Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning

Sylvie C.W. Ong and Surendra Ranganath

Abstract—Research in automatic analysis of sign language has largely focused on recognizing the lexical (or citation) form of sign gestures as they appear in continuous signing, and developing algorithms that scale well to large vocabularies. However, successful recognition of lexical signs is not sufficient for a full understanding of sign language communication. Nonmanual signals and grammatical processes which result in systematic variations in sign appearance are integral aspects of this communication but have received comparatively little attention in the literature. In this survey, we examine data acquisition, feature extraction and classification methods employed for the analysis of sign language gestures. These are discussed with respect to issues such as modeling transitions between signs in continuous signing, modeling inflectional processes, signer independence, and adaptation. We further examine works that attempt to analyze nonmanual signals and discuss issues related to integrating these with (hand) sign gestures. We also discuss the overall progress toward a true test of sign recognition systems—dealing with natural signing by native signers. We suggest some future directions for this research and also point to contributions it can make to other fields of research. Web-based supplemental materials (appendicies) which contain several illustrative examples and videos of signing can be found at www.computer.org/publications/dlib.

Index Terms—Sign language recognition, hand tracking, hand gesture recognition, gesture analysis, head tracking, head gesture recognition, face tracking, facial expression recognition, review.

1 Introduction

In taxonomies of communicative hand/arm gestures, sign language (SL) is often regarded as the most structured of the various gesture categories. For example, different gesture categories have been considered as existing on a continuum, where gesticulation that accompanies verbal discourse is described as the least standardized and SL as the most constrained in terms of conventional forms that are allowed by the rules of syntax ([76], [94], Fig. 1a). In Quek's taxonomy ([112], Fig. 1b), gestures are divided into *acts* and *symbols*, and SL is regarded as largely symbolic, and possibly also largely referential since modalizing gestures are defined as those occuring in conjunction with another communication mode, such as speech. In this view, SL appears to be a small subset of the possible forms of gestural communication. Indeed SL is highly structured and most SL gestures are of a symbolic nature (i.e., the meaning is not transparent from observing the form of the gestures), but these taxonomies obscure the richness and sophistication of the medium. SL communication involves not only hand/arm gestures (i.e., manual signing) but also nonmanual signals (NMS) conveyed through facial expressions, head movements, body postures and torso movements.

Recognizing SL communication therefore requires simultaneous observation of these disparate body articulators and their precise synchronization, and information integration, perhaps utilizing a multimodal approach ([11], [155]). As such, SL communication is highly complex and understanding it involves a substantial commonality with research in machine analysis and understanding of human action and behavior; for example, face and facial expression recognition [84], [104], tracking and human motion analysis [53], [148], and gesture recognition [106]. Detecting, tracking and identifying people, and interpreting human behavior are the capabilities required of pervasive computing and wearable devices in applications such as smart environments and perceptual user interfaces [31], [107]. These devices need to be context-aware, i.e., be able to determine their own context in relation to nearby objects and humans in order to respond appropriately without detailed instructions. Many of the problems and issues encountered in SL recognition are also encountered in the research areas mentioned above; the structured nature of SL makes it an ideal starting point for developing methods to solve these problems.

Sign gestures are not all purely symbolic, and some are in fact *mimetic* or *deictic* (these are defined by Quek as *act* gestures where the movements performed relate directly to the intended interpretation). Mimetic gestures take the form of pantomimes and reflect some aspect of the object or activity that is being referred to. These are similar to *classifier signs* in American Sign Language (ASL) which can represent a particular object or person with the handshape and then act out the movements or actions of that object. Kendon [77] described one of the roles of hand gesticulations that accompany speech as providing images of the shapes of objects, spatial relations between objects or their paths of movement through space. These are in fact some of the same functions of classifier signs in ASL. A form of pantomime called constructed actions (role-playing or pespective shifting [23]) is also regularly used in SL discourse to relate stories about other people or places. Deictic or pointing gestures are extensively used in SL as

• *The authors are with the Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Singapore 117576. E-mail: {engp0560, elesr}@nus.edu.sg.*

Manuscript received 25 May 2004; revised 24 Sept. 2004; accepted 11 Oct. 2004; published online 14 Apr. 2005.
Recommended for acceptance by R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0257-0504.

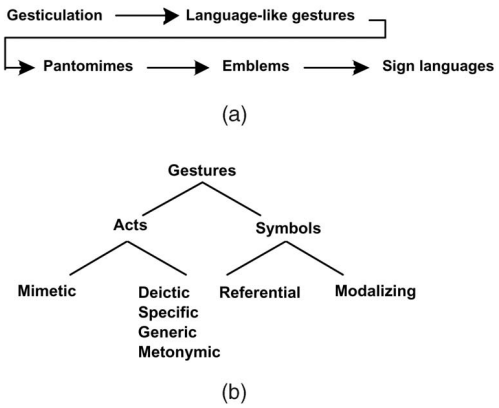

Fig. 1. Two different gesture taxonomies ([112] © 1994 World Scientific Publishing Co., reproduced with permission): (a) Kendon's continuum [94] and (b) Quek's taxonomy [112].

pronouns or to specify an object or person who is present or to specify an absent person by pointing to a previously established referrant location. Hence, designing systems that can automatically recognize classifier signs, pointing gestures, and constructed actions in signing would be a step in the direction of analyzing gesticulation accompanying speech and other less structured gestures. SL gestures also offer a useful benchmark for evaluating hand/arm gesture recognition systems. Non-SL gesture recognition systems often deal with small, limited vocabularies which are defined to simplify the classification task. SL(s), on the other hand, are naturally developed languages as opposed to artificially defined ones and have large, well-defined vocabularies which include gestures that are difficult for recognition systems to disambiguate.

One of the uses envisioned for SL recognition is in a sign-to-text/speech translation system. The complete translation system would additionally require machine translation from the recognized sequence of signs and NMS to the text or speech of a spoken language such as English. In an ideal system, the SL recognition module would have a large and general vocabulary, be able to capture and recognize manual information and NMS, perform accurately in real-time and robustly in arbitrary environments, and allow for maximum user mobility. Such a translation system is not the only use for SL recognition systems however, and other useful applications where the system requirements and constraints may be quite different, include the following:

- Translation or complete dialog systems for use in specific transactional domains such as government offices, post offices, cafeterias, etc. [2], [95], [116], [119]. These systems may also serve as a user interface to PCs or information servers [9]. Such systems could be useful even with limited vocabulary and formulaic phrases, and a constrained data input environment (perhaps using direct-measure device gloves [49], [116] or colored gloves and constrained background for visual input [2]).
- Bandwidth-conserving communication between signers through the use of avatars. Sign input data recognized at one end can be translated to a notational system (like HamNoSys) for transmission and synthesized into animation at the other end of the channel. This represents a great saving in bandwidth as compared to transmitting live video of a human signer. This concept is similar to a system for computer-generated signing developed under the Visicast project ([78]) where text content is translated to SiGML (Signing Gesture Markup Language, based on HamNoSys) to generate parameters for sign synthesis. Another possibility is creating SL documents for storage of recognized sign data in the form of sign notations, to be played back later through animation.
- Automated or semiautomated annotation of video databases of native signing. Linguistic analyses of signed languages and gesticulations that accompany speech require large-scale linguistically annotated corpora. Manual transcription of such video data is time-consuming, and machine vision assisted annotation would greatly improve efficiency. Head tracking and handshape recognition algorithms [99], and sign word boundary detection algorithms [83] have been applied for this purpose.
- Input interface for augmentative communication systems. Assistive systems which are used for human-human communication by people with speech-impairments often require keyboard or joystick input from the user [14]. Gestural input involving some aspects of SL, like handshape for example, might be more user friendly.

In the following, Section 2 gives a brief introduction to ASL, illustrating some aspects relevant to machine analysis. ASL is extensively used by the deaf communities of North America and is also one of the most well-researched among sign languages—by sign linguists as well as by researchers in machine recognition. In Section 3, we survey work related to automatic analysis of manual signing. Hand localization and tracking, and feature extraction in vision-based methods are considered in Sections 3.1 and 3.2, respectively. Classification schemes for sign gestures are considered in Section 3.3. These can be broadly divided into schemes that use a single classification stage or those that classify components of a gesture and then integrate them for sign classification. Section 3.3.1 considers classification methods employed to classify the whole sign or to classify its components. Section 3.3.2 considers methods that integrate component-level results for sign-level classification. Finally, Section 3.4 discusses the main issues involved in classification of sign gestures. Analysis of NMS is examined in Section 4. The issues are presented in Section 4.1 together with works on body pose and movement analysis, while works related to facial expression analysis, head pose, and motion analysis are examined in Appendix D (which can be found at www.computer.org/publications/dlib). The integration of these different cues is discussed in Section 4.2. Section 5 summarizes the state-of-the-art and future work, and Section 6 concludes the paper.

2 American Sign Language—Issues Relevant to Automatic Recognition

Most research work in SL recognition has focused on classifying the lexical meaning of sign gestures. This is understandable since hand gestures do express the main information conveyed in signing. For example, from obser-

Fig. 2. Video stills from the sentence translated into English as: "Are you studying very hard?" (a) is from the sign YOU. (c), (d), (e), and (f) are from the sign STUDY. (b) is during the transition from YOU to STUDY.

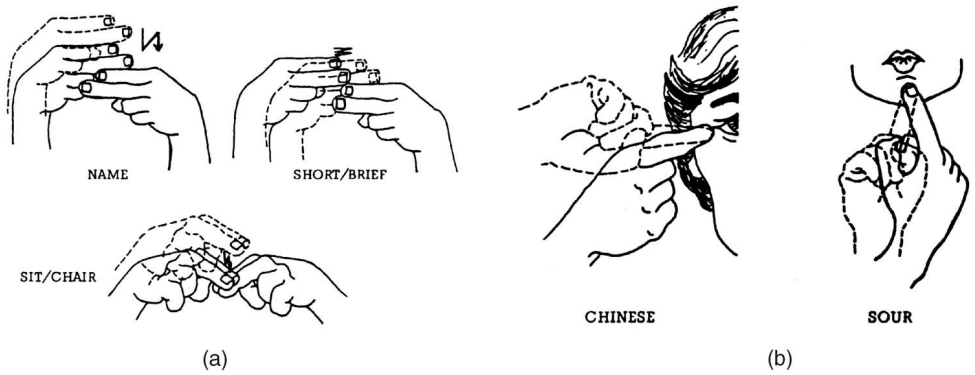

Fig. 3. (a) Minimal pairs of signs: NAME and SHORT are alike except for their movement trajectories, NAME and SIT are alike except for their hand orientations. (b) Two signs whose movement involves a twisting of the wrist without any gross whole-hand motion ([7] © Linstook Press, reprinted by permission).

ving the hand gestures in the sequence of Fig. 2, we can decipher the lexical meaning conveyed as "YOU STUDY."[1] However, without observing NMS and inflections in the signing, we cannot decipher the full meaning of the sentence as: "Are you studying very hard?" The query in the sentence is expressed by the body leaning forward, head thrust forward and raised eyebrows toward the end of the signed sequence (e.g., in Figs. 2e and 2f). To refer to an activity performed with great intensity, the lips are spread wide with the teeth visible and clenched; this co-occurs with the sign STUDY. In addition to information conveyed through these NMS, the hand gesture is performed repetitively in a circular contour with smooth motion. This continuous action further distinguishes the meaning as "studying" instead of "study." In the following sections, we will consider issues related to the lexical form of signs and point out some pertinent issues with respect to two important aspects of signing, viz; modifications to gestures that carry grammatical meaning, and NMS.

2.1 Manual Signing Expressing Lexical Meaning

Sign linguists generally distinguish the basic components (or phoneme subunits) of a sign gesture as consisting of the handshape, hand orientation, location, and movement. Handshape refers to the finger configuration, orientation to the direction in which the palm and fingers are pointing, and location to where the hand is placed relative to the body. Hand movement traces out a trajectory in space. The first phonological model, proposed by Stokoe [124], emphasized the simultaneous organization of these subunits. In contrast, Liddell and Johnson's Movement-Hold model [90] emphasized sequential organization. Movement segments were defined as periods during which some part of the sign is in transition, whether handshape, hand location, or orientation. Hold segments are brief periods when all these parts are static. More recent models ([22], [108], [118], [150]) aim to represent both the simultaneous and sequential structure of signs and it would seem that the computational framework adopted for SL recognition must similarly be able to model both structures. There are a limited number of subunits which combine to make up all the possible signs, for e.g., 30 handshapes, 8 hand orientations, 20 locations, and 40 movement trajectory shapes [90] (different numbers are proposed according to the phonological model adopted). Breaking down signs into their constituent parts has been used by various researchers for devising classification frameworks (Section 3.3.2). All parts are important as evidenced by the existence of minimal signs which differ in only one of the basic parts (Fig. 3a).

