Notice of Violation of IEEE Publication Principles

"Hand Gesture Recognition Framework for Recognizing Sign Gestures and Handling Movement Epenthesis Using Level Building Nested Dynamic Programming Approach"
by Elakkiya A., Selvamani K., Kanimozhi S.
in the Proceedings of the IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), May 2014

After careful and considered review of the content and authorship of this paper by a duly constituted expert committee, this paper has been found to be in violation of IEEE's Publication Principles.

This paper duplicated extensive amounts of text from the paper cited below. The original text was copied without attribution (including appropriate references to the original author(s) and/or paper title) and without permission. A. Elakkiya was solely responsible for the copied material.

Due to the nature of this violation, reasonable effort should be made to remove all past references to this paper, and future references should be made to the following article:

"Enhanced Level Building Algorithm for the Movement Epenthesis Problem in Sign Language Recognition"
by Ruiduo Yang, Sudeep Sarkar, Barbara Loeding
in the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2007

Hand Gesture Recognition Framework for Recognizing Sign Gestures and Handling Movement Epenthesis Using Level Building Nested Dynamic Programming Approach

Elakkiya R,
Assistant Professor,
Agni College of Technology,
Chennai

Selvamani K,
Assistant Professor,
Anna University,
Chennai

Kanimozhi S
Research Scholar,
Anna University,
Chennai

***Abstract*- In this research paper, two crucial problems in continuous sign language recognition from unaided video sequences are considered. At the feature level, the problem of hand segmentation and grouping is considered and at the sentence level, the movement epenthesis problem is considered. A framework that can handle both of these problems based on an enhanced, nested version of the dynamic programming approach is constructed. To handle movement epenthesis problem, a nested version of a dynamic programming framework, called Level Building is used to simultaneously segment and to match signs from continuous sign language sentences. This approach is then coupled with a trigram grammar model to optimally segment and label sign language sentences. This approach will show improvement over past approaches in terms of the frame labeling rate and also our approach shows the flexibility when handling a changing context. The proposed approach is novel since it does not need explicit any models for movement epenthesis.**

Keywords— ***Sign language recognition, Movement epenthesis, Classification, Segmentation***

I. INTRODUCTION

Human computer interface leads the task of sign language recognition and it offers an unique opportunity for the development of motion recognition algorithms. In particular, it lets us easily get beyond just single gestures or signs. In practice HCI would involve composition of individual gestures just as sign sentences are compositions of individual signs. When signs appear in sentence contexts, variations appear; sentences are not the concatenation of individual signs. These frames do not correspond to any sign and can involve change in hand shape, movement, and can be over many frames sometimes equal in length to actual signs.

As a dominant communication medium, Sign Language (SL) is used by millions of sensory and gustatory impaired people every day. In recent years, Sign Language Recognition (SLR) has been studied to automatically transcribe signs into text or speech, so that the information exchange between deaf and hearing communities becomes much easier. To achieve precise skin segmentation, this paper introduce a novel skin colour model for integrating Support Vector Machine (SVM) active learning with boosting algorithm and region segmentation. This model consists of two stages namely the training stage and the segmentation stage. In the training stage, a generic skin colour model is applied to several frames to obtain the initial skin areas for the given gesture video. Later, a binary classifier based on boosting algorithm is used train the system to classify the weakly classified initial skin pixels.

In the segmentation stage, the classifier is incorporated with Region of Interest (ROI) to yield the final skin colour pixels. The main contribution in this proposed research paper is to design a twofold skin colour model namely skin segmentation using SVM classifier and active learning for most informative subunit. First, the SVM classifier is trained using the data collected in dynamic for every signing video sequence, which is adaptive to different racial skin colours and illumination conditions. Second, active learning is employed to select the most informative training subset for SVM, which leads to fast convergence and better performance. Moreover, ROI is adopted to reduce the noise and illumination variation.

