Chapter 27
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.

27.1 Motivation

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

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. So research has thus far focussed

H. Cooper (✉) · B. Holt · R. Bowden
University of Surrey, Guildford, GU2 7XH, UK
e-mail: H.M.Cooper@surrey.ac.uk

B. Holt
e-mail: B.Holt@surrey.ac.uk

R. Bowden
e-mail: R.Bowden@surrey.ac.uk

T.B. Moeslund et al. (eds.), *Visual Analysis of Humans*,
DOI 10.1007/978-0-85729-997-0_27,

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. Due to the expense of Human translation and 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.

27.2 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 (NMFs) 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. For instance, British Sign Language (BSL) grammatical structure 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 [10]. 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 ASL by Liddell and Johnson [67] model sign language on the movement-hold system. Signs are broken into sections where an aspect is changing and sections where a state is held steady. This

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

is in contrast to the work of Stokoe [102] 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.

1. *Adverbs modifying verbs*; signers would not use two signs for 'run quickly' they would modify the sign for run by speeding it up.
2. *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.
3. *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.
4. *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.
5. *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.
6. *Positional signs*; where a sign acts on the part of the body descriptively. e.g. 'bruise' or 'tattoo'.
7. *Body shift*; represented by twisting the shoulders and gaze, often used to indicate role-shifting when relating a dialogue.
8. *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.
9. *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 it is true, that continuous SLR is to become reality.

27.3 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 [88].

Many early SLR systems used datagloves 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 [115] or DataGlove [57, 108]. More often than not, the sensor input was of sufficient discriminatory power that feature extraction was bypassed and the measurements used directly as features [35]. 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 [46], attempted to address this problem. However, due to 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 [129], stereo [50], orthogonal [97]) or other non-invasive sensors. Segen and Kumar [94] used a camera and calibrated light source to compute depth, and Feris et al. [32] 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. [98] 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. [76] or by using side/vertical mounted cameras as with Vogler and Metaxas [111] or the Boston ASL data set [79]. Most recently the Microsoft Kinect™ has offered an affordable depth camera which has made depth a viable option for more researchers. However, at present there are no data sets available and as such the results are limited. 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 15 yr period [21] or the similar project on a smaller scale in the UK by the BSL Corpus Project [11]. However, these data sets are directed at linguistic research, whereas the cross domain European project DictaSign [22] aims to produce a multi-lingual data set suitable for both linguists and computer vision scientists.

Once data have been acquired they are described via features; the features chosen often depend on the elements of sign language being detected.

27.3.1 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 gesture recognition was performed by Mitra and Acharya [74] giving an overview of the field as it stood in 2007. While many gesture recognition techniques are applicable, Sign language offers a more complex challenge than the traditionally more confined domain of gesture recognition.

27.3.1.1 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 [56]. 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 [49] or different coloured fingers [47]. Zhang et al. [128] 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 [4, 52] where a common restriction is long sleeves. Skin colour detection is also used to identify the face position such as in [122]. Often this task is further simplified by restricting the background to a specific colour (chroma keying) [51] or at the very least keeping it uncluttered and static [97]. Zieren and Kraiss [130] explicitly modelled the background which aids the foreground segmentation task. Depth can be used to allow simplification of the problem. Hong et al. [50] and Grzeszcuk et al. [39] 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 [34] and Hadfield and Bowden [41] 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 Imagawa et al. [52] who skin segmented the head and hands before applying a Kalman filter during tracking. Han et al. [43] 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. [48] 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 [1] improved on the original colour based skin segmented tracker, by using a combination of skin segmentation and MHIs to find the hands for tracking. Awad et al. [4] 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 [72] proposed an alternative to colour segmentation, detecting the component parts of the body using Ong and Bowden's detector [86] 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 used labelled data to initialise colour models, head/torso detector and 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 constrained adjacent frames and as such several passes could be made before the final trajectory is extracted. An alternative to this is the solution proposed by Zieren and Kraiss [130] 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, at the cost of processing time. Maintaining a trajectory after the hands have interacted also poses a problem. Shamaie and Sutherland [95] 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.

