Automatic Recognition of Head Movement Gestures in Sign Language Sentences

Daniel Kelly, Jane Reilly Delannoy, John Mc Donald, Charles Markham
Computer Science Department, National University of Ireland, Maynooth, Ireland
dankelly@cs.nuim.ie

***Abstract*—A novel system for the recognition of head movement gestures used to convey non-manual information in sign language is presented. We propose a framework for recognizing a set of head movement gestures and identifying head movements outside of this set. Experiments show our proposed system is capable of classifying three different head movement gestures and identifying 15 other head movements as movements which are outside of the training set. In this paper we perform experiments to investigate the best feature vectors for discriminating between positive a negative head movement gestures and a ROC analysis of the systems classifications performance showed an area under the curve measurement of 0.936 for the best performing feature vector.**

***Keywords*-Sign Language, Non Manual Signals, HMM**

I. Introduction

Sign Language is a form of non-verbal communication where information is mainly conveyed through hand gestures. Since sign language communication is multimodal, it involves not only hand gestures (i.e., manual signing) but also non-manual signals. Non-manual signal are conveyed through facial expressions, head movements, body postures and torso movements. Recognizing Sign Language communication therefore requires simultaneous observation of manual and non-manual signals and their precise synchronization and signal integration. Thus understanding sign language involves research in areas of face tracking, facial expression recognition, human motion analysis and gesture recognition.

Over the past number of years there has been a significant amount of research investigating each of these non-manual signals attempting to quantify their individual importance. Works such as [1], [2], [3] focused on the role of head pose and body movement in sign language. These researchers found evidence which strongly linked head tilts and forwards movements to questions, or affirmations. The analysis of facial expressions for the interpretation of sign language has also received a significant amount of interest [4], [5]. Computer-based approaches which model facial movement using *Active Appearance Models* (AAMs) have been proposed [6], [7], [8].

The development of a system combining manual and non-manual signals is a non-trivial task [9]. This is demonstrated by the limited amount of work dealing with the recognition of multimodal communication channels in sign language. Ma et al [10] used Hidden Markov Models (HMMs) to model multimodal information in sign language, but lip motion was the only non-manual signal used. Their work was based on the assumption that the information portrayed by the lip movement directly coincided with that of the manual signs. While this is a valid assumption for mouthing, it cannot be generalised to other non-manual signals as they often span multiple manual signs and thus should be treated independently.

In this paper we evaluate techniques for the automatic recognition head movement gestures used to convey non-manual information in Irish Sign Language (ISL) sentences. We propose a framework for the automatic recognition of head movement gestures, building on the techniques proposed by Kelly et al [11] who use a HMM threshold model system to recognize manual signals.

II. Feature Extraction

The focus of this work is to evaluate the HMM threshold model framework as a system for recognizing head movement gestures. For completeness, we briefly describe the feature tracking techniques used, though we do not consider it to be the novel part of our work.

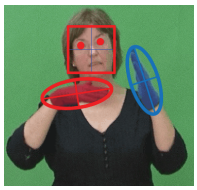

Figure 1. Extracted Features from Image

Face and eye positions are used as features for head movement recognition. Face and eye detection is carried out using a cascade of boosted classifiers working with haar-like features proposed by Viola and Jones [12]. A set of public domain classifiers [13], for the face, left eye and right eye, are used in conjunction with the OpenCV implementation of the haar cascade object detection algorithm. We define the raw features extracted from each image as follows; face position (FC_x, FC_y), left eye position (LE_x, LE_y) and right eye position (RE_x, RE_y).

III. Hidden Markov Models

Hidden Markov Models (HMMs) are a type of statistical model and can model spatiotemporal information in a natural way. HMMs have efficient algorithms for learning and recognition, such as the Baum-Welch algorithm and Viterbi

search algorithm [14]. A HMM is a collection of states connected by transitions. Each transition (or time step) has a pair of probabilities: a transition probability (the probability of taking a particular transition to a particular state) and an output probability (the probability of emitting a particular output symbol from a given state). We use the compact notation $\lambda = \{A, B, \pi\}$ to indicate the complete parameter set of the model where A is a matrix storing transitions probabilities and a_{ij} denotes the probability of making a transition between states s_i and s_j. B is a matrix storing output probabilities for each state and π is a vector storing initial state probabilities. HMMs can use either a set of discrete observation symbols or they can be extended for continuous observations signals. In this work we use continuous multidimensional observation probabilities calculated from a multivariate probability density function.

