IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 6, 2019, accepted June 25, 2019, date of publication June 28, 2019, date of current version August 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925654

Multimodal Fusion Based on LSTM and a Couple Conditional Hidden Markov Model for Chinese Sign Language Recognition

QINKUN XIAO, MINYING QIN, PENG GUO, AND YIDAN ZHAO
Department of Electronics Information and Engineering, Xi'an Technological University, Xi'an 710032, China
Corresponding author: Qinkun Xiao (xiaoqinkun10000@163.com)

This work was supported in part by the NSFC under Grant 61671362, and in part by the Shaanxi Province Natural Science Foundation under Grant 2017JM6041.

ABSTRACT A novel multimodal fusion approach is proposed for Chinese sign language (CSL) recognition. This framework, the LSTM2+CHMM model, uses dual long short-term memory (LSTM) and a couple hidden Markov model (CHMM) to fuse hand and skeleton sequence information. Novel contributions, first, include a unique hand segmentation algorithm using power rate transforms and the RGB-D image fusion. This approach effectively overcomes common limitations, such as complex backgrounds, inconsistent lighting, and variable skin tones. Then, as a result, the proposed skeleton-hand fusion framework can be used for the vision-based sign language recognition (SLR) of non-specific people in non-specific environments. Finally, this LSTM2+CHMM model combines the probability theory with a neural network to provide a unified methodology for multiple-sequence fusion. The proposed SLR framework was tested using the two CSL datasets, and the experimental results showed it to be effective.

INDEX TERMS Multimodal fusion, LSTM, CHMM, CNN, CSL recognition.

I. INTRODUCTION

Vision-based sign language recognition (SLR) is currently an active area of research in the field of artificial intelligence [1]–[18]. SLR is challenging because critical technologies needed for high accuracy identification, such as human-computer interfacing, are still being developed. In addition, existing techniques are often designed for specific people or environments, limiting their robustness. As such, there is a need for precise SLR with non-specific conditions.

SLR involves multiple complex problems, such as human-computer interactions and pattern recognition, which have attracted the attention of experts in multiple fields [18], [19]. Other challenges include variations in data collection and interpretation, such as subtle changes in gestures between individual people that make it difficult to establish a uniform SLR model [19]. In addition, hands are relatively small in videos and their movements are complicated [20]. Differing cultural and personal habits also affect SLR accuracy [21], [22]. Furthermore, robust real-time SLR requires expensive hardware and software for processing complex scenes and rapidly changing backgrounds [5].

Previous SLR algorithms have encountered multiple issues [18]. Primarily, a simple combination of features is not certain to produce better results than a single feature [22]. For example, a histogram of oriented gradients (HOG) has been used to describe local hand information, but some feature combinations have actually reduced the overall performance [22]. To solve this problem, we introduced a convolutional neural network (CNN) to extract hand features and avoid the negative effects of HOG. In previous HMM-based SLR methods, model parameters needed to be adaptively set, which introduced too many system parameters and reduced both the model training and calculation speeds [22]. In this study, we combined LSTM with a couple hidden Markov model (CHMM) for sequence signal processing [23]. Unlike existing HMM-based methods [22], this approach avoids the need to model signs one at a time, while reducing the total number of parameters. In previous studies, adaptive multi-modal signals were typically fused by adaptive fractional fusion methods to achieve higher recognition accuracy than for individual features [24], [25]. However, there is

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

typically not a theoretical explanation for assigning appropriate weights to each feature. In this paper, the CHMM model was used for multi-modal fusion and a probabilistic theoretical derivation is provided.

We propose a hierarchical fusion approach, based on the combination of dual LSTM and CHMM, to resolve the issues discussed above. This two-level fusion framework effectively utilizes low-level sequence classification capabilities and advanced semantic decision-making mechanisms. The motivation for this approach can be described as follows. (1) Our two-level fusion mechanism assumes that if unreliable features reduce classification accuracy in the first level, they can be repaired in a higher level. (2) In previous studies, HMM-based SLR methods required the establishment of an adaptive HMM classifier for each sign word. Our framework uses the LSTM model to automatically extract system parameters, which can be used in the CHMM to improve classifier learning. (3) A CNN was used for automated hand feature extraction, avoiding the negative effects of local features (such as HOG) that are adaptive or unsupervised.

The rest of this paper is organized as follows. Related work is reviewed in section II and our model is introduced in section III. Test results are analyzed in section IV and summarized in section V.

II. RELATED WORK

A variety of classification algorithms have been proposed for SLR, which can generally be divided into HMM-based and NN-based methods. HMM techniques typically involve a weighting scheme that is used for gesture recognition. However, matching times can be prohibitively long and unsuitable for real-time SLR [25]. Curve matching techniques have been proposed for manifold analysis of gesture trajectories [26] and light-HMM methods have also been used to select key frames using a low rank approximation, to improve recognition efficiency [27].

HMM-based approaches typically require fewer training data but tend to exhibit lower recognition accuracy. Neural network algorithms have become popular in recent years, partly because they offer higher recognition accuracy. For example, Huang *et al.* incorporated a convolutional neural network in their algorithm [5], Liu *et al.* developed a long short-term memory (LSTM) model [23], Neverova *et al.* utilized a recurrent neural network (RNN) structure [24], Wu *et al.* used a 3D-CNN-based approach [28], and Molchanov *et al.* proposed a deep dynamic neural network [29]. These algorithms were effective for sequence information processing, such as active recognition. However, the primary disadvantage of deep learning methods like these is they typically require large training sets.

Classical multimodal fusion, an emerging SLR technology, can be divided into early and late fusion techniques. Early fusion takes place at the feature level [30], while late fusion takes place at the decision or scoring level [31]. Late fusion still suffers from certain limitations, such as inflexible parameter learning, longer runtimes, and invalid feature results. As such, neural networks are often used to assist in the multimodal fusion process. As a critical component needed for accurate training, this fusion step has received increased attention from researchers in recent years [27], [32], [33]. For example, Wang *et al.* proposed a combined mode feature for skeleton and hand HOG features [27]. Wu *et al.* input skeleton data and RGB-D images into an HMM in order to fuse multimodal gesture data streams [32]. Neverova *et al.* used a multi-scale, multi-modal neural network called ModDrop to learn cross-modal correlations between multi-modal channel representations [24].

Dynamic sign language recognition (DSLR) systems have been proposed for smart home interactive applications in which a k-means++ method was used to cluster features and train the system. A nonlinear support vector machine (SVM) was then utilized to classify hand movements [34]. However, the testing stage only considered six simple dynamic gestures. Wang *et al.* proposed a multi-view parameter-free framework (MPF) [35] and Yuan *et al.* focused on designing a robust feature description for optical flow frames [36]. A lightweight deep learning model, based on the convolutional 3D (C3D) network, and a recurrent neural network (RNN) were used for complicated action recognition. 3D spatio-temporal information for each sign has also been interpreted using joint angular displacement maps (JADMs), which encode the sign as a color texture image [7]. A new color-coded feature map, called a joint angular velocity map, was recently proposed to accurately model 3D joint motion [8]. These studies provide a variety of options for multi-modal fusion.