The desire for tracked data means that the Kinect™ device has offered the sign recognition community a short-cut to real-time performance. Doliotis et al. have shown that using the Kinect™ tracking in place of their previous skin based results boosts performance from 20% to 95% on a complex data set of ten number gestures [23]. In the relatively short time since its release several proof of concept demonstrations have emerged. Ershaed et al. have focussed on Arabic sign language and have created a system which recognises isolated signs. They present a system working for 4 signs and recognise some close up handshape information [29]. At ESIEA they have been using Fast Artificial Neural Networks to train a system which recognises two French signs [117]. This small vocabulary is a proof of concept but it is unlikely to be scalable to larger lexicons. One of the first videos to be uploaded to the web came from Zafrulla et al. and was an extension of their previous CopyCat game for deaf children [124]. The original system uses coloured gloves and accelerometers to track the hands, this was replaced by tracking from the Kinect™. They use solely the upper part of the torso and normalise the skeleton according to arm length. They have an internal dataset containing 6 signs; 2 subject signs, 2 prepositions and 2 object signs. The signs are used in 4 sentences (subject, preposition, object) and they have recorded 20 examples of each. They list under further work that signer independence would be desirable which suggests that their dataset is single signer but this is not made clear. By using a cross validated system they train Hidden Markov Models (HMMs) (Via the Georgia Tech Gesture Toolkit [71]) to recognise the signs. They perform 3 types of tests, those with full grammar constraints getting 100%, those where the number of signs is known getting 99.98% and those with no restrictions getting 98.8%. Cooper et al. have also extended their previous work [18] to use the 3D tracking capabilities of the Kinect™ [19]. By employing hard coded sign phonemes for both motion and location they show results on two data sets of 20 Greek and 40 German signs respectively. They use sequential pattern boosting [84], which combined with the robust phoneme features, produces

a solution capable of signer independent recognition. This system has been received some evaluation by the Deaf community, the preliminary results of which are shown in [27].

27.3.1.2 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 [118] used 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 [126]. They then combined skin segmentation with five types of differencing for each frame in a sequence, all are down sampled to obtain features [127]. 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 [125]. This system is able to achieve 64% accuracy rates on a more complex subset of the Boston data set [79] including continuous sign from three signers. Cooper and Bowden [16] 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 [56] to use appearance and spatio-temporal features [15].

The variability of the signers also introduces problems, the temporal inconsistencies between signs are a good example of this. Corradini [20] 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 data set of exaggerated gestures which resemble traffic controls. It is unclear how well the DTW will port to the challenge of natural, continuous SLR.

27.3.1.3 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 datagloves the hand shape can be described in terms of joint angles and more generically finger openness as shown by Vogler and Metaxas [110]. 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 [55]. Others achieved

good results using vision based approaches. Ong and Bowden presented a combined hand detector and shape classifier using a boosted cascade classifier [83]. 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 [14] 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 Eigenspaces 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 [90] though the system is unlikely to be tractable. Fillbrandt et al. used 2D appearance models to infer 3D posture and shape of the hands [33]. 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 [42], 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 [69] 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 [3]. 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. [100] 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. [93] 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 [92]. They computed both loose point correspondences and 3D motion estimation, in order to create the full 3D motion trajectory and pose of the hands. In contrast

Guan et al. [40] 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. Oikonomidis et al. use the Kinect™ to acquire real-time depth information about the hand pose [82]. They then optimise the hand model parameters using a variant of Particle Swarm Optimization in order to match the current pose to a model. When using GPU coding they show promising results and are able to achieve frame rates of 15 fps. While this model fitting approach can give the parameters of the hand accurately it requires a further step to extract a known sign hand shape.

27.3.2 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 [53] 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 [70]. They combined HOG features with a 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 [37], which combined optical flow features with an HMM but which only encoded the co-articulation present in the data set lexicon. Jennings [54] 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 four cameras, stereo from two and colour from one; seven channels in total. The channels are combined using Bayesian framework that reduces to a sum of squared differences equation. Stenger et al. [101] 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. [32] 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.

Recently, Pugeault and Bowden have exploited the Kinect™ to create a real-time, interactive American Sign Language (ASL) finger spelling system [91]. Using a combination of both depth and colour streams the hand is segmented and Gabor features extracted. A random forest learnt to distinguish between letter hand shapes, gaining an average of 75%. They address the ambiguity between certain hand shapes with an interface allowing the user to choose between plausible letters.

27.3.3 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. [78] 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 [119]. 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 [75]. Reported accuracies are high but the approach is unlikely to be scalable to larger data sets 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) [68]. 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. [60] presented methods for recognising these individual action units using both extended MHIs 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 [85]) to recognise more abstract facial expressions, such as 'Understanding' or 'Thinking' [96]. They note that their more complex data set, 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 [73]. The training data were 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 [80]. In later work, they combined this with HMMs and an ANN to recognise four sign language expressions [81]. 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 [105–107]. Vogler and Goldstein approach the explicit problem of sign language facial expressions, using a deformable face model [105, 106]. 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 [62]. 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 [5]. 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) [6]. They combined the responses of block differencing weak classifiers with an ANN. They boosted the final classifiers by rebuilding the ANN 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 [45]. 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 [87]. This is then extended to include HMMs to recognise phonemes from the lips [63].

