Incorporating Facial Features into a Multi-Channel Gesture Recognition System for the Interpretation of Irish Sign Language Sequences

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

Abstract

In this paper we present a novel gesture recognition system for the interpretation of Irish Sign Language sequences which incorporates manual and non-manual information. We implement a set of independent Hidden Markov Model networks to recognize hand gestures, head movements and facial features into a single framework for interpreting Irish Sign Language. This framework is not specific to any particular type of gesture and we demonstrate this by showing that manual and non manual signals can be robustly spotted and classified from with continuous sign sequences.

1. Introduction

In sign language, information is mainly conveyed through hand gestures. These hand gestures can be classified into several categories such as conversational gestures, controlling gestures, manipulative gestures and communicative gestures [32]. One of the main difficulties with recognizing a gesture within a continuous sequence of gestures is that the hand(s) must move from the end point of the previous gesture to the start point of the next gesture. These inter gesture transition periods are called movement epenthesis [16] and are not part of either of the signs. As such, an accurate recognition system must be able to distinguish between valid sign segments and movement epenthesis. Extending isolated recognition to continuous signing requires automatic detection of movement epenthesis segments so that the recognition algorithm can be applied on the segmented signs. A proposed solution to movement epenthesis detection is an explicit segmentation model where subsets, of features from gesture data, are used as cues for valid gesture start and end point detection [21, 15]. The limitation of this explicit segmentation model arises from the difficulty in creating general rules for sign boundary detection that could be applied to all types of gestures [19].

[1]† - Joint First Authorship

The problems associated with explicit segmentation can be overcome by implementing Hidden Markov Models (HMMs) for implicit sentence segmentation. Starner et al. [22] and Bauer and Kraiss [4] model each word or subunit with a HMM and then train the HMMs with data collected from full sentences. A downside to this is that training on full sentence data may result in a loss in valid sign recognition accuracy due to the large variations in the appearance of all the possible movement epenthesis that could occur between two signs. Wang et al. [31] also use HMMs to recognize continuous signs sequences with 92.8% accuracy, although signs were assumed to end when no hand motion occurred. Assan et al. [1] model the HMMs such that all transitions go through a single state, while Gao et al. [8] create separate HMMs that model the transitions between each unique pair of signs that occur in sequence. Vogler at al. [28] also use an explicit epenthesis modeling system where one HMM is trained for every two valid combinations of signs.

While these works have had promising results in gesture recognition and movement epenthesis detection, the training of such systems involves a large amount of extra data collection, model training and recognition computation due to the extra number of HMMs required to detect movement epenthesis. In this paper we build on the works of Kelly *et al.* [13], using a HMM based gesture recognition framework which accurately spots and classifies gestures within a continuous sequence of sign language, as one of a number of pre trained gestures as well as calculating the likelihood that the given gesture sequence is or is not a movement epenthesis. Sign language is a multimodal form of communication. It involves not only hand gestures (i.e., manual signing) but also non-manual signals (NMS) 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 and facial ex-

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

pression recognition tracking and 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 [2, 24, 3] focused on the role of head pose and body movement in sign language, where they reported a strong correlation linking 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 [10, 9]. Computer-based approaches which model facial movement using *Active Appearance Models* (AAMs) have been proposed [29, 30, 26]. Of particular interest are the works of Grossman *et al.* on American Sign Language, where they linked eyebrow movement and the degree of eye aperture movement to emotions and questions [10]. They reported that anger, wh-questions (who, where, what, when, why, how) and quizzical questions exhibited lowered brows and squinted eyes, while surprise and y/n questions showed raised brows and widened eyes. The development of a system combining manual and non-manual signals is a non-trivial task [5]. This is demonstrated by the limited amount of work dealing with the recognition of multimodal communication channels in sign language. Ma *et al.* [18] 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 generalized to other non-manual signals as they often span multiple manual signs and thus should be treated independently.

