Pers Ubiquit Comput
DOI 10.1007/s00779-012-0615-1

ORIGINAL ARTICLE

Non-manual cues in automatic sign language recognition

George Caridakis · Stylianos Asteriadis ·
Kostas Karpouzis

Received: 30 September 2011 / Accepted: 23 June 2012

Abstract Present work deals with the incorporation of non-manual cues in automatic sign language recognition. More specifically, eye gaze, head pose, and facial expressions are discussed in relation to their grammatical and syntactic function and means of including them in the recognition phase are investigated. Computer vision issues related to extracting facial features, eye gaze, and head pose cues are presented and classification approaches for incorporating these non-manual cues into the overall Sign Language recognition architecture are introduced.

Keywords Automatic sign language recognition · Facial expressions · Head pose · Eye gaze

1 Introduction

Non-manual cues are extremely crucial in Sign Language (SL), but this importance is not depicted in Automatic Sign Language Recognition (ASLR) approaches. These cues can operate as indicators, provide additional information, as intonation functions in spoken languages, add semantic properties and many other grammatical or syntactical functions. Although sign language recognition is sometimes considered similar, in terms of concepts and obstacles to be tackled, to speech recognition and translation, Dreuw [14] identifies a number of differences mainly related to the linguistic and representation aspects of Sign Language. More specifically, he mentions that, since there is no complete and standardized written form for SLs, there exists an inevitable trade-off between recognition accuracy and generality. HamNoSys [26] is one of the reference "phonetic" transcription systems used to transform functional characteristics of each sign (e.g. handshape, start, and direction of motion) using a pre-defined set of symbols. In theory, one can use the HamNoSys symbols to represent the vast majority of individual signs in a reusable and interoperable manner; however, support for non-manual sign characteristics is still minimal in HamNoSys and by no means covers all possible options (for example, there is no support for facial expressions, which modify the magnitude conveyed by a specific sign, for example, the speed of a car passing by). In addition, this notation supports the representation of atomic signs and not context or syntactic features: for example, it is possible to encode the sign for the verb form "I give" but there is no support for encoding a phrase where "I give to person A", which is usually noted by the signer gazing at a specific direction (where "person A" exists in the virtual signing space—[34]) or by slightly turning his/her body to that point. In the latter case, it is possible to notate a concept such as "movement of the head 45° to the right," but there is no connection to the person or object that the signer "placed" at that particular virtual position (Dreuw refers to that syntactic concept as "discourse entities"). This brings up the issue of word

G. Caridakis (✉) · S. Asteriadis · K. Karpouzis
Image, Video and Multimedia Systems Lab,
National Technical University of Athens, Athens, Greece
e-mail: gcari@image.ntua.gr

S. Asteriadis
e-mail: stiast@image.ntua.gr

K. Karpouzis
e-mail: kkarpou@image.ntua.gr

Published online: 26 October 2012

flexion: in most Sign Languages, facial expressions [29] and signer body stance and movement are used as modifiers for a specific sign. Besides the previous example on changing the context of a verb or action, the difference between a "big" or "small" object is signed by puffing lips or shrugging shoulders, possibly at the same time. Although HamNoSys does cater for representing those non-verbal signs, they are hardly ever annotated in context (that is, with respect to a particular object or person that the signer refers to—[12]). In that case, the "phonetic" transcription approach of HamNoSys would prove ineffective and a "tier" approach would have to be followed for annotation and subsequent recognition: Crasborn effectively divides all activity to different tiers, each representing one of the manual or non-manual characteristics of the sign: *repetition*, which in the non-verbal case would represent stress or magnitude, the *eye brows, eye aperture, and mouth* tier which is useful when encoding facial expressions and the *head* and *eye gaze* tiers for encoding focus on specific objects or action/verb context.

The rest of the article is organized as follows: Sect. 2 introduces the research area discussing related work and challenges, while the following sections present the proposed architecture. Computer vision issues related to extracting eye gaze, head pose cues, and facial features are discussed in Sects. 3.1, 3.2, and 3.3, respectively. Section 4 introduces the adopted classification approach for recognizing facial expressions and methods to incorporate facial expressions and their syntactic and grammatical functions into the overall automatic sign language recognition architecture. Finally, Sect. 5 concludes the article and presents future directions of the presented research work.

2 Automatic sign language recognition

2.1 Sign language definition

Sign language is the linguistic system used by the hearing-disabled group in order to communicate between the members of the group and also with non-impaired people. Unlike spoken languages, sign languages are heavily based on iconicity to convey meaning. A morphosyntactic structure is employed to express linguistic relations in 3D space and is organized much differently than orally articulated languages. Concepts are represented by signs, the basic grammatical unit of a sign language, forming a visual natural language.

With only a few, mostly situation-dependent exceptions, signs are articulated in a notional cube in front of the signer's head and body, the so-called signing space. By exploring the possibilities of signing space and iconicity as well as the productive use of sublexical features such as classifiers, many meanings can be conveyed without relying on established lexemes. Sign languages can therefore cope satisfactorily with much smaller numbers of lexemes in their lexicons than in most spoken languages. Consequently, sign language models and sign language grammars must be able to adequately account for the pronominal systems of sign language based upon positioning in three-dimensional space around a signer and, thus, intelligently inform avatar synthesis and sign recognition components of this unique aspect of signing.