In addition, some studies have investigated SLR based on machine translations, such as the variational auto-encoder (VAE). Huang *et al.* proposed a hierarchical attention network with latent space (LS-HAN) for continuous CSL recognition [6]. A recent coding challenge introduced the sign language translation (SLT) problem [4], in which the objective was to generate spoken translations from sign language videos by taking into account both the order of words and grammar. However, machine translation methods cannot recognize semantic details in each frame of a video. As such, we propose the use of CSL words corresponding to individual frames, necessitating the selection of an algorithm for the automatic grouping of sentences to recognize frames in real-time. This approach combines the advantages of existing fusion techniques and introduces a novel multimodal fusion technique for SLR.

The quantity of available training data is often insufficient for practical applications of large-scale SLR [20]. Public sign language datasets, such as the MSRC-12 Kinect gesture dataset [37], the 73 ASL mark datasets [38], the 12 American sign language datasets [19], the 10-Gesture dataset [39], and the 24 static ASL mark word data set [20] are relatively small. One well-known gesture dataset, the ChaLearn database, only contains 20 gestures [21]. In 2016, Wan *et al.* released a new dataset containing 249 gestures [27] and in 2018 Huang *et al.* conducted a series of experiments on the 500 CSL dataset [5]. Sentence datasets, such as Sun's 63-sentence database, are

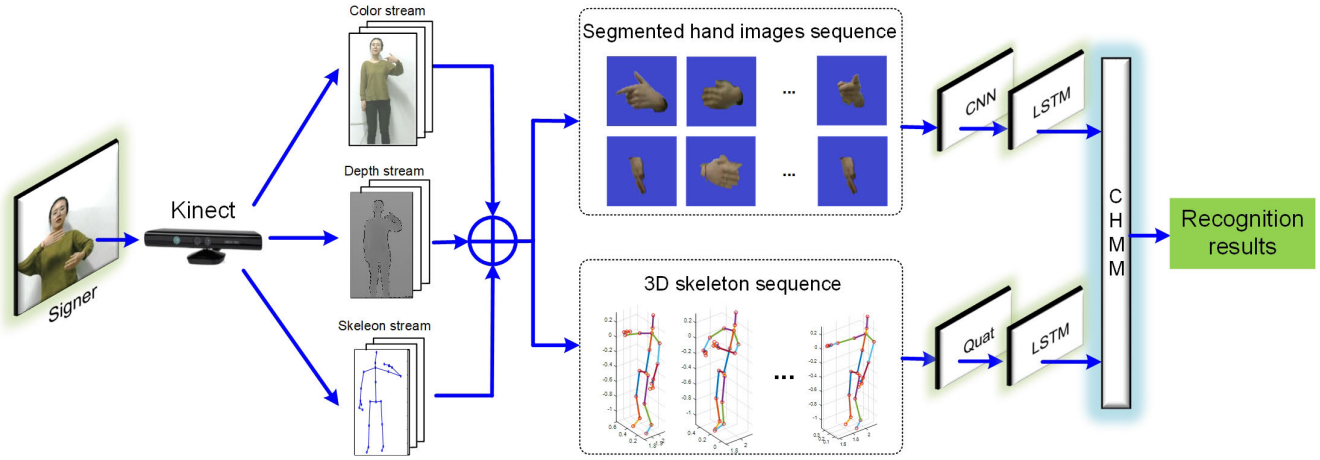

FIGURE 1. Multi-sensor data fusion based on LSTM2+CHMM for SLR.

also available and typically consist of two or four isolated words [40], [41]. This study focuses on CSL recognition and utilizes two datasets. The first includes a series of sentences and was developed in-house. The second is the largest known CSL dataset, developed by Jie *et al.* [5]. The primary objective for this study was improving recognition accuracy through a combination of NN-based and HMM-based methods in a limited sample environment.

III. RECOGNITION METHODOLOGY

As shown in Fig. 1, a Microsoft Kinect device was used to acquire sign language information, including color, depth, and skeleton streams simultaneously. The 3D skeleton data and segmented hand sequences were then combined for continuous CSL recognition using the proposed LSTM2+CHMM model. First, segmented hand images were input to the hand-related LSTM algorithm (denoted as $LSTM_h$). 3D skeleton sequences were simultaneously input to the skeleton-related LSTM algorithm (denoted as $LSTM_s$). The output from these two LSTM sequences were input to a CHMM model. Continuous CSL recognition results were then acquired using graph model probability inferences, the details of which are discussed below.

A. FEATURE EXTRACTION

1) HAND FEATURE EXTRACTION

Input video from the Kinect was used to acquire a color stream $\mathbf{I}^c = (I_i^c)_{i=1}^T$, a depth stream $\mathbf{I}^d = (I_i^d)_{i=1}^T$, and a 3D skeleton stream $\mathbf{K}^{3D} = (k_i^{3D})_{i=1}^T$. At time t, RGB-D images, which are a composition of I_i^c and I_i^d, were used to segment hands from the background.

RGB-D images (I_i^c, I_i^d), based on transfer learning and power rate transforms, were used to segment hands from the background [42]. Two different techniques were used to negate the effects of skin color on segmentation. First, depth information was used to establish a threshold in I_i^d, separating the face from the hands. This approach is viable because the face is always behind the hands during active CSL. Secondly, a hand detector was used in I_i^c to distinguish foreground from background. The active contour method was then applied.

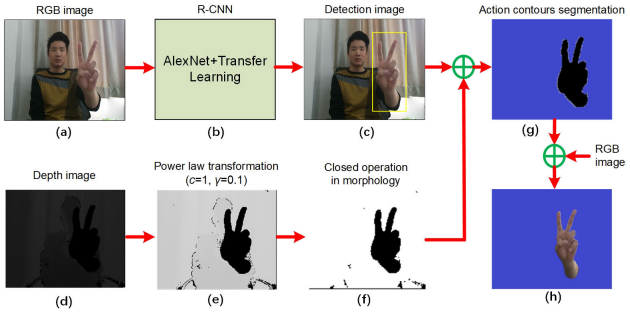

FIGURE 2. Hand segmentation using RGB-D image fusion.

This framework is represented in Fig. 2. A faster R-CNN, trained with a CSL dataset, was used to detect hands in RGB frames as shown in Fig. 2(c) [43], [44]. Depth images I_i^d (Fig. 2(d)) were processed using power rate transforms (Fig.2(e)) and morphological closing operations (Fig.2(f)) to acquire hand masks. The environment used for collecting gesture data is shown in Fig.3(a), where the Kinect is 0.5 meters away from the signer and 1.2 meters from the ground. The grayscale is deeper for shorter distances.

