Sign Language Recognition

Helen Cooper, Brian Holt and Richard Bowden

Abstract This chapter covers the key aspects of Sign Language Recognition (SLR), starting with a brief introduction to the motivations and requirements, followed by a précis of sign linguistics and their impact on the field. The types of data available and the relative merits are explored allowing examination of the features which can be extracted. Classifying the manual aspects of sign (similar to gestures) is then discussed from a tracking and non-tracking viewpoint before summarising some of the approaches to the non-manual aspects of sign languages. Methods for combining the sign classification results into full SLR are given showing the progression towards speech recognition techniques and the further adaptations required for the sign specific case. Finally the current frontiers are discussed and the recent research presented. This covers the task of continuous sign recognition, the work towards true signer independence, how to effectively combine the different modalities of sign, making use of the current linguistic research and adapting to larger more noisy data sets.

1 Motivation

While automatic speech recognition has now advanced to the point of being commercially available, automatic SLR is still in its infancy. Currently all commercial translation services are human based, and therefore expensive, due to the experienced personnel required.

Helen Cooper e-mail: H.M.Cooper@surrey.ac.uk
Brian Holt e-mail: B.Holt@surrey.ac.uk
Richard Bowden e-mail: R.Bowden@surrey.ac.uk
University Of Surrey, Guildford, GU2 7XH, UK

These researchers have received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231135 - DictaSign

SLR aims to develop algorithms and methods to correctly identify a sequence of produced signs and to understand their meaning. Many approaches to SLR incorrectly treat the problem as Gesture Recognition (GR). So research has thus far focused on identifying optimal features and classification methods to correctly label a given sign from a set of possible signs. However, sign language is far more than just a collection of well specified gestures.

Sign languages pose the challenge that they are multi-channel; conveying meaning through many modes at once. While the studies of sign language linguistics are still in their early stages, it is already apparent that this makes many of the techniques used by speech recognition unsuitable for SLR. In addition, publicly available data sets are limited both in quantity and quality, rendering many traditional computer vision learning algorithms inadequate for the task of building classifiers. However, even given the lack of translation tools, most public services are not translated into sign. There is no commonly-used, written form of sign language, so all written communication is in the local spoken language.

This chapter introduces some basic sign linguistics before covering the types of data available and their acquisition methods. This is followed by a discussion on the features used for SLR and the methods for combining them. Finally the current research frontiers and the relating work is presented as an overview of the state of the art.

2 Sign Linguistics

Sign consists of three main parts: Manual features involving gestures made with the hands (employing hand shape and motion to convey meaning), Non-manual features such as facial expressions or body posture, which can both form part of a sign or modify the meaning of a manual sign, and Finger spelling, where words are spelt out gesturally in the local verbal language. Naturally this is an oversimplification, Sign language is as complex as any spoken language, each sign language has many thousands of signs, each differing from the next by minor changes in hand shape, motion, position, non-manual features or context. Since signed languages evolved alongside spoken languages, they do not mimic their counterparts. *e.g.* British Sign Language (BSL) loosely follows the sequence of time-line, location, subject, object, verb and question. It is characterised by topic-comment structure where a topic or scene is set up and then commented on [13]. It uses its own syntax which makes use of both manual and non-manual features, simultaneous and sequential patterning and spatial as well as linear arrangement.

Signs can be described at the sub-unit level using phonemes.[1] These encode different elements of a sign. Unlike speech they do not have to occur sequentially, but can be combined in parallel to describe a sign. Studies of American Sign Language (ASL) by Liddell and Johnson [64] model sign language on the movement-hold sys-

[1] Sometimes referred to as visemes, signemes, cheremes or morphemes. Current linguistic usage suggests phonemes is the accepted term.

tem. Signs are broken into sections where an aspect is changing and sections where a state is held steady. This is in contrast to the work of Stokoe [95] where different components of the sign are described in different channels; the motion made by the hands, the place at which the sign is performed, the hand shapes, the relative arrangement of the hands and finally the orientation of both the hands and fingers to explain the plane in which the hands sit. Both of these models are valid in their own right and yet they encode different aspects of sign. Within SLR both the movement-hold, sequential information from Liddell and Johnson and the parallel forms of Stokoe are desirable annotations.

Below are described a small subset of the constructs of sign language. There is not room here to fully detail the entire structure of the language, instead the focus is on those that pose significant challenges to the field of SLR:

(a) *Adverbs modifying verbs*; signers would not use two signs for ‘run quickly’ they would modify the sign for run by speeding it up.
(b) *Non-manual features (NMFs)*; facial expressions and body posture are key in determining the meaning of sentences, *e.g.* eyebrow position can determine the question type. Some signs are distinguishable only by lip shape, as they share a common manual sign.
(c) *Placement*; pronouns like ’he’, ’she’ or ’it’ do not have their own sign, instead the referent is described and allocated a position in the signing space. Future references point to the position, and relationships can be described by pointing at more than one referent.
(d) *Classifiers*; these are hand shapes which are used to represent classes of objects, they are used when previously described items interact. *e.g.* to distinguish between a person chasing a dog and vice versa.
(e) *Directional verbs*; these happen between the signer and referent(s), the direction of motion indicates the direction of the verb. Good examples of directional verbs are ‘give’ and ‘phone’. The direction of the verb implicitly conveys which nouns are the subject and object.
(f) *Positional Signs*; where a sign acts on the part of the body descriptively. *e.g.* ‘bruise’ or ‘tattoo’.
(g) *Body Shift*; represented by twisting the shoulders and gaze, often used to indicate role-shifting when relating a dialogue.
(h) *Iconicity*; when a sign imitates the thing it represents, it can be altered to give an appropriate representation. *e.g.* the sign for getting out of bed can be altered between leaping out of bed with energy to a recumbent who is reluctant to rise.
(i) *finger spelling*; Where a sign is not known, either by the signer or the recipient, the local spoken word for the sign can be spelt explicitly by finger spelling.

Although SLR and speech recognition are drastically different in many respects, they both suffer from similar issues; co-articulation between signs means that a sign will be modified by those either side of it. Inter-signer differences are large; every signer has their own style, in the same way that everyone has their own accent or handwriting. Also similar to handwriting, signers can be either left hand or right hand dominant. For a left handed signer, most signs will be mirrored, but time line specific ones will be kept consistent with the cultural ‘left to right’ axis. While it is not obvious how best to include these higher level linguistic constructs of the language, it is obviously essential if true, continuous SLR is to become reality.

3 Data Acquisition and Feature Extraction

Acquiring data is the first step in a SLR system. Given that much of the meaning in sign language is conveyed through manual features, this has been the area of focus of the research up to the present as noted by Ong and Ranganath in their 2005 survey [82].

Many early SLR systems used data gloves and accelerometers to acquire specifics of the hands. The measurements (x,y,z, orientation, velocity etc) were measured directly using a sensor such as the Polhemus tracker [103] or DataGlove [54, 99]. More often than not, the sensor input was of sufficient discriminatory power that feature extraction was bypassed and the measurements used directly as features [34]. While these techniques gave the advantage of accurate positions, they did not allow full natural movement and constricted the mobility of the signer, altering the signs performed. Trials with a modified glove-like device, which was less constricting [43], attempted to address this problem. However, due to the the prohibitive costs of such approaches, the use of vision has become more popular. In the case of vision input, a sequence of images are captured from a combination of cameras (e.g. monocular [115], stereo [47], orthogonal [90]) or other non-invasive sensors. Segen and Kumar [87] used a camera and calibrated light source to compute depth, and Feris *et al*. [30] used a number of external light sources to illuminate a scene and then used multi-view geometry to construct a depth image. Starner *et al*. [91] used a front view camera in conjunction with a head mounted camera facing down on the subject’s hands to aid recognition. Depth can also be inferred using stereo cameras as was done by Munoz-Salinas *et al*. [72] or by using side/vertical mounted cameras as with Vogler and Metaxas [100] or the Boston ASL data set [75]. There are several projects which are creating sign language data sets; in Germany there is the DGS-Korpus dictionary project collecting data across the country over a 15yr period [22] or the similar project on a smaller scale in the UK by the BSL Corpus Project [14]. However, these data sets are directed at linguistic research, whereas the cross domain European project DictaSign [23] aims to produce a multi-lingual data set suitable for both linguists and computer vision scientists.