Although sign languages have emerged naturally in deaf communities, they are unrelated to national spoken languages but culturally close communities or countries influence their respective SLs. For example, the British SL has influenced greatly Australian and New Zealand SL and is now considered as one known as BANZSL (BSL, Auslan, and NZSL). The case of the Scandinavian SLs is quite similar. Additionally, as in spoken languages, dialects from different parts of a country may also appear. Language-specific sign components entail a close set of attribute-value pairs, which comprise the language's phonology. Various combinations of sign components may generate every possible existing or new sign. Thus, it is essential that phonetic/morphological notations are sufficiently rich in features appropriate to sign language to sufficiently inform and assist virtual human animation and image processing technologies. A distinctive feature of sign languages is the extensive use they make of classifiers, a set of markers for the indication of class, shape, order, etc. (e.g., human, animal, vehicle, round, square), which complete or modify the concept conveyed by a sign. Classifiers provide a powerful mechanism for generating new concept representations (lexical items) or for modifying the meaning of existing ones. Signs are further formed either with one or with two hands.

2.2 Automatic SL recognition challenges

Gestures used for sign languages are often considered independent of other gesture styles since they are based on linguistics and are performed using a series of individual signs or gestures that combine to form grammatical structures for conversational style interfaces. In some instances such as finger spelling, sign languages can be considered semaphoric in nature. However, the gestures in sign languages are based on their linguistic components, and although they are communicative in nature, they differ from gesticulation in that the gestures correspond to symbols stored in the recognition system. Sign languages are grammatically and lexically complete, and are often compared to speech in terms of the processing required for their recognition.

Signs vary in time and space. Even if a person tries to perform the same sign twice, slight changes of speed and position of the hands will occur between instances of the same sign. Each sign varies in time and space and the signing speed and duration may differ significantly. Movements of the signer, like shifting in one direction or rotating around the body axis, must be considered. Additionally, signing fingers can be occluded, as they are hidden behind other parts of the hands or arms. Unlike the recognition of isolated signs, where the start and endpoint of the signs are known, the system also has to detect the transitions between the signs when recognizing connected signs. For a sequence of connected signs, the performance of each sign is affected by the preceding and the subsequent sign (coarticulation). The structure of a sentence in spoken language is linear, one word followed by another, whereas in sign language, a simultaneous structure exists with a parallel temporal and spatial configuration. The sign can begin and end at any instance of an observed sequence, since a temporal restriction in the execution of a sign does not exist and the number of the signs in a phrase is not fixed. The processing of a large amount of data is time-consuming, so real-time recognition is difficult. Additionally, in the case of frontal monocular vision systems which uses only a single camera, 3D space is projected on a 2D plane, resulting in the loss of depth information and reconstruction of the 3D trajectory of the hand is not always possible. Finally, the position of the signer in front of the camera may vary.

2.3 Vocabulary

Another important issue in the wide research area of automatic sign language recognition is the vocabulary size of the experimental corpora used to verify the robustness and generalization capabilities of the proposed systems. Most articles construct their experimental dataset using a quite restricted number of signs, varying from 10 to 65, while others extend their vocabulary, to what would be a more representative sample of the respective sign language, but still use between 164 and 274 signs. The articles that approximate a universal recognizer are those who reach impressive vocabulary sizes that enumerate up to more than 5,000 signs. Publications belonging to the last group focus mainly on large vocabulary recognition and issues related to a complete real-time system that could support automatic sign language recognition. The number of repetitions performed for each vocabulary entry is also important, as is the training/testing sample ratio. Typically, each sign is repeated 5–10 times, for example, [5, 10, 20, 32, 39], while there are cases where more [16, 17, 21] or less [1, 15, 18, 28, 33] repetitions are performed for each lemma in the restricted, experimental vocabulary.

2.4 Signer dependence

Signer dependence is a decisive aspect of sign language recognition, in the perspective of generalization of the proposed architecture into an actual system. Many articles [5, 10, 20, 33, 38, 37] propose approaches that have only been tested on the same signer as the one used for training or modeling. This, single signer-dependant constraint, cannot be the case for a generic automatic sign language recognizer since it is not possible to obtain training data from all the candidate users of the system. Signer independence, and ways to tackle intersigner variation in the performance of the signs or grammatical idioms of signer groups, is vital to achieve good recognition rates in an arbitrary setting by an unregistered user. Several works [11, 16, 27, 36] have tested their classification schemes on multiple signers, but this is not adequate since the system needs to be tested against signs performed by signers that have not been included in the training dataset, to achieve true signer independence.

2.5 Corpora

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. Furthermore, the definition of computer corpus is a corpus which is encoded in a standardized and homogenous way for open-ended retrieval tasks. The design and construction of sign language corpora, as is the case for most corpora, is not a trivial issue. Design aspects have to be taken under consideration in order for the final database to be useful for analysis and drawing conclusions on the sign language itself. Applicability and reusability, independently of the feature extraction or recognition scheme used, has to be ensured. Suitable transcription and annotation is crucial since it is a prerequisite for supervised classification or a multilayer approach.