To represent a gesture sequence such that it can be modeled by a HMM, the gesture sequence must be defined as a set of observations. An observation O_t, is defined as an observation vector made at time t, where $O_t = \{o_1, o_2, ..., o_M\}$ and M is the dimension of the observation vector. A particular gesture sequence is then defined as $\Theta = \{O_1, O_2, ..., O_T\}$. To calculate the probability of a specific observation O_t, we implement probability density function of an M-dimensional multivariate gaussian (see Equation 1). Where μ is the mean vector and Σ is the covariance matrix.

$$\aleph(O_t; \mu, \Sigma) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} exp\big(-\tfrac{1}{2}(O_t-\mu)^T \Sigma^{-1}(O_t-\mu)\big) \quad (1)$$

A. HMM Threshold Model

Lee and Kim [15] proposed a HMM threshold model to handle non-gesture patterns. The threshold model was implemented to calculate the likelihood threshold of an input pattern and provide a confirmation mechanism for provisionally matched gesture patterns. We build on this work carried out by Lee and Kim to create a framework for calculating a probability distribution of head movement input gesture using continuous multidimensional observations. The computed probability distribution will include probability estimates for each pre-trained sign as well as a probability estimate that the input sign is a non head movement gesture.

In general, a HMM recognition system will choose a model with the best likelihood as the recognized gesture if the likelihood is higher than a predefined threshold. However, this simple likelihood threshold often does not work, thus, Lee and Kim proposed a dynamic threshold model to define the threshold of a given gesture sequence.

A property of the left-right HMM model implies that a self transition of a state represents a particular segment of a target gesture and the outgoing state transition represents a sequential progression of the segments within a gesture sequence. With this property in mind, an ergodic model, with the states copied from all gesture models in the system, can be constructed as shown in Figure III-A and III-A, where dotted lines in Figure III-A denote null transitions (i.e. no observations occur between transitions).

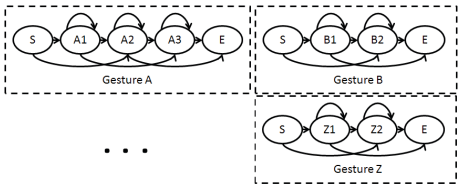

Figure 2. Dedicated Gesture Models

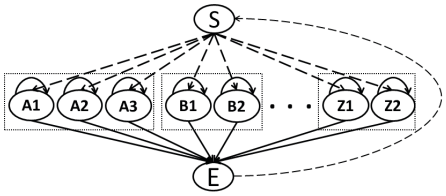

Figure 3. Threshold Model

States are copied such that output observation probabilities and self transition probabilities are kept the same, but all outgoing transition probabilities are equally assigned as defined in Equation 2 where N is the number of states excluding the start and end states (The start and end states produce no observations).

$$a_{ij} = \frac{1 - a_{ij}}{N - 1}, \quad \forall j, i \neq j, \quad (2)$$

As each state represents a subpattern of a pre-trained gesture, constructing the threshold model as an ergodic structure makes it match well with all patterns generated by combining any of the gesture sub-patterns in any order. The likelihood of the threshold model, given a valid gesture pattern, would be smaller than that of the dedicated gesture model because of the reduced outgoing transition probabilities. However, the likelihood of the threshold model, given an arbitrary combination of gesture sub-patterns, would be higher than that of any of the gesture models, thus the threshold model, denoted as $\overline{\lambda}$, can be used as a non head movement gesture measure.

B. HMM Threshold Model & Gesture Recognition

Kelly et al [11] expand on the work of Lee and Kim [15] to develop a HMM threshold model system which models continuous multidimensional sign language observations within a parallel HMM network to recognize two hand signs and identify movement epenthesis. In this paper, we expand on the work of Kelly et al to create a framework for recognizing head movement gestures.