When signs occur in a continuous sequence to form sentences, the hand(s) need to move from the ending location of one sign to the starting location of the next. Simultaneously, the handshape and hand orientation also change from the ending handshape and orientation of one sign to the starting handshape and orientation of the next. These intersign transition periods are called movement epenthesis [90] and are not part of either of the signs. Fig. 2b shows a frame within the movement epenthesis—the right hand is transiting from performing the sign YOU to the sign STUDY. In continuous signing, processes with effects similar to co-articulation in speech do also occur, where the appearance of a sign is affected by the preceding and succeeding signs (e.g., hold deletion, metathesis, and assimilation [137]). However, these processes do not necessarily occur in all signs; for example, hold deletion is variably applied depending on whether the hold involves

1. Words in capital letters are sign glosses which represent signs with their closest meaning in English.

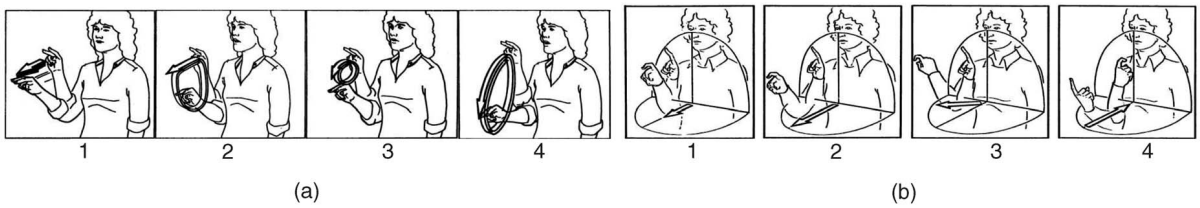

Fig. 4. Grammatical inflections of the sign "ASK": (a) Inflections for various temporal aspects ([109] © 1983 ACM, Inc., reprinted by permission): 1) habitual, i.e., "ask regularly," 2) iterative, i.e., "ask over and over again," 3) durational, i.e., "ask continuously," 4) continuative, i.e., "ask for a long time." (b) Inflections for person agreement (Reprinted by permission of the publisher from *The Signs of Language* by Edward Klima and Ursula Beluggi, Fig 12.2, Cambridge, Mass.: Harvard University Press, Copyright © 1979 by the President and Fellow of Harvard College): 1) Lexical form of "ASK," 2) "I ask you," 3) "I ask her/him," 4) "you ask me."

contact with a body part [90]. Hence, movement epenthesis occurs most frequently during continuous signing and should probably be tackled first by machine analysis, before dealing with the other phonological processes.

Some aspects of signing impact the methods used for feature extraction and classification, especially for vision-based approaches. First, while performing a sign gesture, the hand may be required to be at different orientations with respect to the signer's body and, hence, a fixed hand orientation from a single viewpoint cannot be assumed. Second, different types of movements are involved in signing. Generally, movement refers to the whole hand tracing a global 3D trajectory, as in the sign STUDY of Fig. 2 where the hand moves in a circular trajectory. However, there are other signs which involve local movements only, such as changing the hand orientation by twisting the wrist (e.g., CHINESE and SOUR, Fig. 3b) or moving the fingers only (e.g., COLOR). This imposes conflicting requirements on the field of view; it must be large enough to capture the global motion, but at the same time, small local movements must not be lost. Third, both hands often touch or occlude each other when observed from a single viewpoint and, in some signs, the hands partially occlude the face, as in the signs CHINESE, SOUR, and COLOR. Hence, occlusion handling is an important consideration.

2.2 Grammatical Processes in Sign Gestures

The systematic changes to the sign appearance during continuous signing described above (addition of movement epenthesis, hold deletion, metathesis, assimilation) do not change the sign meaning. However, there are other systematic changes to one or more parts of the sign which affect the sign meaning, and these are briefly described in this section.

In the sentence of Fig. 2, the sign STUDY is inflected for temporal aspect. Here, the handshape, orientation, and location of the sign are basically the same as in its lexical form but the movement of the sign is modified to show how the action (STUDY) is performed with reference to time. Examples of other signs that can be inflected in this way are WRITE, SIT, and SICK (Klima and Bellugi [81] lists 37 such signs). Fig. 4a shows examples of the sign ASK with different types of aspectual inflections. Generally, the meanings conveyed through these inflections are associated with aspects of the verbs that involve frequency, duration, recurrence, permanence, and intensity, and the sign's movement can be modified through its trajectory shape, rate, rhythm, and tension [81], [109]. Klima and Bellugi [81] list 8-11 different types of possible inflections for temporal aspect.

Another type of inflection that can occur is person agreement (first person, second person, or third person). Here, the verb indicates its subject and object by a change in the movement direction, with corresponding changes in its start and end locations, and hand orientation. Fig. 4b shows the sign ASK with different subject-object pairs. Other signs that can be similarly inflected include SHOW, GIVE, and INFORM (Padden [103] lists 63 such verbs). These signs can also be inflected to show the number of persons in the subject and/or object, or show how the verb action is distributed with respect to the individuals participating in the action ([81] lists 10 different types of number agreement and distributional inflections, including dual, reciprocal, multiple, exhaustive, etc.). Verbs can be simultaneously inflected for person and number agreement.

Other examples of grammatical processes which result in systematic variations in sign appearance include emphatic inflections, derivation of nouns from verbs, numerical incorporation, and compound signs. Emphatic inflections are used for the purpose of emphasis and are expressed through repetition in the sign's movement, with tension throughout. Appendix A (which can be found at www.computer.org/publications/dlib) has more details with illustrative photos and videos and discusses some implications for machine understanding. Classifier signs which can be constructed with innumerable variations are also discussed.

2.3 Nonmanual Signals—NMS

In the example of Fig. 2, two facial expressions were performed, with some overlap in their duration. Spreading the lips wide (Figs. 2c and 2d) is an example of using lower facial expressions, which generally provide information about a particular sign through use of the mouth area (lips, tongue, teeth, cheek) [23], [151]. In other examples, tongue through front teeth indicates that something is done carelessly, without paying attention; this can co-occur with a variety of signs like SHOP, DRIVING. Cheeks puffed out describes an object (e.g., TREE, TRUCK, MAN) as big or fat. The other facial expression shown in Fig. 2 depicts raised eyebrows and widened eyes (Figs. 2e and 2f), and is an example of using upper face expressions ([5], [137]), which often occur in tandem with head and body movements (in Figs. 2e and 2f the head and body are tilted forward). They generally convey information indicating emphasis on a sign or different sentence types (i.e., question, negation, rhetorical, assertion, etc.), and involve eye blinks, eye gaze direction, eyebrows, and nose. The eyebrows can be raised in surprise or to ask a question, contracted for emphasis or to show anger, or be drawn down in a frown. The head can tilt up with chin pressed forward, nod, shake or be thrust

forward. The body can lean forward or back, shift and turn to either side. Please refer to Appendix A (www.computer.org/publications/dlib) for more examples of NMS.

Although the description above has focused on ASL, similar use of NMS and grammatical processes occur in SL(s) of other countries, e.g., Japan [157], Taiwan [127], Britain [130], Australia [71], Italy [29], and France [19].

SL communication uses two-handed gestures and NMS; understanding SL therefore involves solving problems that are common to other research areas and applications. This includes tracking of the hands, face and body parts, feature extraction, modeling and recognition of time-varying signals, multimodal integration of information, etc. Due to the interconnectedness of these areas, there is a vast literature available, but our intention here is to only provide an overview of research specific to SL recognition.

3 Automatic Analysis of Hand Gestures in Signing

Hand gesture data is mainly acquired using cameras or direct-measure (glove-based) devices (as surveyed in [125]). Appendix B (www.computer.org/publications/dlib) gives some of the considerations in using the two data acquisition methods.

3.1 Vision-Based Hand Localization and Tracking

In order to capture the whole signing space, the entire upper body needs to be in the camera's field-of-view (FOV). The hand(s) must be located in the image sequence and this is generally implemented by using color, motion, and/or edge information. If *skin-color* detection is used, the signer is often required to wear long-sleeved clothing, with restrictions on other skin-colored objects in the background ([1], [67], [68], [120], [123], [158], [162]). Skin-color detection was combined with motion cues in Akyol and Alvarado [1], Imagawa et al. [67], Yang et al. [158], and combined with edge detection in Terrillon et al. [136]. The hands were distinguished from the face with the assumption that the head is relatively static in [1], [67], [123], and that the head region is bigger in size in [158]. A multilayer perceptron neural network-based frontal face detector was used in [136] for the same purpose. Color cue has also been used in conjunction with colored gloves ([4], [8], [10], [131], [132]).

Motion cues were used in [32], [33], [65], [66], with the assumption that the hand is the only moving object on a stationary background and that the signer's torso and head are relatively still. Another common requirement is that the hand must be constantly moving. In Chen et al. [25] and Huang and Jeng [66], the hand was detected by logically ANDing difference images with edge maps and skin-color regions. In Cui and Weng's system [32], [33], an outline of the motion-detected hand was obtained by mapping partial views of the hand to previously learned hand contours, using a hierarchical nearest neighbor decision rule. This yielded 95 percent hand detection accuracy, but at a high computational cost (58.3s per frame).

Ong and Bowden [101] detected hands with 99.8 percent accuracy in grey scale images with *shape* information alone, using a boosted cascade of classifiers [140]. Signers were constrained to wear long-sleeved dark clothing, in front of mostly dark backgrounds. Tanibata et al. [135] extracted skin, clothes, head, and elbow region by using a very restrictive *person-specific template* that required the signer to be seated in a known initial position/pose. Some of the other works also localized the body torso ([4], [8], [10], [133]), elbow and shoulder ([61]), along with the hands and face, using color cues and knowledge of the body's geometry. This allowed the position and movement of the hands to be referenced to the signer's body.

Two-dimensional tracking can be performed using blob-based ([67], [123], [135]), view-based ([66]), or hand contour/boundary models ([25], [33], [65]), or by matching motion segmented regions ([158]). Particularly challenging is tracking in the presence of occlusion. Some works avoid the occurrence of occlusion entirely by their choice of camera angle ([158]), sign vocabulary ([25], [65], [66], [136]), or by having signs performed unnaturally so as to avoid occluding the face ([33]). In these and other works, the left hand and/or face may be excluded from the image FOV ([25], [65], [66], [133], [136]). Another simplification is to use colored gloves, whereby face/hand overlap becomes straightforward to deal with. In the case of unadorned hands, simple methods for tracking and dealing with occlusions are generally unsatisfactory. For example, prediction techniques are used to estimate hand location based on the model dynamics and previously known locations, with the assumption of small, continuous hand movement ([25], [67], [123], [135], [158]). Starner et al.'s [123] method of subtracting the (assumed static) face region from the merged face/hand blob can only handle small overlaps. Overlapping hands were detected, but, for simplicity, features extracted from the merged blob were assigned to both hands. In addition, the left/right hand labels were always assigned to the left and right-most hand blobs, respectively. Imagawa et al. [67] also had problems dealing with complex bimanual hand movements (crossing, overlapping and bouncing back) as Kalman filters were used for each hand without data association. Tracking accuracy of 82-97 percent was obtained in a lab situation but this degraded to as low as 74 percent for a published videotape [121] with realistic signing at natural speed and NMS (this violated their assumptions of small hand movement between adjacent frames and a relatively static head). Their later work [68] dealt with face/hand overlaps by applying a sliding observation window over the merged blob and computing the likelihood of the window subimage belonging to one of the possible handshape classes. Hand location was correctly determined with 85 percent success rate. Tanibata et al. [135] distinguished the hands and face in cases of overlap by using texture templates from previously found hand and face regions. This method was found to be unsatisfactory when the interframe change in handshape, face orientation, or facial expression was large.

The more robust tracking methods that can deal with fast, discontinous hand motion, significant overlap, and complex hand interactions do not track the hands and face separately, but rather apply probabilistic reasoning for simultaneous assignment of labels to the possible hand/face regions [162], [120]. In both these works, the assumption is that only the two largest skin-colored blobs other than the head could be hands (thus restricting other skin-colored objects in the background and requiring long-sleeved clothing). Zieren et al. [162] tracked (with 81.1 percent accuracy) both hands and face in video sequences of 152 German Sign Language (GSL) signs. Probabilistic reasoning using heuristic rules (based on multiple features such as relative positions of hands, sizes

TABLE 1
Imaging Restrictions and Constraints Used in Vision-Based Approaches Listed in Tables 2 and 3

(a)	long-sleeved clothing
(b)	colored-gloves
(c)	uniform background
(d)	complex but stationary background
(e)	head/face required to be stationary or have less movement than hands
(f)	constant movement of hands required
(g)	fixed body location and pose or specific initial hand location
(h)	left hand and/or face excluded from field-of-view
(i)	vocabulary restricted or unnatural signing to avoid overlapping hands or hand over face
(j)	field-of-view restricted to the hand which is kept at fixed orientation and distance to camera

of skin-colored blobs, and Kalman filter prediction) was applied for labeling detected skin-colored blobs. Sherrah and Gong [120] demonstrated similarly good results while allowing head and body movement with the assumption that the head can be tracked reliably [18]. Multiple cues (motion, color, orientation, size and shape of clusters, distance relative to other body parts) were used to infer blob identities with a Bayesian Network whose structure and node conditional probability distributions represented constraints of articulated body parts.

In contrast to the above works which use 2D approaches, Downton and Drouet [39] used a 3D model-based approach where they built a hierarchical cylindrical model of the upper body, and implemented a project-and-match process with detected edges in the image to obtain kinematic parameters for the model. Their method failed to track after a few frames due to error propagation in the motion estimates. There are also a few works that use multiple cameras to obtain 3D measurements, however at great computational cost. Matsuo et al. [93] used stereo cameras to localize the hands in 3D and estimate the location of body parts. Vogler and Metaxas [141] placed three cameras orthogonally to overcome occlusion, and used deformable models for the arm/hand in each of the three camera views.

With regard to background complexity, several works use uniform backgrounds ([4], [8], [10], [65], [93], [133], [158], [162]). Even with nonuniform background, background subtraction was usually not used to segment out the signer. Instead, the methods focused on using various cues to directly locate the hands, face, or other body parts with simplifying constraints. In contrast, Chen et al. [25] used background modeling and subtraction to extract the foreground within which the hand was located. This eases some imaging restrictions and constraints; [25] did not require colored gloves and long-sleeved clothing, and allowed complex cluttered background that included moving objects. However, the hand was required to be constantly moving.