Recognition rates of several classifiers can be considered as a reliable comparative quantitative measurement only under the restriction that the experiments were performed by testing recognition architectures against uniform and multiple datasets and corpora. Such an approach ensures that classification rates truly reflect the robustness and generalization capabilities of each system being evaluated. Most databases used in sign language processing so far do not provide or include what is important for the evaluation of sign language processing algorithms [35]. The need for creation of benchmark databases that can be used for investigating linguistic problems, and evaluating automatic sign language recognition systems or statistical machine translation systems including individual utterances, narratives, and dialogues pronunciation information is crucial. Other features that are a prerequisite for facial expression analysis and 3D depth estimation is that signing is performed from multiple angles, including a close-up of the face. Databases so far have not been produced with sign language recognition primarily in mind [13]; thai is, no suitable transcription is available. To use these data for the training or for performance evaluation in sign language recognition systems, the necessary transcriptions have to be created, which is a costly process mainly in terms of human resources.

2.6 Isolated versus continuous recognition

While the majority of articles address isolated word recognition, there is still a fair number of work focused on continuous, sentence recognition. Additionally some researchers explore both isolated and continuous recognition, extending their word-level methodologies to sentence-level sign language recognition. The latter is significantly more complex and challenging, since sign boundaries have to be detected automatically and each sign, composing the sentence, is affected by the preceding and the subsequent sign, which consists the coarticulation phenomenon. There are also some works [21, 25] that deal with static letter recognition for fingerspelling scenarios. The simplistic approach to tackle the problem of continuous signing is to extend isolated recognition with sign boundary recognition, thus, enabling the system to recognize sentences as a sequence of signs.

2.7 Classification schemes

Residing in the heart of each recognition attempt, the classification architecture is considered the most decisive among the other architecture aspects and in many cases it decides on or influences in a major degree the design decisions of the other components. Although a plethora of articles propose an established, off the shelf classifier (e.g., HMM), there is also a significant number of approaches that utilize a combination of classification schemes either by pre/postprocessing the input/output of each classifier in the sequential architecture of the overall classifier or by modifying and/or enhancing the internal operation of one type of learning and/or evaluation process of a classification architecture.

Architectures for classifying an unknown sign language lemma or sentence into one of the predefined candidate categories are:

- Neural networks
- HMM and variants
- Linear models
- Tree structures
- Clustering
- State sequence comparison

2.8 Incorporation of non-manual cues

Besides the basic sign language components of location, movement, handshape, and palm orientation, sign language is enriched with non-manual features and a complete grammatical structure. Both aspects are very sparsely dealt with and certainly not fully investigated in the recognition chain. For some sign languages, this is the case also for linguistic studies, since grammatical analysis is incomplete and facial expressions used in conjunction with manual features are not fully recorded and analyzed.

Grammatical phenomena are usually dealt as noise or signer variation and are not separately processed. Of course, incorporating grammatical models into the recognition chain would have to be assisted by some Natural Language Processing module adding another discipline in the already multidisciplinary area of automatic sign language recognition. To date, sign language recognition research has also mostly ignored facial expressions that arise as part of a natural sign language discourse, even though they carry important grammatical and prosodic information. The clear correspondence between the head angles and the head rotation and tilt labels holds great promise for future systems that recognize the non-manual markings of signed languages [6]. The ability to extract and plot the trajectories of various facial parameters may well prove invaluable for research into sign language prosody [29].

Non-manual cues related to present's research work scope could be summarized as:

- Temporal inflections
  - Frequency
  - Duration
  - Recurrence
  - Permanence
- Trajectory modifications
  - Shape
  - Rate
  - Rhythm
  - Tension
- Grammatical
  - Movement epenthesis
  - Emphatic inflections
  - Derivation of nouns from verbs
  - Numerical incorporation
  - Compound signs
  - Non-manual signals

3 Feature extraction

The gaze detector employs facial feature analysis of images captured from a standard web camera in order to determine the direction of the user's gaze. The purpose of the gaze module is to detect the raw user gaze direction details from the web camera in real time. It is based on facial feature detection and tracking, as reported in [4], and follows a variant of this method for head pose and eye gaze estimation. More specifically, starting from the eye centers, which are easy to be detected [3], eye corners and eyelids are detected, as well as two points on each eyebrow, the nostrils' midpoint, and four points on the mouth. These features are subsequently tracked using an iterative, 3-pyramid Lucas-Kanade tracker [22]. Lucas-Kanade tracking is one of the most widespread and used trackers in bibliography, and the choice of this tracker was based on the fact that it can accurately and effectively track features under a large variety of circumstances. However, as is the case in real life conditions, a series of rules has to be adopted in order to tackle constraints imposed by natural lighting and motion conditions: By assuming an orthographic projection at successive frames, the motion vectors of all features for such small periods of time can be considered to be almost equal. Features whose motion vector length m_i at frame i is much larger or smaller than the mean motion of all features m_{mean} ($m_i > t_1 \times m_{\mathrm{mean}}$, $m_i < t_2 \times m_{\mathrm{mean}}$; here, we considered $t_1 = 1.5$ and $t_2 = 0.5$) are considered as outliers, and their position is re-calculated based on their previous position and the re-calculated mean motion of the other features. The above step proved to be very important at improving the tracker's performance under difficult lighting conditions and occlusions.