Once data has been acquired it is described via features, the features chosen often depend on the elements of sign language being detected.

3.1 Manual Features

Sign language involves many features which are based around the hands, in general there are hand shape/orientation (pose) and movement trajectories, which are similar in principle to gestures. A survey of GR was performed by Mitra and Acharya [70] giving an overview of the field as it stood in 2007. While many GR techniques are applicable, Sign language offers a more complex challenge than the traditionally more confined domain of gesture recognition.

3.1.1 Tracking Based

Tracking the hands is a non-trivial task since, in a standard sign language conversation, the hands move extremely quickly and are often subject to motion blur. Hands are deformable objects, changing posture as well as position. They occlude each other and the face, making skin segmented approaches more complex. In addition as the hands interact with each other, tracking can be lost, or the hands confused. In early work, the segmentation task was simplified considerably by requiring the subjects to wear coloured gloves. Usually these gloves were single coloured, one for each hand [53]. In some cases, the gloves used were designed so that the hand pose could be better detected; employing coloured markers such as Holden and Owens [46] or different coloured fingers [44]. Zhang *et al*. [114] made use of multicoloured gloves (where the fingers and palms of the hands were different colours) and used the hands geometry to detect both position and shape. Using coloured gloves reduces the encumbrance to the signer but does not remove it completely. A more natural, realistic approach is without gloves, the most common detection approach uses a skin colour model [49, 7] where a common restriction is long sleeves. Skin colour detection is also used to identify the face position such as in [109]. Often this task is further simplified by restricting the background to a specific colour (chroma keying) [48] or at the very least keeping it uncluttered and static [90]. Zieren and Kraiss [116] explicitly modelled the background which aids the foreground segmentation task. Depth can be used to allow simplification of the problem. Hong *et al*. [47] and Grzeszcuk *et al*. [37] used a stereo camera pair from which they generated depth images which were combined with other cues to build models of the person(s) in the image. Fujimura and Liu [32] and Hadfield and Bowden [38] segmented hands on the naive assumption that hands will be the closest objects to the camera.

It is possible to base a tracker solely on skin colour as shown by Imgawa *et al*. [49] who skin segmented the head and hands before applying a Kalman filter during tracking. Han *et al*. [40] also showed that the Kalman filter enabled the skin segmented tracking to be robust to occlusions between the head and hands, while Holden *et al*. [45] considered snake tracking as a way of disambiguating the head from the hands. They initialised each snake as an ellipse from the hand position on the previous frame, using a gradient based optical flow method and shifted the ellipse to the new object position, fitting from that point. This sort of tracker tends to be non-robust to cluttered or moving backgrounds and can be confused by signers wearing short sleeved clothes. Akyol and Alvarado [4] improved on the original colour based skin segmented tracker, by using a combination of skin segmentation and motion history images (MHIs) to find the hands for tracking. Awad *et al*. [7] presented a face and hand tracking system that combined skin segmentation, frame differencing (motion) and predicted position (from a Kalman filter) in a probabilistic manner. These reduced the confusion with static background images but continued to suffer problems associated with bare forearms.

Micilotta and Bowden [68] proposed an alternative to colour segmentation, detecting the component parts of the body using Ong and Bowden's detector [80] and using these to infer a model of the current body posture, allowing the hand positions to be tracked across a video sequence. Buehler *et al*. implemented a robust tracker, which labelled data to initialise colour models, head/torso detector and Histogram of Oriented Gradients (HOG) pictorial descriptors. It used the distinctive frames in a sequence in much the same way that key frames are used in video encoding, they constrain adjacent frames and as such several passes can be made before the final trajectory is extracted. An alternative to this is the solution proposed by Zieren and Kraiss [116] who tracked multiple hypotheses via body modelling, disambiguating between these hypotheses at the sign level. These backward/forward methods for determining the hand trajectories offer accurate results but at the cost of processing time. Maintaining a trajectory after the hands have interacted also poses a problem. Shamaie and Sutherland [88] tracked bi-manual gestures using a skin segmentation based hand tracker, which calculated bounding box velocities to aid tracking after occlusion or contact. While adaptable to real time use, it suffers from the same problems as other colour only based approaches. Dreuw *et al*. used dynamic programming to determine the path of the head and hands along a whole video sequence, avoiding such failures at the local level [24] but negating the possibility of real-time application.

3.1.2 Non-Tracking Based

Since the task of hand tracking for sign language is a non-trivial problem, there has been work where signs are detected globally rather than tracked and classified. Wong and Cippola [105] used Principal Component Analysis (PCA) on motion gradient images of a sequence, obtaining features for a Bayesian classifier. Zahedi *et al*. investigated several types of appearance based features. They started by using combinations of down-sampled original images, multiplied by binary skin-intensity images and derivatives. These were computed by applying sobel filters [112]. They then combined skin segmentation with five types of differencing for each frame in a sequence, all are down sampled to obtain features [113]. Following this, their appearance based features were combined with the tracking work of Dreuw *et al*. [24] and some geometric features in the form of moments. Creating a system which fuses both tracking and non-tracking based approaches [111]. This system is able to achieve 64% accuracy rates on a more complex subset of the Boston dataset [75] including continuous sign from three signers. Cooper and Bowden [19] proposed a method for sign language recognition on a small sign subset that bypasses the need for tracking entirely. They classified the motion directly by using volumetric Haar-like features in the spatio-temporal domain. They followed this by demonstrating that non-tracking based approaches can also be used at the sub-unit level by extending the work of [53] to use appearance and spatio-temporal features [18].

The variability of the signers also introduces problems, the temporal inconsistencies between signs are a good example of this. Corradini [21] computed a series of moment features containing information about the position of the head and hands before employing Dynamic Time Warping (DTW) to account for the temporal difference in signs. Results are shown on a small dataset of exaggerated gestures which resemble traffic controls. It is unclear how well the DTW will port to the challenge of natural, continuous SLR.

3.1.3 Hand Shape

In systems where the whole signer occupies the field of view, the resolution of video is typically not high enough, and the computing power not sufficient for real time processing, so details of the specific hand shape tend to be ignored, or are approximated by extracting geometric features such as the centre of gravity of the hand blob. Using data gloves the hand shape can be described in terms of joint angles and more generically finger openness as shown by Vogler and Metaxas [102]. Jerde *et al.* combined this type of information with the known constraints of movement of the hands, in order to reduce the complexity of the problem [52]. Others achieved good results using vision based approaches. Ong and Bowden presented a combined hand detector and shape classifier using a boosted cascade classifier [79]. The top level of which detects the deformable model of the hand and the lower levels classified the hand shape into one of several image clusters, using a distance measure based on shape context. This offers 97.4% recognition rate on a database of 300 hand shapes. However, the hand shapes were assigned labels based on their shape context similarity. This means that the labels did not necessarily correspond to known sign hand shapes, nor did a label contain shapes which are actually the same, only those which look the same according to the clustering distance metric. Coogan and Sutherland [17] used a similar principle when they created a hierarchical decision tree, the leaf nodes of which contained the exemplar of a hand shape class, defined by fuzzy k-means clustering of the Eigenvalues resulting from performing PCA on the artificially constructed training images. Using gloved data to give good segmentation of the hands allowed Pahlevanzadeh *et al.* to use a generic cosine detector to describe basic hand shapes [84] though the system is unlikely to be tractable. Fillbrandt *et al.* used 2D appearance models to infer 3D posture and shape of the hands [31]. Each appearance model is linked to the others via a network which encodes the transitions between hand shapes, *i.e.* a model is only linked to another model if the transition between them does not require passage through another model. They tested their solution on a subset of hand shapes and postures but comment that for SLR a more complex model will be required. Hamada *et al*. used a similar transition principle [39], they matched the part of the hand contour which is not affected by occlusion or background clutter. These methods, while producing good results require large quantities of labelled data to build accurate models. Liu and Fujimura [66] analysed hand shape by applying a form of template matching that compared the current hand outline to a gradient image of a template using a Chamfer Distance. Athitsos and Sclaroff used a method for matching binary edges from cluttered images, to edges produced by model hand shapes [6]. Each of the possibilities was given a quantitative match value, from which they computed a list of ranked possible hand shapes for the input image. While the method worked well for small angles of rotation it did not perform so well when large variations were introduced. This is unsurprising given the appearance based approach used. Stenger *et al*. [93] employed shape templates in a degenerate decision tree, which took the form of a cascaded classifier to describe the position of the hands. The posture of the hands could then be classified using a set of exemplar templates, matched using a nearest neighbour classifier. The use of a decision tree improved scalability over previous individual classifier approaches but results in the entire tree needing to be rebuilt should a new template need to be incorporated. Roussos *et al*. [86] employ an Affine-invariant Modelling of hand Shape-Appearance images, offering a compact and descriptive representation of the hand configuration. The hand shape features extracted via the fitting of this model are used to construct an unsupervised set of sub-units.