The imaging restrictions and constraints encountered in vision-based approaches are listed in Table 1.

3.2 Feature Extraction and Parameter Estimation in the Vision-Based Approaches

Research has focused on understanding hand signing in SL or, in the more restrictive case, classification of fingerspelled alphabets and numbers. For the former, the FOV includes the upper body of the signer, allowing the hands the range of movement required for signing. For fingerspelling, the range of hand motion is very small and consists mainly of finger configuration and orientation information. For full signing scenarios, features that characterize whole hand location and movement as well as appearance features that result from handshape and orientation are extracted, whereas for fingerspelling only the latter features are used. Thus, for works where the goal is classification of fingerspelling or handshape ([3], [12], [17], [36], [55], [56], [156]), the entire FOV only contains the hand. In these works (with the exception of [156]), the hand is generally restricted to palm facing the camera, against a uniform background.

For full signing scenarios, a commonly extracted positional feature is the *center-of-gravity* of the hand blob. This can be measured in absolute image coordinates ([123]), relative to the face or body ([4], [8], [10], [82], [133], [135]), relative to the first gesture frame ([33]), or relative to the previous frame ([82]). Alternatively, *motion* features have been used to characterize hand motion, e.g., motion trajectories of hand pixels [158] or optical flow [25]. The above approaches extract measurements and features in 2D. In an effort to obtain 3D measurements, Hienz et al. [61] proposed a simple geometric model of the hand/arm to estimate the hand's distance to camera using the shoulder, elbow, and hand's 2D positions. Approaches which directly measure 3D position using multiple cameras provide better accuracy but at the cost of higher computational complexity. Matsuo et al.'s [93] stereo camera system found the 3D position of both hands in a body-centered coordinate frame. Volger and Metaxas' [141] orthogonal camera system extracted the 3D wrist position coordinates and orientation parameters relative to the signer's spine.

The variety of hand appearance features include: segmented hand images, binary hand silhouettes or hand blobs, and hand contours. *Segmented hand images* are usually normalized for size, in-plane orientation, and/or illumination ([33], [131]), and principal component analysis (PCA) is often applied for dimensionality reduction before further processing ([12], [36], [68], [156]). In Starner et al. [123] and Tanibata et al. [135], geometric moments were calculated from the *hand blob*. Assan and Grobel [4], Bauer and Kraiss [8], [10], calculated the sizes, distances, and angles between distinctly colored fingers, palm, and back of the hand. *Contour-based* representations include various translation, scale, and/or in-plane rotation invariant features such as, Fourier descriptors (FD) [25], [65], [132], size functions [56], the lengths of vectors from the hand centroid to the fingertips region [3], and localized contour sequences [55]. Huang and Jeng [66] represented hand contours with Active Shape Models [27], and extracted a modified Hausdorff distance measure [40] between the prestored

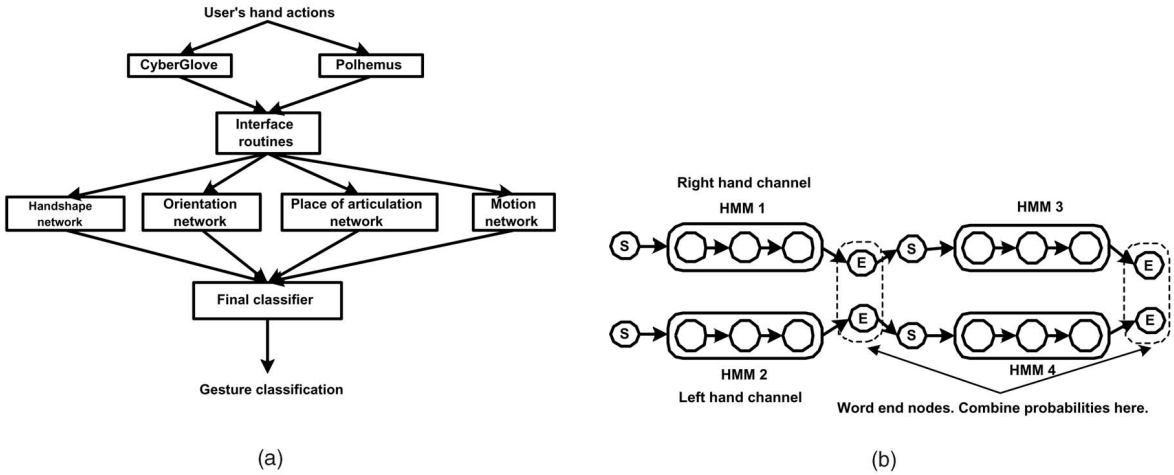

Fig. 5. Schemes for integration of component-level results: (a) System block diagram of a two-stage classification scheme by Vamplew ([138], reproduced with permission). (b) Parallel HMMs—tokens are passed independently in the left and right hand channels, and combined in the word end nodes (E). S denotes word start nodes (reprinted from [143] Copyright (2001), with permission from Elsevier).

shape models and the hand contour in the input test image. Bowden and Sahardi [17] used PCA on training hand contours, but constructed nonlinear Point Distribution Models by piecewise linear approximation with clusters. Hand contour tracking was applied on a fingerspelling video sequence, and the model transited between clusters with probabilities that reflected information about shape space and alphabet probabilities in English. Though contour-based representations use invariant features, they may generally suffer from ambiguities resulting from different handshapes with similar contours.

All of the above methods extracted 2D hand appearance features. In contrast, Holden and Owens [63] and Dorner [38] employed a 3D model-based approach to estimate finger joint angles and 3D hand orientation. In both works, finger joints and wrist were marked with distinct colors, and a 3D hand model was iteratively matched to the image content by comparing the projections of the hand model's joints with the corresponding joint markers detected in the image. Holden and Owens [63] could deal with missing markers due to the hand's self-occlusion by Kalman filter prediction. However, hand orientation was restricted to palm facing the camera. Dorner [38] estimated the hand model state based on constraints on the possible range of joint angles and state transitions, to successfully track in presence of out-of-plane rotations. However, processing speed was quite slow, requiring 5-6s per frame. In these and other works using 3D hand models ([100], [113]), the image FOV is assumed to contain only the hand with high resolution. In a sign recognition system however, the image FOV would contain the entire upper body; hence, the hand size would be small. In addition, these works do not consider situations when the hand is partially occluded (for example, by the other hand).

Fillbrandt et al. [48] attempt to address the shortcomings of the above approaches which directly find correspondence between image features and the 3D hand model. They used a network of 2D Active Appearance Models [28] as an intermediate representation between image features, and a simplified 3D hand model with 9 degrees-of-freedom. Experimental results with high-resolution images of the hand against uniform background yielded an average error of 10 percent in estimating finger parameters, while error for estimating the 3D hand orientation was 10°-20°. The system ran at 4 *fps* on a 1GHz Pentium III and they obtained some good results with low resolution images and partly missing image information. However, further work is needed before the model can be applied to a natural signing environment.

In terms of processing speed, methods that operate at near real-time for tracking and/or feature extraction (roughly 4-16 *fps*) include [1], [4], [8], [10], [61], [67], [123], [162]. Some of the other methods were particularly slow, for example: 1.6s per frame (PII-330M) for tracking in Sherrah and Gong [120], several seconds per frame for feature extraction in Tamura and Kawasaki [133], 58.3s per frame (SGI INDIGO 2 workstation) for hand segmentation in Cui and Weng [32], 60s for hand segmentation, and 70s for feature estimation in Huang and Jeng [66].

Direct-measure devices use trackers to directly measure the 3D position and orientation of the hand(s), and gloves to measure finger joint angles. More details on feature estimation from direct-measure devices can be found in Appendix C (www.computer.org/publications/dlib).

3.3 Classification Schemes for Sign Gestures

The two main approaches in sign gesture classification either employ a single classification stage, or represent the gesture as consisting of simultaneous components which are individually classified and then integrated together for sign-level classification. Fig. 5 shows examples of the latter approach. Fig. 5a ([138]) is a block diagram of the two-stage classification process while Fig. 5b ([144]) shows gesture components modeled as separate hidden Markov model (HMM) channels. Works listed in Tables 2 and 3 indicate the variety of classification schemes and features used under the two broad approaches. In each approach, methods which use both, direct-measure devices and vision are included. In Section 3.3.1, we summarize the main methods employed for classification. These same classification methods can either classify the sign directly or serve to classify any of the sign components. In Section 3.3.2, we see how results from component-level classification can be integrated to produce a final sign label.

TABLE 2
Selected Sign Gesture Recognition Systems Using Sign-Level Classification

Works	Sign vocab.	I/C	S/B	Classification methods	Subjects Train	Subjects Test	Rec. rate%
Direct-measure device approaches							
Fang[45]	208 CSL	C[1]	B	HMM,Self-organizing map,Recurrent NN	5	2	92.1
					5	1*	85
Kadous[72]	95 Auslan	I	S	Instance-based learning,Decision tree learning	5	5	80
					4	1*	12-15
Murakami[96]	10 JSL	I	S	Recurrent NN	1	1	96
Wang[147]	5119 CSL	I	B	HMMs model sequential subunits	1	1	92.8
		C[2]	B		1	1	86.2
Wu[153]	274 CSL	I	B	Semi-Continuous Dynamic Gaussian Mixture Model	1	1	97.4
Vision-based approaches							
Works	Sign vocab.	I/C	S/B	Restrictions† | Features extracted	Classification methods	Subjects Train	Subjects Test	Rec. rate%
Assan[4]	262 Nether.	I.	B	a,b,c | 2D moment-based	HMM	2	2	91.3
	26 Nether.	C[3]	B			1	1	72.8
Bauer[10]	100 GSL	I	B	a,b,c | 2D moment-based	HMMs model sequential	1	1	92.5
	50 GSL‡				subunits	1	1	81.0
Cui[33]	28 ASL	I	S	d,e,f,h,i | 2D segmented hand,position	Recursive PCA+MDA	?	?	93.2
Huang[65]	15 TWL	I	S	a,c,e,h,i | 2D FD, hand orientation,	3D Hopfield NN	?	?	96
				motion vector between frames		?	?*	91
Kobayashi[82]	6 JSL	I	S	unclear | 2D moment-based	Partly-HMM	20	20	98.8
Matsuo[93]	38 JSL	I	B	b,c | 3D hand position	Rule-based	1	1	76-79
Starner[123]	40 ASL	C[4]	B	a,d,e,g | 2D moment-based	HMM	1	1	92-98
Tanibata[135]	65 JSL	I	B	a,d,e,g | 2D moment-based+protrusions	HMMs model LH,RH	1	1	100#
Yang[158]	40 ASL	I	B	a,c | 2D pixel motion trajectories	Time-Delay NN	?	?	96.2

I/C:**I**solated or **C**ontinuous signing. S/B:**S**ingle hand or **B**oth hands. †Refer to Table 1. *Testing on unregistered signer(s). ‡This vocabulary set was not used to train subunits. #Used only test sequences with correctly extracted face and hands. [1]100 sentences each were used in registered and unregistered test sets. [2]Test set was 200 sentences, each consisting of 2-10 signs. [3]Testing with 3 sets of 14 different sentences, each consisting of 3-5 signs. [4]Test set of up to 100 sentences, consisting of 5 signs, a strongly constrained syntax was utilized in some of the experiments.

3.3.1 Classification Methods

Neural Networks and Variants. Multilayer perceptrons (MLP) are often employed for classifying handshape ([44], [50], [56], [96], [139], [145], [154]). Waldron and Kim [145], and Vamplew and Adams [139] additionally used MLPs to classify the hand location, orientation, and movement type from tracker data (see Fig. 5a). Other neural network (NN) variants include: Fuzzy Min-Max NNs ([122]) in [80], Adaptive Neuro-Fuzzy Inference System Networks ([69]) in [3], and Hyperrectangular Composite NNs in [126], all for handshape classification; and 3D Hopfield NN in [65] for sign classification.

Time-series data, such as movement trajectories and sign gestures, consist of many data points and have variable temporal lengths. NNs designed for classifying static data often do not utilize all the information available in the data points. For example, in classifying movement type, [145] used the displacement vectors at the start and midpoint of a gesture as input to the MLP, while [139] used only the accumulated displacement in each of the three primary axes of the tracker. Yang et al. [158] used Time-Delay NNs which were designed for temporal processing, to classify signs from hand pixel motion trajectories. As a small moving window of gesture data from consecutive time frames is used as input, only a small number of weights need to be trained (in contrast, HMMs often require estimation of many model parameters). The input data window eventually covers all the data points in the sequence, but a standard temporal length is still required. Murakami and Taguchi [96] used Recurrent NNs which can take into account temporal context without requiring a fixed temporal length. They considered a sign word to be recognized when the output node values remain unchanged over a heuristically determined period of time.