3.1 Eye gaze estimation

For eye gaze estimation, relative displacements of the iris center with regard to the points around the eyes give a good indication of the directionality of the eyes with regard to a frame where the user faces the agent frontally. These displacements correspond to the eye gaze vectors (see Fig. 1). To re-enforce correct eye center tracking, the tracked eye centers' positions are updated by searching for the darkest neighborhoods around them and placing the eye center in the midpoint of this neighborhood, which helps tackle blinking and saccadic eye movements. Again, these displacements are normalized by the interocular distance at start-up and, thus, are scale independent. The computational complexity of the method permits real-time applications and requires only a simple web camera to operate. Tracking the features takes 13 ms per frame on average for a resolution 288 × 352 pixels of the input video, using a Pentium 4 CPU, running at 2.80 GHz, while re-initializations, whenever occurring, require 330 ms. Further details on the architecture and experimental results of the adopted eye gaze estimation module can be found in [2].

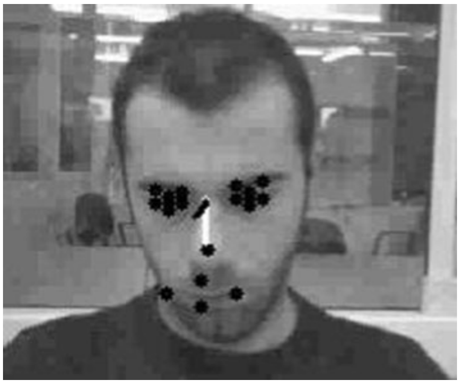

Fig. 1 Gaze direction detection is based on a number of tracked features (shown here as *black dots*) in order to calculate a final head pose (*white line*) and eye gaze (*black line*) vector

3.2 Head pose

Head pose is estimated by calculating the displacement of the eye centers' midpoint, with regard to its position at a frame where the user faces the camera frontally. This displacement produces the head pose vector which is a good index of where the user's head is turned toward (see Fig. 1). Normalization with the interocular distance at start-up (in pixels) guarantees that the head pose vector is scale independent. In order to distinguish between displacements caused by head rotations and by translations, the triangle formed by the triplet of the eyes and the mouth is monitored and head pose vector is only calculated when the fraction of the interocular distance to the eyes-mouth vertical distance changes significantly with regard to a frame where the person is looking frontally. To further suppress error accumulation, the system re-initializes when certain conditions regarding head pose vector length are met: In cases of rapid head rotations that may cause some features to be occluded, when the person comes back to a frontal position, one of the two eye centers might be erroneously tracked, while the other follows the movement of the head. In such cases, the head pose reduces in length and stays fixed when the person is facing the camera frontally. In this case, the system can re-initialize by re-detecting the facial features and restart the tracker. The above step can be seen in Fig. 2, with $\|hpv_i\|$ being the head pose vector length at current frame i, $a = 0.7$, $b = 0.07$, $n = 10$. As face detection and facial feature detection run slower than the tracker, video streaming continues normally and the second frame to be processed is the one caught by the camera at real time. However, initialization normally runs at $\sim$3 fps and, thus, pose and expressions practically do not change significantly after initialization.

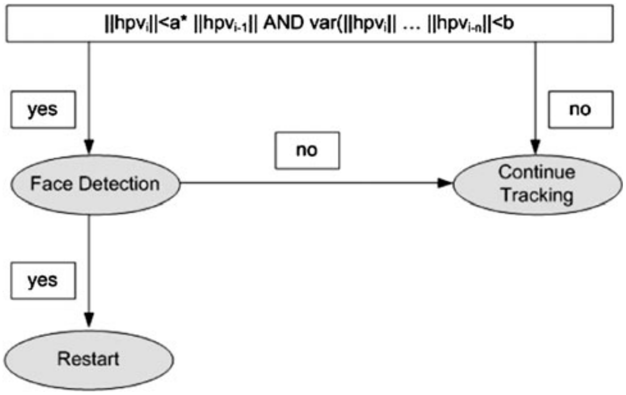

Fig. 2 Diagram depicting conditions under which system re-initializes

The head pose vector coordinates are further smoothed using Kalman filtering. The state variables of the filter are the horizontal and vertical coordinates $\mathrm{hpv}_i = (\mathrm{hpv}_{x,i}, \mathrm{hpv}_{y,i})$ of head pose vector, as well as the corresponding first derivative components (velocity), $u_i = (u_{x,i}, u_{y,i})$. Consequently, the state vector at frame i will be $x_i = (\mathrm{hpv}_{x,i}, \mathrm{hpv}_{y,i}, u_{x,i}, u_{y,i})$. According to Kalman theory [23], the state vector x_{i+1} corresponding to frame $i+1$ is linearly related with the current state x_i, with the system model defined in 1.

$$x_{i+1} = \Phi x_i + w_i \tag{1}$$

where Φ is the state transition matrix and w_i system noise, of gaussian distribution $w_i \sim (0,Q)$.