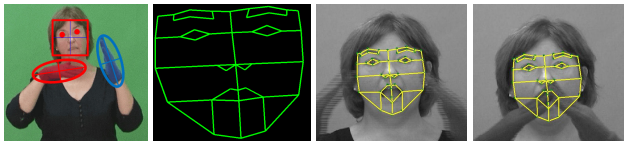
In this paper we propose an accessible approach towards multi-modal human-computer interaction (HCI) for the interpretation of Irish Sign Language (ISL) sequences. In ISL, like most other sign languages, the key information is conveyed using manual signs while non-manual signals are used to convey grammatical structure, syntax and emotional context, as such we process these two elements independently. The goal of the work described in this paper is to develop automatic methods of interpreting both manual and non manual signals in order to extract all the information expressed in sign language sentences. We extend the works of Kelly *et al.* [13] where hand gestures are recognized from continuous manual signals, and [11] which investigated the role of head tilt in ISL, to incorporate facial features such as the eyebrow movement into a multi-channel gesture recognition system.

2. Feature Extraction

From the definition of a spatiotemporal gesture [23], we must track the position and movement of the hands in order to described a hand gesture sequence. We expand on the work a of hand posture recognition system proposed Kelly *et al.*[12] to build a computer vision based feature extraction system for spatiotemporal gesture recognition. For completeness, prior to discussing our framework for continuous spotting of multimodal gestures in sign language, we briefly describe the feature tracking techniques implemented. Tracking of the hands is performed by tracking colored gloves using the Mean Shift algorithm [6]. Face and eye positions are used as features for head movement recognition and also used as hand gesture cues. Face and eye detection is carried out using a cascade of boosted classifiers working with haar-like features proposed by Viola and Jones [25]. A set of public domain classifiers [17], for the face, left eye and right eye, are used in conjunction with the OpenCV implementation of the haar cascade object detection algorithm. The features which we extract from each image are shown in 1(a), from these we define the following raw features: right hand position (RH_x, RH_y), left hand position (LH_x, LH_y), face position (FC_x, FC_y), face width (FW), left eye position (LE_x, LE_y) and right eye position (RE_x, RE_y).

2.1. Facial Feature Extraction

In this paper, we locate the facial features of interest using Cootes' implementation of *Active Shape Models* (ASMs) [7]. In the context of facial feature localization, ASMs can be viewed as statistical models of the shapes of the face which deform iteratively to fit to new images. Since the ASM is constrained by a statistical shape model, the range of possible deformations is constrained by the variance which exists in the training set. As a consequence, the accuracy of the ASM depends on the range of facial movements included in the training set. For the experiments included in this paper, our data set consisted of 3500 images in total. From which 300 key frames representing the variance in the data set were manually labeled with 46 points. Figure 1(b) shows the ASM which was trained on these image-points pairs.



(a) (b) (c) (d)

Figure 1. (a) features extracted from the image (b) sample ASM which was fitted to each image (c) sample of an un-occluded image (d) example of an occluded image

During sign language communication, the face is frequently occluded by the hands. Our approach to overcoming this particular problem was to fit the ASM to the parts

of the face that were visible, and use the previous points for occluded parts of the face. This can be seen in Figure 1 where the position of the mouth from Figure 1(c) is used in Figure 1(d) when the mouth is occluded. This is a valid approach as the hands move rapidly and rarely cover the same portion of the face for multiple frames.

3. 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 [20]. 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. Lee and Kim [14] proposed a single channel HMM threshold model using discrete observations to recognize a set of distinct gesture. We expand on their work by developing a multichannel HMM threshold model system using continuous multidimensional observation vectors. This is an important advancement as using continuous multidimensional observation vectors allows further expansion of our framework into different feature vectors without the loss of information through vector quantization which is required when using discrete observations. 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.

3.1. HMM Threshold Model

Lee and Kim [14] proposed a single channel HMM threshold model using discrete observations to recognize a set of distinct gesture. We expand on the work of Lee and Kim 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. A specific HMM, called a threshold model, is created to model movement epenthesis by calculating the likelihood threshold of an input gesture and provide a confirmation mechanism for provisionally matched gesture patterns. We denote each dedicated gesture HMM as λ_y. Each λ_y is used to calculate the likelihood that the input gesture is belonging to gesture class y. For a network of HMMs $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_Y\}$, a single threshold model $\overline{\lambda}$ is created. The threshold model $\overline{\lambda}$ is used to calculate the likelihood threshold for each of the dedicated gesture HMMs. It is not in the scope of this paper to describe the threshold model in detail and readers should consult the works of Lee and Kim [14] and Kelly *et al.*[13] for a more detailed discussion on the HMM threshold model technique.