Hidden Markov models (HMMs) and variants. Several works classify sign gestures using HMMs which are widely used in continuous speech recognition. HMMs are able to process time-series data with variable temporal lengths and discount timing variations through the use of skipped-states and same-state transitions. HMMs can also implicitly segment continuous speech into individual words—trained word or phoneme HMMs are chained together into a branching tree-structured network and Viterbi decoding is used to find the most probable path through the network,

TABLE 3
Selected Sign Gesture Recognition Systems Using Component-Level Classification

Direct-measure device approaches

Works	Sign vocab.	No. of component categories HS	Mov.	Loc.	Orien.	I/C	S/B	Classification of component-level & sign-level	Subjects Train	Test	Rec. rate%
Hernandez[60]	176 ASL	42	7	11	6	I	S	Decision trees & Dict. lookup	17	1	94
Liang[89]	72-250	51	8	2	6	I	S	HMM & Dynamic prog.	1	1	78.4
	TWL					C[1]	S	w/ stochastic grammar	1	1	84.7
Sagawa[115]	17 JSL	not stated				C[2]	B	Matching with probs. & Dict. lookup	?	?	86.6
Su[126]	90 TWL	34	none	none	none	I	B	Hyperrectangular Composite NNs	2	2	94.1
								& Sum of similarity meas.	2	2*	91.2
Vamplew[139]	52 Auslan	30	13	19	15	I	S	Multilayer perceptron NN	7	4	94.2
								& Nearest-neighbor lookup	7	3*	85.3
Vogler[144]	22 ASL	71	140		none	C[3]	B	HMM & Parallel HMM	1	1	95.5

Vision-based approaches

Works	Sign vocab.	No. of component categories HS	Mov.	Loc.	Orien.	I/C	S/B	Restrictions† | Features extracted	Classification comp. & sign	Subjects Train	Test	Rec. rate%
Holden[63]	22 signs‡	22	none	none	none	I	S	b,c,j |3D hand model,21 DOF	Fuzzy rules w/ adaptive distribs.	1	1	95
Imagawa[68]	33 JSL	?	3		?	I	B	a,d,e |2D segmented hand	PCA+clustering & Dict. lookup	6	6	72-94
Tamura[133]	10 JSL	2	5	5	?	I	S	a,c,e,h |2D hand contour,position	Rule-based & Dict.lookup	?	?	45

I/C:**I**solated or **C**ontinuous signing. S/B:**S**ingle hand or **B**oth hands. †Refer to Table 1. *Testing on unregistered signer(s). ‡Included Auslan signs and artificial signs. [1]Test set was 345 sentences averaging 4.7 words in length. [2]100 sentences were used as test set.[3]Test set of 99 sentences,each consisting of 2-7 signs,in unconstrained (but grammatical) word order.

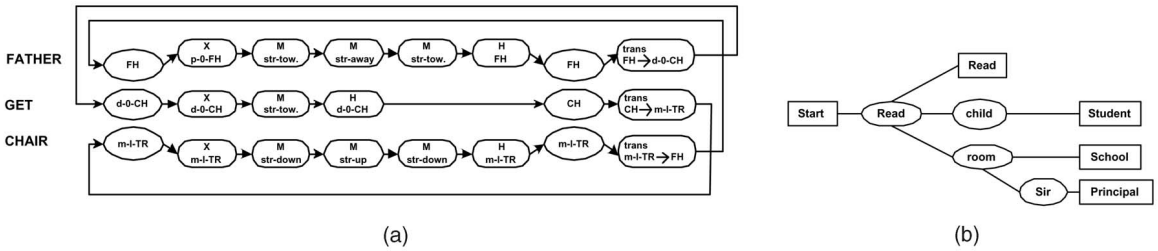

Fig. 6. (a) Network for the signs FATHER, GET and CHAIR. Epenthesis is modeled explicitly with HMMs (labeled with "trans"). The oval nodes are body locations e.g., FH (forehead), TR (trunk) (reprinted from [143] Copyright (2001), with permission from Elsevier). (b) Tree-structured network for the CSL signs READ, STUDENT, SCHOOL, PRINCIPAL, beginning with the phoneme (in Chinese) "READ" ([147] © 2002 IEEE, reproduced with permission).

thereby recovering both the word boundaries and the sequence. This idea has also been used for recognition of continuous signs, using various techniques to increase computational efficiency (some of which originated in speech recognition research [160]). These techniques include language modeling, beam search and network pruning ([8], [10], [50], [147]), N-best pass ([147]), fast matching ([50]), frame predicting ([50]), and clustering of Gaussians ([147]). Language models that have been used include unigram and bigram models in [50], [144], [147], as well as a strongly constrained parts-of-speech grammar in [95], [123]. As an alternative to the tree-structured network approach, Liang and Ouhyoung [89] and Fang et al. [45] explicitly segmented sentences before classification by HMMs (Section 3.4.1).

To reduce training data and enable scaling to large vocabularies, some researchers define sequential subunits, similar to phonetic acoustic models in speech, making every sign a concatentation of HMMs which model subunits. Based on an unsupervised method similar to one employed in speech recognition ([70]), Bauer and Kraiss [8] defined 10 subunits for a vocabulary of 12 signs using k-means clustering. Later, a bootstrap method [10] was introduced to get initial estimates for subunit HMM parameters and obtain the sign transcriptions. Recognition accuracy on 100 isolated signs using 150 HMM subunits was 92.5 percent. Encouragingly, recognition accuracy of 50 new signs without retraining the subunit HMMs was 81.0 percent. Vogler [144] (Fig. 6a), Yuan et al. [161] and Wang et al. [147] defined subunits linguistically instead of using unsupervised learning. [147] achieved 86.2 percent word accuracy in continuous sign recognition for a large vocabulary of 5,119 signs with 2,439 subunit HMMs. Fig. 6b ([147]) shows a tree structure built from these subunits to form sign words.

Kobayashi and Haruyama [82] argue that HMMs, which are meant to model piecewise stationary processes, are ill-suited for modeling gesture features which are always transient and propose the Partly hidden Markov model. Here the observation node probability is dependent on two states, one hidden and the other observable. Experimental results for isolated sign recognition showed a 73 percent improvement in error rate over HMMs. However, the vocabulary set of six Japanese Sign Language (JSL) signs is too small to draw concrete conclusions.

Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA). Birk et al. [12] and Imagawa et al. [68] both reduced dimensionality of segmented hand images by PCA before classification. Imagawa et al. [68] applied an unsupervised approach where training images were clustered in eigenspace and test images were classified to the cluster identity which gave the maximum likelihood score. Kong and Ranganath [85] classified 11 3D movement trajectories by performing periodicity detection using Fourier analysis, followed by Vector Quantization Principal Component Analysis [74]. Cui and Weng [33] used a recursive partition tree and applied PCA and MDA operations at each node. This method was able to achieve nonlinear classification boundaries in the feature space of 28 ASL signs. Deng and Tsui [36] found that when the entire data set is used for MDA, the performance degrades with increasing number of classes. To overcome this and to avoid strict division of data into partitions (as in [33]), they applied PCA and then performed crude classification into clusters with Gaussian distributions before applying MDA locally. The final classification of an input vector into one of 110 ASL signs took into account the likelihood of being in each of the clusters. Wu and Huang [156] aimed to overcome the difficulty of getting good results from MDA without a large labeled training data set. A small labeled data set and a large unlabeled data set were both modeled by the same mixture density, and a modified Discriminant-EM algorithm was used to estimate the mixture density parameters. A classifier trained with 10,000 unlabeled samples and 140 labeled samples of segmented hand images classified 14 handshapes with 92.4 percent accuracy, including test images where the hands had significant out-of-plane rotations. The above works mainly dealt with handshape classification ([12], [156]) or classification of signs based on just the beginning and ending handshape ([36], [68]). In [85] and [33] which classified movement trajectory and signs, respectively, mapping to a fixed temporal length was required.

Other methods. Some of the other methods that have been applied for classification of handshape are: decision trees ([59], [60]), nearest-neighbor matching ([86]), image template matching ([55], [131]), and correlation with phase-only filters from discrete Fourier transforms ([136]). Rule-based methods based on dictionary entries or decision trees have also been applied to classifying motion trajectories or signs ([61], [72], [73], [80], [93], [127]). Classification is by template matching with the ideal sequence of motion directions, or finding features (like concativity, change in direction) that characterize each motion type. The rules are usually hand-coded and, thus, may not generalize well. Wu and Gao [153] presented the *Semicontinuous Dynamic Gaussian Mixture Model* as an alternative to HMMs for processing temporal data, with the advantage of faster training time and fewer model parameters. This model was applied to recognizing sign words from a vocabulary of 274, but only using finger joint angle data (from two Cybergloves). They achieved fast recognition (0.04s per sign) and 97.4 percent accuracy.

3.3.2 Schemes for Integrating Component-Level Results

A common approach is to hand-code the categories of handshape, hand orientation, hand location, and movement type that make up each sign in the vocabulary, forming a lexicon of sign definitions. Classifying the sign label from component-level results is then performed by comparing the ideal lexicon categories with the corresponding recognized components ([60], [68], [80], [115], [126], [133], [139]). Various methods of performing this matching operation have been implemented; for example, Vamplew and Adams [139] employed a nearest-neighbor algorithm with a heuristic distance measure for matching sign word candidates. In Sagawa and Takeuchi [115], the dictionary entries defined the mean and variance (which were learned from training examples) of handshape, orientation, and motion type attributes as well as the degree of overlap in the timing of these components. Candidate sign words were then given a probability score based on the actual values of the component attributes in the input gesture data. In Su [126], scoring was based on an accumulated similarity measure of input handshape data from the first and last 10 sample vectors of a gesture. A major assumption was that signs can be distinguished based on just the starting and ending handshapes. Liang and Ouhyoung [89] classified all four gesture components using HMMs. Classification at the sign and sentence level was then accomplished using dynamic programming, taking into account the probability of the handshape, location, orientation, and movement components according to dictionary definitions as well as unigram and bigram probabilities of the sign gestures.

Methods based on HMMs include Gao et al. [50], where HMMs model individual sign words while observations of the HMM states correspond to component-level labels for position, orientation, and handshape, which were classified by MLPs. Vogler [144] proposed the *Parallel HMM* algorithm to model gesture components and recognize continuous signing in sentences. The right hand's shape, movement, and location, along with the left hand's movement and location were represented by separate HMM channels which were trained with relevant data and features. For recognition, individual HMM networks were built in each channel and a modified Viterbi decoding algorithm searched through all the networks in parallel. Path probabilities from each network that went through the same sequence of words were combined (Fig. 5b). Tanibata et al. [135] proposed a similar scheme where output probabilities from HMMs which model the right and left hand's gesture data were multiplied together for isolated word recognition.

Waldron and Kim [145] combined component-level results (from handshape, hand location, orientation, and movement type classification) with NNs—experimenting with MLPs as well as Kohonen self-organizing maps. The self-organizing map performed slightly worse than the MLP (83 percent versus 86 percent sign recognition accuracy), but it was possible to relabel the map to recognize new signs without requiring additional training data (experimental results were given for relabeling to accomodate two new signs). In an adaptive fuzzy expert

system ([30]) by Holden and Owens [63], signs were classified based on start and end handshapes and finger motion, using triangular fuzzy membership functions, whose parameters were found from training data.

An advantage of decoupling component-level and sign-level classification is that fewer classes would need to be distinguished at the component level. This conforms with the findings of sign linguists that there are a small, limited number of categories in each of the gesture components which can be combined to form a large number of sign words. For example, in Liang and Ouhyoung [89], the most number of classes at the component-level was 51 categories (for handshape), which is smaller than the 71 to 250 sign words that were recognized. Though some of these works may have small vocabularies (e.g., 22 signs in [144]), their focus, nevertheless, is on developing frameworks that allow scaling to large vocabularies. In general, this approach enables the component-level classifiers to be simpler, and with fewer parameters to be learned, due to the fewer number of classes to be distinguished and to the reduced input dimensions (since only the relevant component features are input to each classifier). In the works where sign-level classification was based on a lexicon of sign definitions, only training data for component-level classification was required and not at the whole-sign level ([60], [80], [89], [126], [133], [139], [144]). Furthermore, new signs can be recognized without retraining the component-level classifiers, if they cover all categories of components that may appear in signs. For example, the system in Hernandez-Rebollar et al. [60] trained to classify 30 signs, can be expanded to classify 176 new signs by just adding their descriptions into the lexicon.

In addition, approaches that do not require any training at the sign-level may be the most suitable for dealing with inflections and other grammatical processes in signing. As described in Section 2.2 and Appendix A (which can be found at www.computer.org/publications/dlib), the citation form of a sign can be systematically modified in one or more of its components to result in an inflected or derived sign form. This increases the vocabulary size to many more times than the number of lexical signs, with a correspondingly increased data requirement if training is required at the sign level. However, there is a limited number of ways in which these grammatical processes occur; hence, much less training data would be required if these processes could be recognized at the component level.

3.4 Main Issues in the Classification of Sign Gestures

The success of the works reported in the literature should not be measured just in terms of recognition rate but also in terms of how well they deal with the main issues involved in classification of sign gestures. In the following, we consider issues which apply to both vision-based and direct-measure device approaches. For a discussion of imaging environment constraints and restrictions, and feature estimation issues pertaining to vision-based approaches, the reader is referred to Sections 3.1 and 3.2.

