordings with the minimum and maximum cumulative DTW distance, respectively, to all other recordings within a category. The centroid recording is the most representative example from a given category since it characterizes the central (median) tendency of recordings within that category. Conversely, the outlier characterizes the least representative recording in a given category.

As suggested by prior work, e.g. (Gillian et al., 2011), preprocessing is necessary for DTW if either (a) the source range of the N-dimensional data varies or (b) if invariance to spatial variability and variability of signal magnitude is desired. To address the first case, we scaled all the features to the range [-1, 1] by dividing by the largest maximum value in each feature. For DTW analysis, such scaling is suitable for data that is already centered at zero, which is the case for our extracted MPEG-4 data, since 0 denotes a neutral pose for each feature. To address the second case, we performed z-normalization, so that the time-series for each FAP would have a zero mean and unit variance. Since DTW only performs alignment in the time dimension, series with very different amplitudes may not allow for proper comparisons when using the DTW similarity measure. Normalizing the amplitude values brings all series to a standard scale and allows for better similarity measures to be determined.

Table 2 provides details about the centroid and outlier recordings that were identified using the above procedure. The table includes information about the length of each NME, as indicated by the number of video frames in duration and the number of manual glosses that occur in parallel with each NME.

NME Category	Centroid		Outlier	
	v. frames	glosses	v. frames	glosses
Topic	19	2	13	2
Negation	13	3	30	3
Y/N-question	39	4	20	2
Wh-question	31	2	15	1
Rhetorical	12	1	40	4

Table 2: Number of frames and glosses for the centroids and outliers that were selected.

For example, for the "topic" category of NME, the recording that was selected by the algorithm as the centroid contained the utterance: "CAR BREAK-DOWN WAVE-NO," in which the topic NME occurred during the two glosses "CAR BREAK-DOWN."

3.3 Visualizing Centroids Versus Outliers

Figure 2a visualizes the DTW distance among all pairs of recordings in the set of "topic" ASL NME recordings. Each node in the graph represents a recording, and each edge, the DTW distance between the nodes. Nodes are numbered based on their listing in the dataset. Lighter colors for nodes and edges denote smaller DTW distances, thus more similar time-series values. The graphs were produced using the Python package NetworkX (Hagberg et al., 2008) with the Fruchterman Reingold layout and the Viridis color-map. Since the software default layout locates the nodes with the highest degree in the center, the input for the algorithm was the DTW distance matrix for a comparison set where each of the DTW distances is replaced with its absolute difference with the max distance. Thus the node in the center is the centroid with smallest total DTW distance to its neighbors. Figure 2b represents the cumulative distances for each node, with arrows indicating the centroid and outlier for this dataset.

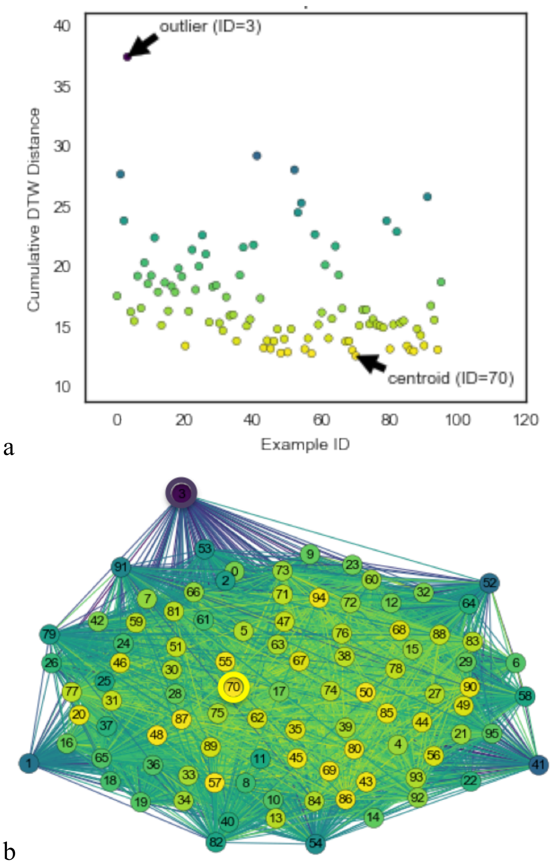

Figure 2: Visualization of (a) cumulative DTW-distances and (b) DTW-distance graph for the video recordings in the ASL Topic NME dataset. The centroid and the outlier are indicated in both graphs, by the brightest and the darkest color, respectively.