Rezaei *et al*. used stereo cameras to reconstruct a 3D model of the hand [85]. They computed both loose point correspondences and 3-D motion estimation, in order to create the full 3D motion trajectory and pose of the hands. In contrast Guan *et al*. used multiple cameras, not to create a 3D model, but instead for a contour based 2D matching approach, they then fused results from across each of the cameras.

3.2 Finger Spelling

Manual features are also extended to finger spelling, a subset of sign language. Recognising finger spelling requires careful description of the shapes of the hands and for some languages the motion of the hands.

Isaacs and Foo [50] worked on finger spelling using wavelet features to detect static hand shapes. This approach limited them to non-dynamic alphabets. Liwicki and Everingham also concentrated on BSL finger spelling [67]. They combined HOG features with an HMM to model individual letters and non-letters. This allowed a scalable approach to the problem; unlike some of the previous work by Goh and Holden [35], which combined optical flow features with an Hidden Markov Model (HMM) but which only encoded the co-articulation present in the dataset lexicon. Jennings [51] demonstrates a robust finger tracking system that uses stereo cameras for depth, edges and colour. The system works by attempting to detect and track the finger using many different approaches and then by combining the approaches into a model, and the model which best explains the input data is taken as the solution. The approaches (or channels) are edges from 4 cameras, stereo from 2 and colour from 1, 7 channels in total. The channels are combined using Bayesian framework that reduces to a sum of squared differences equation. Stenger *et al*. [94] presented a model-based hand tracking system that used quadrics to build the underlying 3D model from which contours (handling occlusion) were generated that could be compared to edges in the image. Tracking is then done using an Unscented Kalman Filter. Feris *et al*. [30] generated an edge image from depth which is then used to generate a scale and translation invariant feature set very similar to Local Binary Patterns. This method was demonstrated to achieve high recognition rates, notably where other methods failed to discriminate between very similar signs.

3.3 Non-Manual Features

In addition to the manual features, there is a significant amount of information contained in the non-manual channels. The most notable of these are the facial expressions, lip shapes (as used by lip readers), as well as head pose which was recently surveyed by Murphy-Chutorian and Trivedi. [74] Little work has currently been performed on body pose, which plays a part during dialogues and stories.

Facial expression recognition can either be explicitly construed for sign language, or a more generic human interaction system. Some expressions, described by Ekman [26], are culturally independent (fear, sadness, happiness, anger, disgust and surprise). Most non-sign related expression research has been based on these categories, resulting in systems which do not always transfer directly to sign language expressions. In this field Yacoob and Davies used temporal information for recognition. They computed optical flow on local face features, to determine which regions of the face move to create each expression [106]. This reliance solely on the motion of the face works well for isolated, exaggerated expressions but will be easily confused by mixed or incomplete expressions as found in the real world. In contrast Moore and Bowden worked in the appearance domain. They used boosted classifiers on chamfer images to describe the forms made by a face during a given expression [71]. Reported accuracies are high but the approach is unlikely to be scalable to larger datasets due to its classifier per expression architecture.

Other branches of emotion detection research use a lower level representation of expression, such as Facial Action Coding System (FACS) [65]. FACS is an expression encoding scheme based on facial muscle movement. In principle, any facial expression can be described using a combination of facial action units (AUs). Koelstra *et al*. [57] presented methods for recognising these individual action units using both extended motion history images and Non-rigid Registration using Free-Form Deformations, reporting accuracies over 90%.

Recently the non-sign facial expression recognition community has begun work with less contrived data sets. These approaches are more likely to be applicable to sign expressions, as they will have fewer constraints, having been trained on more natural data sets. An example of this is the work by Sheerman-Chase *et al.* who combined static and dynamic features from tracked facial features (based on Ong's facial feature tracker [78]) to recognise more abstract facial expressions, such as 'Understanding' or 'Thinking' [89]. They note that their more complex dataset, while labelled, is still ambiguous in places due to the disagreement between human annotators. For this reason they constrain their experiments to work on data where the annotators showed strong agreement.

Ming and Ranganath separated emotions and sign language expressions explicitly. Their work split these into lower and upper face signals [69]. The training data was separated by performing Independent Component Analysis (ICA) on PCA derived feature vectors. This was then compared to results from Gabor Wavelet Networks. They showed that while the networks out performed the component analysis, this was only the case for high numbers of wavelets and as such, the required processing time was much higher.

Nguyen and Ranganath then tracked features on the face using a Kanade-Lucas-Tomasi Feature Tracker, commenting on the difficulties posed by inter-signer differences. They proposed a method to cluster face shape spaces from probabilistic PCA to combat these inconsistencies [76]. In later work, they combined this with HMMs and a Neural Network (NN) to recognise four sign language expressions [77]. They concentrate mainly on the tracking framework as a base for recognition, resulting in scope for further extensions to the work at the classification level.

Vogler worked on facial feature tracking within the context of SLR [96, 98, 97]. Vogler and Goldstein approach the explicit problem of sign language facial expressions, using a deformable face model [96, 97]. They showed that by matching points to the model and categorising them as inliers or outliers, it is possible to manage occlusions by the hands. They propose that tracking during full occlusion is not necessary, but that instead a 'graceful recovery' should be the goal. This is an interesting and important concept as it suggests that when the signer's mouth is occluded it is not necessary to know the mouth shape. Instead they believe that it can be inferred by the information at either side, in a similar manner to a human observer. While the theory is correct, the implementation may prove challenging.

Krinidis *et al.* used a deformable surface model to track the face [59]. From the parameters of the fitted surface model at each stage, a characteristic feature vector was created, when combined with Radial Basis Function Interpolation networks it can be used to accurately predict the pan, tilt and roll of the head. Ba and Odobez used appearance models of the colour and texture of faces, combined with tracking information, to estimate pose for visual focus of attention estimation [8]. They learn their models from the Prima-Pointing database of head poses, which contains a wide range of poses. Bailey and Milgram used the same database to present their regression technique, Boosted Input Selection Algorithm for Regression (BISAR) [9]. They combined the responses of block differencing weak classifiers with a NN. They boosted the final classifiers by rebuilding the NN after each weak classifier is chosen, using the output to create the weights for selection of the next weak classifier.

Some signs in BSL are disambiguated solely by the lip shapes accompanying them. Lip reading is already an established field, for aiding speech recognition or covert surveillance. It is known that human lip readers rely heavily on context when lip reading and also have training tricks, which allow them to set a baseline for a new subject, such as asking them questions where the answers are either known or easily inferred. Heracleous *et al*. showed that using the hand shapes from cued speech (where hand gestures are used to disambiguate vowels in spoken words for lip readers) improved the recognition rate of lip reading significantly [42]. They modelled the lip using some basic shape parameters, however it is also possible to track the lips, as shown by Ong and Bowden who use rigid flocks of linear predictors to track 34 points on the contour of the lips [81]. This is then extended to include HMMs to recognise phonemes from the lips [60].