Since hands remain mostly in front of the body during signing, grayscale values for the hand are lower in the depth image. In this study, the power rate transform was used to stretch the image grayscale histogram in order to highlight hand information. This process can be expressed as follows:

$$s = cr^{\gamma}, \tag{1}$$

where c and γ are control parameters, r is the initial grayscale value, and s is the transferred grayscale. According to the formula, when $\gamma > 1$, image stretching is focused primarily in the high grayscale range and image details are highlighted. When $\gamma < 1$, image stretching is focused primarily in

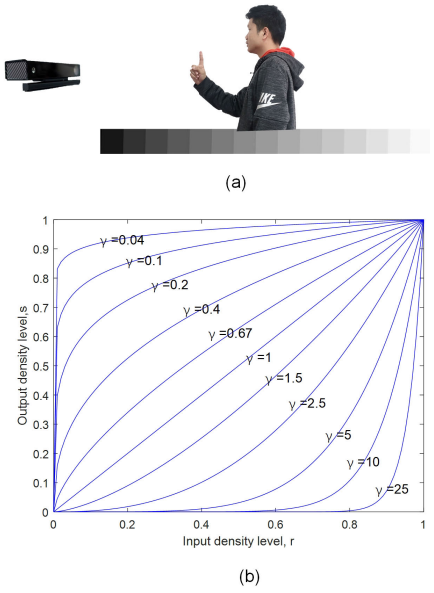

FIGURE 3. (a) The CSL data collection environment and (b) the power rate transform function for varying parameters.

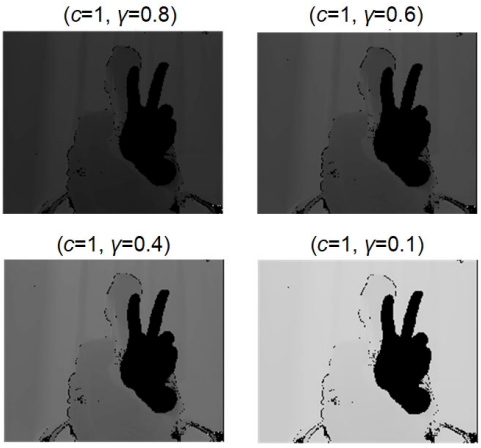

FIGURE 4. A comparison of results using power rate transforms in depth images for multiple values of c and γ.

the low grayscale range. A function image for the power rate transform is shown in Fig. 3(b). According to the histogram of I_i^d, all values are concentrated in a small range and grayscale stretching occurred for the parameters $c = 1$, $\gamma = 0.8, 0.6, 0.2$, and 0.1. The power rate transform of I_i^d was calculated as shown in Fig. 4. Experiments demonstrated that smaller values of γ produced better grayscale stretching effects. As such, values of $c = 1, \gamma = 0.1$ were chosen for the power rate transform of I_i^d. As seen in Fig. 2(f), the hand silhouette is more obvious after the power rate transform, though some noise contours are still visible. This small-area noise can be removed using a morphological closing operation, in which the region is expanded and the image is then etched.

Action contours, calculated from the detected field (Fig.2(c)) and the processed depth image (Fig.2(f)) were used to segment the hands (Fig.2(g)) [45]. RGB images (Fig.2(a)) were then combined to produce the final hand segmentation (Fig.2(h)). Since the hand area is small, it is inconvenient to directly extract features using trained large-scale networks such as AlexNet or VGG. As such, we have established a small-scale CNN to extract image features.

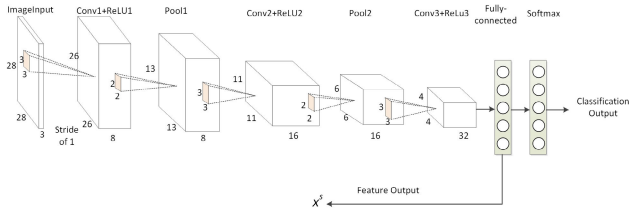

FIGURE 5. An illustration of hand feature extraction using the CNN.

Fig. 5 shows the 11-layer CNN model used to extract features from the segmented hands. This model includes an input layer ($28 \times 28 \times 3$), 3 convolution layers, 2 max pooling layers, 3 nonlinear layers, 1 fully-connected layer, and 1 classification output layer. The input to the CNN is an RGB image containing both left-hand and right-hand data. The 1^{st} convolution layer includes 8 different 3×3 kernels, while the 1^{st} pooling layer contains 8 different 2×2 neighborhood domains. The 2^{nd} convolutional layer includes 16 different 3×3 kernels, while the 2^{nd} pooling layer contains 16 different 2×2 neighborhood domains. The 3^{rd} convolution layer includes 32 different 3×3 kernels. The x^s hand features were extracted from the fully-connected layer.

2) 3D SKELETON FEATURE EXTRACTION

The skeletal stream for the ith skeleton contained 25 nodes and was represented by $K^{3D} = (k_1^{3D}, \cdots, k_n^{3D})$. A quaternion representation was selected for 3D skeletal features as human body gestures can be described by angles between bones. Two skeletal vectors were defined as $e_n = (a_1, b_1, c_1)$ and $e_m = (a_2, b_2, c_2)$. The cosine of the angle θ between two bones is then given by:

$$\cos\theta = \frac{a_1a_2 + b_1b_2 + c_1c_2}{\sqrt{a_1^2 + b_1^2 + c_1^2} \cdot \sqrt{a_2^2 + b_2^2 + c_2^2}}. \tag{2}$$

The rotation axis for non-parallel vectors e_n and e_m in a plane are given by:

$$\begin{aligned} \mathbf{r} = \mathbf{e}_n \times \mathbf{e}_m &= \begin{vmatrix} i & j & k \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix} \\ &= \left(\begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}, - \begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}, \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \right), \end{aligned} \tag{3}$$

where $\mathbf{r}$ represents a rotation from $\mathbf{e}_n$ to $\mathbf{e}_m$ and the direction of the normal vector is determined by the right-hand criterion. This implies:

$$\begin{cases} \mathbf{v} = (\mathbf{a}, \mathbf{b}, \mathbf{c}) = \sin\left(\dfrac{\theta}{2}\right) \cdot \mathbf{r}, \\ \mathbf{w} = \cos\left(\dfrac{\theta}{2}\right). \end{cases} \tag{4}$$

The unit quaternion $\mathbf{q}$ can be expressed as:

$$\mathbf{q} = (\mathbf{v}, \mathbf{w}) = (\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{w}) \tag{5}$$

and a complete 3D skeletal gesture is represented as a $4 \times 25 = 100$ dimensional vector with 25 nodes: $x^g = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{25})$.

B. CSL RECOGNITION

1) THE LSTM2+CHMM FRAMEWORK

In this study, two LSTMs and a CHMM algorithm were used for CSL data modeling. The included LSTM was a type of recurrent neural network (RNN), while $LSTM_h$ was used to calculate hand-related sequence data and $LSTM_s$ was used to calculate skeletal data. These two models had the same structure but different input feature vector types. LSTM output was input to the CHMM model and the two data streams were fused using CHMM inference to produce CSL recognition results.

Calculation of the LSTM models shown in Fig. 6 can be described as follows [23]:

$$\begin{cases} f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i[h_{t-1}, x] + b_i) \\ d_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \\ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \\ C_t = f_t * C_{t-1} + i_t * d_t \\ h_t = o_t * \tanh(C_t) \end{cases} \tag{6}$$

where x_t is a hand-related or skeleton-related feature vector input signal. The h_{t-1} term is a middle hidden variable, h_0 is the initial value, $\sigma(\cdot)$ and $\tanh(\cdot)$ are active functions, W is the network weight, and b is the deviation. The output is given by $y_t = \mathrm{Softmax}(W_y h_t + b_y)$.