In Figure 2, the reader should note that the node with the brightest yellow color and most central location of the graph image visually indicates the "centroid" of this group, and the node with the darkest color and most remote location in the graph is the "outlier." In addition, to aid visibility, the centroid and outlier have been surrounded by a thick outline in Figure 2b.

4. Initial Confirmation of Centroid Quality Using Intra-Signer Data

As a preliminary assessment of the validity of the above procedure, we compared the "centroid" and the "outlier" we identified, using a methodology inspired by prior DTW research on template-based classification (Gillian et al., 2001). After identifying the centroid and the outlier for each ASL NME dataset, we constructed a classifier for assigning a label (wh-question, yes/no-question, rhetorical question, negation, or topic) to a given recording of unknown NME category. We treated the centroid for each ASL NME category (topic, negation, etc.) as an "exemplar" of that category. To classify some given recording of unknown category, we compared its distance to each of the five exemplars. The unknown recording was labeled with the category of the exemplar to which it had the minimum DTW distance. For sake of comparison, we also constructed a second classifier based on using the "outliers" identified for each NME category as if they were an exemplar of that category.

To evaluate these two classifiers, we removed the five centroids and the five outliers from the five NME datasets. Then, we used each of our two classifiers to assign a label to each of the recordings, and we calculated the accuracy of each classifier at this task. The results are shown in Figure 3.

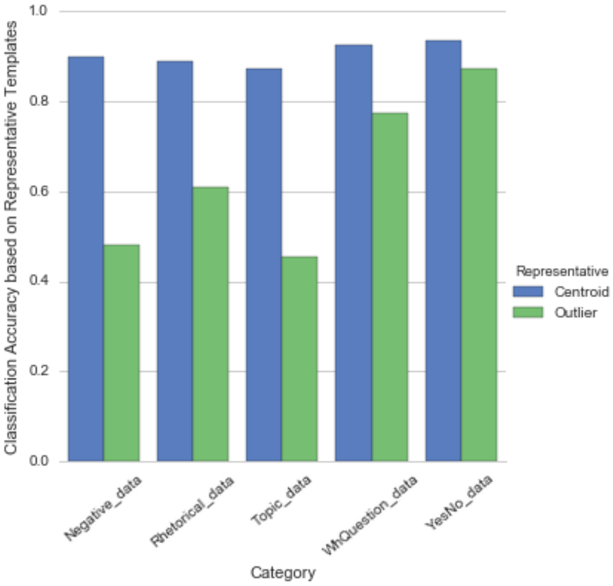

Figure 3: Accuracy results when using centroid versus outliers as exemplars for classification.

While we are not primarily interested in building an NME classifier in this paper, this form of comparison allows us to see how well the "centroids" identified using our above procedure are serving as representatives of each category of NME. We found that when the centroids are used as exemplars, a higher accuracy is achieved for all the NME types.

5. Inter-Signer Evaluation

While the preliminary assessment above suggested that the centroids were serving as effective exemplars of our ASL NME categories, the goal of this study is to determine whether our new centroid selection method, based on multidimensional DTW and MPEG4 facial features, would be useful for selecting an exemplar of human performance that could serve as the basis for animating a virtual human character. We note two key challenges in such a usage scenario:

- A virtual human signer may have different facial proportions than the human in the datasets.
- The specific sentence we wish to synthesize as animation may not have been performed by the human in the original ASL NME datasets.