4. HMM Threshold Model For Gesture Recognition

We develop a HMM threshold model system which we use to recognize hand gestures, head movement gestures and eyebrow gestures from continuous image sequences of sign language sentences being performed by a fluent signer.We now briefly describe this system.

4.1. HMM Training

We implement and train a dedicated HMM for each gesture to be recognized. We denote each dedicated HMM as λ_y where $y \in Y$ and Y is the set gesture labels. Each HMM is trained using an automated HMM initialization and training technique, utilizing an iterative clustering, Baum Welch and Viterbi realignment process, proposed by Kelly *et al.*[13]. A HMM threshold model, $\overline{\lambda}$ is then created using the network of trained HMMs λ_y (where $y \in Y$). The set of HMMs, to recognize the Y pre-trained gestures, is then denoted as $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_I, \overline{\lambda}\}$.

4.2. HMM Gesture Classification

Given an unknown sequence of gesture observations Θ, the goal is to accurately classify the gesture as a non-gesture or as one of the Y trained gestures. To classify the observations, the Viterbi algorithm is run on each model given the unknown observation sequences Θ, calculating the most likely state paths through each model y. The likelihoods of each state path, which we denote as $P(\Theta|\lambda_y)$, are also calculated. The sequence of observations can then be classified as y if the maximum likelihood $P_{ML}(\Theta|\lambda_y) \geq \Phi_y$, where the $P_{ML}(\Theta|\lambda_y) = \max_y P(\Theta|\lambda_y)$, $\Phi_y = P(\Theta|\overline{\lambda})\Gamma_y$ and Γ_y is a constant scalar value used to tune the sensitivity of the system movement epenthesis gestures.

4.3. Parallel HMM Training

When recognizing two handed spatiotemporal gestures, a parallel HMM is required to model the left and right hands [27]. We implement a parallel HMM Threshold Model system which initializes and trains a dedicated parallel HMM denoted as $\lambda_y^{'} = \{\lambda_{Ly}, \lambda_{Ry}\}$ where λ_{Ly} and λ_{Ry} are HMMs which model the left and right hand gestures respectively. The parallel HMMs are also trained using the same automated HMM initialization and training technique, utilizing an iterative clustering, Baum Welch and Viterbi realignment process, proposed by Kelly *et al.*[13]. A weighting of ω_{Ly} and ω_{Ry}, where $\omega_{Ly} + \omega_{Ry} = 1$, is applied to the left hand HMM and right hand HMM respectively, to account for variations in information held in each of the hands for a particular sign. A parallel HMM threshold model, $\overline{\lambda'} = \{\overline{\lambda_L}, \overline{\lambda_R}\}$ is then created using the network of trained parallel HMMs λ_y ($y \in Y$).

4.4. Parallel HMM Gesture Classification

To classify the parallel observations $\Theta^{'} = \{\Theta_L, \Theta_R\}$, the Viterbi algorithm is run on each model given the unknown observation sequences Θ_L and Θ_R, calculating the most likely state paths through each model y. The likelihoods of each state path, which we denote as $P(\Theta_L|\lambda_{Ly})$ and $P(\Theta_L|\lambda_{Ry})$, are also calculated. We calculate the overall likelihoods of a dedicated gesture and a movement epenthesis with the equations defined in Equations 1 and 2.

$$P(\Theta^{'}|\lambda_y^{'}) = P(\Theta_L|\lambda_{Ly})\omega_{Ly} + P(\Theta_R|\lambda_{Ry})\omega_{Ry} \quad (1)$$

$$\Phi_y^{'} = \frac{P(\Theta_L|\overline{\lambda_L})\Gamma_{Ly} + P(\Theta_R|\overline{\lambda_R})\Gamma_{Ry}}{2} \quad (2)$$

Where Γ_{Ly} and Γ_{Ry} are constant scalar values used to tune the sensitivity of the system to movement epenthesis. The sequence of observations can then be classified as y if $P_{ML}(\Theta^{'}|\lambda_y^{'}) \geq \Phi_y^{'}$, where $P_{ML}(\Theta^{'}|\lambda_y^{'})$ is the maximum likelihood defined as $\max_y P(\Theta^{'}|\lambda_y^{'})$.