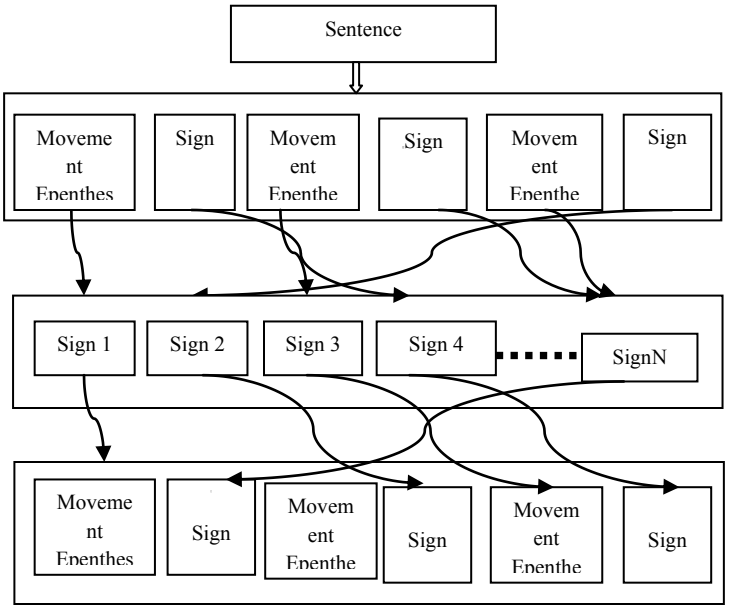

Figure 1. Nested Approach to handle Movement Epenthesis

II. RELATED WORK

MOST approaches [4], [5] to continuous sign language recognition or continuous gesture recognition use hidden Markov models (HMM) [3] or dynamic time warping (DTW) [1], [2]. These matching processes were popularized by their effectiveness in speech recognition. HMM-based approaches are also popular for other types of sequences, such as text sequences [7]. Although a speech or text sequence can be considered to be similar to a sign language or gesture sequence in the sense that both of them can also be represented as a sequence of feature vectors, a video-based continuous sign language sequence does have vital differences. These differences make it hard to simply apply the successful approaches in speech recognition to sign language recognition.

One such differentiating aspect is the importance of movement epenthesis. During the phonological processes in sign language, sometimes a movement segment needs to be added between two consecutive signs to move the hands from the end of one sign to the beginning of the next [6]. This is called movement epenthesis. These frames do not correspond to any sign and can involve changes in hand shape, movement, and can be over many frames, sometimes equal in length of actual signs. Consequently, automated sign recognition systems need a way to ignore or identify and remove the movement epenthesis frames prior to translation of the true signs.

The earliest work of which we are aware that explicitly modeled movement epenthesis in a continuous sign language recognition system with dedicated HMMs is by Vogler and Metaxas [8]. In another work [9], they also used context-dependent signs to model movement epenthesis and signs together. In a similar application of this approach, Yuan et al. [10] and Gao et al. [11] explicitly modeled movement epenthesis and matched with both sign and movement epenthesis models. The difference was that they used an automatic approach to precluster the movement epenthesis in the training data. More recently, Yang and Sarkar [12] used conditional random fields (CRF) to segment a sentence by removing movement epenthesis segments. However, this approach does not result in sign recognition, but just the segmentation of the sentence.

Although experimental results have shown that approaches that explicitly model movement epenthesis yield results superior to those ignoring movement epenthesis effects and context-dependent modeling [8], the question of scalability still remains. To obtain enough training data to model movement epenthesis is a real issue. With N signs, one may expect the number of movement epenthesis models to be N2, i.e., quadratic in the number of signs. Also, to build movement epenthesis models, one has to label the associated frames in the training data, most likely manually and, hence, the model can be easily biased to this set of sentences. So, it is important during experimentation to separate the train and test data with respect to sentences as well, and not just with respect to instances of the same sentences. Unlike previous approaches, we take a dynamic programming approach to address the problem of movement epenthesis, building upon the idea in [13]. Dynamic programming-based matching does not place demands on the training data as much as probabilistic models such as HMMs do.

Due to these complex low-level segmentation issues, previous continuous American Sign Language (ASL) recognition has mostly relied on assistive tools to obtain clean feature vectors. For example, Volger et al. [9], [15], [16] used a 3D tracking system and Cyber gloves, Wang et al. [17] used cyber gloves and 3D tracker, Starner et al. [18], [19] used color gloves, accelerometers, and head/shoulder mounted cameras, and Kadous [20] used power gloves. Although using assistive tools can yield better results, they also place added burden on the signer and can feel unnatural enough to even change the appearance of a normal sign.