Thus, we conducted a more rigorous form of evaluation to determine whether these centroids would be effective exemplars when considering data from a different signer performing a different sentence. Specifically, we compared the centroid and the outlier for each ASL NME category to two "gold standard" recordings of an ASL performance from a male native ASL signer performing the same category (topic, wh-question, etc.) of ASL NME.

Notably, the (male) human in the gold standard recordings is different from the (female) human in the recordings in the dataset used as the basis for centroid and outlier selection. Furthermore, the specific sentences used in the gold standard recordings did not appear in the original data set. Thus, this inter-signer evaluation is a more rigorous method for evaluating whether the centroid recordings identified in our original ASL NME datasets would be effective for animation synthesis.

NME Dataset	Example 1		Example 2	
	v. frames	glosses	v. frames	glosses
Topic	55	2	31	1
Negation	63	4	33	2
Y/N-question	73	5	45	2
Wh-question	35	1	55	3
Rhetorical	25	1	102	4

Table 3: Number of video frames and glosses for the two examples per NME that serve as gold standards.

The source of these gold standard recordings is the collection of ASL videos that we previously released to the research community as a standard evaluation dataset in (Huenerfauth and Kacorri, 2014). In that paper, we defined codenames for the individual recordings in that dataset; using that nomenclature, the recordings selected as gold standards were: W1 and W2 (wh-question), Y3 and Y4 (yes/no-question), N2 and N5 (negation), T3 and T4 (topic), and R3 and R9 (rhetorical question).

Figure 4 presents the difference between our gold standards and (a) the centroid selected using the method in section 3 or (b) the outlier selected using the method in section 3. Since we identified two gold standard recordings, the height of the bars indicates the average of the DTW distance from both of the gold standard recordings. As shown in the figure, the centroid outperforms the outlier.

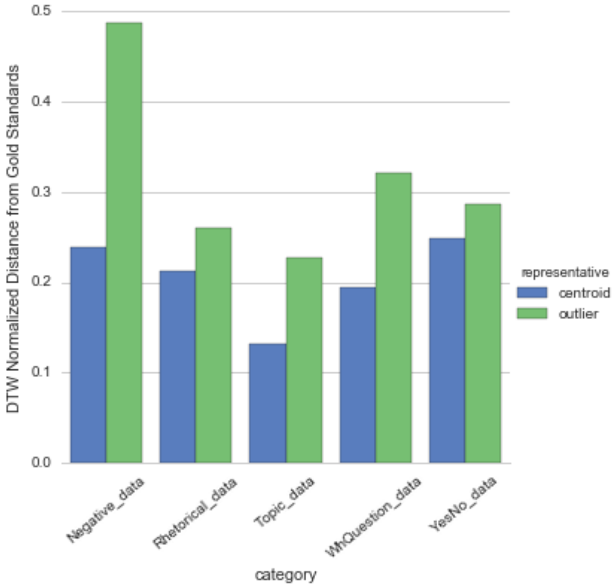

Figure 4: Average DTW distance of the centroid and outlier from two gold-standard recordings of a different signer performing ASL sentences not found in the original datasets, with a smaller distance indicating higher similarity.

6. Conclusions and Future Work

The results of this study indicate that centroid-based exemplar selection using multidimensional DTW operating in the space of MPEG4 FAP recordings is a viable method for selecting exemplar recordings of ASL NMEs. We have demonstrated a methodology for selecting human recordings that can serve as a basis for later stages of sign language animation synthesis. In so doing, this study has investigated a new methodology for utilizing sign language corpus data to advance research in sign language avatar technology, specifically for the selection of NME movements.

While this paper presented our results from analyzing nine MPEG4 facial action parameters (representing eyebrow height and head orientation), we plan to further investigate the utility of modeling additional facial parameters for components of ASL NMEs , such as brow furrowing or eyelid aperture. In addition, this paper presented a preliminary evaluation of the effectiveness of this exemplar selection algorithm. In future work, we intend to conduct a user study to evaluate the quality of animations of sign language generated using this technique, with ASL signers evaluating the animations.