Considering very small state changes between consecutive frames, a linear model can be adapted, and the state transition matrix can be parameterized as follows:

$$\Phi = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Considering that the system's observable variables are the first two state variables (head pose vector components), and that the observation model is described in Eq. 2

$$z_i = Hx_i + v_i \tag{2}$$

then

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

where H is the matrix correlating current state with observation vector and v_i being measurements noise, which is considered to be of gaussian distribution $v_i \sim (0,R)$.

Adopting that state variables and transition probabilities follow gaussian distributions, Kalman algorithm offers the capability to adapt iterative formulas for estimating state vectors from observations. Initial values for the first two state vectors are zero-valued head pose components and, as velocity state variables, initial values are set to their values at the second frame (the algorithm is launched at the second frame of each sequence): $\mathrm{hpv}_{x,0} = 0$, $\mathrm{hpv}_{y,0} = 0$, $u_{x,0} = \mathrm{hpv}_{x,1} - \mathrm{hpv}_{x,0}$, $\mathrm{u}_{\mathrm{y},0} = \mathrm{hpv}_{\mathrm{y},1} - \mathrm{hpv}_{\mathrm{y},0}$.

As the error covariance matrix of the state vector has no significant impact on the results, initially it can be set equal to the unity matrix 4 × 4. Furthermore, considering that system noise is about 0.2° and, dividing it with a scale factor (in order to be consistent with ground truth data), final system noise is about 10^{-2}. Similar noise was considered for state variables corresponding to velocity components. Thus, the noise covariance matrix we consider is the following:

$$Q = \begin{pmatrix} 10^{-4} & 0 & 0 & 0 \\ 0 & 10^{-4} & 0 & 0 \\ 0 & 0 & 10^{-4} & 0 \\ 0 & 0 & 0 & 10^{-4} \end{pmatrix}$$

Measurements error matrix is calculated based on the head pose algorithm estimates and the ground truth head pose angle variables (divided by a scale factor). Thus, in our experiments, for the two observable states, the maximum of the variances of measurements error was considered, and the final measurements error matrix was the following:

$$R = \begin{pmatrix} 2.38 & 0 \\ 0 & 0.46 \end{pmatrix}$$

Figure 3 shows typical examples of estimated head pose vector horizontal and vertical components and the corresponding ground truth values throughout example sequences. Estimates' variances are marked at certain points. It is obvious that, during the sequences, the variance values tend to take fixed values.

In order to test the system's validity to estimate head rotation values accurately, a series of experiments were conducted on Boston University dataset (BU) [9]. This dataset offers ground truth data related to head rotation around the horizontal and vertical axis (*pitch* and *yaw*, respectively), as well as *roll* angle data (head rotation parallel to the image plane). The five participants of the dataset were asked to make non-pretending, spontaneous movements, both rotational and translational. Furthermore, the dataset was taken in an office environment, with complex background and typical office lighting conditions. Table 1 shows results on head pose estimation for every participant, independently. Mapping between head pose vector, as described above, and actual head rotation angles, was done by multiplying with a factor, common for all videos. It is obvious that the system has the ability to follow head rotation with a high degree of reliability but the main advantage of this technique is that it is completely

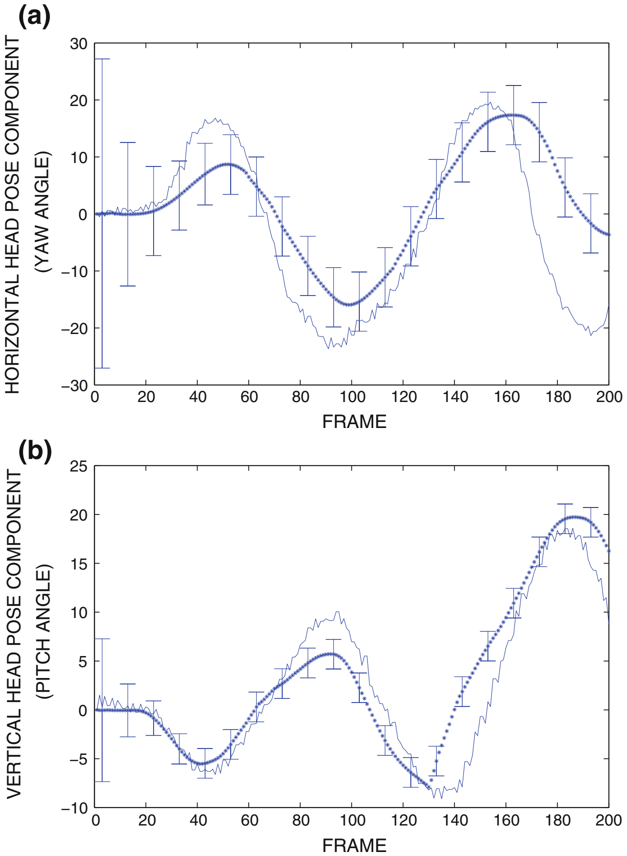

Fig. 3 **a** *Horizontal* head rotation, **b** *vertical* head rotation: true (*continuous lines*), estimated values (*dotted lines*)

Table 1 Mean absolute error at estimating horizontal (*yaw*), vertical (*pitch*) and *roll* angles on the BU dataset, for each participant

Participant ID	Yaw angle	Pitch angle	Roll angle
jam	7.43°	3.80°	4.73°
jim	9.15°	3.94°	5.37°
llm	10.5°	7.22°	4.73°
ssm	6.63°	4.71°	7.80°
vam	8.01°	5.69°	7.62°
Average	8.35°	5.07°	6.05°

non-intrusive and does not necessitate a pre-processing calibration or training phase. Both the head pose and eye gaze methods described here are based on solely image processing modules, not requiring intrusive or highly specialized hardware. Such as approach eliminates the requirement for person-specific calibration stage and is, thus, ideal for the purposes of ASLR.