III. PROPOSED MOVEMENT EPENTHESIS FORMULATION

Let the set of V model signs in the training database be represented as:

$$Si = \_s1 , s2 ..., sN \quad (1)$$

where $1 \leq i \leq V$, and Ni is the number of frames in the ith sign model. In addition to these signs, we will use symbols to represent movement epenthesismovement epenthesislabels of various lengths. We, of course, do not have explicit models corresponding to these symbols. We use these symbols for the convenience of expressing the problem mathematically.

$$SV+k = \_c1, c2, \ldots . , ck \quad (2)$$

where $1 \leq k \leq Nmax$ and $Nmax$ is the maximummovement epenthesislength. $c1, \cdots, ck$ are the dummy frames in themovement epenthesislabeled signs. Let the test sequence T of length$M$ be denoted by:

$$T = t1, t2,...., tM \quad (3)$$

In terms of dynamic warping term, we seek an optimal path to match T and the candidate sign *sequences* in order to compute the distance scores. Mathematically, considering the optimal warping path $P(u)$ as a multi-valued function such that $P(u) = (T(u), S(u))$ where $1 \leq u \leq Nu$ is the index, Nu is the length of the warping path, $(T(u), S(u))$ represents the sequence of coordinates of the warping path, that is, the $T(u)$th frame of the test sequence is matched with the $S(u)$th frame of the candidate sign sequences $S* i$. $S(u)$ can be represented as the combination of a sign coordinate and a sub-sign coordinate such as: $S(u) = (Q(u),K(u))$ that is, the $T(u)$th frame of the test sequence is matched with the $K(u)$th frame in the $Q(u)$th sign in the candidate signs sequence.

The function $d(.)$ is the distance between a test frame and a frame from the model sequences, included the dummymovement epenthesissymbols. For distances with the frames in the V model signs this would depend on the choice of the low-level features and the distance measure used. We denote this by $M(ti, skj)$.

The cost of amovement epenthesislabel is denoted by $\alpha.d(ti, skj) = \_M(ti, skj)$, if $j \leq V\alpha$, if $j > V$. The use of themovement epenthesislabel cost, α, is the essential difference between the classical problem formulation for recognizing connected words in speech and our formulation for recognition of connected signs in sign languages.

A. Nested Dynamic Programming Using Enhanced Level Building Algorithm

One naive way to obtain the solution is to enumerate among all the possible sign sequence candidates Si , compute the warping distance score between $S* I$ and T, find the Si with minimum score. Clearly the computational complexity of such an approach is prohibitive. Hence, we adopt an sequential approach to build this optimal sign sequence using a framework called Level Building and enhance it to allow formovement epenthesislabels. Each level corresponds to the possible order of signs ormovement epenthesisin the test sentence. Thus, the first level is concerned with the first possible label in the sentence, and so on. Each level is associated with a set of possible start and end locations within the sequence. And at each level we store the best possible match for each combination of end point from the previous level. The optimal sequence of signs andmovement epenthesislabels is constructed by backtracking. For each level l, we store the optimal cost for matching between sign Si and with the ending frame as m using a 3 dimensional array A.

Tmj denotes a subsequence of the test sequence that starts at the jth frame and ends at the mth frame. Hence Ail(m) gives us the minimum cumulative score for matching the ith model sign, Si to the test sequence upto them-th frame, for the lth sign label in the sequence. The choice of the cost for labeling a frame as movement epenthesis is a crucial one. We choose this by considering the distribution of match and non-match scores between signs in the training set. A match score is defined to the cost of matching different instances of the same sign and a non-match score is cost of matching instances of different signs. These scores are computed using dynamic warping and using the same frame to frame distance function used in the Level Building algorithm. They are normalized by the length of the warping path.

We then search of a threshold value that one can use to classify these scores into match and non-match ones. We choose the optimal α to be the optimal Bayesian decision boundary to accomplish this. However, instead of parametrically modeling each distribution (match and non-match) and then choosing the threshold, we empirically find the optimal value by sequentially searching for it. In essence, we are choosing the movement epenthesis labeling cost to be near the boundary of the match and non-match values. The proposed approach is illustrated in figure 2.