3.4.1 Continuous Signing in Sentences

Tables 2 and 3 reveal that most of the works deal with isolated sign recognition where the user either performs the signs one at a time, starting and ending at a neutral position, or with exaggerated pauses, or while applying an external switch between each word. Extending isolated recognition to continuous signing requires automatic detection of word boundaries so that the recognition algorithm can be applied on the segmented signs. As such, valid sign segments where the movement trajectory, handshape, and orientation are meaningful parts of the sign need to be distinguished from movement epenthesis segments, where the hand(s) are merely transiting from the ending location and hand configuration of one sign to the start of the next sign. The general approach for explicit segmentation uses a subset of features from gesture data as cues for boundary detection. Sagawa and Takeuchi [115] considered a minimum in the hand velocity, a minimum in the differential of glove finger flexure values and a large change in motion trajectory angle as candidate points for word boundaries. Transition periods and valid word segments were further distinguished by calculating the ratio between the minimum acceleration value and maximum velocity in the segment—a minimal ratio indicated a word, otherwise a transition. In experiments with 100 JSL sentences, 80.2 percent of the word segments were correctly detected, while 11.2 percent of the transition segments were misjudged as words. In contrast, Liang and Ouhyoung [89] considered a sign gesture as consisting of a sequence of handshapes connected by motion and assumed that valid sign words are contained in segments where the time-varying parameters in finger flexure data dropped below a threshold. The handshape, orientation, location, and movement type in these segments were classified, while sections with large finger movement were ignored. The limitation of these methods which use a few gesture features as cues arises from the difficulty in specifying rules for determining sign boundaries that would apply in all instances. For example, [115] assumed that sign words are contained in segments where there is significant hand displacement and finger movement while boundary points are characterized by a low value in those parameters. However, in general, this may not always occur at sign boundaries. On the other hand, the method in [89] might miss important data for signs that involve a change in handshape co-occuring with a meaningful movement trajectory. A promising approach was proposed in Fang et al. [45] where the appropriate features for segmentation cues were automatically learned by a self-organizing map from finger flexure and tracker position data. The self-organizing map output was input to a Recurrent NN, which processed data in temporal context to label data frames as the left boundary, right boundary, or interior of a segment with 98.8 percent accuracy. Transient frames near segment boundaries were assumed to be movement epenthesis and ignored.

A few researchers considered segmentation in fingerspelling sequences, where the task is to mark points where valid handshapes occur. Kramer and Leifer [86] and Wu and Gao [154] performed handshape recognition during segments where there was a drop in the velocity of glove finger flexure data. Erenshteyn et al. [44] extracted segments by low-pass filtering and derivative analysis and discarded transitions and redundant frames by performing recognition only at the midpoint of these segments. Segmentation accuracy was 88-92 percent. Harling and Edwards [57] used the sum of finger tension values as a cue—a maximum indicated a valid handshape while a minimum indicated a transition. The finger tension values were calculated as a

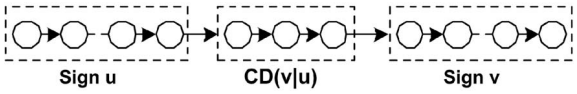

Fig. 7. Transitions between each unique pair of signs are explicitly modeled with a HMM ([50] © 2000 World Scientific Publishing Co., reproduced with permission).

function of finger-bend values. Birk et al. [12] recognized fingerspelling from image sequences and used frame differencing to discard image frames with large motion.

A popular approach for dealing with continuous signs without explicit segmentation as above is to use HMMs for implicit sentence segmentation (as mentioned in Section 3.3.1). In continuous speech recognition, coarticulation effects due to neighboring phonemes predominantly result in pronunciation variations. This is usually accounted for by modeling sounds in context—for example, triphones model a phoneme in the context of its preceding and succeeding phonemes, thereby tripling the number of HMM models required. The various methods that have been employed in dealing with sign transitions are generally different from the context-dependent models in speech. For example, Starner et al. [123] and Bauer and Kraiss [8] used one HMM to model each sign word (or subunit, in [8]) and trained the HMMs using data from entire sentences in an embedded training scheme ([159]), in order to incorporate variations in sign appearance during continuous signing. This would result in a large variation in the observations of the initial and ending states of a HMM due to the large variations in the appearance of all the possible movement epenthesis that could occur between two signs. This may result in loss of modeling accuracy for valid sign words. Wang et al. ([146], [147]) used a different approach where they trained HMMs on isolated words and subunits and chained them together only at recognition time, while employing measures to detect and discount possible movement epenthesis frames—signs were assumed to end in still frames, and the following frames were considered to be transition frames. This method of training with isolated sign data would not be able to accomodate processes where the appearance of a sign is affected by its context (e.g., hold deletion). Other works accounted for movement epenthesis by explicitly modeling it. In Assan and Grobel [4], all transitions between signs go through a single state, while in Gao et al. [50] separate HMMs model the transitions between each unique pair of signs that occur in sequence (Fig. 7). In more recent experiments [51], the number of such transition HMMs was reduced by clustering the transition frames. In Vogler [144], separate HMMs model the transitions between each unique ending and starting location of signs (Fig. 6a). In [50], [51] and [144], all HMM models are trained on data from entire sentences and, hence, in principle, variations in sign appearance due to context are accounted for. Volger [144] also assessed the advantage of explicit epenthesis modeling by making experimental comparisons with context-independent HMMs (as used in [123], [8]), and context-dependent biphone HMMs (one HMM is trained for every two valid combination of signs). On a test set of 97 sentences constructed from a 53-sign vocabulary, explicit epenthesis modeling was shown to have the best word recognition accuracy (92.1 percent) while context-independent modeling had the worst (87.7 percent versus 89.9 percent for biphone models). Yuan et al. [161] used HMMs for continuous sign recognition without employing a language model. They alternated word recognition with movement epenthesis detection. The ending data frame of a word was detected when the attempt to match subsequent frames to the word's last state produced a sharp drop in the probability scores. The next few frames were regarded as movement epenthesis if there was significant movement of a short duration and were discarded. Word recognition accuracy for sentences employing a vocabulary of 40 CSL signs was 70 percent.

3.4.2 Grammatical Processes in Sign Gestures

Generally, there have been very few works that address inflectional and derivational processes that affect the spatial and temporal dimensions of sign appearance in systematic ways (as described in Section 2.2 and Appendix A at www.computer.org/publications/dlib). HMMs, which have been applied successfully to lexical sign recognition, are designed to tolerate variability in the timing of observation features which are the essence of temporal aspect inflections. The approach of mapping each isolated gesture sequence into a standard temporal length ([33], [158]) causes loss of information on the movement dynamics. The few works that address grammatical processes in SL generally deal only with spatial variations. Sagawa and Takeuchi [114] deciphered the subject-object pairs of JSL verbs in sentences by learning the (Gaussian) probability densities of various spatial parameters of the verb's movement from training examples and, thus, calculated the probabilities of spatial parameters in test data. Six different sentences constructed from two verbs and three different subject-object pairs, which were tested on the same signer that provided the training set, was recognized with an average word accuracy of 93.4 percent. Braffort [19] proposed an architecture where HMMs were employed for classifying lexical signs using all the features of the sign gesture (glove finger flexure values, tracker location and orientation), while verbs which can express person agreement were classified by their movement trajectory alone and classifier signs were classified by their finger flexure values only. Sentences comprising seven signs from the three different categories were successfully recognized with 92-96 percent word accuracy. They further proposed a rule-based interpreter module to establish the spatial relationship between the recognized signs, by maintaining a record of the sign articulations around the signing space. Although they were not applied to sign recognition, Parametric HMMs were proposed in [152] to estimate parameters representing systematic variations such as the distance between hands in a two-handed gesture and movement direction in a pointing gesture. However, it is unclear whether the method is suitable for larger vocabularies that exhibit multiple simultaneous variations.

The works above only deal with a subset of possible spatial variations, with no straightforward extension to modeling systematic speed and timing variations. In Watanabe [149], however, both spatial size and speed information were extracted from two different musical conducting gestures with 90 percent success. This method first recognized the basic gesture using min/max points in the gesture trajectory and then measured the change in hand center-of-gravity between successive images to obtain gesture magnitude and speed information. In contrast, Ong

and Ranganath [102] proposed an approach which simultaneously recognized the lexical meaning and the inflected meaning of gestures using Bayesian Networks. Temporal and spatial movement aspects that exhibit systematic variation (specifically movement size, direction, and speed profile) were categorized into distinct classes. Preliminary experimental results on classification of three motion trajectory shapes (straight line, arc, circle) and four types of systematic temporal and spatial modifications (increases in speed and/or size, even and uneven rhythms) often encountered in ASL yielded 85 percent accuracy for eight test subjects.

3.4.3 Signer Independence

Analogous to speaker independence in speech recognition, an ideal sign recognition system would work "right out of the box," giving good recognition accuracy for signers not represented in the training data set (unregistered signers). Sources of interperson variations that could impact sign recognition accuracy include different personal signing styles, different sign usage due to geographical or social background ([137]), and fit of gloves in direct-measure device approaches. In this area, sign recognition lags far behind speech—many works report signer-dependent results where a single signer provided both training and test data ([10], [50], [63], [89], [93], [96], [123], [131], [133], [135], [144], [147], [153]), while other works have only 2 to 10 signers in the training and test set ([12], [36], [55], [56], [59], [61], [72], [126], [127], [139], [145]). The most number of test subjects was 20 in [25], [66], [82] and 60 for alphabet handshape recognition in [3]. This is still significantly less than the number of test speakers for which good results were reported in speech systems.

When the number of signers in the training set is small, results on test data from unregistered signers can be severely degraded. In Kadous [72], accuracy decreased from an average of 80 percent to 15 percent when the system that was trained on four signers was tested on an unregistered signer. In Assan and Grobel [4], accuracy for training on one signer and testing on a different signer was 51.9 percent compared to 92 percent when the same signer supplied both training and test data. Better results were obtained when data from more signers was used for training. In Vamplew and Adams [139], seven signers provided training data; test data from these same (registered) signers was recognized with 94.2 percent accuracy versus 85.3 percent accuracy for three unregistered signers. Fang et al. [45] trained a recognition system for continuous signing on five signers and obtained test data accuracy of 92.1 percent for these signers, compared to 85.0 percent for an unregistered signer. Classification accuracy for unregistered signers is also relatively good when only handshape is considered, perhaps due to less interperson variation as compared to the other gesture components. For example, [56] and [126] reported 93-96 percent handshape classification accuracy for registered signers versus 85-91 percent accuracy for unregistered signers. Interestingly, Kong and Ranganath [85] showed similarly good results for classifying 3D movement trajectories. Test data from six unregistered signers were classified with 91.2 percent accuracy versus 99.7 percent for test data from four registered signers.

In speech recognition, performance for a new speaker can be improved by using a small amount of data from the new speaker to adapt a prior trained system without retraining the system from scratch. The equivalent area of signer adaptation is relatively new. Some experimental results were shown in Ong and Ranganath [102] where speaker adaptation methods were modified to perform maximum a posteriori estimation [52] on component-level classifiers and Bayesian estimation of Bayesian Network parameters [58]. This gave 88.5 percent gesture recognition accuracy for test data from a new signer by adapting a system that was previously trained on three other signers —a 75.7 percent reduction in error rate as compared to using the unadapted system.

4 Analysis of Nonmanual Signals (NMS)

4.1 Issues

Broadly, the main elements of NMS in SL involve facial expressions, head and body pose, and movement. Often body and especially head movements co-occur with facial expressions (e.g., a question is asked by thrusting the head forward while simultaneously raising the eyebrows). The head could also tilt to the side or rotate left/right. This is further complicated by hand gestures being performed on or in front of the face/head region. Thus, tracking of the head is required while it is undergoing rigid motion, with possible out-of-plane rotation and occlusion by hands. Further, the face has to be distinguished from the hands. Recent surveys [47], [104] show much research interest in automatic analysis of facial expressions. However, these works generally cannot be directly applied to facial expressions in NMS due to their limited robustness and inability to characterize the temporal evolution of expressions. Most facial expression recognition approaches constrain faces to be fairly stationary and frontal to the camera. On the other hand, works that consider head tracking in less constrained environments do not include facial expression recognition. Black and Yacoob's local parametric model [13] is a notable exception—they successfully tracked facial features under significant rigid head motion and out-of-plane rotation and recognized six different expressions of emotions in video sequences.

Though facial expressions in NMS involve articulators that include the cheeks, tongue, nose and chin, most local feature-based approaches only consider the mouth, eyes and eyebrows (e.g., [13]). Facial expression has often been analyzed on static images of the peak expression, thereby ignoring the dynamics, timing, and intensity of the expression. This is not a good fit for NMS where different facial expressions are performed sequentially, and sometimes repetitively, evolving over a period of time. Thus, the timing of the expression in relation to the hand gestures produced, as well as the temporal evolution of the expression's intensity need to be determined. There are very few works that measure the intensity of facial expressions or which model the dynamics of expressions (examples of exceptions are [13], [46]).

In many works, facial expression recognition is limited to the six basic emotions as defined by Ekman [42]—happiness, sadness, surprise, fear, anger, disgust—plus the neutral expression, which involve the face as a whole. This is too constrained for NMS where the upper and lower face expressions can be considered to be separate, parallel channels of information that carry different grammatical

information or semantic meaning [151]. In this respect, the more promising approaches use a mid-level representation of facial action either defined by the researchers themselves ([13]) or which follow an existing coding scheme (MPEG-4 or Facial Action Coding System [37]). The recognition results of the mid-level representation code could in turn be used to interpret NMS facial expressions, in a fashion similar to ruled-based approaches which interpret recognized codes as emotion classes [13], [105].

A few works that consider facial expression analysis [24], [79], [117], [132] and head motion and pose analysis [43], [157] in the context of SL are described in Appendix D (www.computer.org/publications/dlib).

The body movements and postures involved in NMS generally consists of torso motion (without whole-body movement), for example, body leaning forwards/backwards or turning to the sides. So far, no work has specifically considered recognition of this type of body motion. Although there has been much work done in tracking and recognition of human activities that involve whole body movements, e.g., walking or dancing (as surveyed in [148]), these approaches may have difficulty in dealing with the subtler body motions exhibited in NMS.