27.4 Recognition

While some machine learning techniques were covered briefly in Sect. 27.3.1.3, 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 sub-unit of sign.

27.4.1 Classification Methods

The earliest work on SLR applied ANNs. 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 [77] published one of the first papers on SLR. Their idea was to train an ANN 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 continuous recognition. Kim et al. [59] used datagloves to provide x,y,z coordinates as well as angles, from which they trained a Fuzzy Min Max ANN to recognise 25 isolated gestures with a success rate of 85%. Lee et al. [64] used a Fuzzy Min Max ANN to recognise the phonemes of continuous Korean Sign Language (KSL) with a vocabulary of 131 words as well as fingerspelling without modelling a grammar. Waldron and Kim [115] presented an isolated SLR system using ANNs. They trained a first layer ANN for each of the four sub-unit types present in the manual part of ASL, and then combined the results of the first layer in a second layer ANN that actually recognises the isolated words. Huang et al. [51] presented a simple isolated sign recognition system using a Hopfield ANN. Yamaguchi et al. [120] recognised a very small number of words using associative memory (similar to ANNs). Yang et al. [122] 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 [89]. The seminal work of Starner and Pentland [98] demonstrated that HMMs present a strong technique for recognising sign language and Grobel and Assan [38] presented a HMM based isolated sign (gesture) recognition system that performed well given the restrictions that it applied.

Vogler and Metaxas [108] 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. modelling transient movements) yield different results, with explicit modelling 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 are 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 [111]. In their following work, Vogler and Metaxas [109] 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. [58] 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 [65] presented a sign language recognition system that used data captured from a single DataGlove. A feature vector was constructed 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 [57] presented a sign language recognition system that used instance based learning k-Nearest Neighbors (KNNs) and decision tree learning to classify isolated signs using dataglove features. The results were not as high as ANN 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. [30] 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 [118], 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.

27.4.2 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 [108] borrowed heavily from the studies of sign language by Liddell and Johnson [67], splitting signs into motion and pause sections. While their later work [109], used PaHMMs on both hand shape and motion sub-units, as proposed by the linguist Stokoe [102]. Work has also concentrated on learning signs from low numbers of examples. Lichtenauer et al. [66] 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. [9] 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. [56] extended this work with head and hand detection based on boosting (cascaded weak classifiers), a body-centred 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 [15] 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 first 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. [123] 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. [61] and Han et al. [44] 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.

27.5 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.

27.5.1 Continuous Sign Recognition

The majority of work on SLR has been focussed 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 [108] and Gao et al. [36] modelled the movement epenthesis explicitly. Gao et al. [36] used datagloves 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. [121] 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.

27.5.2 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 [130] 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 [114] 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 [112], 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 MAP estimation approach, and finally showed how they can be combined to yield superior results.

27.5.3 Fusing Multi-modal Sign Data

From the review of SLR by Ong and Ranganath [88], 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 27.3.3, 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. [113] 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 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 Gebärdensprache/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 [114], 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. [2] 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.

27.5.4 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 [9, 48, 98, 116]. N-Gram grammars are often used to improve recognition rates, most often bi-gram [35, 47, 89] but also uni-gram [7]. Bungeroth and Ney [13] demonstrated that statistical sign language translation using the Bayes rule is possible and has the potential to be developed into a real-world translation tool. Bauer et al. [8] 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.

27.5.5 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 [31] 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. [12] 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 [17] 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. [99] 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.

27.6 Conclusion

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 gesture recognition 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 who have [2, 113] concentrate solely on the facial expressions of sign. There is still much to be explored in the field 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 [25, 114]. Sign video data once recognised can be compressed using SLR into an encoded form (e.g. Signing Gesture Markup Language (SiGML) [28]) 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.

27.6.1 Further Reading

Recommendations for further reading on sign-language recognition include: [88, 103, 104].

Acknowledgements This research was supported by funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231135 – DictaSign.