4.5. Manual Sign Feature Processing

A spatiotemporal gesture is defined by the hands' position and movement, where the position refers to the hands' location relative to the body and movement traces out a trajectory in space. Kelly *et al.*[13] perform a number of experiments on isolated spatiotemporal gestures and movement epenthesis to find the best performing feature vector. Results showed that the best performing feature vector was a five dimensional vector describing the position of the hand relative to the eyes (RP_x, RP_y), the direction the hand was moving (V_x, V_y) and the distance between the two hands (D_H). For manual signs, we define O_t as the observation vector made at time t, where

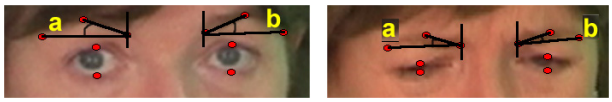


Figure 2. Example of subject performing a raised brow gestures (left) and a lowered brow gesture (right). a and b represent the angles ϕ_L and ϕ_R respectively

$O_t = \{RP_x, RP_y, V_x, V_y, D_H\}$. A particular hand sign sequence is then defined as $\Theta = \{O_1, O_2, ..., O_T\}$.

4.6. Head Movement Feature Processing

In a separate work, Kelly *et al.*[11] also perform experiments on isolated head movement gestures to find the best performing feature vector. The experiments showed that the best feature vector, with an AUC of 0.936, was a two dimensional vector, (V_x^H, V_y^H), describing the directional movement of the head in the x and y directions. To calculate the directional vector of the head the mid point between the eyes was used to calculated the direction the head moved from frame to frame. A sliding window was used to average the directional vector and the experiments showed that a window size of 12 frames achieved the best results.

4.7. Eyebrow Movement Feature Processing

Research on American Sign Language conducted by Grossman *et al.* [10] has linked eyebrow gestures to certain affective states and questions. Anger, wh-questions (who, where, what, when, why, how) and quizzical questions exhibited lowered brows and squinted eyes, while surprise and y/n questions showed raised brows and widened eyes. In this paper we focus on identifying these lowered brow gestures and raised brow gestures.

To test the discriminative performance of different feature vectors, we recorded a set of videos where each of the two eyebrow gestures occurred a total of 20 times. A fluent ISL signer performed the these gestures within different sign language sentences. The start and end points of the gestures were then labeled and isolated observation sequences Θ_i^{τ} were extracted. An additional set of 20 other brow gesture sequences, outside of the training set, were also labeled in the video sequences to test the performance of the system when identifying negative gestures.

We use the HMM classification techniques, described in Section 4.2, to classify eyebrow observation sequences. 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 1 shows the AUC measurement of different features which were evaluated during our experiments. We used a sliding window to average different features 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 the best performing window sizes for each feature vector in Table 1. The experiments show that the best performing feature vector was the vector (ϕ_{LR}, D^{Δ}). The value ϕ_{LR} is computed by calculating the average angle betweenϕ_L and ϕ_R shown in Figure 4.7.

Table 1. AUC Measurements for Different Feature Combinations

Features	$W^{\dagger}$	AUC
F_1 - Eye Brow Angle ϕ_{LR} + Distance Between Eye D	4	0.911
F_2 - Change Eye Brow Angle ϕ^{Δ}_{LR} + Change Distance Between Eye D^{Δ}	2	0.723
F_3 - **Eye Brow Angle** ϕ_{LR} **+ Change Distance Between Eye** D^{Δ}	**2**	**0.948**
F_4 - Change Eye Brow Angle ϕ^{Δ}_{LR} + Distance Between Eye D	2	0.812
F_5 - Distance Bottom Brow To Bottom Eye Aperture D_{BE} + Change Distance Between Eye D^{Δ}	6	0.933
F_6 - Distance Bottom Brow To Bottom Eye Aperture D_{BE} + Distance Between Eye D	6	0.903
F_7 - Eye Squint Size (D_{ES}) +Change Distance Between Eye D^{Δ}	6	0.776