4.2 Integration of Manual Signing and Nonmanual Signals

Results from the analysis of NMS need to be integrated with recognition results of the hand gestures in order to extract all the information expressed. Our search for works in automatic NMS analysis revealed none that capture the information from all the nonmanual cues of facial expression, head and body posture and movement. Some classify facial expression only [79], [117], [132], while others classify head movement only [43], [157]. Of these, there are only a couple of works which consider combining information extracted from nonmanual cues with results of gesture recognition. Ma et al. [92] modeled features extracted from lip motion and hand gestures with separate HMM channels using a modified version of Bourlard's multistream model [16] and resembling Vogler's Parallel HMM [144]. Viterbi scores from each channel are combined at sign boundaries where synchronization occurs. The different time scales of hand gestures and lip motion were accounted for by having different number of states for the same phrase/sign in each channel. In experiments where the lip motion expressed the same word (in spoken Chinese) as the gestured sign, 9 out of 10 phrases which were incorrectly recognized with hand gesture modeling alone, were correctly recognized when lip motion was also modeled.

There are several issues involved in integrating information from NMS with sign gesture recognition. In [92], the assumption was that each phrase uttered by the lips coincides with a sign/phrase in the gesture. However, in general NMS may co-occur with one or more signs/phrases, and hence a method for dealing with the different time scales in such cases is required. Also, in [92], the lip motion and hand gesturing convey identical information, while in general, NMS convey independent information, and the recognition results of NMS may not always serve to disambiguate results of hand gesture recognition. In fact, NMS often independently convey information in multiple channels through upper and lower face expressions, and head and body movements. Multiple cameras may be required to capture the torso's movement and still obtain good resolution images of the face for facial expression analysis. While some of the schemes employed in general multimodal integration research might be useful for application to this domain, we note that most of these schemes involve at most two channels of information, one of which is generally speech/voice ([11], [155]). It remains to be seen whether these can be applied to the multiple channels of information conveyed by NMS and hand gesturing in SL.

5 Discussion

In the Gesture Workshop of 1997, Edwards [41] identified two aspects of SL communication that had often been overlooked by researchers—facial expression and the use of space and spatial relationships in signing, especially with regard to classifier signs. In the ensuing period, although there has been some work to tackle these aspects, the focus of research continues to be elsewhere and hence progress has been limited. Among the facial expression recognition works surveyed, none were capable of recognizing and interpreting upper face and lower face expressions from video sequences, while simultaneously modeling the dynamics and intensity of expressions. A few works recognize head movements, particularly nods and shakes, but none interpret the body movements in NMS. Apart from [92] which sought to improve sign gesture recognition results by combining with lip reading, we are not aware of other work reporting results of integrating NMS and hand gestures. Works that interpret sign gestures whose form and manner of movement convey grammatical information mostly focused on spatial variations of the sign's movement. None of the works surveyed gave experimental results for intepretation of the mimetic classifier signs mentioned by Edwards [41] and Bossard et al. [15], [19]. It is obvious from the discussion in Section 3.4.2 that this aspect of signing has not received attention. Current systems that only consider the citation form of signs would miss important information conveyed in natural signing, such as movement dynamics that convey temporal aspect and spatial variations that convey subject-object agreement. Worse still, since current systems do not account for spatial relationships between signs, some signs would be completely undecipherable, for example classifier signs that describe spatial relationships between objects, or signs that point to a location that had previously been established as a referrant position. Noun-verb pairs like SEAT and SIT would be confused since the only difference between them is in the repetitive motion of the noun.

Two issues that have received much attention are recognition of continuous signing in sentences (Section 3.4.1) and scaling to large sign vocabularies. To handle large vocabularies with limited training data, some researchers used the idea of sequential subunits ([8], [10], [147], [161]), while others decomposed a sign gesture into its simultaneous components (Table 3). Notably, Vogler [144] did both—sign gestures were modeled as simultaneous, parallel channels of information which were each in turn modeled with sequential subunits. The largest vocabulary reported in experiments was 5,119 CSL signs in Wang et al. [147]. In contrast, many of the

other works are limited in the vocabulary size they can handle due to only using a subset of the information necessary for recognizing a comprehensive vocabulary. For example, it is common for input data to be from one hand only ([19], [25], [33], [36], [60], [63], [65], [66], [72], [82], [89], [96], [133], [139], [145]). Matsuo et al. [93] and Yang et al. [158] used input from both hands but only measured position and motion data. A few of the works used only hand appearance features as input without any position or orientation data ([126], [153], [36], [63], [66]). Even though all these works reported good results for sign recognition (possibly arising from either choice of vocabulary or some inherent information redundancy in gesture components), the existence of minimal sign pairs means that recognition of a comprehensive sign vocabulary is not possible without input from all the gesture components.

From Tables 2 and 3, we see that vision-based approaches have tended to experiment with smaller vocabulary sizes as compared to direct-measure device approaches. The largest vocabulary size used was 262 in the recognition of isolated signs of the Netherlands SL [4]. This could be due to the difficulty in simultaneously extracting whole hand movement features and detailed hand appearance features from images. Most works that localize and track hand movement, extract gross local features derived from the hand silhouette or contour. Thus, they may not be able to properly distinguish handshape and 3D hand orientation. Furthermore, handshape classification from multiple viewpoints is very difficult to achieve—Wu and Huang [156] being one of the few to do so, although on a limited number (14) of handshapes. Many of the vision-based approaches achieved fairly good recognition results but at the expense of very restrictive image capture environments and, hence, robustness is a real problem. An interesting direction to overcome this limitation was taken in the wearable system of Brashear et al. [21], where features from both vision and accelerometer data were used to classify signs. Signing was done in relatively unconstrained environments, i.e., while the signer was moving about in natural everyday settings. Continuous sentences constructed from a vocabulary of five signs were recognized with 90.5 percent accuracy, an improvement over using vision only data (52.4 percent) and accelerometer only data (65.9 percent). Low accuracy and precision in direct-measure devices can also affect recognition rate, a possibility in Kadous [72] as PowerGloves which have coarse sensing were used.

At present, it is difficult to directly compare recognition results reported in the literature. Factors that could influence results include restrictions on vocabulary (to avoid minimal pairs or signs performed near the face), slower than normal signing speed, and unnatural signing to avoid occlusion. Unfortunately, this kind of experimental information is usually not reported. Another important issue is that very few systems have used data from native signers. Some exceptions are Imagawa et al. [67] and Tamura and Kawasaki [133]. Tanibata et al. [135] used a professional interpreter. Braffort [20] made the point that the goal of recognizing natural signing requires close collaboration with native signers and SL linguists. Also, as the field matures, it is timely to tackle the problem of reproducibility by establishing standard databases. There are already some efforts in this direction. Neidle et al. [99] describe a corpus of native ASL signing that is being collected for the purpose of linguistic research as well as for aiding vision-based sign recognition research. Other efforts in this direction include [54], [75], [83].

We mentioned in the introduction that methods developed to solve problems in SL recognition can be applied to non-SL domains. An example of this is Nam and Wohn's work ([97]) on recognizing deictic, mimetic and pictographic gestures. Each gesture was broken down into attributes of handshape, hand orientation, and movement in a manner similar to decomposing sign gestures into their components. They further decomposed movement into sequential subunits of movement primitives and HMMs were employed to explicitly model connecting movements, similar to the approach in [144]. In [142], Vogler et al. applied the framework of decomposing movement into sequential subunits for the analysis of human gait. Three different gaits (walking on level terrain, up a slope, down a slope) were distinguished by analyzing all gaits as consisting of subunits (half-steps) and modeling the subunits with HMMs.

6 Conclusion

Automatic analysis of SL gestures has come a long way from its initial beginnings in merely classifying static signs and alphabets. Current work can successfully deal with dynamic signs which involve movement and which appear in continuous sequences. Much attention has also been focused on building large vocabulary recognition systems. In this respect, vision-based systems lag behind those that acquire gesture data with direct-measure devices. Robustness to the image capture environment is also an issue. Two aspects of gesture recognition that have not received much attention are building signer independent recognition systems and addressing the more difficult aspects of signing, such as grammatical inflections and mimetic signs. Furthermore, NMS have received scant attention. Understanding NMS and interpreting them in conjunction with gesture recognition is vital for understanding SL communication. Finding solutions for problems related to this would have application to research in recognition of other types of natural human activity or multimodal communication. Human activity and nonverbal communication involve visual cues which are similar to those involved in sign language communication in that these visual cues are a result of human body articulators moving in space. These cues are characterized by simultaneous occurrence with specific spatial relationships to each other, and interactions with various degrees of interdependencies. We as humans can grasp many things at once about our surroundings and other people's behavior—we need to teach computers how to do the same, and sign language recognition research appears to be a good place to start.

Acknowledgments

The authors are grateful for detailed and insightful comments from the anonymous reviewers which have been invaluable in revising the initial manuscript. They would like to thank Judy L.Y. Ho[2] from the Deaf and Hard-of-Hearing Federation (Singapore), for the example video sequences. Finally, the support of NUS research grant RP-263-000-214-112 is gratefully acknowledged.

2. Judy Ho is a graduate of Gallaudet University, Washington D.C., and has been using ASL and living within the deaf community for 16 years.

REFERENCES

[1] S. Akyol and P. Alvarado, "Finding Relevant Image Content for mobile Sign Language Recognition," *Proc. IASTED Int'l Conf. Signal Processing, Pattern Recognition and Application,* pp. 48-52, 2001.

[2] S. Akyol and U. Canzler, "An Information Terminal Using Vision Based Sign Language Recognition," *Proc. ITEA Workshop Virtual Home Environments,* pp. 61-68, 2002.

[3] O. Al-Jarrah and A. Halawani, "Recognition of Gestures in Arabic Sign Language Using Neuro-Fuzzy Systems," *Artificial Intelligence,* vol. 133, pp. 117-138, Dec. 2001.

[4] M. Assan and K. Grobel, "Video-Based Sign Language Recognition Using Hidden Markov Models," *Proc. Gesture Workshop,* pp. 97-109, 1997.

[5] C. Baker and C.A. Padden, "Focusing on the Nonmanual Components of American Sign Language," *Understanding Language through Sign Language Research,* P. Siple, ed., pp. 27-57, 1978.

[6] M.S. Bartlett, H.M. Lades, and T.J. Sejnowski, "Independent Component Representations for Face Recognition," *Proc. SPIE Conf. Human Vision and Electronic Imaging III,* vol. 3299, pp. 528-539, 1998.

[7] R. Battison, *Lexical Borrowing in American Sign Language.* Silver Spring, Md.: Linstok Press, 2003.

[8] B. Bauer and K.-F. Kraiss, "Towards an Automatic Sign Language Recognition System Using Subunits," *Proc. Gesture Workshop,* pp. 64-75, 2001.

[9] B. Bauer and K.-F. Kraiss, "Towards a 3rd Generation Mobile Telecommunication for Deaf People," *Proc. 10th Aachen Symp. Signal Theory Algorithms and Software for Mobile Comm.,* pp. 101-106, Sept. 2001.

[10] B. Bauer and K.F. Kraiss, "Video-Based Sign Recognition Using Self-Organizing Subunits," *Proc. Int'l Conf. Pattern Recognition,* vol. 2, pp. 434-437, 2002.

[11] M. Billinghurst, "Put that Where? Voice and Gesture at the Graphics Interface," *ACM SIGGRAPH Computer Graphics,* vol. 32, no. 4, pp. 60-63, Nov. 1998.

[12] H. Birk, T.B. Moeslund, and C.B. Madsen, "Real-Time Recognition of Hand Alphabet Gestures Using Principal Component Analysis," *Proc. Scandinavian Conf. Image Analysis,* pp. 261-268, 1997.

[13] M.J. Black and Y. Yacoob, "Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using Local Parametric Model of Image Motion," *Proc. Int'l Conf. Computer Vision,* pp. 374-381, 1995.

[14] *Augmentative Communication: An Introduction,* S. Blackstone, ed. Rockville, Md.: Am. Speech-Language-Hearing Assoc., 1986.

[15] B. Bossard, A. Braffort, and M. Jardino, "Some Issues in Sign Language Processing," *Proc. Gesture Workshop,* pp. 90-100, 2003.

[16] H. Bourlard, "Nonstationary Multi-Channel (Multi-Stream) Processing Towards Robust and Adaptive ASR," *Proc. Tampere Workshop Robust Methods for Speech Recognition in Adverse Conditions,* pp. 1-10, 1995.

[17] R. Bowden and M. Sarhadi, "A Nonlinear Model of Shape and Motion for Tracking Fingerspelt American Sign Language," *Image and Vision Computing,* vol. 20, pp. 597-607, 2002.

[18] G. Bradski, "Computer Vision Face Tracking for Use in Perceptual User Interface," *Intel Technical J.,* second quarter 1998.

[19] A. Braffort, "ARGo: An Architecture for Sign Language Recognition and Interpretation," *Proc. Gesture Workshop,* pg. 17-30, 1996.

[20] A. Braffort, "Research on Computer Science and Sign Language: Ethical Aspects," *Proc. Gesture Workshop,* pp. 1-8, 2001.

[21] H. Brashear, T. Starner, P. Lukowicz, and H. Junker, "Using Multiple Sensors for Mobile Sign Language Recognition," *Proc. Int'l Symp. Wearable Computers,* pp. 45-52, Oct. 2003.

[22] D. Brentari, "Sign Language Phonology: ASL," *The Handbook of Phonological Theory,* J.A. Goldsmith, ed., pp. 615-639, 1995.