4 Recognition

While some machine learning techniques were covered briefly in the section 3.1.3, this section focusses on how they have been applied to the task of sign recognition. The previous section looked at the low level features which provide the basis for SLR. In this section it is shown how machine learning can create combinations of these low level features to accurately describe a sign, or a subunit of sign.

4.1 Classification Methods

The earliest work on SLR applied NNs. However, given the success enjoyed by HMMs in the field of speech recognition, and the similarity of the problem of speech recognition and SLR, HMM based classification has dominated SLR since the mid 90's.

Murakami and Taguchi [73] published one of the first papers on SLR. Their idea was to train a NN given the features from their dataglove and recognise isolated signs, which worked even in the person independent context. Their system failed to address segmentation of the signs in time and is trained at a sign level, meaning that it is not extendible to continous recognition. Kim *et al*. [56] used datagloves to provide x,y,z coordinates as well as angles, from which they trained a Fuzzy Min Max NN to recognise 25 isolated gestures with a success rate of 85%. Lee *et al*. [61] used a Fuzzy Min Max NN to recognise the phonemes of continuous Korean Sign Language (KSL) with a vocabulary of 131 words as well as fingerspelling without modeling a grammar. Waldron and Kim [103] presented an isolated SLR system using NNs. They trained a first layer NN for each of the four subunit types present in the manual part of ASL, and then combined the results of the first layer in a second layer NN that actually recognises the isolated words. Huang *et al*. [48] presented a simple isolated sign recognition system using a Hopfield NN. Yamaguchi *et* *al*. [107] recognised a very small number of words using associative memory (similar to NNs). Yang *et al*. [109] presented a general method to extract motion trajectories, and then used them within a Time Delay Neural Network (TDNN) to recognise ASL. Motion segmentation is performed, and then regions of interest were selected using colour and geometry cues. The affine transforms associated with these motion trajectories were concatenated and used to drive the TDNN which classifies accurately and robustly. They demonstrated experimentally that this method achieved convincing results.

HMMs are a technique particularly well suited to the problem of SLR. The temporal aspect of SLR is simplified because it is dealt with automatically by HMMs [83]. The seminal work of Starner and Pentland [91] demonstrated that HMMs present a strong technique for recognising sign language and Grobel and Assan [36] presented a HMM based isolated sign (gesture) recognition system that performed well given the restrictions that it applied.

Vogler and Metaxas [99] show that word-level HMMs are SLR suitable, provided that the movement epenthesis is also taken into consideration. They showed how different HMM topologies (context dependent vs modeling transient movements) yield different results, with explicit modeling of the epenthesis yielding better results, and even more so when a statistical language model is introduced to aid classification in the presence of ambiguity and co-articulation. Due to the relative disadvantages of HMMs (poor performance when training data is insufficient, no method to weight features dynamically and violations of the stochastic independence assumptions), they coupled the HMM recogniser with motion analysis using computer vision techniques to improve combined recognition rates [100]. In their following work, Vogler and Metaxas [101] demonstrated that Parallel Hidden Markov Models (PaHMMs) are superior to regular HMMs, Factorial HMMs and Coupled HMMs for the recognition of sign language due the intrinsic parallel nature of the phonemes. The major problem though is that regular HMMs are simply not scalable in terms of handling the parallel nature of phonemes present in sign. PaHMMs are presented as a solution to this problem by modelling parallel processes independently and combining output probabilities afterwards.

Kim *et al*. [55] presented a KSL recognition system capable of recognising 5 sentences from a monocular camera input without a restricted grammar. They made use of a Deterministic Finite Automaton (DFA) to model the movement-stroke back to rest (to remove the epenthesis), and recognise with an DFA. Liang and OuhYoung [62] presented a sign language recognition system that used data captured from a single DataGlove. A feature vector was contructed that comprised posture, position, orientation, and motion. Three different HMMs were trained, and these are combined using a weighted sum of the highest probabilities to generate an overall score. Results were good on constrained data but the method is unlikely to generalise to real-world applications.

Kadous [54] presented a sign language recognition system that used instance based learning k-Nearest Neighbours (KNNs) and decision tree learning to classify isolated signs using dataglove features. The results were not as high as NN systems or HMM based systems, therefore given the relatively simple nature of the task it suggests that recognition using instance based learning such as KNN may not be a suitable approach.

Fang *et al*. [28] used a cascaded classifier that classified progressively one or two hands, hand shape and finally used a Self Organizing Feature Map (SOFM)/HMM to classify the words. The novelty of their approach was to allow multiple paths in the cascaded classifier to be taken, allowing for 'fuzziness'. Their approach was fast and robust, delivering very good classification results over a large lexicon, but it is ill-suited to a real-life application.

Other classifiers are suitable when using alternative inputs such as Wong and Cippola [105], who used a limited data set of only 10 basic gestures and require relatively large training sets to train their relevance vector machine (RVM). It should also be noted that their RVM requires significantly more training time than other vector machines but in return for a faster classifier which generalises better.

4.2 Phoneme Level Representations

Work in the field of sign language linguistics has informed the features used for detection. This is clearly shown in work which classifies in two stages; using first a sign sub-unit layer, followed by a sign level layer. This offers SLR the same advantages as it offered speech recognition. Namely a scalable approach to large vocabularies as well as a more robust solution for time variations between examples.

The early work of Vogler and Metaxas [99] borrowed heavily from the studies of sign language by Liddell and Johnson [64], splitting signs into motion and pause sections. While their later work [101], used PaHMMs on both hand shape and motion sub-units, as proposed by the linguist Stokoe [95]. Work has also concentrated on learning signs from low numbers of examples. Lichtenauer *et al*. [63] presented a method to automatically construct a sign language classifier for a previously unseen sign. Their method works by collating features for signs from many people then comparing the features of the new sign to that set. They then construct a new classification model for the target sign. This relies on a large training set for the base features (120 signs by 75 people) yet subsequently allows a new sign classifier to be trained using one shot learning. Bowden *et al*. [12] also presented a sign language recognition system capable of correctly classifying new signs given a single training example. Their approach used a 2 stage classifier bank, the first of which used hard coded classifiers to detect hand shape, arrangement, motion and position sub-units. The second stage removed noise from the 34 bit feature vector (from stage 1) using ICA, before applying temporal dynamics to classify the sign. Results are very high given the low number of training examples and absence of grammar. Kadir *et al*. [53] extended this work with head and hand detection based on boosting (cascaded weak classifiers), a body-centered description (normalises movements into a 2D space) and then a 2 stage classifier where stage 1 classifier generates linguistic feature vector and stage 2 classifier uses Viterbi on a Markov chain for highest recognition probability. Cooper and Bowden [18] continued this work still further with an approach to SLR that does not require tracking. Instead, a bank of classifiers are used to detect phoneme parts of sign activity by training and classifying (AdaBoost cascade) on certain sign sub-units. These were then combined into a second stage word-level classifier by applying a 1st order Markov assumption. The results showed that the detection rates achieved with a large lexicon and few training examples were almost equivalent to a tracking based approach.

Alternative methods have looked at data driven approaches to defining sub-units. Yin *et al*. [110] used an accelerometer glove to gather information about a sign, before applying discriminative feature extraction and similar state tying algorithms, to decide sub-unit level segmentation of the data. Kong *et al*. [58] and Han *et al*. [41] have looked at automatic segmentation of the motions of sign into sub-units, using discontinuities in the trajectory and acceleration, to indicate where segments begin and end, these are then clustered into a code book of possible exemplar trajectories using either DTW distance measures, in the case of Han *et al*. or PCA features by Kong *et al*.

5 Research Frontiers

There are many facets of SLR which have attracted attention in the computer vision community. This section serves to outline the areas which are currently generating the most interest due to the challenges they propose. While some of these are recent topics, others have been challenging computer vision experts for many years. Offered here is a brief overview of the seminal work and the current state of the art in each area.