Finally, we note an application of this research for researchers who are seeking to mine a sign language corpus. With the growth of sign language corpora, one challenge faced by linguistic researchers is visualizing patterns in this data to support hypothesis development. We note that the visualizations of "distance" shown in Figure 2 may have potential for assisting linguistic researchers in exploring the NMEs within recordings in a sign language corpus. For instance, graphs or plots of multidimensional DTW similarity of MPEG4 facial features may suggest neighbors, outliers, or clusters of similar recordings in a corpus. In fact, as investigated in (Schmidt et al., 2003), automated statistical clustering techniques could be used to identify variants within a set of recordings, or these graph-like visualizations could support discovery by linguistic researchers. Such graphs might suggest the existence sub-variants in a set of recordings of some ASL NME, which could be investigated in further linguistic work.

Furthermore, the type of visualizations in Figure 2 may be useful by researchers who are collecting sign language corpora so that they may quickly visualize the diversity of their collection during the recording and annotation process; this may indicate how diverse the corpus is. Therefore, such visualizations may help researchers to determine, during the collection of a corpus, whether further data are needed.

7. Acknowledgements

This material is based upon work supported by the National Science Foundation under award number 1059218, 1065009, 1065013, and 1506786. This material is also based upon work supported by the Science Fellowship and Dissertation Fellowship programs of The Graduate Center, CUNY. We are grateful for support from Dimitris Metaxas at Rutgers University.

8. Bibliographical References

Gillian, N., Knapp, R.B., O'Modhrain, S. (2011). Recognition of multivariate temporal musical gestures using n-dimensional dynamic time warping. In *Proc of the 11th Int'l conference on New Interfaces for Musical Expression.*

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the DTW package. *Journal of Statistical Software* 31(7), pp. 1–24.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proc of the 7th Python in Science Conference*, Pasadena, CA, pp. 11–15.

Huenerfauth, M. (2004). Spatial and planning models of ASL classifier predicates for machine translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI.*

Huenerfauth, M. (2008). Generating American Sign Language animation: overcoming misconceptions and technical challenges. *Univ. Access. Inf. Soc.* 6(4), pp. 419–434.

Huenerfauth, M, Kacorri, H. (2014). Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language. In *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, LREC 2014,* Reykjavik, Iceland.

Kacorri, H. (2015). TR-2015001: A survey and critique of facial expression synthesis in sign language animation. *Computer Science Technical Reports*. Paper 403. The Graduate Center, CUNY, New York, NY, USA.

Kacorri, H., Huenerfauth, M. (2015). Evaluating a dynamic time warping based scoring algorithm for facial expressions in ASL animations. In *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), INTERSPEECH 2015*, Dresden, Germany.

Mana, N., Pianesi, F. (2006). HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads. In *Proc. of the 8th Int'l Conf on Multimodal Interfaces*, pp. 380–387.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., Lee R.G. (2000). *The Syntax of American Sign Language: Functional categories and hierarchical structure*, Cambridge, MA: The MIT Press.

Neidle, C., Liu, J., Liu, B., Peng, X., Vogler, C., Metaxas, D. (2014). Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in American Sign Language (ASL). In *Proc. 6th workshop on representation and processing of sign languages, LREC 2014.*

Ouhyoung, M., Lin, H.S., Wu, Y.T., Cheng, Y.S., Seifert, D. (2012). Unconventional approaches for facial animation and tracking. In *SIGGRAPH Asia*, pp. 24.

Pandzic, I. S., and Forchheimer, R. (2003). *MPEG-4 facial animation: the standard, implementation and applications*. Wiley.

Sakoe, H., Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing* 26(1), pp. 43–49.

Schmidt, C., Koller, O., Ney, H., Hoyoux, T., Piater, J. (2013). Enhancing gloss-based corpora with facial features using active appearance models. In *3rd Int'l Symp on Sign Language Translation and Avatar Technology (SLTAT)*. Chicago, IL, USA.

Visage Technologies. (2016). FaceTrack. Retrieved January 29, 2016, from http://www.visagetechnologies.com/products-and-services/visagesdk/facetrack/.