sensors

MDPI

Article

Recognition of Signed Expressions in an Experimental System Supporting Deaf Clients in the City Office

Tomasz Kapuscinski * and Marian Wysocki

Department of Computer and Control Engineering, Faculty of Electrical and Computer Engineering, Rzeszow University of Technology, W. Pola 2, 35-959 Rzeszow, Poland; mwysocki@kia.prz.edu.pl

* Correspondence: tomekkap@kia.prz.edu.pl; Tel.: +48-17-865-1614

Received: 30 March 2020; Accepted: 11 April 2020; Published: 13 April 2020

Abstract: The paper addresses the recognition of dynamic Polish Sign Language expressions in an experimental system supporting deaf people in an office when applying for an ID card. A method of processing a continuous stream of RGB-D data and a feature vector are proposed. The classification is carried out using the k-nearest neighbors algorithm with dynamic time warping, hidden Markov models, and bidirectional long short-term memory. The leave-one-subject-out protocol is used for the dataset containing 121 Polish Sign Language sentences performed five times by four deaf people. A data augmentation method is also proposed and tested. Preliminary observations and conclusions from the use of the system in a laboratory, as well as in real conditions with an experimental installation in the Office of Civil Affairs are given.

Keywords: human–computer interface; computer vision; sign language recognition

1. Introduction

According to the World Federation of the Deaf, there are about 70 million deaf people in the world. They cannot easily articulate words and have great difficulty in understanding written content and in expressing thoughts in writing. Sign language is their primary means of communication.

Sign language skills in the hearing community are negligible. Moreover, most technical communication systems use written or spoken languages. Therefore, deaf people face barriers in social contacts, and they find it difficult to function alone.

That is why works are undertaken to build technical devices supporting the communication of deaf people with their environment. One of the tasks that such systems must perform is automatic sign language recognition.

This work presents the recognition of Polish Sign Language (PSL) expressions used in the experimental vision-based system supporting deaf people in the office when applying for an ID card.

The main contributions of the paper are:

1. A method of processing the sequences of depth images and skeletons, acquired using the RGB-D sensor, to determine feature vectors independent of small changes in the user position and rotation,
2. Hand segmentation in the depth image by a modified version of the seeded region growing algorithm,
3. A feature vector proposal inspired by linguistic research on the so-called three main phonological features of a sign (shape, place of articulation, and movement),
4. The experimental selection of parameters for classifiers commonly used in time series recognition and evaluation of their effectiveness while recognizing the considered sign expressions,

5. The assessment of the impact of using augmented data in the training phase obtained by averaging the original sequences aligned by dynamic time warping,
6. The recognition of signed expressions taking into account division into thematic subgroups resulting from established conversation schemes,
7. The preliminary observations and conclusions from the use of the system in real conditions.

The structure of this paper is as follows. Section 2 provides the research background and relevant literature references. Section 3 defines the problem. Section 4 describes the data acquisition and processing, as well as the proposed feature vector. The classification methods are presented in Section 5. The experiments and the resulting conclusions are described in Section 6. A summary along with a proposal for further work are provided in Section 7.

2. Recent Works

Comprehensive overviews of the literature on sign language recognition can be found in [1–6]. The described solutions can be divided into methods using special data gloves or computer vision.

The first glove equipped with sensors transforming selected handshapes into electrical signals was patented in 1983 [7]. Since then, many solutions have been developed, where users have to wear special gloves, clothing, or other wearable sensors [8–12]. The advantage of using data gloves is precision. However, they restrain the user and limit his/her freedom.

As modern solutions should strive to ensure that human-computer communication occurs naturally, the use of color cameras has become the dominant trend. In the literature, there are publications on American sign language recognition (ASL), e.g., [13–15], Japanese (JSL), e.g., [16–18], German (GSL), e.g., [19], Chinese (CSL), e.g., [20–22], Taiwanese (TSL), e.g., [23,24], Dutch (DSL), e.g., [25,26], Australian (Auslan), e.g., [27], Polish (PSL), e.g., [28], and many others. Most often, these solutions were based on the detection of human skin color to extract the user's hands and face [29]. Classification was most often carried out using: hidden Markov models (HMM), e.g., [30–32], artificial neural networks (ANN), e.g., [33–35], dynamic time warping (DTW), e.g., [27,36], and other methods. Vision methods allow natural interaction and inclusion of non-manual features [37,38], but they are dependent on lighting conditions, background colors, and the user's clothing. Therefore, such solutions work only in controlled laboratory conditions.