References

1. Akyol, S., Alvarado, P.: Finding relevant image content for mobile sign language recognition. In: Procs. of IASTED Int. Conf. on Signal Processing, Pattern Recognition and Application, pp. 48–52, Rhodes, Greece (3–6 July 2001) [543]
2. Aran, O., Burger, T., Caplier, A., Akarun, L. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. Pattern Recognit. Lett. **42**(5), 812–822 (2009) [554,556]
3. Athitsos, V., Sclaroff, S.: Estimating 3D hand pose from a cluttered image. In: Procs. of CVPR, vol. 2, Madison WI, USA (June 2003) [546]
4. 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, Hong Kong, China, pp. 239–242 (August 2006) [543]
5. Ba, S.O., Odobez, J.M.: Visual focus of attention estimation from head pose posterior probability distributions. In: Procs. of IEEE Int. Conf. on Multimedia and Expo, pp. 53–56 (2008) [549]
6. Bailly, K., Milgram, M.: Bisar: Boosted input selection algorithm for regression. In: Procs. of Int. Joint Conf. on Neural Networks, pp. 249–255 (2009) [549]

7. Bauer, B., Hienz, H., Kraiss, K. Video-based continuous sign language recognition using statistical methods. In: Procs. of ICPR, Barcelona, Spain, vol. 15, pp. 463–466 (September 2000) [555]
8. 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) [555]
9. 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, Prague, Czech Republic. LNCS, pp. 390–401, Springer, Berlin (11–14 May 2004) [552,555]
10. British Deaf Association: Dictionary of British Sign Language/English. Faber & Faber, London (1992) [540]
11. BSL Corpus Project. Bsl corpus project site (2010) [542]
12. Buehler, M., Everingham, P., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: Procs. of CVPR, Miami, FL, USA, pp. 2961–2968 (20–26 June 2009) [555]
13. Bungeroth, J., Ney, H.: Statistical sign language translation. In: Procs. of LREC: Wkshp: Representation and Processing of Sign Languages, Lisbon, Portugal, pp. 105–108 (26–28 May 2004) [555]
14. Coogan, T., Sutherland, A.: Transformation invariance in hand shape recognition. In: Procs. of ICPR, Hong Kong, China, pp. 485–488 (August 2006) [546]
15. Cooper, H., Bowden, R.: Large lexicon detection of sign language. In: Procs. of ICCV: Wkshp: Human–Computer Interaction, Rio de Janeiro, Brazil, pp. 88–97 (16–19 October 2007) [545,552]
16. Cooper, H., Bowden, R.: Sign language recognition using boosted volumetric features. In: Procs. of IAPR Conf. on Machine Vision Applications, Tokyo, Japan, pp. 359–362 (16–18 May 2007) [545]
17. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: Procs. of CVPR, Miami, FL, USA, pp. 2568–2574 (20–26 June 2009) [555]
18. Cooper, H., Bowden, R.: Sign language recognition using linguistically derived sub-units. In: Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Languages Technologies, Valetta, Malta, (17–23 May 2010) [544]
19. Cooper, H., Ong, E.-J., Bowden, R.: Give me a sign: An interactive sign dictionary. Technical report, University of Surrey (2011) [544]
20. 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, Vancouver, BC, pp. 82–90. IEEE Comput. Soc., Los Alamitos (9–12 July 2001) [545]
21. DGS-Corpus. Dgs-corpus website (2010) [542]
22. DictaSign Project. Dictasign project website (2010) [542]
23. Doliotis, P., Stefan, A., Mcmurrough, C., Eckhard, D., Athitsos, V.: Comparing gesture recognition accuracy using color and depth information. In: Conference on Pervasive Technologies Related to Assistive Environments (PETRA) (May 2011) [544]
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, Southampton, UK, pp. 293–298 (10–12 April 2006) [544,545]
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, San Diego, CA, USA, vol. 1, pp. 21–30, Springer, Berlin (19–24 July 2009) [556]
26. Ekman, P.: Basic emotions. In: Dalgleish, T., Power, T. (eds.) The Handbook of Cognition and Emotion, pp. 45–60. Wiley, New York (1999) [548]