3.3 Facial expressions

Facial features are detected in order to model facial expressions. Our approach, as described in detail in [19], achieves robust extraction of facial feature points for nose, eyebrows, eyes, and mouth. The face is first located, so that approximate facial feature locations can be estimated from the head position and rotation. Face roll rotation is estimated and corrected, and the head is segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose, and mouth. Each of those areas, called feature-candidate areas, contains the features whose boundaries need to be extracted for our purposes. Inside the corresponding feature-candidate areas, precise feature extraction is performed for each facial feature, that is, eyes, eyebrows, mouth, and nose, using a multi-cue approach, generating a small number of intermediate feature masks. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria to perform validation and weight assignment on each intermediate mask; each feature's weighted masks are then fused to produce a final mask along with confidence level estimation. The edges of the final masks are considered to be the extracted feature points as depicted in Fig. 4, which in turn are used to calculate MPEG-4 FAPs.

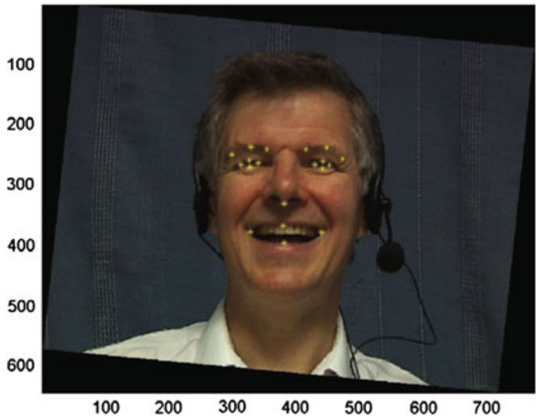

Fig. 4 Detected prominent facial features

A detailed description of the facial feature detection and tracking procedure and evaluation results are included in [19].

3.4 Expressivity

Expressivity of behavior is an integral part of the communication process as it can provide information on the current emotional state, mood, and personality of a person [31]. Many researchers have investigated human motion characteristics and encoded them into dual categories such as slow/fast, small/expansive, weak/energetic, small/large, unpleasant/pleasant.

To model expressivity, in our work, we use the six dimensions of behavior [8], as a more accomplished way to describe the expressivity, since it tackles all the parameters of expression of emotion [30]. Five parameters modeling behavior expressivity have been defined at the analysis level, as a subset of features derived from the field of expressivity synthesis:

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power

Overall activation is considered as the quantity of movement during a conversational turn. Spatial extent is modeled by expanding or condensing the entire space in front of the agent that is used for gesturing. The temporal expressivity parameter of the gesture signifies the duration of the movement while the speed expressivity parameter refers to the arm movement during the gesture's stroke phase (e.g., quick versus sustained actions). Gestures have three phases: preparation, stroke, and retraction. The real message is in the stroke, while the preparation and retraction elements consist of moving the arms to and from the rest position, to and from the start and end of the stroke. Fluidity differentiates smooth/graceful from sudden/jerky ones. This concept seeks to capture the continuity between movements, as such, it seems appropriate to modify the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs.

Gesture expressivity in automatic sign language recognition, to our knowledge, is completely ignored by researchers. Although gesture expressivity cues cannot be considered as non-manual features, with a strict definition of the term, they provide significant qualitative information about the gesture or sign lemma. This qualitative aspect of information conveyed by the signer could be employed in the automatic SL recognition chain in order to provide information about the agent or the recipient of an action, or even the semantic class of the object involved in it. Repetitiveness is strongly correlated to repetition of movement which may further declare frequency, plurality, or grammatical category differentiation, for example, between verb and noun (a single-movement sign may indicate verb function, while repetition of the same movement in a single sign may indicate the respective deverbal noun). Spatial extension is also related to the expansion of the movement that may indicate size or volume, whereas speed or vigor in combination with the appropriate non-manual signs may express a range of adverbial properties. Moving to an extremely ambitious application domain, quite different from automatic recognition, expressivity cues could be utilized in SL poetry transcription [24].