[23] B. Bridges and M. Metzger, *Deaf Tend Your: Non-Manual Signals in American Sign Language.* Calliope Press, 1996.

[24] U. Canzler and T. Dziurzyk, "Extraction of Non Manual Features for Videobased Sign Language Recognition," *Proc. IAPR Workshop Machine Vision Application,* pp. 318-321, 2002.

[25] F.-S. Chen, C.-M. Fu, and C.-L. Huang, "Hand Gesture Recognition Using a Real-Time Tracking Method and Hidden Markov Models," *Image and Vision Computing,* vol. 21, no. 8, pp. 745-758, 2003.

[26] H.-I. Choi and P.-K. Rhee, "Head Gesture Recognition Using HMMs," *Expert Systems with Applications,* vol. 17, pp. 213-221, 1999.

[27] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models—Their Training and Application," *Computer Vision Image Understanding,* vol. 61, no. 1, pp. 38-59, 1995.

[28] T. Cootes, K. Walker, and C. Taylor, "View-Based Active Appearance Models," *Proc. Int'l Conf. Auto. Face & Gesture Recognition,* pp. 227-232, 2000.

[29] S. Corazza, "The Morphology of Classifier Handshapes in Italian Sign Language (LIS)," *Sign Language Research: Theoretical Issues,* C. Lucas, ed., Washington, D.C.: Gallaudet Univ. Press, 1990.

[30] E. Cox, "Adaptive Fuzzy Systems," *IEEE Spectrum,* pp. 27-31, Feb. 1993.

[31] J.L. Crowley, J. Coutaz, F. Berard, "Things That See," *Comm. ACM,* vol. 43, no. 3, pp. 54-64, Mar. 2000.

[32] Y. Cui and J. Weng, "A Learning-Based Prediction-and-Verification Segmentation Scheme for Hand Sign Image Sequence," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 21, no. 8, pp. 798-804, Aug. 1999.

[33] Y. Cui and J. Weng, "Appearance-Based Hand Sign Recognition from Intensity Image Sequences," *Computer Vision Image Understanding,* vol. 78, no. 2, pp. 157-176, 2000.

[34] *CyberGlove User's Manual.* Virtual Technologies Inc., 1995.

[35] *DataGlove Model 2 User's Manual.* Redwood City, Calif.: VPL Research Inc., 1987.

[36] J.-W. Deng and H.T. Tsui, "A Novel Two-Layer PCA/MDA Scheme for Hand Posture Recognition," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 283-286, 2002.

[37] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 21, no. 10, pp. 974-989, Oct. 1999.

[38] B. Dorner, "Chasing the Colour Glove: Visual Hand Tracking," Master's thesis, Simon Fraser Univ., 1994.

[39] A.C. Downton and H. Drouet, "Model-Based Image Analysis for Unconstrained Human Upper-Body Motion," *Proc. Int'l Conf Image Processing and Its Applications,* pp. 274-277, Apr. 1992.

[40] M.-P. Dubuisson and A.K. Jain, "A Modified Hausdorff Distance for Object Matching," *Proc. Int'l Conf. Pattern Recognition,* pp. 566-568, 1994.

[41] A.D.N. Edwards, "Progress in Sign Language Recognition," *Proc. Gesture Workshop,* pp. 13-21, 1997.

[42] P. Ekman, *Emotion in the Human Face.* Cambridge Univ. Press, 1982.

[43] U.M. Erdem and S. Sclaroff, "Automatic Detection of Relevant Head Gestures in American Sign Language Communication," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 460-463, 2002.

[44] R. Erenshteyn, P. Laskov, R. Foulds, L. Messing, and G. Stern, "Recognition Approach to Gesture Language Understanding," *Proc. Int'l Conf. Pattern Recognition,* vol. 3, pp. 431-435, 1996.

[45] G. Fang, W. Gao, X. Chen, C. Wang, and J. Ma, "Signer-Independent Continuous Sign Language Recognition Based on SRN/HMM," *Proc. Gesture Workshop,* pp. 76-85, 2001.

[46] B. Fasel and J. Luettin, "Recognition of Asymmetric Facial Action Unit Activities and Intensities," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 1100-1103, 2000.

[47] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition,* vol. 36, pp. 259-275, 2003.

[48] H. Fillbrandt, S. Akyol, and K.-F. Kraiss, "Extraction of 3D Hand Shape and Posture from Images Sequences from Sign Language Recognition," *Proc. Int'l Workshop Analysis and Modeling of Faces and Gestures,* pp. 181-186, 2003.

[49] W. Gao, J. Ma, S. Shan, X. Chen, W. Zheng, H. Zhang, J. Yan, and J. Wu, "HandTalker: A Multimodal Dialog System Using Sign Language and 3-D Virtual Human," *Proc. Int'l Conf. Advances in Multimodal Interfaces,* pp. 564-571, 2000.

[50] W. Gao, J. Ma, J. Wu, and C. Wang, "Sign Language Recognition Based on HMM/ANN/DP," *Int'l J. Pattern Recognition Artificial Intelligence,* vol. 14, no. 5, pp. 587-602, 2000.

[51] W. Gao, G. Fang, D. Zhao, and Y. Chen, "Transition Movement Models for Large Vocabulary Continuous Sign Language Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 553-558, 2004.

[52] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chain," *IEEE Trans. Speech and Audio Processing,* vol. 2, pp. 291-298, Apr. 1994.

[53] D. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision Image Understanding,* vol. 73 pp. 82-98, no. 1, Jan. 1999.

[54] S. Gibet, J. Richardson, T. Lebourque, and A. Braffort, "Corpus of 3D Natural Movements and Sign Language Primitives of Movement," *Proc. Gesture Workshop,* 1997.

[55] L. Gupta and S. Ma, "Gesture-Based Interaction and Communication: Automated Classification of Hand Gesture Contours," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Application Rev.,* vol. 31, no. 1, pp. 114-120, Feb. 2001.

[56] M. Handouyahia, D. Ziou, and S. Wang, "Sign Language Recognition Using Moment-Based Size Functions," *Proc. Int'l Conf. Vision Interface,* pp. 210-216, 1999.

[57] P.A. Harling and A.D.N. Edwards, "Hand Tension as a Gesture Segmentation Cue," *Proc. Gesture Workshop,* pp. 75-88, 1996.

[58] D. Heckerman, "A Tutorial on Learning with Bayesian Networks," technical report, Microsoft Research, Mar. 1995.

[59] J.L. Hernandez-Rebollar, R.W. Lindeman, and N. Kyriakopoulos, "A Multi-Class Pattern Recognition System for Practical Finger Spelling Translation," *Proc. Int'l Conf. Multimodal Interfaces,* pp. 185-190, 2002.

[60] J.L. Hernandez-Rebollar, N. Kyriakopoulos, and R.W. Lindeman, "A New Instrumented Approach for Translating American Sign Language into Sound and Text," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 547-552, 2004.

[61] H. Hienz, K. Grobel, and G. Offner, "Real-Time Hand-Arm Motion Analysis Using a Single Video Camera," *Proc Int'l Conf. Automatic Face and Gesture Recognition,* pp. 323-327, 1996.

[62] H. Hienz and K. Grobel, "Automatic Estimation of Body Regions from Video Image," *Proc. Gesture Workshop,* pp. 135-145, 1997.

[63] E.-J. Holden and R. Owens, "Visual Sign Language Recognition," *Proc. Int'l Workshop Theoretical Foundations of Computer Vision,* pp. 270-287, 2000.

[64] G. Hommel, F.G. Hofmann, and J. Henz, "The TU Berlin High-Precision Sensor Glove," *Proc. Fourth Int'l Scientific Conf.,* vol. 2, pp. F47-F49, 1994.

[65] C.-L. Huang and W.-Y. Huang, "Sign Language Recognition Using Model-Based Tracking and a 3D Hopfield Neural Network," *Machine Vision and Application,* vol. 10, pp. 292-307, 1998.

[66] C.-L. Huang and S.-H. Jeng, "A Model-Based Hand Gesture Recognition System," *Machine Vision and Application,* vol. 12, no. 5, pp. 243-258, 2001.

[67] K. Imagawa, S. Lu, and S. Igi, "Color-Based Hand Tracking System for Sign Language Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 462-467, 1998.

[68] K. Imagawa, H. Matsuo, R.-i. Taniguchi, D. Arita, S. Lu, and S. Igi, "Recognition of Local Features for Camera-Based Sign Language Recognition System," *Proc. Int'l Conf. Pattern Recognition,* vol 4, pp. 849-853, 2000.

[69] J.-S.R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 23, no. 3, pp. 665-685, May-June 1993.

[70] F. Jelinek, *Statistical Methods For Speech Recognition.* MIT Press, 1998.

[71] T. Johnston, "Auslan: The Sign Language of the Australian Deaf Community," PhD thesis, Dept. of Linguistics, Univ. of Sydney, 1989.

[72] M.W. Kadous, "Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language," *Proc. Workshop Integration of Gestures in Language and Speech,* pp. 165-174, 1996.

[73] M.W. Kadous, "Learning Comprehensible Descriptions of Multivariate Time Series," *Proc. Int'l Conf. Machine Learning,* pp. 454-463, 1999.

[74] N. Kambhatla and T.K. Leen, "Dimension Reduction by Local Principal Component Analysis," *Neural Computation,* vol. 9, no. 7, pp. 1493-1516, Oct. 1997.

[75] K. Kanda, A. Ichikawa, Y. Nagashima, Y. Kato, M. Terauchi, D. Hara, and M. Sato, "Notation System and Statistical Analysis of NMS in JSL," *Proc. Gesture Workshop,* pp. 181-192, 2001.

[76] A. Kendon, "How Gestures Can Become Like Words," *Cross-Cultural Perspectives in Nonverbal Comm.,* F. Poyatos, ed., pp. 131-141, 1988.

[77] A. Kendon, "Human Gesture," *Tools, Language, and Cognition in Human Evolution,* K.R. Gibson and T. Ingold, eds., pp. 43-62, Cambridge Univ. Press, 1993.

[78] R. Kennaway, "Experience with and Requirements for a Gesture Description Language for Synthetic Animation," *Proc. Gesture Workshop,* pp. 300-311, 2003.

[79] K.W. Ming and S. Ranganath, "Representations for Facial Expressions," *Proc. Int'l Conf. Control Automation, Robotics and Vision,* vol. 2, pp. 716-721, Dec. 2002.

[80] J.-S. Kim, W. Jang, and Z. Bien, "A Dynamic Gesture Recognition System for the Korean Sign Language (KSL)," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 26, no. 2, pp. 354-359, Apr. 1996.

[81] E.S. Klima and U. Bellugi, *The Signs of Language.* Harvard Univ. Press, 1979.

[82] T. Kobayashi and S. Haruyama, "Partly-Hidden Markov Model and Its Application to Gesture Recognition," *Proc Int'l Conf. Acoustics, Speech and Signal Processing,* vol. 4, pp. 3081-3084, 1997.

[83] A. Koizumi, H. Sagawa, and M. Takeuchi, "An Annotated Japanese Sign Language Corpus," *Proc. Int'l Conf. Language Resources and Evaluation,* vol. III, pp. 927-930, 2002.

[84] S.G. Kong, J. Heo, B.R. Abidi, J. Paik, and M.A. Abidi, "Recent Advances in Visual and Infrared Face Recognition—A Review," *Computer Vision Image Understanding,* 2004.

[85] W.W. Kong and S. Ranganath, "3-D Hand Trajectory Recognition for Signing Exact English," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 535-540, 2004.

[86] J. Kramer and L. Leifer, "The Talking Glove: An Expressive and Receptive Verbal Communication Aid for the Deaf, Deaf-Blind, and Nonvocal," *Proc. Third Ann. Conf. Computer Technology, Special Education, Rehabilitation,* pp. 335-340, Oct. 1987.

[87] V. Krüger, A. Happe, and G. Sommer, "Affine Real-Time Face Tracking Using Gabor Wavelet Networks," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 127-130, Sept. 2000.

[88] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, Reliable Head Tracking Under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 22, no. 4, pp. 322-336, Apr. 2000.

[89] R.-H. Liang and M. Ouhyoung, "A Real-Time Continuous Gesture Recognition System for Sign Language," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 558-565, 1998.

[90] S.K. Liddell and R.E. Johnson, "American Sign Language: The Phonological Base," *Sign Language Studies,* vol. 64, pp. 195-277, 1989.

[91] S.K. Liddell, *Grammar, Gesture, and Meaning in American Sign Language.* Cambridge Univ. Press, 2003.

[92] J. Ma, W. Gao, and R. Wang, "A Parallel Multistream Model for Integration of Sign Language Recognition and Lip Motion," *Proc. Int'l Conf. Advances in Multimodal Interfaces,* pp. 582-589, 2000.

[93] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, and T. Teshima, "The Recognition Algorithm with Non-Contact for Japanese Sign Language Using Morphological Analysis," *Proc. Gesture Workshop,* pp. 273-285, 1997.

[94] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought.* Univ. of Chicago Press, 1992.

[95] R.M. McGuire, J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear, and D.S. Ross, "Towards a One-Way American Sign Language Translator," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 620-625, 2004.

[96] K. Murakami and H. Taguchi, "Gesture Recognition Using Recurrent Neural Networks," *Proc. SIGCHI Conf. Human Factors in Computing Systems,* pp. 237-242, 1991.

[97] Y. Nam and K.Y. Wohn, "Recognition and Modeling of Hand Gestures Using Colored Petri Nets," *IEEE Trans. Systems, Man, and Cybernetics, Part A,* vol. 29, no. 5, pp. 514-521, Sept. 1999.