For a network of HMMs $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_C\}$, where λ_c is a dedicated gesture HMM used to calculate the likelihood that the input gesture is belonging to gesture class c, a single

threshold model $\overline{\lambda}$ is created to calculate the likelihood threshold for each of the dedicated gesture HMMs.

IV. Non Manual Signal Recognition

While hand gestures do play central grammatical roles, movements of the head, torso and face are used to express certain aspects of ISL. In this work we will focus on a single non-manual signal, the head movement, to evaluate our techniques when recognizing non-manual features.

A. Model Training

Our system initializes and trains a dedicated HMM for each head movement gesture to be recognized. In this work we evaluate our techniques using three different head movement gestures; a left head movement, a right head movement and a left-forward movement. A visual example of a signer performing each of the three different head movement gesture is in shown in Figure IV-A.

Figure 4. Example of the three different head movement gestures the system was tested on (a) Right Movement (b) Left Movement (c) Left Forward Movement

To train the head movement HMMs, we recorded 18 different videos of a fluent ISL signer performing the head movements naturally within full sign language sentences. Six videos where recorded for each head movement gesture. Each head movement HMM λ_i^H (where $0 < i < I$ and I is the total number of head gestures) was then trained on the observation sequences extracted from the corresponding videos.

The start and end point of each of the head movement gestures were labeled, the observation sequences Θ_i were extracted and each HMM was then trained using the iterative HMM training model proposed by Kelly et al [11]. A HMM threshold model, $\overline{\lambda^H}$ is then created using the network of trained HMMs λ_i^H (where $0 < i < I$). The set of HMMs, to recognize the I pre-trained head movement gestures, is then denoted as $\Lambda^H = \{\lambda_1^H, \lambda_2^H, ..., \lambda_I^H, \overline{\lambda^H}\}$.

B. Head Movement Recognition

Given an unknown sequence of head movement observations Θ^H, the goal is to accurately classify the head movement as one of the I trained gestures or as a movement which is not a trained gesture. To classify the observations, the Viterbi algorithm is run on each model given the unknown observation sequences Θ^H, calculating the most likely state paths through each model i. The likelihoods of each state path, which we denote as $P(\Theta^H|\lambda_i^H)$, are also calculated. The sequence of observations can then be classified as i if Equation 3 evaluates to be true.

$$P(\Theta^H|\lambda_i^H) \geq \Psi_i^H \tag{3}$$

$$\Psi_i^H = P(\Theta^H|\overline{\lambda^H})\Gamma_i^H \tag{4}$$

Where Γ_i^H is a constant scalar value used to tune the sensitivity of the system to head movement which the system was not trained on.

C. Experiments

An accurate head movement gesture recognition system must be able to discriminate between positive and negative head movement gesture samples, therefore, we perform a set of experiments to find the best feature set when discriminating between isolated positive and negative head gestures.

To test the discriminative performance of different feature vectors, we recorded an additional 7 videos for each head gesture (21 in total), where a fluent ISL signer performed the head movement gestures within different sign language sentences. The start and end points of the head gestures were then labeled and isolated observation sequences Θ_i^τ were extracted. An additional set of 15 other head gesture sequence, outside of the training set, were also labeled in the video sequences to test the performance of the system when identifying negative gestures.

The classification of a gesture is based on a comparison of a weighted threshold model likelihood with the weight denoted as Γ_i^H. In our ROC analysis of the system, we vary the weight, Γ_i^H, over the range $0 \leq \Gamma_i^H \leq 1$ and then create a confusion matrix for each of the weights.

To evaluate the performance of different features, we performed a ROC analysis on the models generated from the different feature combinations and calculated the area under the curve (AUC) for each feature vector model. Table I shows the AUC measurement of four different features which were evaluated during our experiments. To calculate the directional vector of the head, (V_x^H, V_y^H), we used the mid point between the eyes and calculated the direction the midpoint moved from frame to frame. We used a sliding window to average the directional vector and in our experiments we evaluated the best performing window size for each feature vector. Although we evaluated each feature

vector with a range of different window sizes, we report only the best performing window sizes for each feature vector in Table I.