† - *Window Size*

5. Continuous Recognition

In order to spot and classify manual signs, head movement gestures and eyebrow gestures we must extract four observation channels from the video streams. The four observation channels correspond to the left hand observations Θ_L, the right hand observations Θ_R, the head movement observations Θ^H and eyebrow observations Θ^B. The observations Θ_L and Θ_R are combined into a parallel observation sequence $\Theta^{'}$ which will be processed by the set of parallel HMMs. Since manual and non-manual signals are independent, the recognition of Θ, Θ^H and Θ^B will be processed independently and will be combined after the independent spotting and recognition of gestures within each of the three independent channels.

5.1. Continuous Manual Sign Recognition

We will now describe our system for spotting and classifying manual signs within a continuous sequence of parallel observations, $\Theta^{'}$, extracted from natural sign language sentences. The first step in our spotting algorithm is gesture end point detection. To detect a gesture end point in a continuous stream of gesture observations $\Theta^{'} = \{O^{'}_1, O^{'}_2, ..., O^{'}_T\}$, we calculate the model likelihoods of observation sequence $\theta^{'} = \{O^{'}_{T-F}, O^{'}_{T-F-1}, ..., O^{'}_T\}$ where $\theta^{'}$ is a subset of $\Theta^{'}$ and F defines the length of the observation (no. of frames) subset used. In this paper we set F to the average length of the observation sequences used to train the system.

A candidate hand gesture, κ, with end point, $\kappa_e = T$, is flagged when $\exists y : P(\Theta^{'}|\lambda^{'}_y) \geq \Psi^{'}_y$.

$$\Phi_y(\Theta^{'}) = \frac{P(\Theta^{'}|\lambda^{'}_y)}{P(\Theta^{'}|\lambda^{'}_y)+\Psi^{'}_y} \tag{3}$$

For each candidate end point we calculate a corresponding start point κ_s. Different candidate start points are evaluated using the measurement shown in Equation 3 where $\Phi_y(\Theta^{'})$ is normalized metric (between 0 and 1) which measures the strength of gesture y given observations $\Theta^{'}$. To find a candidate start point, the metric $\Phi_y(\Theta^{'}_{s\kappa_e})$ is calculated over different values of s, where $\Theta^{'}_{s\kappa_e} = \{O^{'}_s, O^{'}_{s+1}, ..., O^{'}_{\kappa_e}\}$ and $(\kappa_e - F^2) \leq s < \kappa_e$. The candidate gesture start point κ_s, is then found using Equation 4.

$$\kappa_s = \underset{s}{\text{argmax}} \Phi_y(\Theta^{'}_{s\kappa_e}) \tag{4}$$

The start and end point detection algorithm may flag candidate gestures which overlap and for this reason we expand on our continuous sign recognition algorithm with a candidate selection algorithm. The purpose of the candidate selection algorithm is to remove overlapping candidate gestures such that the single most likely gesture is the remaining gesture for a particular time frame.

The first step in the candidate selection algorithm is to cluster overlapping gestures, with the same gesture classification, together. Each of these candidate gestures, within the cluster, have an associated metric $\kappa_p = \Phi_y(\Theta^{'}_{\kappa_s\kappa_e})$. We remove all but one candidate gesture from this cluster leaving the candidate gesture, κ^B, with the highest κ_p value. We repeat this step for each cluster to produce a set of candidate gestures $\Upsilon = \{\kappa^{B1}, \kappa^{B2}, ..., \kappa^{BK}\}$, where K is the total number of clusters created from grouping overlapping gestures, with the same gesture classification, together. The second step in the candidate selection algorithm is an iterative selection step to remove the least probable candidate gestures as shown in Algorithm 1.

Input: Set of Candidate Gestures Υ
Output: Set of Recognized Gestures
Sort(Υ) by In Order of Increasing κ^B_P
for $i \leq K$ **do**
  if $\exists j \in J = \{i+1, i+2, ..., K\}, such\ that\ \Upsilon[j]\ overlaps\ with\ \Upsilon[i]$ **then**
    Remove $\Upsilon[i]$ from Υ;
  end
end

Algorithm 1: Second Step of Candidate Selection Algorithm

5.2. Continuous Non Manual Signal Recognition

The spotting and classifying of single channel observations sequences Θ^H and Θ^B is then conducted using the methods described in Section 5.1 above, however to keep the notation consistent with the techniques described in Section 4.2, the notation $\Theta^{'}$, $\lambda_c^{'}$ and $\Psi_c^{'}$ should be substituted with Θ, λ_i and Ψ_i respectively.