5.1 Continuous Sign Recognition

The majority of work on SLR has been focused on recognising isolated instances of signs, this is not applicable to a real world sign language recognition system. The task of recognising continuous sign language is complicated primarily by the problem that in natural sign language, the transition between signs is not clearly marked because the hands will be moving to the starting position of the next sign. This is referred to as the *movement epenthesis* or co-articulation (which borrows from speech terminology). Both Vogler [99] and Gao *et al*. [33] modelled the movement epenthesis explicitly. Gao *et al*. [33] used data gloves and found the end points and starting points of all signs in their vocabulary. Clustering these movement transitions into three general clusters using a temporal clustering algorithm (using DTW), allowed them to recognise 750 continuous sign language sentences with an accuracy of 90.8%. More recently, Yang *et al*. [108] presented a technique by which signs could be isolated from continuous sign data by introducing an adaptive threshold model (which discriminates between signs in a dictionary and non sign patterns).

Applying a short sign detector and an appearance model improved sign spotting accuracy. They then recognise the isolated signs that have been identified.

5.2 *Signer Independence*

A major problem relating to recognition is that of applying the system to a signer on whom the system has not been trained. Zieren and Kraiss [116] applied their previous work to the problem of signer independence. Their results showed that the two problems are robust feature selection and interpersonal variation in the signs. They have shown that their system works very well with signer dependence, but recognition rates drop considerably in real world situations. In [3] Von Agris *et al.* presented a comprehensive SLR system using techniques from speech recognition to adapt the signer features and classification, making the recognition task signer independent. In other work [1], they demonstrated how three approaches to speaker adaptation in speech recognition can be successfully applied to the problem of signer adaptation for signer independent sign language recognition. They contrasted a PCA based approach, a maximum likelihood linear regression approach and a maximum a posteriori probability (MAP) estimation approach, and finally showed how they can be combined to yield superior results .

5.3 *Fusing Multi-Modal Sign Data*

From the review of SLR by Ong and Ranganath [82], one of their main observations is the lack of attention that non-manual features has received in the literature. This is still the case several years on. Much of the information in a sign is conveyed through this channel, and particularly there are signs that are identical in respect of the manual features and only distinguishable by the non-manual features accompanying the sign. The difficulty is identifying exactly which elements are important to the sign, and which elements are coincidental. For example, does the blink of the signer convey information valuable to the sign, or was the signer simply blinking? This problem of identifying the parts of the sign that contains information relevant to the understanding of the sign makes SLR a complex problem to solve. Non-manual features can broadly be divided into Facial Features which may consist of lip movement, eye gaze and facial expression; and Body Posture, e.g. moving the upper body forward to refer to the future, or sideways to demonstrate the change of the subject in a dialogue. While, as described in section 3.3, there has been some work towards the facial features, very little work has been done in the literature regarding the role of body posture in SLR. The next step in the puzzle is how to combine the information from the manual and non-manual streams.

Von Agris *et al.* [2] attempted to quantify the significance of non-manual features in SLR, finding that the overall recognition rate was improved by including non-manual features in the recognition process. They merged manual features with (facial) non-manual features that are modelled using an Active Appearance Model (AAM). After showing how features are extracted from the AAM, they presented results of both continuous and isolated sign recognition using manual features and non-manual features. Results showed that some signs of Deutsche Gebrdensprache/German Sign Language (DGS) can be recognised based on non-manual features alone, but generally the recognition rate increases by between 1.5% and 6% upon inclusion of non-manual features. In [3], Von Agris *et al*. present a comprehensive sign language recognition system using images from a single camera. The system was developed to use manual and non-manual features in a PaHMM to recognise signs, and furthermore, statistical language modelling is applied and compared.

Aran *et al*. [5] compared various methods to integrate manual features and non-manual features in a sign language recognition system. Fundamentally they have identified a two step classification process, whereby the first step involves classifying based on manual signs. When there was ambiguity, they introduced a second stage classifier to use non-manual signs to resolve the problem. While this might appear a viable approach, it is not clear from sign language linguistics that it is scalable to the full SLR problem.

5.4 Using Linguistics

The task of recognition is often simplified by forcing the possible word sequence to conform to a grammar which limits the potential choices and thereby improves recognition rates [91, 104, 12, 45]. N-Gram grammars are often used to improve recognition rates, most often bi-gram [83, 44, 34] but also uni-gram [10]. Bungeroth and Ney [16] demonstrated that statistical sign language translation using Bayes rule is possible and has the potential to be developed into a real-world translation tool. Bauer *et al*. [11] presented a sign language translation system consisting of a SLR module which fed a translation module. Recognition was done on word level HMMs (high accuracy rate, but not scalable), and the translation was done using statistical grammars developed from the data.

5.5 Generalising to More Complex Corpora

Due to the lack of adequately labelled data sets, research has turned to weakly supervised approaches. Several groups have presented work aligning subtitles with signed TV broadcasts. Farhadi and Forsyth [29] used HMMs with both static and dynamic features, to find estimates of the start and end of a sign, before building a discriminative word model to perform word spotting on 31 different words over an 80000 frame children's film. Buehler *et al*. [15] used 10.5 hours of TV data, showing detailed results for 41 signs with full ground truth, alongside more generic results for a larger 210 word list. They achieve this by creating a distance metric for signs, based on the hand trajectory, shape and orientation and performing a brute force search. Cooper and Bowden [20] used hand and head positions in combination with data mining to extract 23 signs from a 30 minute TV broadcast. By adapting the mining to create a temporally constrained implementation they introduced a viable alternative to the brute force search. Stein *et al*. [92] are collating a series of weather broadcasts in DGS and German. This data set will also contain the DGS glosses which will enable users to better quantify the results of weakly supervised approaches.

6 Conclusions

SLR has long since advanced beyond classifying isolated signs or alphabet forms for finger spelling. While the field may continue to draw on the advances in GR the focus has shifted to approach the more linguistic features associated with the challenge. Work has developed on extracting signs from continuous streams and using linguistic grammars to aid recognition. However, there is still much to be learnt from relevant fields such as speech recognition or hand writing recognition. In addition, while some have imposed grammatical rules from linguistics, others have looked at data driven approaches, both have their merits since the linguistics of most sign languages are still in their infancy.

While the community continues to discuss the need for including non-manual features, few have actually done so. Those which have [2, 5], concentrate solely on the facial expressions of sign. There is still much to be explored in the veins of body posture or placement and classifier (hand shape) combinations.

Finally, to compound all these challenges, there is the issue of signer independence. While larger data sets are starting to appear, few allow true tests of signer independence over long continuous sequences. Maybe this is one of the most urgent problems in SLR that of creating data sets which are not only realistic, but also well annotated to facilitate machine learning.

Despite these problems recent uses of SLR include translation to spoken language, or to another sign language when combined with avatar technology [3, 25]. Sign video data once recognised can be compressed using SLR into an encoded form (*e.g*. Signing Gesture Markup Language (SiGML) [27]) for efficient transmission over a network. SLR is also set to be used as an annotation aid, to automate annotation of sign video for linguistic research, currently a time-consuming and expensive task.