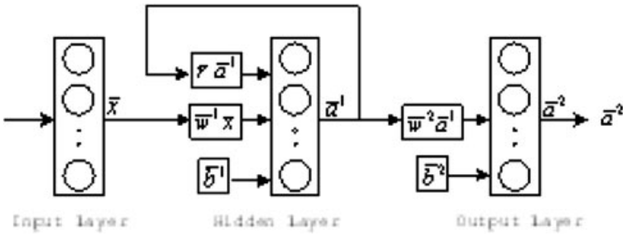

Fig. 5 Elman recurrent neural network used for facial expression recognition

Table 2 Experimental results for facial expression recognition using the adopted approach

Dataset	Neutral	Q1	Q2	Q3	Q4
Overall	79.07	87.10	86.67	66.10	74.70
Selected	100.00	98.29	96.43	100.00	100.00

4 Incorporation of non-manual cues into the recognition architecture

In order to recognize facial expressions, we need to utilize a classification model that is able to model and learn dynamics, such as a Hidden Markov Model or a recurrent neural network. In this work, we are using a recurrent neural network; see Fig. 5. This type of network differs from conventional feed-forward networks in that the first layer has a recurrent connection. The delay in this connection stores values from the previous time step which can be used in the current time step, thus providing the element of memory.

Although we are following an approach that only comprises a single layer of recurrent connections, in reality the network has the ability to learn patterns of a greater length as well as current values are affected by all previous values and not only by the last one. A two-layer network with feedback from the first-layer output to the first-layer input is adopted. This recurrent connection allows the Elman network to both detect and generate time-varying patterns. The input layer of the utilized network has 25 neurons (FAPs). The hidden layer has 20 neurons, and the output layer has as many neurons as the possible classes corresponding to the facial expressions. Details on the architecture, as well as experimental results on facial expression recognition, can be found in [7]. Partial experimental results related to recognizing affective quadrants, according to an activation-evaluation dimensional emotion representation approach, are illustrated in Table 2.

5 Conclusions

Current work deals with the incorporation of non-manual cues in automatic sign language recognition. Since facial expressions, eye gaze and signer head pose are used as modifiers for specific signs, they should also be included in the automatic recognition phase. Related computer vision methods for extracting low-level features (eye gaze and head pose) are discussed and, in a higher level, a classification approach for recognizing facial expressions is introduced. The grammatical and syntactic function of these cues and means of including them in the recognition phase are investigated.

Although the work presented here provides solid basis for further investigation of incorporation of non-manual features in the automatic sign language incorporation, researching this aspect of sign language recognition is far from complete. Experimental verification of the enhancement of SL recognition with facial expressions, eye gaze, and head pose is needed in order to prove and measure the acquired gain. Synchronization issues with manual features and fusion with classification techniques based on these features need to be addressed and investigated.