B. Sign Representation

Since the major contribution of this work is the enhanced Level Building algorithm, we just sketch the low-level representation used for completeness. Since our test is done based on pure video data, we

developeda segmentation scheme to segment the hands out of the scene to form the feature vectors for each frame. This step is automatic, but has some noise. The assumption that we make is that the hands move faster than other objects in the scene (including the face), and that the hand area can be somehwat localized by skin color detection. We used the mixed Gaussian model and a safe threshold such that non skin pixels can be falsely classified as skin pixels.

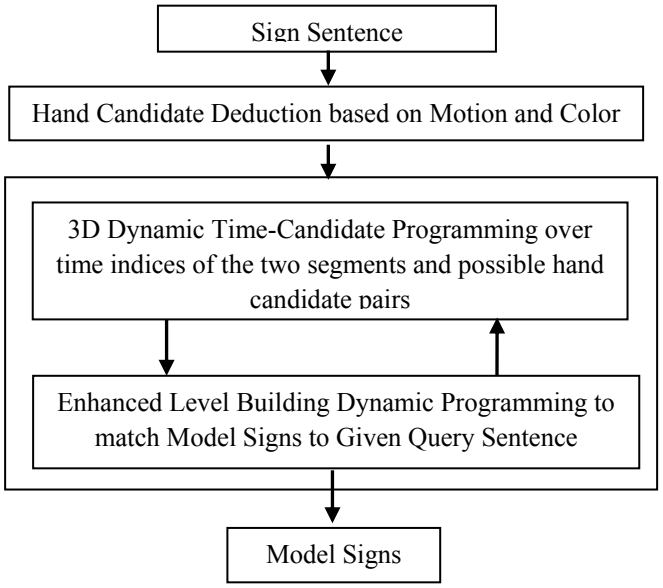

Fig 2: Schematic representation of a nested dynamic programming approach

We represent the possibly changing (but slowly) background, using a set of key frames. These key frames are identified as frames that are sufficiently different from each other. We sequentially search for them, starting from the first frame, which is always chosen to be a key frame. We compute the difference of any frame with previous key frame. If the non-component size in the thresholded difference image is large then the frame is labeled as the next key frame. This process continues until the end of the sequence. Then we compute the difference image of each frame to the key frames.

For each sentence *Si* with frames (*F*1*, F*2*, ... , FN* > repeat

1. Assign *k*1 = 1,*m* = 1, *i* = 2. For frame *F*2*,... , FN* repeat

(a) Compute difference image, *D*, between *Fi* and *Fkm*. Find the largest connected component in *D* in terms of its number of valid pixels *PD*.

(b) If *PD* > *threshold*, set *m* = *m* + 1, set *km* = *i*.

(c) Set *i* = *i* + 1. If *i* > *N* go to next step, else repeat above steps.

2. For each frame *Fi*, repeat (a) Compute a difference image *SD*, where *SD* = (Σ*mj* =1*abs*(*Fi, Fkj*))/(*m* − 1)

(b) Mask *SD* with the skin likelihood image. Do edge detection on *SD* and obtain the edge image *E*.

(c) Apply a dilation filter to *E*.

(d) For each valid pixel in *E*, set the corresponding pixel of *SD* to be 0

(e) Remove the small connected components in *SD*.

(f) Extract the Boundary Image *B*.

Given the hand boundaries, we then capture the spatial structure by considering the distribution of the horizontal and vertical distances between pairs of pixels in it; we compute the joint relational histogram of the displacement between all pairs of coordinates on boundary images. We then represent these relational histograms, normalized to sum to one, as points in a space of probability functions (SoPF), like that used in [9]. The SoPF is constructed by performing principal component analysis of these relational histograms from the training set of images. The coordinates in the SoPF is the feature vector used in the matching process.