References

1. von Agris, U., Blomer, C., Kraiss, K.F.: Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, MLLR, and MAP. In: Procs. of ICPR, pp. 1 – 4. Tampa, Florida, USA (2008). DOI 10.1109/ICPR.2008.4761363
2. von Agris, U., Knorr, M., Kraiss, K.: The significance of facial features for automatic sign language recognition. In: Procs. of FGR, pp. 1 – 6. Amsterdam, The Netherlands (2008)
3. von Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.: Recent developments in visual sign language recognition. Universal Access in the Information Society **6**(4), 323 – 362 (2008)
4. Akyol, S., Alvarado, P.: Finding relevant image content for mobile sign language recognition. In: Procs. of IASTEDInt. Conf. on Signal Processing, Pattern Recognition and Application, pp. 48 – 52. Rhodes, Greece (2001)
5. Aran, O., Burger, T., Caplier, A., Akarun, L.: A belief-based sequential fusion approach for fusing manual signs and non-manual signals. PATTERN RECOGN LETTERS **42**(5), 812 – 822 (2009)
6. Athitsos, V., Sclaroff, S.: Estimating 3D hand pose from a cluttered image. In: Procs. of CVPR, vol. 2. Madison WI, USA (2003)
7. Awad, G., Han, J., Sutherland, A.: A unified system for segmentation and tracking of face and hands in sign language recognition. In: Procs. of ICPR, vol. 1, pp. 239 – 242. Hong Kong, China (2006). DOI 10.1109/ICPR.2006.194
8. Ba, S.O., Odobez, J.M.: Visual focus of attention estimation from head pose posterior probability distributions. In: Procs. of IEEEInt. Conf. on Multimedia and Expo, pp. 53–56 (2008). DOI 10.1109/ICME.2008.4607369
9. Bailly, K., Milgram, M.: Bisar: Boosted input selection algorithm for regression. In: Procs. of Int. Joint Conf. on Neural Networks, pp. 249–255 (2009). DOI 10.1109/IJCNN.2009.5178908
10. Bauer, B., Hienz, H., Kraiss, K.: Video-based continuous sign language recognition using statistical methods. In: Procs. of ICPR, vol. 15, pp. 463 – 466. Barcelona, Spain (2000)
11. Bauer, B., Nießen, S., Hienz, H.: Towards an automatic sign language translation system. In: Procs. of Int. Wkshp : Physicality and Tangibility in Interaction: Towards New Paradigms for Interaction Beyond the Desktop. Siena, Italy (1999)
12. Bowden, R., Windridge, D., Kadir, T., Zisserman, A., Brady, M.: A linguistic feature vector for the visual interpretation of sign language. In: Procs. of ECCV, LNCS, pp. 390 – 401. Springer, Prague, Czech Republic (2004)
13. British Deaf Association: Dictionary of British Sign Language/English. Faber and Faber (1992)
14. BSL Corpus Project: Bsl corpus project site (2010). URL www.bslcorpusproject.org/
15. Buehler P. Everingham, M., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: Procs. of CVPR, pp. 2961 – 2968. Miami, FL, USA (2009)
16. Bungeroth, J., Ney, H.: Statistical sign language translation. In: Procs. of LREC : Wkshp : Representation and Processing of Sign Languages, pp. 105 – 108. Lisbon, Portugal (2004)
17. Coogan, T., Sutherland, A.: Transformation invariance in hand shape recognition. In: Procs. of ICPR, vol. 3, pp. 485 – 488. Hong Kong, China (2006). DOI 10.1109/ICPR.2006.1134
18. Cooper, H., Bowden, R.: Large lexicon detection of sign language. In: Procs. of ICCV : Wkshp : Human-Computer Interaction, pp. 88 – 97. Rio de Janario, Brazil (2007). DOI 10.1007/978-3-540-75773-3_10
19. Cooper, H., Bowden, R.: Sign language recognition using boosted volumetric features. In: Procs. of IAPR Conf. on Machine Vision Applications, pp. 359 – 362. Tokyo, Japan (2007)
20. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: Procs. of CVPR, pp. 2568 – 2574. Miami, FL, USA (2009). DOI DOI10.1109/CVPRW.2009.5206647

21. Corradini, A.: Dynamic time warping for off-line recognition of a small gesture vocabulary. In: Procs. of ICCV: Wkshp : Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 82 – 90. IEEE Computer Society, Vancouver, BC (2001)
22. DGS-Corpus: Dgs-corpus website (2010). URL www.sign-lang.uni-hamburg.de/dgs-korpus
23. DictaSign Project: Dictasign project website (2010). URL www.dictasign.eu
24. Dreuw, P., Deselaers, T., Rybach, D., Keysers, D., Ney, H.: Tracking using dynamic programming for appearance-based sign language recognition. In: Procs. of FGR, pp. 293 – 298. Southampton, UK (2006). DOI 10.1109/FGR.2006.107
25. Efthimiou, E., Fotinea, S.E., Vogler, C., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., Segouat, J.: Sign language recognition, generation, and modelling: A research effort with applications in deaf communication. In: Procs. of Int. Conf. on Universal Access in Human-Computer Interaction. Addressing Diversity, vol. 1, pp. 21 – 30. Springer-Verlag, San Diego, CA, USA (2009). DOI http://dx.doi.org/10.1007/978-3-642-02707-9_3
26. Ekman, P.: Basic emotions. In: T. Dalgleish, T. Power (eds.) The Handbook of Cognition and Emotion, pp. 45–60. John Wiley & Sons, Ltd. (1999)
27. Elliott, R., Glauert, J., Kennaway, J., Parsons, K.: D5-2: SiGML Definition. ViSiCAST Project working document (2001)
28. Fang, G., Gao, W., Zhao, D.: Large vocabulary sign language recognition based on fuzzy decision trees. IEEE SYS MAN CYBERN Part A **34**(3), 305 – 314 (2004)
29. Farhadi, A., Forsyth, D.: Aligning ASL for statistical translation using a discriminative word model. In: Procs. of CVPR, pp. 1471 – 1476. New York, NY, USA (2006). DOI http://dx.doi.org/10.1109/CVPR.2006.51
30. Feris, R., Turk, M., Raskar, R., Tan, K., Ohashi, G.: Exploiting depth discontinuities for vision-based fingerspelling recognition. In: Procs. of CVPR : Wkshp :, vol. 10. IEEE Computer Society Washington, DC, USA, Washington, DC, USA (2004)
31. Fillbrandt, H., Akyol, S., Kraiss, K.F.: Extraction of 3D hand shape and posture from image sequences for sign language recognition. In: Procs. of ICCV : Wkshp : Analysis and Modeling of Faces and Gestures, pp. 181 – 186. Nice, France (2003)
32. Fujimura, K., Liu, X.: Sign recognition using depth image streams. In: Procs. of FGR, pp. 381 – 386. Southampton, UK (2006)
33. Gao, W., Fang, G., Zhao, D., Chen, Y.: Transition movement models for large vocabulary continuous sign language recognition. In: Procs. of FGR, pp. 553 – 558. Seoul, Korea (2004). DOI 10.1109/AFGR.2004.1301591
34. Gao, W., Ma, J., Wu, J., Wang, C.: Sign language recognition based on HMM/ANN/DP. International journal of pattern recognition and artificial intelligence **14**(5), 587 – 602 (2000)
35. Goh, P., Holden, E.J.: Dynamic fingerspelling recognition using geometric and motion features. In: Procs. of ICIP, pp. 2741–2744 (2006). DOI 10.1109/ICIP.2006.313114
36. Grobel, K., Assan, M.: Isolated sign language recognition using hidden markov models. In: Procs. of IEEEInt. Conf. on Systems, Man, and Cybernetics, vol. 1, pp. 162 – 167. Orlando, FL, USA (1997)
37. Grzeszcuk, R., Bradski, G., Chu, M., Bouguet, J.: Stereo based gesture recognition invariant to 3d pose and lighting. In: Procs. of CVPR, vol. 1 (2000)
38. Hadfield, S., Bowden, R.: Generalised pose estimation using depth. In: Procs. of ECCVInt. Wkshp : Sign, Gesture, Activity". Heraklion, Crete (2010)
39. Hamada, Y., Shimada, N., Shirai, Y.: Hand shape estimation under complex backgrounds for sign language recognition. In: Procs. of FGR, pp. 589 – 594. Seoul, Korea (2004). DOI 10.1109/AFGR.2004.1301597
40. Han, J., Awad, G., Sutherland, A.: Automatic skin segmentation and tracking in sign language recognition. IET Computer Vision **3**(1), 24 – 35 (2009). DOI 10.1049/iet-cvi:20080006
41. Han, J., Awad, G., Sutherland, A.: Modelling and segmenting subunits for sign language recognition based on hand motion analysis. PATTERN RECOGN LETTERS **30**(6), 623 – 633 (2009)