[98] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee, *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure.* MIT Press, 2000.

[99] C. Neidle, S. Sclaroff, and V. Athitsos, "SignStream: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data," *Behavior Research Methods, Instruments and Computers,* vol. 33, no. 3, pp. 311-320, 2001.

[100] C. Nolker and H. Ritter, "Visual Recognition of Continuous Hand Postures," *IEEE Trans. Neural Networks,* vol. 13, no. 4, pp. 983-994, July 2002.

[101] E.-J. Ong and R. Bowden, "A Boosted Classifier Tree for Hand Shape Detection," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 889-894, 2004.

[102] S.C.W. Ong and S. Ranganath, "Deciphering Gestures with Layered Meanings and Signer Adaptation," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 559-564, 2004.

[103] C.A. Padden, "Interaction of Morphology and Syntax in American Sign Language, " doctoral dissertation, Univ. of Calif., San Diego, 1983.

[104] M. Pantic and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[105] M. Pantic and L.J.M. Rothkrantz, "Expert System for Automatic Analysis of Facial Expressions," *Image and Vision Computing J.*, vol. 18, no. 11, pp. 881-905, 2000.
[106] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 19, no. 7, pp. 677-695, July 1997.
[107] A. Pentland, "Looking at People: Sensing for Ubiquitous and Wearable Computing," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 22, no. 1, pp. 107-119, Jan. 2000.
[108] D. Perlmutter, "Sonority and Syllable Structure in American Sign Language," *Phonetics and Phonology: Current issues in ASL phonology,* G. Coulter, ed., vol. 3, pp. 227-261, 1993.
[109] H. Poizner, E.S. Klima, U. Bellugi, and R.B. Livingston, "Motion Analysis of Grammatical Processes in a Visual-Gestural Language," *Proc. ACM SIGGRAPH/SIGART Interdisciplinary Workshop,* pp. 271-292, 1983.
[110] *Polhemus 3Space User's Manual.* Colchester, Vt.: Polhemus, 1991.
[111] *Power Glove Serial Interface (2.0 ed.),* Student Chapter of the ACM, Univ. of Illinois at Urbana-Champaign, 1994.
[112] F. Quek, "Toward a Vision-Based Hand Gesture Interface," *Proc. Virtual Reality Sofware and Technical Conf.,* pp. 17-29, 1994.
[113] J.M. Rehg and T. Kanade, "Visual Tracking of High DOF Articulated Structures: An Application to Human Hand Tracking," *Proc. European Conf. Computer Vision,* vol. 2, pp. 35-46, 1994.
[114] H. Sagawa and M. Takeuchi, "A Method for Analyzing Spatial Relationships between Words in Sign Language Recognition," *Proc. Gesture Workshop,* pp. 197-210, 1999.
[115] H. Sagawa and M. Takeuchi, "A Method for Recognizing a Sequence of Sign Language Words Represented in a Japanese Sign Language Sentence," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 434-439, 2000.
[116] H. Sagawa and M. Takeuchi, "Development of an Information Kiosk with a Sign Language Recognition System," *Proc. ACM Conf. Universal Usability,* pp. 149-150, 2000.
[117] H. Sako and A. Smith, "Real-Time Facial Expression Recognition Based on Features Position and Dimension," *Proc. Int'l Conf. Pattern Recognition,* pp. 643-648, 1996.
[118] W. Sandler, *Phonological Representation of the Sign.* Dordrecht: Foris, 1989.
[119] S. Lu, S. Igi, H. Matsuo, and Y. Nagashima, "Towards a Dialogue System Based on Recognition and Synthesis of Japanese Sign Language," *Proc. Gesture Workshop,* pp. 259-271, 1997.
[120] J. Sherrah and S. Gong, "Resolving Visual Uncertainty and Occlusion through Probabilistic Reasoning," *Proc. British Machine Vision Conf.,* pp. 252-261, 2000.
[121] *Sign Language J.,* Sign Factory, Japan, Spring 1996.
[122] P. Simpson, "Fuzzy Min-Max Neural Networks-Part 1: Classification," *IEEE Trans. Neural Networks,* vol. 3, pp. 776-786, 1992.
[123] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 20, no. 12, pp. 1371-1375, Dec. 1998.
[124] W.C. Stokoe, "Sign Language Structure: An Outline of the Visual Communication System of the American Deaf," *Studies in Linguistics: Occasional Papers 8,* 1960.
[125] D.J. Sturman and D. Zeltzer, "A Survey of Glove-Based Input," *IEEE Computer Graphics and Applications,* vol. 14, pp. 30-39, 1994.
[126] M.-C. Su, "A Fuzzy Rule-Based Approach to Spatio-Temporal Hand Gesture Recognition," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Application Rev.,* vol. 30, no. 2, pp. 276-281, May 2000.
[127] M.-C. Su, Y.-X. Zhao, H. Huang, and H.-F Chen, "A Fuzzy Rule-Based Approach to Recognizing 3-D Arm Movements," *IEEE Trans. Neural Systems Rehabilitation Eng.,* vol. 9, no. 2, pp. 191-201, June 2001.
[128] T. Supalla and E. Newport, "How Many Seats in a Chair? The Derivation of Nouns and Verbs in American Sign Language," *Understanding Language through Sign Language Research,* P. Siple, ed., pp. 91-133, 1978.
[129] T. Supalla, "The Classifier System in American Sign Language," *Noun Classes and Categorization,* C. Craig, ed., pp. 181-214, 1986.
[130] R. Sutton-Spence and B. Woll, *The linguistics of British Sign Language: An Introduction.* Cambridge Univ. Press, 1998.
[131] A. Sutherland, "Real-Time Video-Based Recognition of Sign Language Gestures Using Guided Template Matching," *Proc. Gesture Workshop,* pp. 31-38, 1996.
[132] G.J. Sweeney and A.C. Downton, "Towards Appearance-Based Multi-Channel Gesture Recognition," *Proc. Gesture Workshop,* pp. 7-16, 1996.
[133] S. Tamura and S. Kawasaki, "Recognition of Sign Language Motion Images," *Pattern Recognition,* vol. 21, no. 4, pp. 343-353, 1988.
[134] J. Tang and R. Nakatsu, "A Head Gesture Recognition Algorithm," *Proc. Int'l Conf. Advances in Multimodal Integration,* pp. 72-80, 2000.
[135] N. Tanibata, N. Shimada, and Y. Shirai, "Extraction of Hand Features for Recognition of Sign Language Words," *Proc. Int'l Conf. Vision Interface,* pp. 391-398, 2002.
[136] J.-C. Terrillon, A. Piplr, Y. Niwa, and K. Yamamoto, "Robust Face Detection and Japanese Sign Language Hand Posture Recognition for Human-Computer Interaction in an "Intelligent" Room," *Proc. Int'l Conf. Vision Interface,* pp. 369-376, 2002.
[137] C. Valli and C. Lucas, *Linguistics of American Sign Language: A Resource Text for ASL Users.* Washington, D.C.: Gallaudet Univ. Press, 1992.
[138] P. Vamplew, "Recognition of Sign Language Using Neural Networks," PhD thesis, Dept. of Computer Science, Univ. of Tasmania, May 1996.
[139] P. Vamplew and A. Adams, "Recognition of Sign Language Gestures Using Neural Networks," *Australian J. Intelligence Information Processing Systems,* vol. 5, no. 2, pp. 94-102, 1998.
[140] P. Viola and M. Jones, "Robust Real-Time Object Detection," *Proc. IEEE Workshop Statistical and Computational Theories of Vision,* 2001.
[141] C. Vogler and D. Metaxas, "Adapting Hidden Markov Models for ASL Recognition by Using Three-Dimensional Computer Vision Methods," *Proc. Int'l Conf. Systems, Man, Cybernetics,* vol. 1, pp. 156-161, 1997.
[142] C. Vogler, H. Sun, and D. Metaxas, "A Framework for Motion Recognition with Applications to American Sign Language and Gait Recognition," *Proc. IEEE Workshop Human Motion,* pp. 33-38, 2000.
[143] C. Vogler and D. Metaxas, "A Framework for Recognizing the Simultaneous Aspects of American Sign Language," *Computer Vision Image Understanding,* vol. 81, pp. 358-384, 2001.
[144] C. Vogler, "American Sign Language Recognition: Reducing the Complexity of the Task with Phoneme-Based Modeling and Parallel Hidden Markov Models," PhD thesis, Univ. of Pennsylvania, 2003.
[145] M.B. Waldron and S. Kim, "Isolated ASL Sign Recognition System for Deaf Persons," *IEEE Trans. Rehabilitation Eng.,* vol. 3, no. 3, pp. 261-271, Sept. 1995.
[146] C. Wang, W. Gao, and Z. Xuan, "A Real-Time Large Vocabulary Continuous Recognition System for Chinese Sign Language," *Proc. IEEE Pacific Rim Conf. Multimedia,* pp. 150-157, 2001.
[147] C. Wang, W. Gao, and S. Shan, "An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 393-398, 2002.
[148] L. Wang, W. Hu, and T. Tan, "Recent Developments in Human Motion Analysis," *Pattern Recognition,* vol. 36, pp. 585-601, 2003.
[149] T. Watanabe and M. Yachida, "Real Time Gesture Recognition Using Eigenspace from Multi Input Image Sequence," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 428-433, 1998.
[150] R.B. Wilbur, "Syllables and Segments: Hold the Movement and Move the Holds!" *Phonetics and Phonology: Current Issues in ASL Phonology,* G. Coulter, ed., vol. 3, pp. 135-168, 1993.
[151] R.B. Wilbur, "Phonological and Prosodic Layering of Nonmanuals in American Sign Language," *The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima,* H. Lane and K. Emmorey, eds., pp. 213-241, 2000.
[152] A.D. Wilson and A.F. Bobick, "Parametric Hidden Markov Models for Gesture Recognition," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 21, no. 9, pp. 885-900, Sept. 1999.
[153] J. Wu and W. Gao, "A Fast Sign Word Recognition Method for Chinese Sign Language," *Proc. Int'l Conf. Advances in Multimodal Interfaces,* pp. 599-606, 2000.
[154] J. Wu and W. Gao, "The Recognition of Finger-Spelling for Chinese Sign Language," *Proc. Gesture Workshop,* pp. 96-100, 2001.
[155] L. Wu, S.L. Oviatt, and P.R. Cohen, "Multimodal Integration—A Statistical View," *IEEE Trans. Multimedia,* vol. 1, no. 4, pp. 334-341, 1999.
[156] Y. Wu and T.S. Huang, "View-Independent Recognition of Hand Postures," *Proc. Conf. Computer Vision Pattern Recognition,* vol. 2, pp. 88-94, 2000.

[157] M. Xu, B. Raytchev, K. Sakaue, O. Hasegawa, A. Koizumi, M. Takeuchi, and H. Sagawa, "A Vision-Based Method for Recognizing Non-Manual Information in Japanese Sign Language," *Proc. Int'l Conf. Advances in Multimodal Interfaces,* pp. 572-581, 2000.
[158] M.-H. Yang, N. Ahuja, and M. Tabb, "Extraction of 2D Motion Trajectories and Its Application to Hand Gesture Recognition," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 24, no. 8, pp. 1061-1074, Aug. 2002.
[159] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3).* Cambridge Univ., 1995.
[160] S. Young, "A Review of Large-Vocabulary Continuous-Speech Recognition," *IEEE Signal Processing Magazine,* pp. 45-57, Sept. 1996.
[161] Q. Yuan, W. Gao, H. Yao, and C. Wang, "Recognition of Strong and Weak Connection Models in Continuous Sign Language," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 75-78, 2002.
[162] J. Zieren, N. Unger, and S. Akyol, "Hands Tracking from Frontal View for Vision-Based Gesture Recognition," *Proc. 24th DAGM Symp.,* pp. 531-539, 2002.
[163] T.G. Zimmerman, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvill, "Hand Gesture Interface Device," *Proc SIGCHI/GI Conf. Human Factors in Computing Systems and Graphics Interface,* pp. 189-192, 1986.

Sylvie C.W. Ong received the BSc (honors) degree in electrical engineering from Queen's University, Kingston, Canada, in 1993. From 1994 to 1996, she was with the Information Products Division, at Chartered Electronics Industries, Singapore, where she was working on the design and development of video and audio capture cards. From 1997 to 1999, she worked on firmware development and automated testing systems for computer monitors at Singapore Technologies Electronics Systems Assembly. She is currently a PhD candidate at the Department of Electrical and Computer Enginering at the National University of Singapore. Her research interests include human-computer interaction, statistical modeling of time-series and other sequential data, and machine learning, with focus on developing computational frameworks required for interpreting and understanding human behavior.

Surendra Ranganath received the BTech degree in electrical engineering from the Indian Institute of Technology, Kanpur, the ME degree in electrical communication engineering from the Indian Institute of Science, Bangalore, and the PhD degree in electrical engineering from the University of California (Davis). From 1982 to 1985, he was with the Applied Research Group at Tektronix, Inc., Beaverton, Oregon, where he was working in the area of digital video processing for enhanced and high definition TV. From 1986 to 1991, he was with the medical imaging group at Philips Laboratories, Briarcliff Manor, New York. In 1991, he joined the Department of Electrical and Computer Engineering at the National University of Singapore, where he is currently an associate professor. His research interests are in digital signal and image processing, computer vision, and neural networks with focus on human-computer interaction and video understanding applications.