IV. EXPERIMENTAL RESULTS

We have conducted extensive experimentation of the approach in the context of the task of recognizing continuous American Sign Language sentences from image sequences. We present not only visual results of labeling continuous ASL sentences, but also quantify the performance. We compare the performance with that obtained by classical Level Building, which does not account for movement epenthesis. We were not able to compare with other explicit model based approaches to handling movement epenthesis since they require large training data, which, as far as we are aware of, is not available; we would need about 1000 labeled ASL sentences for the vocabulary size comparable to that used in this paper. In the results, we also present empirical evidence of the optimality of the choice of the α parameter is used to decide on the me mapping cost and present the impact of the grammar model.

A. Dataset

The vocabulary consists of signs that a deaf person would need to communicate with security personnel at airports. The video data is taken at 30 fps, with an image resolution of 460 by 290. There are 39 different signs that are articulated in 25 different continuous sentences. (Note that for approaches that explicitly model me we would need around 1000 sentences to capture the variations between signs.) Some signs appear more than once in some sentences. The total number of individual sign instances in the dataset is 73. There are 5 instances of

each sentence. Some sentences have significant variations between multiple instances of the same sentence. We manually labeled the frames corresponding to the signs in the sentences for the training partition. The grammar is trained based on a text corpus of 150 sentences that is independent of the video data.

B. Labeling Results

A labeling result for three sentences is diagrammatically presented in Fig 2. Each horizontal bar represents a sentence and is partitioned into signs or me blocks. The size of each block is proportional to the number of frames corresponding to that label. For each sentence we present the ground truth as determined by an ASL expert and the results from the algorithm. It is obvious that the signer is signing at different speeds for each sign. For instance, the sign I is spread over a large number of frames. The framework can easily handle such case. Apart from a 1 to 2 frame mismatch at the beginning and the end, the labeling match pretty well. To quantitatively evaluate the results, we use errors as advocated.

If the recognized sentence inserted a sign that does not actually exist, one insertion error is spotted; if however the recognized sentence omitted a sign where it actually existed, one deletion error is counted; if the recognized sentence reports a wrong sign, we will consider it as a substitution error. We computed these errors automatically by computing the Levenshtein distance using a dynamic programming approach between the actual results and manually labeled ground truth. Fig. 3 shows the error rates we obtained with the optimal α (more on this later) for each test set in the 5-fold validation experimentation, using a tri-gram model. The sign-level error rate for each test set ranges between 9% and 28%. On average, the error rate is 17%, with a corresponding correct recognition rate of 83%.

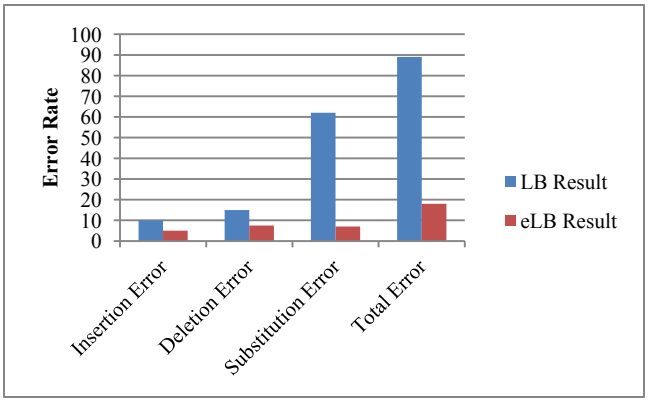

Fig. 3 Sign level error rates set in the 5-fold cross validation experiments with ASL data.