42. Heracleous, P., Aboutabit, N., Beautemps, D.: Lip shape and hand position fusion for automatic vowel recognition in cued speech for french. IEEE Signal Processing Letters **16**(5), 339–342 (2009). DOI 10.1109/LSP.2009.2016011
43. Hernandez-Rebollar, J., Lindeman, R., Kyriakopoulos, N.: A multi-class pattern recognition system for practical finger spelling translation. In: Procs. of IEEEInt. Conf. on Multimodal Interfaces, p. 185. IEEE Computer Society (2002)
44. Hienz, H., Bauer, B., Karl-Friedrich, K.: HMM-based continuous sign language recognition using stochastic grammars. In: Procs. of GW, pp. 185 – 196. Springer, Gif-sur-Yvette, France (1999)
45. Holden, E., Lee, G., Owens, R.: Australian sign language recognition. Machine Vision and Applications **16**(5), 312 – 320 (2005)
46. Holden, E., Owens, R.: Visual sign language recognition. In: Procs. of Int. Wkshp : Theoretical Foundations of Computer Vision, *LNCS*, vol. 2032, pp. 270 – 288. Springer, Dagstuhl Castle, Germany (2000)
47. Hong, S., Setiawan, N., Lee, C.: Real-time vision based gesture recognition for human-robot interaction. In: Procs. of Int. Conf. on Knowledge-Based and Intelligent Information & Engineering Systems : Italian Wkshp : Neural Networks, *LNCS*, vol. 4692, p. 493. Springer, Vietri sul Mare, Italy (2007)
48. Huang, C.L., Huang, W.Y., Lien, C.C.: Sign language recognition using 3D hopfield neural network. In: Procs. of ICIP, vol. 2, pp. 611 – 614 (1995). DOI 10.1109/ICIP.1995.537553
49. Imagawa, K., Lu, S., Igi, S.: Color-based hands tracking system for sign language recognition. In: Procs. of FGR, pp. 462 – 467. Nara, Japan (1998)
50. Isaacs, J., Foo, J.S.: Hand pose estimation for american sign language recognition. In: Procs. of Southeastern Symposium on System Theory, pp. 132 – 136. Atlanta, GA, USA (2004)
51. Jennings, C.: Robust finger tracking with multiple cameras. In: Procs. of ICCV : Wkshp : Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 152 – 160. Corfu, Greece (1999)
52. Jerde, T.E., Soechting, J.F., Flanders, M.: Biological constraints simplify the recognition of hand shapes. IEEE Transactions on Bio-Medical Engineering **50**(2), 265–269 (2003). DOI 10.1109/TBME.2002.807640
53. Kadir, T., Bowden, R., Ong, E., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: Procs. of BMVC, vol. 2, pp. 939 – 948. Kingston, UK (2004)
54. Kadous, M.: Machine recognition of auslan signs using powergloves: Towards large-lexicon recognition of sign language. In: Procs. of Wkshp : Integration of Gesture in Language and Speech (1996)
55. Kim, J., Park, K., Bang, W., Kim, J., Bien, Z.: Continuous korean sign language recognition using automata based gesture segmentation and hidden markov model. In: Procs. of Int. Conf. on Control, Automation and Systems, pp. 822 – 825 (2001)
56. Kim, J.S., Jang, W., Bien, Z.: A dynamic gesture recognition system for the korean sign language (KSL). IEEE SYS MAN CYBERN Part B **26**(2), 354 – 359 (1996). DOI 10.1109/3477.485888
57. Koelstra, S., Pantic, M., Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. IEEE TPAMI **32**(11), 1940 –1954 (2010). DOI 10.1109/TPAMI.2010.50
58. Kong, W.W., Ranganath, S.: Automatic hand trajectory segmentation and phoneme transcription for sign language. In: Procs. of FGR, pp. 1 – 6. Amsterdam, The Netherlands (2008). DOI 10.1109/AFGR.2008.4813462
59. Krinidis, M., Nikolaidis, N., Pitas, I.: 3-d head pose estimation in monocular video sequences using deformable surfaces and radial basis functions. IEEE Transactions on Circuits and Systems for Video Technology **19**(2), 261–272 (2009). DOI 10.1109/TCSVT.2008.2009261
60. Lan, Y., Harvey, R., Theobald, B.J., Ong, E.J., Bowden, R.: Comparing visual features for lipreading. In: Procs. of Int. Conf. Auditory-visual Speech Processing. Norwich, UK (2009)

61. Lee, C.S., Bien, Z., Park, G.T., Jang, W., Kim, J.S., Kim, S.K.: Real-time recognition system of korean sign language based on elementary components. In: Procs. of IEEEInt. Conf. on Fuzzy Systems, vol. 3, pp. 1463 – 1468 (1997). DOI 10.1109/FUZZY.1997.619759
62. Liang, R., Ouhyoung, M.: A real-time continuous gesture recognition system for sign language. In: Procs. of FGR, pp. 558 – 567. Nara, Japan (1998)
63. Lichtenauer, J., Hendriks, E., Reinders, M.: Learning to recognize a sign from a single example. In: Procs. of FGR, pp. 1 – 6. Amsterdam, The Netherlands (2008). DOI 10.1109/AFGR.2008.4813450
64. Liddell, S.K., Johnson, R.E.: American sign language: The phonological base. Sign Language Studies **64**, 195 – 278 (1989)
65. Lien, J.J.J., Kanade, T., Cohn, J., Li, C.C.: Automated facial expression recognition based on facs action units. In: Procs. of FGR, pp. 390–395. Nara, Japan (1998)
66. Liu, X., Fujimura, K.: Hand gesture recognition using depth data. In: Procs. of FGR, pp. 529 – 534. Seoul, Korea (2004). DOI 10.1109/AFGR.2004.1301587
67. Liwicki, S., Everingham, M.: Automatic recognition of fingerspelled words in british sign language. In: Procs. of CVPR, pp. 50–57. Miami, FL, USA (2009). DOI 10.1109/CVPR.2009.5204291
68. Micilotta, A., Bowden, R.: View-based location and tracking of body parts for visual interaction. In: Procs. of BMVC, vol. 2, pp. 849 – 858. Kingston, UK (2004)
69. Ming, K.W., Ranganath, S.: Representations for facial expressions. In: Procs. of Int. Conf. on Control, Automation, Robotics and Vision, vol. 2, pp. 716 – 721 (2002). DOI 10.1109/ICARCV.2002.1238510
70. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE SYS MAN CYBERN Part C **37**(3), 311 – 324 (2007)
71. Moore, S., Bowden, R.: Automatic facial expression recognition using boosted discriminatory classifiers. In: Procs. of ICCV : Wkshp : Analysis and Modeling of Faces and Gestures. Rio de Janario, Brazil (2007)
72. Munoz-Salinas, R., Medina-Carnicer, R., Madrid-Cuevas, F., Carmona-Poyato, A.: Depth silhouettes for gesture recognition. PATTERN RECOGN LETTERS **29**(3), 319 – 329 (2008)
73. Murakami, K., Taguchi, H.: Gesture recognition using recurrent neural networks. In: Procs. of SIGCHI Conf. on Human factors in computing systems: Reaching through technology, pp. 237 – 242. ACM New York, NY, USA (1991)
74. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. IEEE TPAMI **31**(4), 607–626 (2009). DOI 10.1109/TPAMI.2008.106
75. Neidle, C.: National centre for sign language and gesture resources (2006). URL www.bu.edu/asllrp/cslgr/
76. Nguyen, T.D., Ranganath, S.: Towards recognition of facial expressions in sign language: Tracking facial features under occlusion. In: Procs. of ICIP, pp. 3228 – 3231 (2008). DOI 10.1109/ICIP.2008.4712483
77. Nguyen, T.D., Ranganath, S.: Tracking facial features under occlusions and recognizing facial expressions in sign language. In: Procs. of FGR, pp. 1 – 7. Amsterdam, The Netherlands (2008). DOI 10.1109/AFGR.2008.4813464
78. Ong, E.J., Bowden, R.: Robust facial feature tracking using shape-constrained multiresolution selected linear predictors. IEEE TPAMI, Accepted, to Appear
79. Ong, E.J., Bowden, R.: A boosted classifier tree for hand shape detection. In: Procs. of FGR, pp. 889–894. Seoul, Korea (2004). DOI 10.1109/AFGR.2004.1301646
80. Ong, E.J., Bowden, R.: Detection and segmentation of hand shapes using boosted classifiers. In: Procs. of FGR. Seoul, Korea (2004)
81. Ong, E.J., Bowden, R.: Robust lip-tracking using rigid flocks of selected linear predictors. In: Procs. of FGR. Amsterdam, The Netherlands (2008)
82. Ong, S., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE TPAMI **27**(6), 873 – 891 (2005)
83. Ouhyoung, M., Liang, R.H.: A sign language recognition system using hidden markov model and context sensive search. In: Procs. of ACM Virtual Reality Software and Technology Conference, pp. 59 – 66 (1996)