LSTM output y_t^s, y_t^g was then input to the CHMM as an observed signal. The CHMM was divided into two basic HMMs to simplify its calculation, as described by graph model theory [22]. This HMM is a simple dynamic Bayesian network (DBN) that can be determined using Markov chain theory and a Viterbi decoding algorithm [22]. It is defined by the parameter $\lambda = (\pi, A, B)$, where π indicates prior knowledge, A is a state transaction matrix, and B is the observation probability.

As shown in Fig. 6, $Y^s = (y_1^s, \cdots, y_T^s)$ and $Y^g = (y_1^g, \cdots, y_T^g)$, where y_i^s and y_i^g denote the ith hand and skeleton feature vectors, respectively. A mixed-state DBN model was developed to represent continuous CSL recognition systems using gesture observations Y and sign states W. Hand-related contributions were assumed equal to skeleton-related contributions. As a result, the state sequence $W^s = (w_1^s, \cdots, w_T^s)$ could be updated using $W^g = (w_1^g, \cdots, w_T^g)$. Information in these two sequences was fused by probability inference, producing a final state with higher estimation accuracy.

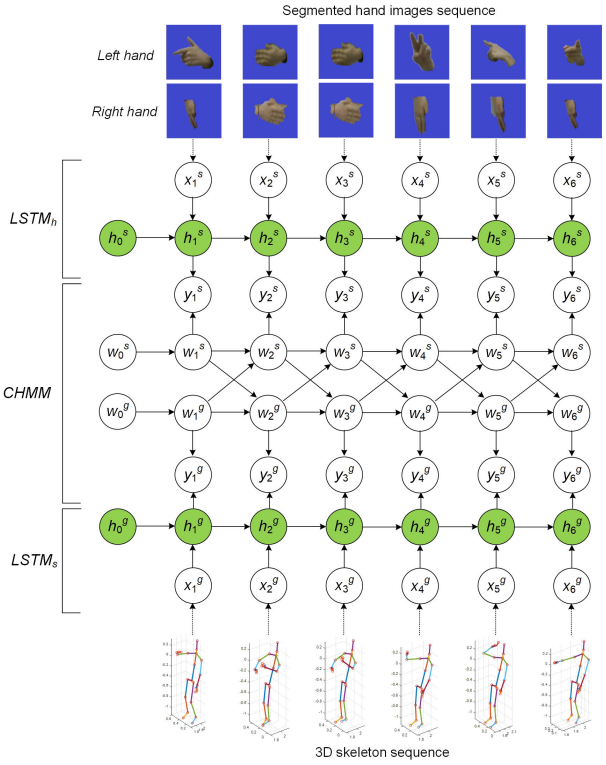

FIGURE 6. The LSTM2+CHMM model for CSL recognition.

2) LSTM2+CHMM CALCULATION

Inference was first performed by dividing the CHMM into two HMMs to calculate optimal hidden state probabilities. Hand-related HMMs included 3 parameters [46]:

$$\begin{cases} \pi^s = [p_i^s]_{1\times n} = P(w_0^s) \\ A^s = [a_{ij}^s]_{n\times n} = P(w_{t+1}^s | w_t^s) \\ B^s = P(y_t^s | w_t^s) = \mathbb{N}(\mu_y^s, \Sigma_y^s)(y_t^s) \end{cases} \tag{7}$$

where π^s is the prior distribution of hand-related states w_0^s. If w_0^s includes n states, then $(s_1, s_2, \cdots, s_n)$, where s_i corresponds to the ith sign. Hence, $P(w_0^s = s_i) = \pi^s(i)$. The term A^s is a state transaction matrix and a_{ij}^s denotes the transaction probability from the ith sign state to the jth sign state. As such, $a_{ij}^s = P(w_{t+1}^s = s_j | w_t^s = s_i)$. B^s is an observation matrix such that y_t^s is a continuous variable. The observation probability $P(y_t^s | w_t^s = s_i)$ is then a Gaussian distribution, where μ_y and Σ_y are the mean and variation, respectively. From Bayesian theory, optimal state sequence estimation for $w_{1:t}^s$ and $P(w_1^s)$ can be calculated as follows [46]:

$$P(w_1^s) = P(w_1^s | w_0^s) P(x_0^s). \tag{8}$$

The observation y_1 then yields:

$$P(w_1^s y_1^s) = \frac{P(y_1^s | w_1^s) P(x_1^s)}{P(y_1^s)}. \tag{9}$$

The state probability at time t can be determined from:

$$P(w_{t+1}^s) = P(w_{t+1}^s | w_t^s) P(w_t^s) \tag{10}$$

and the state can be optimized using an observation sequence:

$$\begin{aligned} &P(w^s_{1+t}|y^s_{1:1+t}) \\ &= P(w^s_{1+t}|y^s_{1+t}, y^s_{1:t}) \\ &= P(w^s_{1+t}|y^s_{1+t}) \cdot P(w^s_{1+t}|y^s_{1:t}) \\ &= \alpha P(y^s_{1+t}|w^s_{1+t}) P(w^s_{1+t}) \cdot P(w^s_{1+t}|y^s_{1:t}) \\ &= \alpha P(y^s_{1+t}|x^s_{1+t}) P(x^s_{1+t}|x^s_t) \\ &\quad \times P(x^s_t) \sum\nolimits_{x_t} P(w^s_{1+t}|w^s_t) P(w^s_t|y^s_{1:t}) \end{aligned} \tag{11}$$

The final fusion inference probability was acquired by assuming the skeleton-related HMM to be the primary network, implying it is more important for recognition. The 3 parameters used in this inference process are given by the following:

$$\begin{cases} \pi^g = [p^g_i]_{1\times n} = P(w^g_0) \\ A^g = [a^g_{ijk}]_{n\times m\times l} = P(w^g_{1+t}|w^g_t, w^s_t) \\ B^g = P(y^g_t|w^g_t) = \mathbb{N}(\mu^g_y, \Sigma^g_y)(y^g_t) \end{cases} \tag{12}$$

where π^g is the prior distribution of skeleton-related states w^g_0. If w^g_0 includes n states, then $(s_1, s_2, \cdots, s_n)$. Here, s_i corresponds to the ith sign and $P(w^g_0 = s_i) = \pi^g(i)$. The term A^g is a state transaction matrix and a_{ijg} denotes the transaction probability from the ith sign state to the jth sign state. As a result, $a^g_{ijk} = P(w^g_{t+1} = s_k|w^g_t = s_i, w^s_t = s_j)$. B^g is an observation matrix and y^g_t is a continuous variable. The observation probability $P(y^g_t|w^g_t = s_i)$ is then a Gaussian distribution, where μ_y and Σ_y are the mean and variation, respectively. The initial hand-related HMM inference state is given by:

$$\begin{aligned} P(w^g_1) &= P(w^g_1|w^g_0, w^g_0) P(w^g_0, w^g_0) \\ &= P(w^g_1|w^g_0, w^s_0) P(w^g_0) P(w^s_0). \end{aligned} \tag{13}$$

w^g_1 can then be updated using y^g_1:

$$\begin{aligned} P(w^g_1 y^g_1) &= \frac{P(y^g_1|w^g_1) P(w^g_1)}{P(y^g_1)} \\ &= \frac{P(y^g_1|w^g_1) P(w^g_1|w^g_0, w^s_0) P(w^g_0) P(w^s_0)}{P(y^g_1)} \end{aligned} \tag{14}$$