References

1. Assan M, Grobel K (1998) Video-based sign language recognition using hidden markov models. In: Proceedings of the international gesture workshop on gesture and sign language in human–computer interaction. Springer, London, UK, pp 97–109
2. Asteriadis S, Karpouzis K, Kollias S (2009) Feature extraction and selection for inferring user engagement in an hci environment. HCI International, San Diego, CA, 19–24 July 2009. http://www.image.ece.ntua.gr/publications.php
3. Asteriadis S, Nikolaidis N, Pitas I, Pardàs M (2007) Detection of facial characteristics based on edge information. In: Second international conference on computer vision theory and applications (VISAPP), vol 2. Barcelona, Spain, pp 247–252
4. Asteriadis S, Tzouveli P, Karpouzis K, Kollias S (2007) Nonverbal feedback on user interest based on gaze direction and head pose. In: 2nd International workshop on semantic media adaptation and personalization (SMAP 2007), London, United Kingdom, December. http://www.image.ece.ntua.gr/publications.php
5. Bowden R, Windridge D, Kadir T, Zisserman A, Brady M (2004) Computer vision, ECCV04, chap. A linguistic feature vector for the visual interpretation of sign language. Springer, New York
6. Canzler U, Dziurzyk T (2002) Extraction of non-manual features for videobased sign language recognition. International Association of Pattern Recognition, pp 318–321
7. Caridakis G, Karpouzis K, Wallace M, Kessous L, Amir N (2010) Multimodal users affective state analysis in naturalistic interaction. J Multimodal User Interfaces 3(1):49–66
8. Caridakis G, Raouzaiou A, Karpouzis K, Kollias S (2006) Synthesizing gesture expressivity based on real sequences. In: Workshop on multimodal corpora: from multimodal behavior theories to usable models, LREC 2006 conference, Genoa, Italy, pp 24–26. Citeseer
9. Cascia ML, Sclaroff S, Athitsos V (2000) Fast, reliable head tracking under varying illumination: an approach based on robust registration of texture-mapped 3d models. IEEE Trans Pattern Anal Mach Intell 22:322–336
10. Cooper HM, Bowden R (2007) Large lexicon detection of sign language. In: IEEE workshop human computer interaction, vol 4796, pp 88–97
11. Cortes G, Garcia L, Benitez C, Segura JC (2006) Hmm-based continuous sign language recognition using a fast optical flow parameterization of visual information. In: INTERSPEECH-2006
12. Crasborn O, Mesch J, Waters D, Nonhebel A, van der Kooij E, Woll B, Bergman B (2007) Sharing sign language data online: experiences from the ECHO project. Int J Corpus Linguistics 12(4):535–562
13. Dreuw P, Neidle C, Athitsos V, Sclaroff S, Ney H (2008) Benchmark databases for video-based automatic sign language recognition. In: International conference on language resources and evaluation. Marrakech, Morocco. http://www-i6.informatik.rwth-aachen.de/dreuw/database.php
14. Dreuw P, Stein D, Deselaers T, Rybach D, Zahedi M, Bungeroth J, Ney H (2008) Spoken language processing techniques for sign language recognition and translation. Technol Disabil 20(2):121–133
15. Fang G, Gao W, Ma J (2001) Signer-independent sign language recognition based on sofm/hmm. In: Proceedings of the IEEE ICCV workshop on recognition, analysis, and tracking of faces and gestures in real-time systems, 2001, pp 90–95. doi:10.1109/RATFG.2001.938915
16. Hernandez-Rebollar J, Kyriakopoulos N, Lindeman R (2004) A new instrumented approach for translating american sign language into sound and text. Proceedings of the sixth IEEE international conference on automatic face and gesture recognition, 2004, pp 547–552. doi:10.1109/AFGR.2004.1301590
17. Imagawa K, Matsuo H, ichiro Taniguchi R, Arita D, Lu, S, Igi S (2000) Recognition of local features for camera-based sign language recognition system. International conference on pattern recognition 04, 4849. doi:http://doi.ieeecomputersociety.org/10.1109/ICPR.2000.903050
18. Infantino I, Rizzo R, Gaglio, S (2007) A framework for sign language sentence recognition by commonsense context. IEEE Trans Syst Man Cybern C Appl Rev 37(5):1034–1039. doi:10.1109/TSMCC.2007.900624
19. Ioannou S, Caridakis G, Karpouzis K, Kollias S (2007) Robust feature detection for facial expression recognition http://www.image.ece.ntua.gr/publications.php
20. Kadir T, Bowden R, Ong EJ, Zisserman A (2004) Minimal training, large lexicon, unconstrained sign language recognition. In: Proceedings of the British Machine Vision Conference, vol 1. BMVA Press, pp 96.1–96.10
21. Lee YH, Tsai CY (2007) Taiwan sign language (tsl) recognition based on 3d data and neural networks. Expert systems with applications
22. Lucas B, Kanade T (1981) An iterative image registration technique with an application to stereo vision (ijcai). In: Proceedings of the 7th international joint conference on artificial intelligence (IJCAI '81), pp 674–679
23. Maybeck PS (1979) Stochastic models, estimation, and control. Academic Press, New York
24. Müller W, Fischer O (2003) From sign to signing: iconicity in language and literature 3, vol 3. John Benjamins Publishing Company, Amsterdam
25. Pashaloudi V, Margaritis K (2004) A performance study of a recognition system for greek sign language alphabet letters. In: International conference "Speech and Computer"
26. Prillwitz S, Leven R, Zienert H, Hanke T, Henning J et al (1989) Hamnosys. version 2.0; hamburg notation system for sign languages. an introductory guide. International studies on sign language and communication of the Deaf 5
27. Shanableh T, Assaleh K, Al-Rousan M (2007) Spatio-temporal feature-extraction techniques for isolated gesture recognition in

arabic sign language. IEEE Trans Syst Man Cybern B 37(3): 641–650. doi:10.1109/TSMCB.2006.889630

28. Su MC (2000) A fuzzy rule-based approach to spatio-temporal hand gesture recognition. IEEE Trans Syst Man Cybern C Appl Rev 30(2):276–281. doi:10.1109/5326.868448
29. Vogler C, Goldenstein S (2008) Facial movement analysis in asl. Univ Access Inf Soc 6:363–374
30. Wallbott H (1998) Bodily expression of emotion. Eur J Social Psychol 28(6):879–896
31. Wallbott H, Scherer K (1986) Cues and channels in emotion recognition. J Pers Social Psychol 51(4):690
32. Wang C, Gao W, Shan S (2002) An approach based on phonemes to large vocabulary chinese sign language recognition. In: Automatic face and gesture recognition, pp 393–398. doi:10.1109/AFGR.2002.1004188
33. Wang Q, Chen X, Wang C, Gao W (2006) Sign language recognition from homography. In: IEEE international conference on multimedia and expo, pp 429–432. doi:10.1109/ICME.2006.262564
34. Wrobel U (2001) Referenz in gebärdensprachen: Raum und person. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München 37:25–50
35. Zahedi M, Dreuw P, Rybach D, Deselaers T, Ney H (2006) Continuous sign language recognition—approaches from speech recognition and available data resources. In: LREC workshop on the representation and processing of sign languages: lexicographic matters and didactic scenarios. Genoa, Italy, pp 21–24
36. Zahedi M, Keysers D, Ney H (2005) Appearance-based recognition of words in american sign language. In: Pattern recognition and image analysis
37. Zhang L, Fang G, Gao W, Chen X, Chen Y (2004) Vision-based sign language recognition using sign-wise tied mixture hmm. Advances in multimedia information processing PCM, pp 1035–1042
38. Zhang LG, Chen Y, Fang G, Chen X, Gao W (2004) A vision-based sign language recognition system using tied-mixture density hmm. ICMI, pp 198–204
39. Zieren J, Kraiss KF (2005) Robust person-independent visual sign language recognition. In: Pattern recognition and image analysis