The solution to these problems became possible with the appearance of RGB-D cameras on the market. Depth information allows separating the background by segmenting the person as a foreground object. It is also possible to incorporate 3D features into vectors describing recognized gestures. Moreover, gestures can be accurately described using multi-dimensional data structures, the so-called point clouds [39], expressed in real-world coordinates. It is worth mentioning that some RGB-D cameras, based on the time-of-flight principle, operate even in a dark room [40]. Solutions using stereovision and multi-camera systems have been known in the literature for a long time, e.g., [41–43]. Currently, most methods use RGB-D cameras, e.g., [44–47].

Deep learning methods are also used to recognize sign languages, e.g., [48–50]. However, the results obtained are not as groundbreaking as in the case of static image recognition. The reason is the lack of publicly available large training datasets described using commonly accepted annotation standards [1].

Most of the works available in the literature concern the recognition of single words or simple expressions. Only a few papers describe continuous sign language recognition, e.g., [48,50]. They require solving the problem of the temporal segmentation of the incoming data stream and the non-trivial task of distinguishing meaningful gestures from unintentional hand movements. An additional difficulty is the phenomenon of co-articulation involving the deformation of the ends and beginnings of adjacent sign characters and the epenthesis effect involving adding some transitions between signs. There are also significant differences in the speed of gesture performance.

In the literature, there are descriptions of several systems that use sign language to support deaf people. Examples of such solutions are: an Internet communicator [51], an application for

sending SMS [52], games for the deaf [53,54], and educational tools [55,56]. In most cases, however, emerging systems are not based on sign language recognition and, therefore, do not meet the natural interaction paradigm.

3. Problem Statement

The SyKoMisystem developed at the Department of Computer and Control Engineering of the Rzeszów University of Technology is a human-computer interface supporting deaf communication with an office clerk in a public institution. The signed expressions of a deaf person are observed by a camera and translated by a computer into written or spoken language for the office clerk. Selected from the database responses of the office clerk are presented to a deaf person in the form of a short movie in sign language.

The system has a modular structure and consists of sign-to-text and text-to-sign modules, as well as dictionaries of expressions adequate for the selected application. The flexible configuration of these modules enables the creation of derivative products that do not always require bi-directional translation.

The current version allows a trained deaf person to submit an application for an ID card or collect an ID card. It is possible to adapt the system to other topics in any public institution. Information boards or interactive "kiosks" can also be built using selected modules. Such devices can support deaf people at stations, in trains, buses, planes, or even supermarkets. It is also possible to create educational tools for learning sign language in the form of games. The experimental version of SyKoMi was installed in the Civil Affairs Department of Rzeszow City Hall.

4. Data Acquisition and Processing

An RGB-D sensor was used, which in addition to the color and depth images, also allowed the acquisition of skeletal data. Skeletal data consisted of a hierarchical set of interconnected segments, which were bone analogs. The developed method required that the skeleton contain at least the joints corresponding to the right hand (*rh*), left hand (*lh*), left shoulder (*ls*), and neck (*ss*) (Figure 1a).