In general, the state at time t is:

$$\begin{aligned} P(w^g_{1+t}) &= P(w^g_{1+t}|w^g_t, w^s_t) P(w^g_t, w^s_t) \\ &= P(w^g_{1+t}|w^g_t, w^s_t) P(w^g_t) P(w^s_t) \end{aligned} \tag{15}$$

An optimized state can be estimated from the observed sequences:

$$\begin{aligned} P(w^g_{1+t}|y^g_{1:1+t}) &= P(w^g_{1+t}|y^g_{1+t}, y^g_{1:t}) \\ &= P(w^g_{1+t}|y^g_{1+t}) \cdot P(w^g_{1+t}|y^g_{1:t}) \\ &= \alpha P(y^g_{1+t}|w^g_{1+t}) P(w^g_{1+t}) \cdot P(w^g_{1+t}|y^g_{1:t}) \\ &= \alpha P(y^g_{1+t}|w^g_{1+t}) P(w^g_{1+t}) \cdot P(w^g_{1+t}|y^g_{1:t}) \\ &= \alpha P(y^g_{1+t}|w^g_{1+t}) P(w^g_{1+t}|w^g_t, w^s_t) P(w^g_t) \\ &\quad \times P(w^s_t) \cdot P(w^g_{1+t}|y^g_{1:t}) \\ &= \alpha P(y^g_{1+t}|w^g_{1+t}) P(w^g_{1+t}|w^g_t, w^s_t) P(w^g_t) \\ &\quad \times P(w^s_t) \sum\nolimits_{w_t} P(w^g_{1+t}|w^g_t) P(w^g_t|y^g_{1:t}) \end{aligned} \tag{16}$$

Bayesian theory then produces [46]:

$$\begin{aligned} &\max_{w_1,\cdots,w_t} P(w^g_{1:1+t}|y^g_{1:1+t}) \\ &= \alpha P(y^g_{1+t}|y^g_t) \max_{w_t} P(w^g_{1+t}|w^g_t) \\ &\quad \times \max_{w_1,\cdots,w_{t-1}} P(w^g_t|y^g_{1:t}) \end{aligned} \tag{17}$$

with an optimal CSL classification prediction of:

$$\begin{aligned} (\hat{w}^g_{1:t})^* &= E[w^g_{1:t}|y^g_{1:t}] \\ &= \sum_w w^g_{1:t} \cdot (\max_{w_1,\cdots,w_{t-1}} P(w^g_{1:t}|y^g_{1:t})). \end{aligned} \tag{18}$$

In contrast, using hand-related HMMs as the primary probability network gives:

$$\begin{aligned} (\hat{w}^s_{1:t})^* &= E[w^s_{1:t}|y^s_{1:t}] \\ &= \sum_w w^s_{1:t} \cdot (\max_{w_1,\ldots,w_{t-1}} P(w^s_{1:t}|y^s_{1:t})). \end{aligned} \tag{19}$$

This leads to the final multimodal fusion CSL classification result:

$$\begin{aligned} (\hat{w}^{fusion}_{1:t})^* &= E[w^{fusion}_{1:t}|y^s_{1:t}, y^g_{1:t}] \\ &= \sum_w \{w^s_{1:t} \cdot (\max_{w_1,\ldots,w_{t-1}} P(w^s_{1:t}|y^s_{1:t})) \\ &\quad + w^g_{1:t} \cdot (\max_{w_1,\ldots,w_{t-1}} P(w^g_{1:t}|y^g_{1:t}))\}. \end{aligned} \tag{20}$$

IV. RESULTS AND DISCUSSION

A. DATASET

Two SLR datasets were used to evaluate the proposed model. As shown in Figs. 7-8 . The first CSL dataset was collected in-house and can be used for daily communication. The results of this analysis, including statistical information, are listed in Tab. 1. This dataset consisted of 50 continuous common CSL sentences, such as "What's your name?", "Don't forget to bring an umbrella", and "Hello, everyone." Each sentence was made up of 3-5 signs, with a total of 150 different isolated signs in the dataset, including human, you, we, ID card, happy, home, etc. In all, there were 500 instances of each isolated sign, for a total of $150 \times 500 = 75,000$ instances. The second CSL dataset, which included RGB-D Kinect images, focused on a large vocabulary [5]. It consisted of 500 different isolated signs for words such as head, body, lady, glass, etc., totaling 125,000 instances.

TABLE 1. Statistics for the 1st CSL dataset.

Modalities	Number of signers	FPS
RGB, depth, skeleton	5	30
RGB resolution	**Depth resolution**	**Number of nodes**
1280 × 720	512 × 424	25
Category	**Video duration**	**Total instances**
50 sentences, 150 isolated signs	2-4 seconds	75,000

B. EVALUATION OF LSTM2+CHMM PERFORMANCE

1) ISOLATED SIGN RECOGNITION

Isolated sign recognition was evaluated by dividing our CSL dataset into three components. Half of these data were used

FIGURE 7. An illustration of CSL sentences composed by isolating CSL words in our dataset.

FIGURE 8. An illustration of isolated words in the 2nd dataset (images reproduced from [5]).

for training and the other half for testing. The proposed method was assessed using an i5 core CPU with 8GB of RAM and the Microsoft Windows 10 operating system.

A subset containing 20 isolated CSL signs was selected to test the proposed algorithm and display the corresponding results. Fig. 9 shows a confusion matrix for subset recognition, indicating recognition accuracy to be highly satisfactory. Some of these signs, such as "Please" and "Happy" were similar, making them difficult to distinguish and decreasing the overall recognition rate. However, this feature could be improved further through increased training. The total recognition accuracy across 150 signs in our CSL dataset was 89.55%. Comparisons with existing techniques are provided in Tab. 2. It is evident the recognition accuracy gradually decreased with increasing database size. The main reasons can be explained as: first, the more data categories, the more factors to affect classification. Second, the more data categories, the more complex of classification surface, it is more difficult to obtain optimization results, resulting in lower classification rates. For example, using the SLR method based on LSTM, a database size of 50 produced an accuracy of 97.12%, while a database size of 150 produced a cognitive accuracy of 77.12%. It is evident the proposed LSTM+HMM approach outperformed conventional LSTM methods, with LSTM2+CHMM achieving even higher accuracy. In addition, data combinations based on color + depth tended

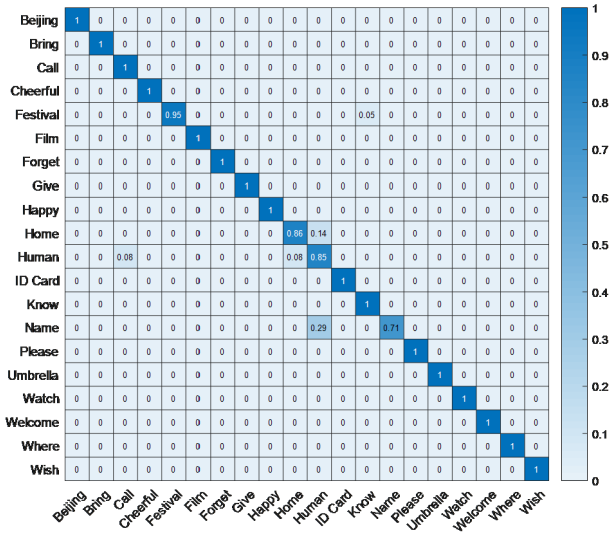

FIGURE 9. A confusion matrix displaying recognition accuracy for our CSL subset.