27. Elliott, R., Cooper, H., Ong, E.-J., Glauert, J., Bowden, R., Lefebvre-Albaret, F.: Search-by-example in multilingual sign language databases. In: ACM SIGACCESS Conference on Computers and Accessibility (ASSETS): Sign Language Translation and Avatar Technology, Dundee, UK (23 October 2011) [545]
28. Elliott, R., Glauert, J., Kennaway, J., Parsons, K.: D5-2: SiGML Definition. ViSiCAST Project working document (2001) [556]
29. Ershaed, H., Al-Alali, I., Khasawneh, N., Fraiwan, M.: An arabic sign language computer interface using the Xbox Kinect. In: Annual Undergraduate Research Conf. on Applied Computing, Dubai, UAE (May 2011) [544]
30. Gaolin, F., Gao, W., Debin, Z.: Large vocabulary sign language recognition based on fuzzy decision trees. IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum., **34**(3), 305–314 (2004) [551]
31. Farhadi, A., Forsyth, D.: Aligning ASL for statistical translation using a discriminative word model. In: Procs. of CVPR, New York, NY, USA, pp. 1471–1476 (June 2006) [555]
32. Feris, R., Turk, M., Raskar, R., Tan, K., Ohashi, G.: Exploiting depth discontinuities for vision-based fingerspelling recognition. In: Procs. of CVPR: Wkshp, Washington, DC, USA vol. 10, IEEE Comput. Soc., Los Alamitos (June 2004) [542,547]
33. 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, Nice, France, pp. 181–186 (14–18 October 2003) [546]
34. Fujimura, K., Liu, X.: Sign recognition using depth image streams. In: Procs. of FGR, Southampton, UK, pp. 381–386 (10–12 April 2006) [543]
35. Gao, W., Ma, J., Wu, J., Wang, C.: Sign language recognition based on HMM/ANN/DP. Int. J. Pattern Recognit. Artif. Intell. **14**(5), 587–602 (2000) [542,555]
36. Gao, W., Fang, G., Zhao, D., Chen, Y.: Transition movement models for large vocabulary continuous sign language recognition. In: Procs. of FGR, Seoul, Korea, pp. 553–558 (17–19 May 2004) [553]
37. Goh, P., Holden, E.-J.: Dynamic fingerspelling recognition using geometric and motion features. In: Procs. of ICIP, pp. 2741–2744 (2006) [547]
38. Grobel, K., Assan, M.: Isolated sign language recognition using hidden Markov models. In: Procs. of IEEE Int. Conf. on Systems, Man, and Cybernetics, Orlando, FL, USA, vol. 1, pp. 162–167 (12–15 October 1997) [550]
39. Grzeszcuk, R., Bradski, G., Chu, M.H., Bouguet, J.Y.: Stereo based gesture recognition invariant to 3d pose and lighting. In: Procs. of CVPR, vol. 1 (2000) [543]
40. Guan, H., Chang, J.S., Chen, L., Feris, R.S., Turk, M.: Multi-view appearance-based 3D hand pose estimation, p. 154 [547]
41. Hadfield, S., Bowden, R.: Generalised pose estimation using depth. In: Procs. of ECCV Int. Wkshp: Sign, Gesture, Activity, Heraklion, Crete (5–11 September 2010) [543]
42. Hamada, Y., Shimada, N., Shirai, Y.: Hand shape estimation under complex backgrounds for sign language recognition. In: Procs. of FGR, Seoul, Korea, pp. 589–594 (17–19 May 2004) [546]
43. Han, J., Awad, G., Sutherland, A.: Automatic skin segmentation and tracking in sign language recognition. IET Comput. Vis. **3**(1), 24–35 (2009) [543]
44. Han, J.W., Awad, G., Sutherland, A.: Modelling and segmenting subunits for sign language recognition based on hand motion analysis. Pattern Recognit. Lett. **30**(6), 623–633 (2009) [552]
45. Heracleous, P., Aboutabit, N., Beautemps, D.: Lip shape and hand position fusion for automatic vowel recognition in cued speech for French. IEEE Signal Process. Lett. **16**(5), 339–342 (2009) [549]
46. Hernandez-Rebollar, J.L., Lindeman, R.W., Kyriakopoulos, N.: A multi-class pattern recognition system for practical finger spelling translation. In: Procs. of IEEE Int. Conf. on Multimodal Interfaces, p. 185. IEEE Comput. Soc., Los Alamitos (2002) [542]
47. Hienz, H., Bauer, B., Karl-Friedrich, K.: HMM-based continuous sign language recognition using stochastic grammars. In: Procs. of GW, Gif-sur-Yvette, France, pp. 185–196. Springer, Berlin (17–19 March 1999) [543,555]