6. Continuous Recognition Experiments

We perform a set of experiments to evaluate our manual and non-manual signal recognition framework. We test our framework on a set of eight different manual signs, a set of three different head movement gestures and a set of two eyebrow gestures.

The set of gestures were not selected to be visually distinct but to represent a suitable cross section of the manual signs and head movement gestures that can occur in sign language. Figure 3 illustrates an example of a signer performing each of the eight manual signs, and Figure 6 illustrates an example of a signer performing each of the three different head movement gestures. Figure 4.7 illustrates the eyebrow gestures.

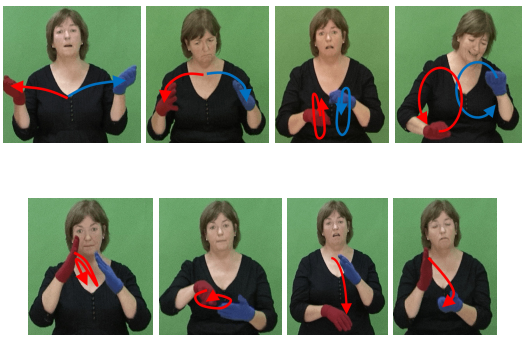


Figure 3. Example of the eight different signs the system was tested on:(Left to Right) Newspaper, A lot, Bike, Clean, Paint, Plate, Lost, Gone

We recorded a total of 160 additional video clips of full unsegmented sign language sentences being performed by a fluent signer to test the performance of our continuous recognition framework. Each video clip contained at least one of the eight chosen manual signs. The three head movement gestures occurred a total of 30 times within the 160 videos while the two eye brow gestures occurred a total of 35 times. Videos were recorded at 25 frames per second with an average length of 5 seconds. Observation sequences Θ_L, Θ_R, Θ^H and Θ^B were extracted from each video clip and our continuous recognition framework, described in Section 5, was used to process the observation sequences to spot gestures within the multiple observation channels.

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

In the gesture spotting and classification task, there are three types of errors: *an insertion error* occurs when the spotter reports a nonexistent gesture, *a deletion error* occurs when the spotter fails to detect a gesture, and *a substitution error* occurs when the spotter falsely classifies a gesture. From these error measures we define two performance metrics shown in Equation 5, where CS is the number of correctly spotted gestures, IG is the number of input gestures and IE is the number of insertion errors.

$$DetectionRatio = \frac{CS}{IG} \quad Reliability = \frac{CS}{IG + IE} \tag{5}$$

Table 2 shows the performance of our system when spotting and classifying signs within continuous sequences of video. The experiment shows an overall detection rate of 95.1% and an overall reliability of 93.4% when independently spotting and classifying manual and non-manual gestures in continuous sign language sentences.

Table 2. Continuous Spotter and Classifier Performance

Gesture	C	D	I	S	Det	Rel	E_S	E_E
Gone	20	0	0	0	1.0	1.0	±2.5	±8.4
Alot	20	0	0	0	1.0	1.0	±1.5	±1.6
Lost	20	0	0	0	1.0	1.0	±1.5	±3.5
Plate	19	0	1	0	0.95	0.90	±8.1	±12.2
Bike	20	0	0	0	1.0	1.0	±12.1	±12.0
Paint	20	0	0	0	1.0	1.0	±26.1	±20.7
Paper	16	0	1	3	0.8	0.76	±5.9	±1.6
Clean	18	0	1	1	0.9	0.85	±4.8	±5.2
Head Left	11	0	1	0	0.91	0.84	±10.1	±7.7
Head Right	10	0	0	0	1.0	1.0	±4.0	±4.3
Head Left Forward	8	0	0	1	0.88	0.88	±12.9 ±	6.5
EyeBrowDown	18	0	0	2	0.9	0.9	±19.2	±15.3
EyeBrowUp	15	0	0	0	1.0	1.0	±17.1	±24.9
Total	**215**	**0**	**4**	**7**	**0.951**	**0.934**	**±9.6**	**±9.5**