84. Pahlevanzadeh, M., Vafadoost, M., Shahnazi, M.: Sign language recognition. In: Procs. of Int. Symposium on Signal Processing and Its Applications, pp. 1 – 4 (2007). DOI 10.1109/ISSPA.2007.4555448
85. Rezaei, A., Vafadoost, M., Rezaei, S., Daliri, A.: 3D pose estimation via elliptical fourier descriptors for deformable hand representations. In: Procs. of Int. Conf. on Bioinformatics and Biomedical Engineering, pp. 1871 – 1875 (2008). DOI 10.1109/ICBBE.2008.797
86. Roussos, A., Theodorakis, S., Pitsikalis, P., Maragos, P.: Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In: Workshop on Sign, Gesture and Activity, 11th European Conference on Computer Vision (ECCV) (2010)
87. Segen, J., Kumar, S.: Shadow gestures: 3D hand pose estimation using a single camera. In: Procs. of CVPR, vol. 1. Fort Collins, CO, USA (1999)
88. Shamaie, A., Sutherland, A.: A dynamic model for real-time tracking of hands in bimanual movements. In: Procs. of GW, pp. 172 – 179. Genova, Italy (2003)
89. Sheerman-Chase, T., Ong, E.J., Bowden, R.: Feature selection of facial displays for detection of non verbal communication in natural conversation. In: Procs. of ICCV : Wkshp : Human-Computer Interaction, pp. 1985 – 1992. Kyoto, Japan (2009)
90. Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden markov models. In: Procs. of Int. Symposium on Computer Vision, pp. 265 – 270 (1995). DOI 10.1109/ISCV.1995.477012
91. Starner, T., Weaver, J., Pentland, A.: Real-time american sign language recognition using desk and wearable computer based video. IEEE TPAMI **20**(12), 1371 – 1375 (1998)
92. Stein, D., Forster, J., Zelle, U., Dreuw, P., Ney, H.: Analysis of the german sign language weather forecast corpus. In: Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, pp. 225–230. Valletta, Malta (2010)
93. Stenger, B.: Template-based hand pose recognition using multiple cues. In: Procs. of ACCV, vol. 2, pp. 551 – 561. Springer, Hyderabad, India (2006)
94. Stenger, B., Mendonca, P., Cipolla, R.: Model-based 3D tracking of an articulated hand. In: Procs. of CVPR, vol. 2. Kauai, HI, USA (2001)
95. Stokoe, W.C.: Sign language structure: An outline of the visual communication systems of the american deaf. Studies in Linguistics: Occasional Papers **8**, 3 – 37 (1960)
96. Vogler, C., Goldenstein, S.: Analysis of facial expressions in american sign language. In: Procs. of Int. Conf. on Universal Access in Human-Computer Interaction. Las Vegas, Nevada, USA (2005)
97. Vogler, C., Goldenstein, S.: Facial movement analysis in ASL. Universal Access in the Information Society **6**(4), 363 – 374 (2008)
98. Vogler, C., Li, Z., Kanaujia, A., Goldenstein, S., Metaxas, D.: The best of both worlds: Combining 3D deformable models with active shape models. In: Procs. of ICCV, pp. 1 – 7. Rio de Janario, Brazil (2007)
99. Vogler, C., Metaxas, D.: Adapting hidden markov models for ASL recognition by using three-dimensional computer vision methods. In: Procs. of IEEEInt. Conf. on Systems, Man, and Cybernetics, vol. 1, pp. 156 – 161. Orlando, FL, USA (1997)
100. Vogler, C., Metaxas, D.: ASL recognition based on a coupling between HMMs and 3D motion analysis. In: Procs. of ICCV, pp. 363 – 369. IEEE Computer Society, Bombay, India (1998)
101. Vogler, C., Metaxas, D.: Parallel hidden markov models for american sign language recognition. In: Procs. of ICCV, vol. 1, pp. 116 – 122. Corfu, Greece (1999)
102. Vogler, C., Metaxas, D.: Handshapes and movements: Multiple-channel american sign language recognition. In: Procs. of GW, pp. 247 – 258. Springer, Genova, Italy (2003)
103. Waldron, M.B., Kim, S.: Isolated ASL sign recognition system for deaf persons. IEEE Transactions on Rehabilitation Engineering **3**(3), 261 – 271 (1995). DOI 10.1109/86.413199
104. Wang, C., Gao, W., Shan, S.: An approach based on phonemes to large vocabulary chinese sign language recognition. In: Procs. of FGR, pp. 411 – 416. Wshington, DC, USA (2002)
105. Wong, S.F., Cipolla, R.: Real-time interpretation of hand motions using a sparse bayesian classifier on motion gradient orientation images. In: Procs. of BMVC, vol. 1, pp. 379 – 388. Oxford, UK (2005)

106. Yacoob, Y., Davis, L.: Recognizing human facial expressions from long image sequences using optical-flow. IEEE TPAMI **18**(6), 636 – 642 (1996)
107. Yamaguchi, T., Yoshihara, M., Akiba, M., Kuga, M., Kanazawa, N., Kamata, K.: Japanese sign language recognition system using information infrastructure. In: Procs. of IEEEInt. Conf. on Fuzzy Systems, vol. 5, pp. 65 – 66 (1995). DOI 10.1109/FUZZY.1995.410043
108. Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. IEEE TPAMI **31**(7), 1264 – 1277 (2009). DOI 10.1109/TPAMI.2008.172
109. Yang, M.H., Ahuja, N., Tabb, M.: Extraction of 2D motion trajectories and its application to hand gesture recognition. IEEE TPAMI **24**, 1061 – 1074 (2002)
110. Yin, P., Starner, T., Hamilton, H., Essa, I., Rehg, J.M.: Learning the basic units in american sign language using discriminative segmental feature selection. In: Procs. of ASSP, pp. 4757 – 4760. Taipei, Taiwan (2009). DOI 10.1109/ICASSP.2009.4960694
111. Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Ney, H.: Geometric features for improving continuous appearance-based sign language recognition. In: Procs. of BMVC, p. III:1019. Edinburgh, UK (2006)
112. Zahedi, M., Keysers, D., Deselaers, T., Ney, H.: Combination of tangent distance and an image based distortion model for appearance-based sign language recognition. In: Procs. of German Association for Pattern Recognition Symposium, *LNCS*, vol. 3663, p. 401. Springer, Vienna, Austria (2005)
113. Zahedi, M., Keysers, D., Ney, H.: Appearance-based recognition of words in american sign language. In: Procs. of IbPRIA, vol. 1, pp. 511 – 519. Estoril, Portugal (2005)
114. Zhang, L., Chen, Y., Fang, G., Chen, X., Gao, W.: A vision-based sign language recognition system using tied-mixture density HMM. In: Procs. of Int. Conf. on Multimodal interfaces, pp. 198 – 204. ACM New York, NY, USA, State College, PA, USA (2004)
115. Zieren, J., Kraiss, K.: Non-intrusive sign language recognition for human computer interaction. In: Procs. of IFAC/IFIP/IFORS/IEA symposium on analysis, design and evaluation of human machine systems (2004)
116. Zieren, J., Kraiss, K.: Robust person-independent visual sign language recognition. In: Procs. of IbPRIA, pp. 520 – 528. Springer, Estoril, Portugal (2005)