48. Holden, E.J., Lee, G., Owens, R.: Australian sign language recognition. Mach. Vis. Appl. **16**(5), 312–320 (2005) [543,555]
49. Holden, E.J., Owens, R.: Visual sign language recognition. In: Procs. of Int. Wkshp: Theoretical Foundations of Computer Vision, Dagstuhl Castle, Germany. LNCS, vol. 2032, pp. 270–288. Springer, Berlin (12–17 March 2000) [543]
50. Hong, S., Setiawan, N.A., 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, Vietri sul Mare, Italy. LNCS, vol. 4692, p. 493. Springer, Berlin (12–14 September 2007) [542,543]
51. 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 (23–26 October 1995) [543,550]
52. Imagawa, K., Lu, S., Igi, S.: Color-based hands tracking system for sign language recognition. In: Procs. of FGR, Nara, Japan, pp. 462–467 (14–16 April 1998) [543]
53. Isaacs, J., Foo, J.S.: Hand pose estimation for American sign language recognition. In: Procs. of Southeastern Symposium on System Theory, Atlanta, GA, USA, pp. 132–136 (March 2004) [547]
54. 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, Corfu, Greece, pp. 152–160 (21–24 September 1999) [547]
55. Jerde, T.E., Soechting, J.F., Flanders, M.: Biological constraints simplify the recognition of hand shapes. IEEE Trans. Biomed. Eng. **50**(2), 265–269 (2003) [545]
56. Kadir, T., Bowden, R., Ong, E.J., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: Procs. of BMVC, Kingston, UK, pp. 939–948 (7–9 September 2004) [543,545,552]
57. Kadous, M.W.: 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) [542,551]
58. Kim, J.B., Park, K.H., Bang, W.C., Kim, J.S., 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) [551]
59. Kim, J.-S., Jang, W., Bien, Z.: A dynamic gesture recognition system for the Korean sign language (KSL). IEEE Trans. Syst. Man Cybern., Part B, Cybern. **26**(2), 354–359 (1996) [550]
60. Koelstra, S., Pantic, M., Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11), 1940–1954 (2010) [548]
61. Kong, W.W., Ranganath, S.: Automatic hand trajectory segmentation and phoneme transcription for sign language. In: Procs. of FGR, Amsterdam, The Netherlands, pp. 1–6 (17–19 September 2008) [552]
62. Krinidis, M., Nikolaidis, N., Pitas, I.: 3-d head pose estimation in monocular video sequences using deformable surfaces and radial basis functions. IEEE Trans. Circuits Syst. Video Technol. **19**(2), 261–272 (2009) [549]
63. Lan, Y., Harvey, R., Theobald, B.-J., Ong, E.-J., Bowdenn R.: Comparing visual features for lipreading. In: Procs. of Int. Conf. Auditory-visual Speech Processing, Norwich, UK (2009) [549]
64. 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 IEEE Int. Conf. on Fuzzy Systems, vol. 3, pp. 1463–1468 (1–5 July 1997) [550]
65. Liang, R.H., Ouhyoung, M.: A real-time continuous gesture recognition system for sign language. In: Procs. of FGR, Nara, Japan, pp. 558–567 (14–16 April 1998) [551]
66. Lichtenauer, J., Hendriks, E., Reinders, M.: Learning to recognize a sign from a single example. In: Procs. of FGR, Amsterdam, The Netherlands, pp. 1–6 (17–19 September 2008) [552]
67. Liddell, S.K., Johnson, R.E.: American sign language: The phonological base. Sign Lang. Stud. **64**, 195–278 (1989) [540,552]