Table I
AUC MEASUREMENTS FOR DIFFERENT FEATURE COMBINATIONS

Features	Window Size	ROC AUC
F_1 - Unit Direction Vector $(\hat{V}_x^H, \hat{V}_y^H)$	6	0.821
F_2 **- Direction Vector** (V_x^H, V_y^H)	**12**	**0.936**
F_3 - Unit Direction Vector $(\hat{V}_x^H, \hat{V}_y^H)$ + Angle Eyes (θ_{eyes})	6	0.863
F_4 - Direction Vector (V_x^H, V_y^H) + Angle Eyes (θ_{eyes})	6	0.868

V. CONCLUSION

In this paper we have discussed current research in the area of automatic recognition of non-manual signals used in sign language. The development of a system to recognize non-manual signals is a non-trivial task and this is demonstrated by the limited number of works dealing with non-manual signals in the context of sign language sentences.

We have presented a framework for recognizing head movement gestures used to convey non-manual information in sign language sentences. We expanded the HMM threshold model technique, proposed by Lee and Kim [15], to develop a system which models continuous multidimensional head movement observations within a HMM network to recognize head movements and identify head movement gestures which the system was not trained on. We perform experiments to investigate possible observation vectors which best discriminate between positive and negative head movement gestures samples. A ROC analysis of different observation vectors showed that the best performing vector, with an AUC measurement of 0.936, was a two dimensional vector describing the movement of the eye midpoint within a sliding window averaged over 12 frames.

The significance of the research presented in this paper is that we have developed a general technique for recognising head movement gestures. With a view to developing an automatic sign language recognition system, identifying non-manual signals such as head movement is an important task. In this paper we have demonstrated that our techniques are capable of identifying typical head movement gestures which occur in sign language sentences, therefore enable us to determine whether or not a question was posed by the signer. Future work will involve incorporating these techniques into a wider framework for automatic recognition of multi-modal continuous sign language.

ACKNOWLEDGMENT

The Authors would like to acknowledge the financial support of the Irish Research Council for Science, Engineering and Technology (IRCSET).

REFERENCES

[1] B. Bahan, "Nonmanual realisation of agreement in american sign language," Ph.D. dissertation, University of California, Berkely, 1996.

[2] E. van der Kooij, O. Crasborn, and W. Emmerik, "Explaining prosodic body leans in sign language of the netherlands: Pragmatics required," *Journal of Pragmatics*, vol. 38, 2006, prosody and Pragmatics.

[3] C. Baker-Shenk, "Factors affecting the form of question signals in asl," *Diversity and Diachrony*, 1986.

[4] R. B. Grossman and J. Kegl, "To capture a face: A novel technique for the analysis and quantification of facial expressions in american sign language," pp. p273–305, 2006.

[5] R. Grossman and J. Kegl, "Moving faces: Categorization of dynamic facial expressions in american sign language by deaf and hearing participants," *Journal of Nonverbal Behavior*, vol. 31, no. 1, pp. 23–38, 2007.

[6] U. von Agris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," pp. 1–6, 2008.

[7] 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*, vol. 6, no. 4, pp. 323–362, 2008.

[8] C. Vogler and S. Goldenstein, "Facial movement analysis in asl," *Universal Access in the Information Society*, vol. 6, no. 4, pp. 363–374, 2008.

[9] S. C., W. Ong, and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Trans. PAMI*, vol. 27, no. 6, pp. 873–891, 2005.

[10] J. Ma, W. Gao, and R. Wang, "A parallel multistream model for integration of sign language recognition and lip motion," in *ICMI '00: Proc of the 3rd Intl Conf on Adv in Multimodal Interfaces*, 2000, pp. 582–589.

[11] D. Kelly, J. McDonald, and C. Markham, "Recognizing spatiotemporal gestures and movement epenthesis in sign language," in *IMVIP 2009*, 2009.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR, IEEE*, vol. 1, p. 511, 2001.

[13] L. A.-C. M. Castrillon-Santana, O. Deniz-Suarez and J. Lorenzo-Navarro, "Performance evaluation of public domain haar detectors for face and facial feature detection," *VISAPP 2008*, 2008.

[14] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[15] H. K. Lee and J. H. Kim, "An hmm-based threshold model approach for gesture recognition," *IEEE PAMI*, vol. 21, no. 10, pp. 961–973, 1999.