TABLE 2. Recognition results for isolated CSL words with various framework settings.

Method (Number of classes)	**Modality**	**Accuracy**
LSTM (50)	color + depth	97.12±0.89%
LSTM (100)	color + depth	88.33±1.59%
LSTM (150)	color + depth	77.12±2.33%
LSTM + HMM (50)	color + depth	98.22±0.21%
LSTM + HMM (100)	color + depth	89.22±1.58%
LSTM + HMM (150)	color + depth	85.22±1.69%
LSTM (50)	skeleton	93.19±0.65%
LSTM (100)	skeleton	83.19±0.99%
LSTM (150)	skeleton	78.19±1.99%
LSTM + HMM (50)	skeleton	98.55±0.22%
LSTM + HMM (100)	skeleton	88.21±1.66%
LSTM + HMM (150)	skeleton	82.55±2.23%
LSTM2 + CHMM (50)	color + depth + skeleton	99.55±0.21%
LSTM2 + CHMM (100)	color + depth + skeleton	91.15±1.89%
LSTM2 + CHMM (150)	color + depth + skeleton	89.55±3.59%

to produce better results than skeleton-based data, primarily because skeletal information does not include hand features. However, skeletons are more independent of environmental factors than color or depth information, making them suitable for SLR with non-specific people and environments. The proposed LSTM2+CHMM model incorporated color, depth, and skeleton data, fusing the images to produce robust segmentation results. Both segmented hands and skeletons are suitable for non-specific people and non-specific environments. As a result, this combined methodology exhibited the highest recognition accuracy.

2) CONTINUOUS CSL RECOGNITION

As most sign language databases are primarily used for isolated sign recognition, continuous sign language data are rarely available. As such, 50 sentences were established, composed of 150 isolated sign words, to test the continuous CSL recognition framework proposed in this paper. These included common phrases such as: "I am very happy to see you", "Do not forget to bring an umbrella", "This is my business card", "Is there a room?", etc.

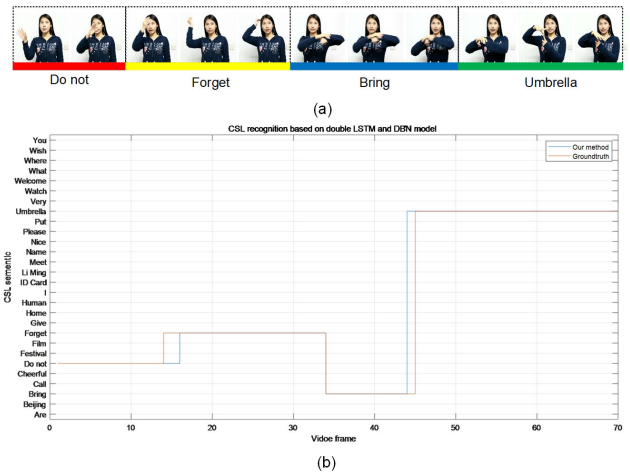

FIGURE 10. Results of the first test. (a) An illustration of continuous CSL recognition for "Don't forget to bring an umbrella." (b) CSL recognition results for the proposed method. There is little difference between the ground truth label and predicted labels. The video included 70 frames and produced a recognition accuracy of 97.3%.

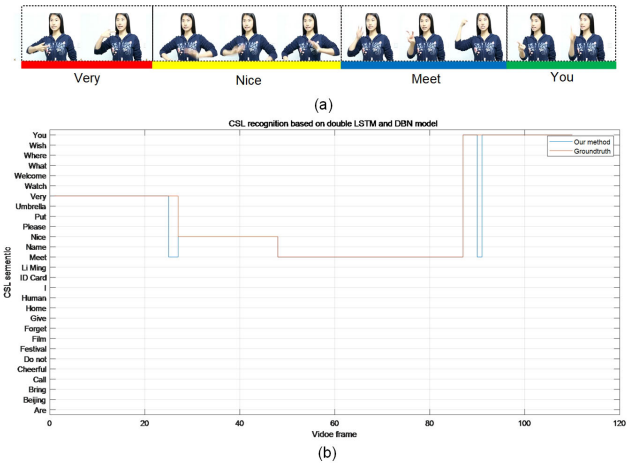

FIGURE 11. Results of the second test. (a) An illustration of continuous CSL recognition for "Nice to meet you." (b) CSL recognition results for the proposed method. There is little difference between the ground truth label and predicted labels. The video included 110 frames and produced a recognition accuracy of 97.1%.

Figs. 10-13 provide four examples of continuous CSL recognition. It is evident from these test results that the proposed algorithm can effectively identify continuous CSL terminology. The 1^{st}- 4^{th} examples produced recognition accuracies of 97.3%, 97.1%, 84.6%, and 98.1%, respectively, with an average value across all 50 sentences of 80.25%. A comparison with previous studies is provided in Tab. 3. It is evident that our method can automatically segment isolated signs, primarily because the framework uses LSTM and Markov chain-based probability algorithms for sequence data modeling.

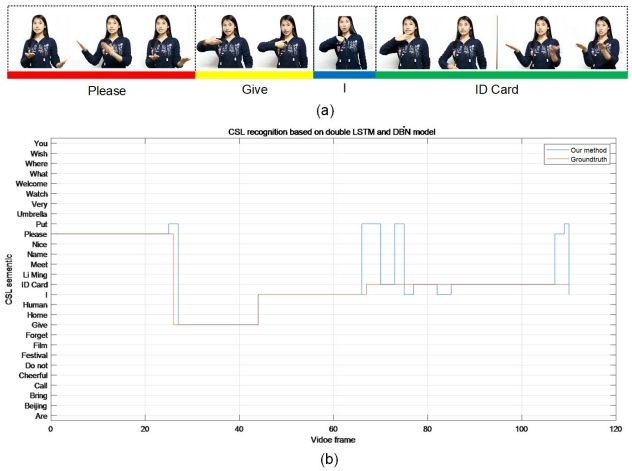

FIGURE 12. Results of the third test. (a) An illustration of continuous CSL recognition for "Please give me your ID card." (b) CSL recognition results for the proposed method. There is little difference between the ground truth label and predicted labels. The video included 110 frames and produced a recognition accuracy of 84.6%.