68. Lien, J.-J.J., Kanade, T., Cohn, J., Li, C.-C.: Automated facial expression recognition based on FACS action units. In: Procs. of FGR, Nara, Japan, pp. 390–395 (14–16 April 1998) [548]
69. Liu, X., Fujimura, K.: Hand gesture recognition using depth data. In: Procs. of FGR, Seoul, Korea, pp. 529–534 (17–19 May 2004) [546]
70. Liwicki, S., Everingham, M.: Automatic recognition of fingerspelled words in British sign language. In: Procs. of CVPR, Miami, FL, USA, pp. 50–57 (20–26 June 2009) [547]
71. Lyons, K., Brashear, H., Westeyn, T.L., Kim, J.S., Starner, T.: GART: The gesture and activity recognition toolkit. In: Procs. of Int. Conf. HCI, Beijing, China, pp. 718–727 (July 2007) [544]
72. Micilotta, A., Bowden, R.: View-based location and tracking of body parts for visual interaction. In: Procs. of BMVC, Kingston, UK, pp. 849–858 (7–9 September 2004) [543]
73. 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) [548]
74. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. **37**(3), 311–324 (2007) [542]
75. 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 Janeiro, Brazil, (16–19 October 2007) [548]
76. Munoz-Salinas, R., Medina-Carnicer, R., Madrid-Cuevas, F.J., Carmona-Poyato, A.: Depth silhouettes for gesture recognition. Pattern Recognit. Lett. **29**(3), 319–329 (2008) [542]
77. 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 (1991) [550]
78. Murphy-Chutorian, E., Trivedi, M. M.: Head pose estimation in computer vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 607–626 (2009) [548]
79. Neidle, C.: National centre for sign language and gesture resources (2006) [542,545]
80. 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 (12–15 October 2008) [549]
81. Nguyen, T.D., Ranganath, S.: Tracking facial features under occlusions and recognizing facial expressions in sign language. In: Procs. of FGR, Amsterdam, The Netherlands, pp. 1–7 (17–19 September 2008) [549]
82. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3D tracking of hand articulations using Kinect. In: Procs. of BMVC, Dundee, UK (August 29 – September 10 2011) [547]
83. Ong, E.-J., Bowden, R.: A boosted classifier tree for hand shape detection. In: Procs. of FGR, Seoul, Korea, pp. 889–894 (17–19 May 2004) [546]
84. Ong, E.-J., Bowden, R.: Learning sequential patterns for lipreading. In: Procs. of BMVC, Dundee, UK (August 29 – September 10 2011) [544]
85. Ong, E.-J., Bowden, R.: Robust facial feature tracking using shape-constrained multiresolution selected linear predictors. IEEE Trans. Pattern Anal. Mach. Intell. **33**(9), 1844–1859 (September 2011). doi:10.1109/TPAMI.2010.205 [548]
86. Ong, E.-J., Bowden, R.: Detection and segmentation of hand shapes using boosted classifiers. In: Procs. of FGR, Seoul, Korea, (17–19 May 2004) [543]
87. Ong, E.-J., Bowden, R.: Robust lip-tracking using rigid flocks of selected linear predictors. In: Procs. of FGR, Amsterdam, The Netherlands (17–19 September 2008) [549]
88. Ong, S.C.W., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6), 873–891 (2005) [542, 554,556]
89. Ouhyoung, M., Liang, R.-H.. A sign language recognition system using hidden Markov model and context sensitive search. In: Procs. of ACM Virtual Reality Software and Technology Conference, pp. 59–66 (1996) [550,555]
90. Pahlevanzadeh, M., Vafadoost, M., Shahnazi, M.: Sign language recognition. In: Procs. of Int. Symposium on Signal Processing and Its Applications, pp. 1–4 (12–15 February 2007) [546]

91. Pugeault, N., Bowden, R.: Spelling it out: Real-time ASL fingerspelling recognition. In: Consumer Depth Cameras for Computer Vision (CDC4CV), Barcelona, Spain (7–11 November, 2011) [547]
92. 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 (16–18 May 2008) [546]
93. 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) [546]
94. Segen, J., Kumar, S.: Shadow gestures: 3D hand pose estimation using a single camera. In: Procs. of CVPR, vol. 1, Fort Collins, CO, USA (23–25 June 1999) [542]
95. Shamaie, A., Sutherland, A.: A dynamic model for real-time tracking of hands in bimanual movements. In: Procs. of GW, Genova, Italy, pp. 172–179 (15–17 April 2003) [544]
96. 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, Kyoto, Japan, pp. 1985–1992 (29 September – 2 October 2009) [548]
97. 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 (21–23 November 1995) [542,543]
98. Starner, T., Weaver, J., Pentland, A.: Real-time American sign language recognition using desk and wearable computer based video. IEEE Trans. Pattern Anal. Mach. Intell. **20**(12), 1371–1375 (1998) [542,550,555]
99. 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, Valletta, Malta, pp. 225–230 (May 2010) [555]
100. Stenger, B.: Template-based hand pose recognition using multiple cues. In: Procs. of ACCV, Hyderabad, India, vol. 2, pp. 551–561. Springer, Berlin (13–16 January 2006) [546]
101. Stenger, B., Mendonca, P.R.S., Cipolla, R.: Model-based 3D tracking of an articulated hand. In: Procs. of CVPR, Kauai, HI, USA, vol. 2 (December 2001) [547]
102. Stokoe, W.C.: Sign language structure: An outline of the visual communication systems of the American deaf. Stud. Linguist., Occas. Pap. **8**, 3–37 (1960) [541,552]
103. Sutton-Spence, R., Woll, B.: The Linguistics of British Sign Language: An Introduction. Cambridge University Press, Cambridge (1999) [556]
104. Valli, C., Lucas, C.: Linguistics of American Sign Language: An Introduction. Gallaudet University Press, Washington (2000) [556]
105. 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) [549]
106. Vogler, C., Goldenstein, S.: Facial movement analysis in ASL. Universal Access in the Information Society **6**(4), 363–374 (2008) [549]
107. 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, Rio de Janeiro, Brazil, pp. 1–7 (16–19 October 2007) [549]
108. Vogler, C., Metaxas, D.: Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. In: Procs. of IEEE Int. Conf. on Systems, Man, and Cybernetics, Orlando, FL, USA, vol. 1, pp. 156–161 (12–15 October 1997) [542,550, 552,553]
109. Vogler, C., Metaxas, D.: Parallel hidden Markov models for American sign language recognition. In: Procs. of ICCV, Corfu, Greece, pp. 116–122 (21–24 September 1999) [551,552]
110. Vogler, C., Metaxas, D.: Handshapes and movements: Multiple-channel American sign language recognition. In: Procs. of GW, Genova, Italy, pp. 247–258 (15–17 April 2003) [545]