C-#Correct Gestures, D-#Deletion Errors, I-#Insertion Errors
S-#Substitution Errors, Det-#Detection Ratio, Rel-#Reliability
E_S-#Absolute Error Start Point, E_S-#Absolute Error End Point

We also evaluate the performance of the start and end point detection relative to ground truth data labeled by a human sign language translator. Table 2 also shows the average absolute difference between the spotters start and end points and the human interpreters start and end points for signs that were correctly spotted and classified. The average start point error was 9.6 frames and the average end point error was 9.5 frames. From this experiment we can conclude

that our spotter is capable of detecting start points, within an average of 384 milliseconds of a human interpreter, and end points, within an average of 380 milliseconds of a human interpreter.

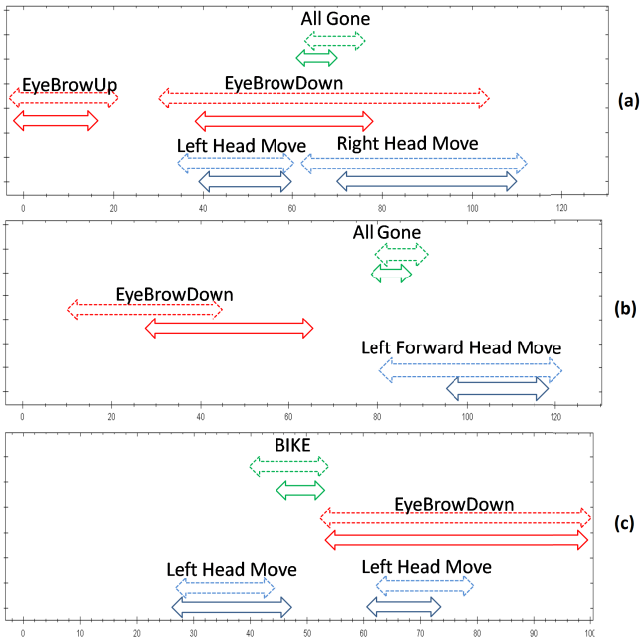


Figure 5. Multimodal gesture labeling comparison of a human interpreter vs. our recognition system (Dotted Arrows Represent Hand Labeled Gestures while solid arrows represent labels generated by our system)

Figure 6 shows the gestures spotted by our system in three different sentences. Figure 6(a) and Figure 6(b) show gestures spotted from two sentences where the signer performs the words "CAR PETROL ALL GONE" in both sentences. In the first sentence the signer is asking a question "CAR PETROL ALL GONE HOW?", but in the second sentence the signer is asking a yes/no question "CAR PETROL ALL GONE?". The manual signs for both these sentences are the same but the difference can only be recognized from the head movement and eyebrow gestures. It can be seen that our system spots an eyebrow down gesture coinciding with a left head movement followed by right head movement. This indicates that the signer is asking a "wh" question. In the second sentence our system spots a eyebrow down gesture at the beginning of the sentence followed by a head forward movement, indicating the signer is asking a yes/no question. Figure 6(c) shows gestures from a sentence "WHO BIKE BROKE?", where our system spots an eyebrow down gesture coinciding with a left head movement. Similar to the gestures in Figure 6(a), the eyebrow down gesture coinciding with a head movement gesture indicates a "wh" question. Also from Figure 6 (a), it was observed that an eyebrow up gesture occurred at the start of the sequence. This is an interesting observation as the eyebrow up gesture is linked to the start of a new sentence or sequence.