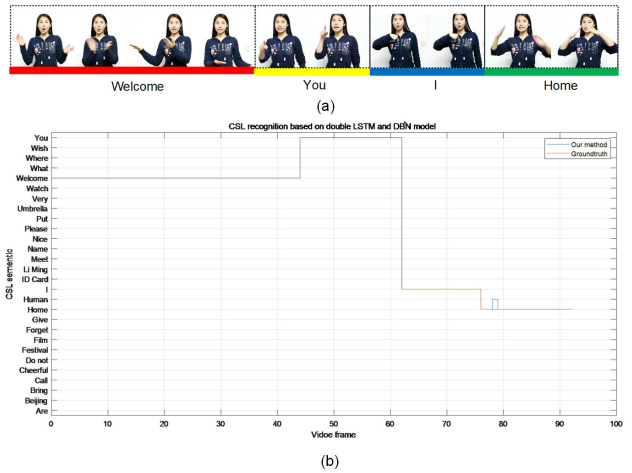

FIGURE 13. Results of the fourth test. (a) An illustration of continuous CSL recognition for "Welcome to my home." (b) CSL recognition results for the proposed method. There is little difference between the ground truth label and predicted labels. The video included 92 frames and produced a recognition accuracy of 98.1%.

We also compared continuous CSL recognition performance for different data sizes and various frameworks. As shown in Tab. 3, recognition accuracy gradually decreased as data increased, making large-scale SL recognition difficult. For example, LSTM recognition accuracy for 10 samples was 96.15%. However, this decreased to 73.19% with 50 samples, which may have been caused by a decline in isolated word recognition or increased sentence complexity. Regardless, the LSTM+HMM approach outperformed the LSTM method, with LSTM2+CHMM producing the highest recognition accuracy (80.25%) across all 50 sentences. These results indicate that multimodal fusion is beneficial for automated segmentation of SL sentences, independent of the data sampling environment.

TABLE 3. The results of continuous CSL recognition for various framework settings.

Method (Number of sentences)	**Modality**	**Accuracy**
LSTM (10)	color + depth	96.15±1.33%
LSTM (30)	color + depth	85.17±2.56%
LSTM (50)	color + depth	73.19±3.11%
LSTM + HMM (10)	color + depth	97.02±1.11%
LSTM + HMM (30)	color + depth	87.42±2.01%
LSTM + HMM (50)	color + depth	78.25±3.03%
LSTM (10)	skeleton	92.48±1.31%
LSTM (30)	skeleton	83.48±2.25%
LSTM (50)	skeleton	69.48±1.89%
LSTM + HMM (10)	skeleton	95.56±1.02%
LSTM + HMM (30)	skeleton	85.33 ± 2.62%
LSTM + HMM (50)	skeleton	75.56±2.02%
LSTM2 + CHMM (10)	color + depth + skeleton	98.36±1.08%
LSTM2 + CHMM (30)	color + depth + skeleton	90.25±2.78%
LSTM2 + CHMM (50)	color + depth + skeleton	80.25±3.01%

C. A COMPARISON OF RECOGNITION ACCURACY

1) BASELINE

In tasks related to time series processing, such as motion or speech recognition, integrating spatial and temporal information is critical for accurate classification. There are two primary strategies used for this process. The first utilizes manually-designed spatio-temporal features to build classifiers. The second directly constructs spatio-temporal models that can simulate hidden sequences, such as HMM. In this study, both techniques were used to design baselines for comparison with our approach.

LSTM2+CHMM is a tool for automatically extracting and processing spatio-temporal features. As such, it was compared with two other manually-developed spatio-temporal models: space-time interest point (STIP) [47] and improved dense trajectories (iDTs) [48]. A traditional Gaussian mixture-hidden Markov model (GMM-HMM) was also used as a baseline for comparison with the proposed technique. STIP is a common spatio-temporal feature and iDT has produced some of the best results to date. STIP calculates the HOG and HOF by detecting Harris corners in videos, while iDT calculates local HOG or HOF features based on optical flow tracking and low-level gradient histograms. After local feature extraction, a support vector machine (SVM) was utilized for CSL recognition in which segmented data and labels were used to train multi-class SVM classifiers. GMM-HMM is a traditional time series pattern classification method, similar to speech recognition or SLR. The baseline was developed using manually-extracted features to express motion sequences, after which statistical pattern recognition was used to train the GMM-HMM. Changes in both hand shape and body skeletal structure were observed to be highly distinguishing features for describing CSL movements and were thus used to train the GMM-HMM. This study followed the methodology presented by Tang *et al.*, who similarly used a Kinect as an input device and proposed an efficient algorithm for hand segmentation, tracking, and cropping hand shapes from a background [49]. However, this algorithm combined RGB and depth information without specific requirements for uniformity or stability. We extracted local

HOG features for the hand and simultaneously used the 3D positions of 25 skeletal joints, producing a 75D trajectory feature vector. After combining these two features, a 111D vector was used as the final representation for training the GMM-HMM. This vector was extracted from each video frame, classified, and used to train the GMM-HMM in CSL recognition.

2) RESULTS AND ANALYSIS

The proposed CSL recognition model was compared with conventional SLR algorithms, including GMM-HMM [27], adaptive HMM [22], 3D-CNN [5], iDTs+SVM [48], and STIP+SVM [47] using the 2^{nd} CSL dataset. The results of this performance comparison are shown in Table 4. It is evident the proposed technique achieved the best results despite differences in processing modes for sequence signals. We also compared the proposed fusion model to existing fusion models, such as SLR Fusion 1 and Fusion 2, each of which outperformed simple feature sequence processing. Although Fusion 1, Fusion 2, and the proposed model functioned similarly, our integration consistently achieved the best results.

TABLE 4. Accuracy comparison on different SLR methods.

SLR method	Fusion 1 [52]	Fusion 2 [22]	Our fusion
GMM-HMM [27]	$48.65 \pm 3.59\%$	$55.91 \pm 1.89\%$	$60.78 \pm 2.01\%$
A-HMM [22]	$61.21 \pm 3.21\%$	$68.33 \pm 3.33\%$	$70.15 \pm 3.25\%$
STIP-SVM [49]	$60.21 \pm 1.28\%$	$61.22 \pm 2.02\%$	$62.33 \pm 2.25\%$
iDT-SVM [50]	$67.75 \pm 1.21\%$	$68.58 \pm 1.18\%$	$70.21 \pm 1.13\%$
3D-CNN [5]	$70.57 \pm 2.98\%$	$75.58 \pm 2.18\%$	$79.33 \pm 1.25\%$
LSTM2+CHMM	$75.22 \pm 3.21\%$	$80.13 \pm 2.25\%$	$82.55 \pm 2.01\%$

Tab. 4 indicates that recognition algorithms based on a neural network generally perform better than HMM-based methods, such as 3DCNN or LSTM2+CHMM, with maximum accuracies of 79.33% and 82.55%, respectively. These values are much higher than single HMM-based methods, such as GMM-HMM or A-HMM (60.78% and 70.15%, respectively). This is likely because deep learning-based methods extract richer spatio-temporal information. It is also evident that our proposed multimodal combination technique includes other advantages. Specifically, our fusion model achieved a higher recognition accuracy than fusion 1 or fusion 2. The proposed fusion method is also completely independent of the data acquisition environment, allowing its application to non-specific people and scenes. In addition, identification methods based on artificially-designed features are often inferior to those acquired using deep learning, such as iDT- or STIP-based methods (recognition accuracies of 62.33% and 70.21%, respectively). Our method not only combines the long-term memory functions of RNNs with the automatic segmentation function of HMMs, but also uses a CNN to extract features automatically. The advantages of these methods are combined while the potential disadvantages are eliminated. As a result, our proposed CSL recognition algorithm achieved the highest accuracy.