111. Vogler, C., Metaxas, D.: ASL recognition based on a coupling between HMMs and 3D motion analysis. In: Procs. of ICCV, Bombay, India pp. 363–369. IEEE Comput. Soc., Los Alamitos (4–7 January 1998) [542,551]
112. 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, Tampa, Florida, USA, pp. 1–4 (8–11 December 2008) [553]
113. von Agris, U., Knorr, M., Kraiss, K.F.: The significance of facial features for automatic sign language recognition. In: Procs. of FGR, Amsterdam, The Netherlands, pp. 1–6 (17–19 September 2008) [554,556]
114. von Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent developments in visual sign language recognition. Univers. Access Inf. Soc. **6**(4), 323–362 (2008) [553,554,556]
115. Waldron, M.B., Kim, S.: Isolated ASL sign recognition system for deaf persons. IEEE Trans. Rehabil. Eng. **3**(3), 261–271 (1995) [542,550]
116. Wang, C., Gao, W., Shan, S.: An approach based on phonemes to large vocabulary Chinese sign language recognition. In: Procs. of FGR, Washington, DC, USA, pp. 411–416 (20–21 May 2002) [555]
117. Wassner, H.: Kinect + Reseau de Neurone = Reconnaissance de Gestes. http://tinyurl.com/5wbteug (May 2011) [544]
118. 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, Oxford, UK, vol. 1, pp. 379–388 (6–8 September 2005) [545,551]
119. Yacoob, Y., Davis, L.S.: Recognizing human facial expressions from long image sequences using optical-flow. IEEE Trans. Pattern Anal. Mach. Intell. **18**(6), 636–642 (1996) [548]
120. Yamaguchi, T., Yoshihara, M., Akiba, M., Kuga, M., Kanazawa, N., Kamata, K.: Japanese sign language recognition system using information infrastructure. In: Procs. of IEEE Int. Conf. on Fuzzy Systems, vol. 5, pp. 65–66 (20–24 March 1995) [550]
121. Yang, H.-D., Sclaroff, S., Lee, S.-W.: Sign language spotting with a threshold model based on conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell. **31**(7), 1264–1277 (2009) [553]
122. Yang, M.-H., Ahuja, N., Tabb, M.: Extraction of 2D motion trajectories and its application to hand gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 1061–1074 (2002) [543,550]
123. 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, Taipei, Taiwan, pp. 4757–4760 (19–24 April 2009) [552]
124. Zafrulla, Z., Brashear, H., Presti, P., Hamilton, H., Starner, T.: Copycat – Center for Accessible Technology in Sign. http://tinyurl.com/3tksn6s, http://www.youtube.com/watch?v=qFH5rSzmgFE&feature=related (2010) [544]
125. Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Ney, H.: Geometric features for improving continuous appearance-based sign language recognition. In: Procs. of BMVC, Edinburgh, UK, pp. 1019–1028 (4–7 September 2006) [545]
126. 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, Vienna, Austria. LNCS, vol. 3663, page 401, Springer, Berlin (31 August – 2 September 2005) [545]
127. Zahedi, M., Keysers, D., Ney, H.: Appearance-based recognition of words in American sign language. In: Procs. of IbPRIA, Estoril, Portugal, pp. 511–519 (7–9 June 2005) [545]
128. Zhang, L.G., 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, State College, PA, USA, pp. 198–204, ACM, New York (13–15 October 2004) [543]
129. Zieren, J., Kraiss, K.F., 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) [542]
130. Zieren, J., Kraiss, K.F., Robust person-independent visual sign language recognition. In: Procs. of IbPRIA, Estoril, Portugal, pp. 520–528 (7–9 June 2005) [543,544,553]