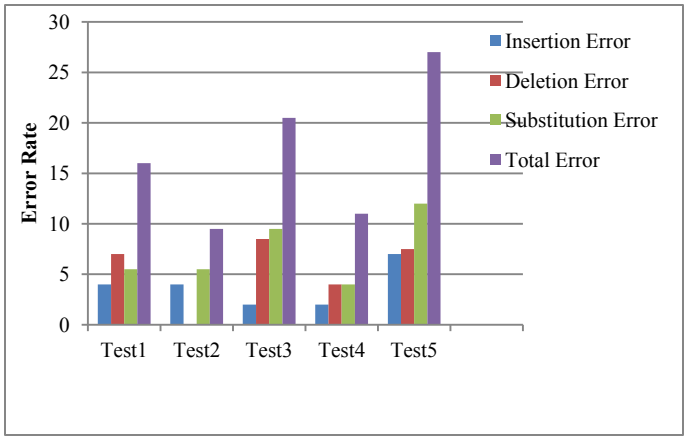

Fig. 4 Comparison of Enhanced Level Building vs classical Level Building

V. CONCLUSION

In this paper, we have presented a framework for Sign Language Recognition based on nested version using dynamic approach for continuous video sequences. The key contribution that can distinguish our proposed work from previous work is that we solved the feature selection problem for Sign Language Recognition by selecting not only the informative subunits, but also the discriminative features (weak classifiers) associated with the signs in subunits. The recognition task is mainly carried out on the subunit-level instead of traditional sign-level that helps to improve the scalability of Sign Language Recognition systems. In future, a portable interface device for Sign Language translation could be developed based on the proposed work and this proposed framework could also be applied to other gesture recognition systems.

References

[1] C. Sylvie and S. Ranganath, "Automatic Sign Language Analysis: A Survey and the Future Beyond Lexical Meaning," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 873-891, June 2005.

[2] B. Loeding, S. Sarkar, A. Parashar, and A. Karshmer, "Progress in Automated Computer Recognition of Sign Language," Lecture Notes in Computer Science, vol. 3118, pp. 1079-1087, Springer, 2004.

[3] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[4] C. Myers and L. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. Acoustics, Speech,

and Signal Processing, vol. 29, no. 2, pp. 284-297, Apr. 1981.
[5] J. Lichtenauer, E. Hendriks, and M. Reinders, "Sign Language Recognition by Combining Statistical DTW and Independent Classification," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 2040-2046, Nov. 2008.
[6] M. Skounakis, M. Craven, and S. Ray, "Hierarchical Hidden Markov Models for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence, 2003.
[7] C. Valli and C. Lucas, Linguistics of American Sign Language: A Resource Text for ASL Users. Gallaudet Univ. Press, 1992.
[8] C. Vogler and D. Metaxas, "A Framework of Recognizing the Simultaneous Aspects of American Sign Language," Computer Vision and Image Understanding, vol. 81, no. 81, pp. 358-384, 2001.
[9] C. Vogler and D. Metaxas, "ASL Recognition Based on a Coupling between HMMs and 3D Motion Analysis," Proc. Int'l Conf. Computer Vision, pp. 363-369, 1998.
[10] 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.
[11] W. Gao, G. Fang, D. Zhao, and Y. Chen, "Transition Movement Models for Large Vocabulary Continuous Sign Language Recognition," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, pp. 553-558, 2004.
[12] R. Yang and S. Sarkar, "Detecting Coarticulation in Sign Language Using Conditional Random Fields," Proc. Int'l Conf. Pattern Recognition, pp. 108-112, 2006.
[13] R. Yang, S. Sarkar, and B.L. Loeding "Enhanced Level Building Algorithm for the Movement Epenthesis Problem in Sign Language Recognition," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1-8, 2007.
[14] L. Rabiner and B. Juang, Fundamentals of Speech Recognition. PTR Prentice Hall, 1993.
[15] C. Vogler and D. Metaxas, "Handshapes and Movements: Multiple- Channel ASL Recognition," Lecture Notes in Artificial Intelligence, vol. 2915, pp. 247-258, Springer, 2004.
[16] C. Vogler, H. Sun, and D. Metaxas, "A Framework for Motion Recognition with Application to American Sign Language and Gait Recognition," Proc. Workshop Human Motion, pp. 33-38, 2000.
[17] C. Wang, W. Gao, and S. Shan, "An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, pp. 393-398, 2002.
[18] H. Brashear, V. Henderson, K.-H. Park, H. Hamilton, S. Lee, and T. Starner, "American Sign Language Recognition in Game Development for Deaf Children," Proc. Int'l ACM SIGACCESS Conf. Computers and Accessibility, pp. 79-86, 2006.
[19] T. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, pp. 189-194, 1995.
[20] M. Kadous, "Machine Translation of AUSLAN Signs Using Powergloves: Towards Large Lexicon-Recognition of Sign Language," Proc. Workshop the Integration of Gesture in Language and Speech, pp. 165-174, 1996.