V. CONCLUSION

A novel LSTM2+CHMM model was proposed for continuous CSL recognition. This method, which was based on multimodal data fusion, probabilistic inference, and deep learning, constructs two LSTM models and uses CHMM to fuse multiple modes. A CNN was used to extract features and two LSTMs were used to mine sequence information. SLR recognition results were acquired based on graph model probability inference. The proposed technique was evaluated using two SLR databases, with experimental results indicating it to be highly effective for accurate real-time CSL recognition.

ACKNOWLEDGMENT

(Qinkun Xiao and Minying Qin contributed equally to this work.)

REFERENCES

[1] A. Shamama, S. S. Kumar, V. Snehanshu, and A. Vishal, "Hand gesture recognition: A survey," in *Nanoelectronics, Circuits and Communication Systems—Proceeding of NCCS* (Lecture Notes in Electrical Engineering), vol. 511. 2019, pp. 365–371.

[2] D. A. Kumar, A. S. C. S. Sastry, P. V. V. Kishore, E. K. Kumar, and M. T. K. Kumar, "S3DRGF: Spatial 3-D relational geometric features for 3-D sign language representation and recognition," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 169–173, Jan. 2019.

[3] D. Avola, M. Bernardi, L. Cinque, G. L. Foresti, and C. Massaroni,, "Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 234–245, Jan. 2019.

[4] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. CVPR*, Jun. 2018, pp. 7784–7793.

[5] J. Huang, W. Zhou, H. Li, and W. Li, "Attention based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[6] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 2257–2264.

[7] E. K. Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. T. K. Kumar, and D. A. Kumar, "Training CNNs for 3-D sign language recognition with color texture coded joint angular displacement maps," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 645–649, May 2018.

[8] E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, D. A. Kumar, and A. S. C. S. Sastry, "Three-dimensional sign language recognition with angular velocity maps and connived feature resnet," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1860–1864, Dec. 2018.

[9] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1880–1891, Jul. 2018.

[10] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1311–1325, Dec. 2018.

[11] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu, "Recognizing american sign language gestures from within continuous videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2018, pp. 2145–2154.

[12] L. Pigou, M. V. Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3086–3093.

[13] G. Liu, Y. Liu, M. Guo, P. Li, and M. Li, "Variational inference with Gaussian mixture model and householder flow," *Neural Netw.*, vol. 109, pp. 43–55, Jan. 2019.

[14] C. L. C. Mattos and A. A. Barreto, "A stochastic variational framework for recurrent Gaussian processes models," *Neural Netw.*, vol. 112, pp. 54–72, Apr. 2019.

[15] W. Gao, J. Ma, J. Wu, and C. Wang, "Sign language recognition based on HMM/ANN/DP," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 14, no. 5, pp. 587–602, 2000.

[16] G. Fang, W. Gao, and D. Zhao, "Large vocabulary sign language recognition based on fuzzy decision trees," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 34, no. 3, pp. 305–314, May 2004.

[17] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1685–1699, Sep. 2009.

[18] H. Cheng, L. Yang, and Z. Liu, "Survey on 3D hand gesture recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 1659–1673, Sep. 2016.
[19] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 1975–1979.
[20] C. Dong, M. Leu, and Z. Yin, "American sign language alphabet recognition using microsoft Kinect," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–52.
[21] S. Escalera, X. Baró, J. Gonzàlez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-LópezHugo J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *Proc. Eur. Conf. Comput. Vis.*, Mar. 2014, pp. 459–473.
[22] D. Guo, W. Zhou, H. Li, and M. Wang, "Online early-late fusion based on adaptive HMM for sign language recognition," *ACM Trans. Multimedia Comput., Commun.*, vol. 14, no. 1, p. 8, Jan. 2017.
[23] T. Liu, W. Zhou, and H. Li, "Sign language recognition with long short-term memory," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2871–2875.
[24] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1692–1706, Aug. 2016.
[25] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping," in *Proc. Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Feb. 2013, pp. 620–625.
[26] Y. Lin, X. Chai, Y. Zhou, and X. Chen, "Curve matching from the view of manifold for sign language recognition," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 233–246.
[27] H. Wang, X. Chai, Y. Zhou, and X. Chen, "Fast sign language recognition benefited from low rank approximation," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–6.
[28] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.
[29] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz,, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4207–4215.
[30] J. Ye, H. Hu, G. J. Qi, and K. A. Hua, "A temporal order modeling approach to human action recognition from multimodal sensor data," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 2, p. 14, May 2017.
[31] F. S. Khan, R. M. Anwer, J. V. D. Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, " Color attributes for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3306–3313.
[32] Z. Wu, Y. G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *Proc. ACM Conf. Multimedia*, Nov. 2014, pp. 167–176.
[33] O. R. Terrades, E. Valveny, and S. Tabbone, "Optimal classifier fusion in a non-Bayesian probabilistic framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1630–1644, Sep. 2009.
[34] M. R. Abid, E. M. Petriu, and E. Amjadian, "Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 3, pp. 596–605, Mar. 2015.
[35] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[36] Y. Yuan, Y. Zhao, and Q. Wang, "Action recognition using spatial-optical data organization and sequential learning framework," *Neurocomputing*, vol. 315, pp. 221–233, Nov. 2018.
[37] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2012, pp. 1737–1746.
[38] C. Sun, T. Zhang, B. Bao, C. Xu, and T. Mei, "Discriminative exemplar coding for sign language recognition with Kinect," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1418–1428, Oct. 2013.
[39] Z. Ren, J. Yuan, and A. Zhengyou Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proc. 19th ACM Int. Conf. Multimedia*, Nov. 2011, pp. 1093–1096.
[40] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun./Jul. 2016, pp. 56–64.
[41] C. Sun, T. Zhang, and C. Xu, "Latent support vector machine modeling for sign language recognition with Kinect," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 2, p. 20, May 2015.
[42] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 2012, pp. 1097–1105.
[44] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.
[45] L. D. Cohen and I. Cohen, "Finite-element methods for active contour models and balloons for 2-D and 3-D images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1131–1147, Nov. 1993.
[46] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2010.
[47] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.
[48] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3169–3176.
[49] A. Tang, K. Lu, Y. Wang, J. Huang, and H. Li, "A real-time hand posture recognition system using deep neural networks," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 2, p. 21, May 2015.
[50] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. CVPR*, Jun. 2015, pp. 1741–1750.

QINKUN XIAO was born in 1974. He received the Ph.D. degree from Northwestern Polytechnic University, in 2007. From 2007 to 2009, he held a postdoctoral position at Tsinghua University. He is currently a Ph.D. Supervisor and a Professor with Xi'an technological University. His research interests include object recognition and information retrieval, dynamic Bayesian networks, and image processing.

MINYING QIN was born in 1994. She is currently a Graduate Student with Xi'an Technological University. Her research interests include motion recognition and video information processing.

PENG GUO was born in 1993. He is currently a pre-admitted Graduate Student with Xi'an Technological University. His research interests include motion recognition and video information processing.

YIDAN ZHAO was born in 1993. She is currently a Graduate Student with Xi'an Technological University. Her research interests include motion recognition and video information processing.