7. Conclusion

In this paper we have discussed current methods of continuous sign recognition, identifying that in general the drawbacks of these methods were that they imposed unnatural constraints on the signer, such as pauses between words, or required explicit training of models to handle movement epenthesis. Building on the techniques of Kelly *et al.*, our technique is capable of recognizing gestures from within unconstrained sign language sequences. Our system requires that a set of dedicated gesture models be trained, and as a result of this training a single threshold model can be created to identify movement epenthesis without explicitly training the model on movement epenthesis samples. We have also discussed the importance of non-manual signals in sign language. We have highlighted there are currently a limited number of works which incorporate both manual and non-manual signals into a single framework for continuous automatic sign language recognition. The main contribution of this work, is that we have presented an accessible approach towards multimodal human-computer interaction (HCI) for the interpretation of Irish Sign Language (ISL) sequences. The gesture recognition framework we propose is not specific to any particular type of gesture and we demonstrate this by showing that manual and non manual signals can be robustly spotted and classified from within continuous sign sequences. By incorporating non-manual signals such as eyebrow gestures and head movement gestures into our framework, we have shown that our technique provides the necessary foundations for differentiating between different types of questions, and also recognizing the start of sign language sentences. Also unlike current works, each manual and non-manual signal is processed independently within our multimodal framework. Experiments conducted demonstrate that our system achieved a detection ratio of 0.951 and a reliability measure of 0.934. Experiments also showed that our gesture spotting system successfully flagged gesture start points and end points within ± 384 milliseconds and ± 380 milliseconds respectively when compared to a human interpreter. Through these experiments we have proved the robustness of our system when recognizing a number of different manual and non-manual signals. Future work will entail, extending upon the results presented here to recognize a larger set of sign language phrases and to incorporate further non-manual signals and hand pose information into our framework.

Acknowledgment

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

References

[1] M. Assan and K. Grobel. Video-based sign language recognition using hidden markov models. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 97–109, London, UK, 1998. Springer-Verlag.

[2] B. Bahan. *Nonmanual Realisation of Agreement in American sign language*. PhD thesis, University of California, Berkely, 1996.

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

[4] B. Bauer and K.-F. Kraiss. Towards an automatic sign language recognition system using subunits. In *GW '01: Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, pages 64–75, London, UK, 2002. Springer-Verlag.

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

[6] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2:142–149 vol.2, 2000.

[7] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, October 2001.

[8] W. Gao, G. Fang, D. Zhao, and Y. Chen. Transition movement models for large vocabulary continuous sign language recognition. *IEEE FG 2004*, pages 553–558, May 2004.

[9] 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*, 31(1):23–38, 2007.

[10] 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, 2006.

[11] D. Kelly, J. R. Delannoy, J. M. Donald, and C. Markham. Automatic recognition of head movement gestures in sign language sentences. *in CIICT 2009*, 2009.

[12] D. Kelly, J. McDonald, T. Lysaght, and C. Markham. Analysis of sign language gestures using size functions and principal component analysis. In *IMVIP 2008*, 2008.

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

[14] H. K. Lee and J. H. Kim. An hmm-based threshold model approach for gesture recognition. *IEEE PAMI*, 21(10):961–973, 1999.

[15] R. H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *IEEE FG 1998*, page 558, Washington, DC, USA, 1998. IEEE Computer Society.

[16] J. R. Liddell, S.K. American sign language: The phonological base. *Sign Langauge Studies*, 64.

[17] 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.

[18] 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*, pages 582–589, 2000.

[19] S. C. W. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):873–891, 2005.

[20] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[21] H. Sagawa and M. Takeuchi. A method for recognizing a sequence of sign language words represented in a japanese sign language sentence. In *IEEE FG 2000*, page 434, Washington, DC, USA, 2000. IEEE Computer Society.

[22] T. Starner, A. Pentland, and J. Weaver. Real-time american sign language recognition using desk and wearable computer based video. *IEEE PAMI*, 20(12):1371–1375, 1998.

[23] J. Stokoe, William C. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of Deaf Studies and Deaf Education, v10 n1 p3-37 Win 2005*, 2005.

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

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

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

[27] C. Vogler and D. Metaxas. Parallel hidden markov models for american sign language recognition. In *In ICCV*, pages 116–122, 1999.

[28] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Computer Vision and Image Understanding*, 81:358–384, 2001.

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

[30] 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.

[31] C. Wang, S. Shan, and W. Gao. An approach based on phonemes to large vocabulary chinese sign language recognition. In *IEEE FG 2002*, page 411, Washington, DC, USA, 2002. IEEE Computer Society.

[32] Y. Wu and T. Huang. Human hand modeling, analysis and animation in the context of hci, 1999.