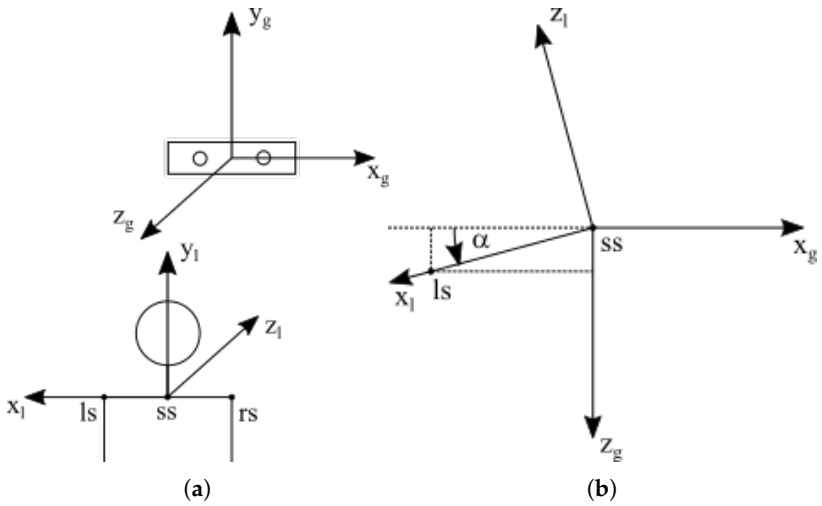

Figure 1. Global *g* and local *l* reference system: (**a**) illustrative view, (**b**) "top view" after translating the center of the *l* system into the center of the *g* system.

The Kinect sensor or any other RGB-D camera with dedicated software can be used to acquire such data.

The global scene reference system g was associated with the sensor, as shown in Figure 1. It was assumed that the deaf person sits in a chair opposite the sensor. Their hands are on their knees. Raising any hand above the plane $y = y_{hi}$ denotes the start of an expression, and lowering both hands below

$y = y_{lo}$, where $y_{hi} - y_{lo} > P_w$ and P_w is the palm width. The values y_{lo}, y_{hi}, and P_w were determined experimentally. The introduction of two limits: h_{lo} and y_{hi}, and the dead-band between them, greater than the palm width, prevented short-term state changes when the y coordinate of the palm oscillated around the thresholds.

The user was expected to sit in a fixed position relative to the sensor, but in fact, it was difficult to ensure this. Therefore, to become independent of minor changes in the user's position and orientation, the coordinates of his/her hands were expressed in the local reference system l associated with the user (Figure 1). Using the homogeneous coordinates, we get:

$$\begin{bmatrix} x_{l,rh} \\ y_{l,rh} \\ z_{l,rh} \\ 1 \end{bmatrix} = \begin{bmatrix} -cos(\alpha) & 0 & sin(\alpha) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(\alpha) & 0 & -cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & a \\ 0 & 1 & 0 & b \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{g,rh} \\ y_{g,rh} \\ z_{g,rh} \\ 1 \end{bmatrix} \tag{1}$$

where $\begin{bmatrix} a & b & c & 1 \end{bmatrix}^T$ and $\alpha = asin\left(\frac{z_{g,ls} - z_{g,ss}}{\sqrt{\left(x_{g,ls} - x_{g,ss}\right)^2 + \left(z_{g,ls} - z_{g,ss}\right)^2}}\right)$ denote the translation vector and rotation angle between the g or l system and the subscripts indicate the reference system and joint name (see Figure 1). The coordinates of the left hand $\begin{bmatrix} x_{l,lh} & y_{l,lh} & z_{l,lh} & 1 \end{bmatrix}^T$ in the local reference system were determined similarly.

The depth image was used to segment both hands because: (i) the use of an RGB image would involve several restrictions on the background color and clothing of the user; (ii) the depth map obtained using time-of-flight technology was more resistant to changes in lighting conditions; (iii) with sudden changes in lighting, some compensation algorithms are activated for the color stream, which reduces the frame rate and increases the blur effect; and (iv) 3D spatial information can be used to segment hands.

The shape of the hand could be determined by thresholding the depth image. However, such an approach required that it was the object closest to the camera. For some sign gestures, this is difficult to fulfill. Besides, when the user is sitting, his/her knees and thighs are closer. Therefore, an alternative solution, based on depth image segmentation by a modified version of the seeded region growing algorithm, was proposed. The initial position of the right hand $P_{d,rh} = [i_{d,rh}, j_{d,rh}]^T$ in the depth image reference system d (the seed of the algorithm) was determined based on its real-world coordinates in the g system and a pinhole camera model:

$$\begin{bmatrix} wi_{d,rh} \\ wj_{d,rh} \\ w \end{bmatrix} = \begin{bmatrix} \frac{1}{n_x} & 0 & \tilde{d_x} \\ 0 & -\frac{1}{n_y} & \tilde{d_y} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_x & 0 & 0 & 0 \\ 0 & f_y & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{g,rh} \\ y_{g,rh} \\ z_{r,rh} \\ 1 \end{bmatrix}, i_{d,rh} = \frac{wi_{d,rh}}{w}, j_{d,rh} = \frac{wj_{d,rh}}{w} \tag{2}$$

where $\tilde{d_x} = \left(\frac{w}{2} + d_x\right)\frac{1}{n_x}$, $\tilde{d_y} = \left(\frac{h}{2} - d_y\right)\frac{1}{n_y}$, f_x, f_y are the focal lengths along the x and y axes, n_x, n_y the pixel size, w, h the image sensor size, and d_x, d_y the coordinates of the point of intersection of the camera optical axis with its sensing plane.

The method `MapCameraPointToDepthSpace`from class `CoordinateMapper` available in the `Kinect for Windows v2 Windows Runtime API` has been used [57]. It uses the intrinsic parameters of the camera determined during the factory calibration. If another sensor is used, whose software does not offer such functionality, the camera calibration process should be performed [58]. Left hand position $P_{d,lh} = [i_{d,lh}, j_{d,lh}]^T$ was determined in an analogous manner.

To segment hand in the depth image D, the following modified version of the seeded region growing algorithm was proposed (Algorithm 1):

Algorithm 1: Hand segmentation in the depth image using the modified version of the seeded region growing.

Data: Depth image D, right hand initial position (seed) $P_{d,rh}$
Result: Right hand binary image B_{rh}
1 Create matrix B_{rh}, dim B_{rh} = dim D, and fill it with zeros;
2 Create an empty stack of points S;
3 Insert $P_{d,rh}$ at the top of S;
4 **while** $S \neq \emptyset$ **do**
5 Get $P_{d,n} = [i_{d,n}, j_{d,n}]^T$ from the top of S;
6 **if** $B_{rh}(i_{d,n}, j_{d,n}) \neq 1$ **then**
7 $B_{rh}(i_{d,n}, j_{d,n}) = 1$;
8 **foreach** *neighbor* $Q_{d,m} = [i_{d,m}, j_{d,m}]^T$ *of* $P_{d,n}$ **do**
9 $d_1 = \left|Q_{g,m}P_{g,n}\right|$;
10 $d_2 = \left|Q_{g,m}P_{g,rh}\right|$;
11 **if** $d_1 \leq maxNeighborDist$ & $d_2 \leq maxSeedDist$ **then**
12 Insert $Q_{d,m}$ at the top of S;
13 **end**
14 **end**
15 **end**
16 **end**

In Step 8, the 8-connectivity was used. The proposed modification consisted of transferring the decision from the image plane to the 3D space (Steps 9 and 10) and adding a new point to the hand when it was close enough to the adjacent point and not too far from the seed point (Condition 11). Transferring the decision to real-world coordinates reduced the method's dependence on the distance between the hand and sensor. Equivalents of points $P_{d,n}$ and $Q_{d,m}$ in the reference system g were determined using the formulas:

$$\begin{aligned} x_g &= \frac{(f_x + z_g)}{f_x}\left(\frac{W}{2} - i_d\right) n_x \\ y_g &= \frac{(f_y + z_g)}{f_y}\left(\frac{H}{2} - j_d\right) n_y \\ z_g &= \mu D(i_d, j_d) \end{aligned} \tag{3}$$

where W, H are respectively the width and height of the image, n_x, n_y denote the pixel size (mm/px), and μ is a scaling factor, which for the Kinect One sensor was equal to 1.

It was necessary to observe almost the entire silhouette of a person to recognize dynamic sign language expressions, so the user's hand occupied a relatively small area in the image. Therefore, its shape was roughly described by the coefficient of compactness:

$$\gamma_{rh} = \frac{P_{rh}^2}{4\pi S_{rh}} \tag{4}$$

and the eccentricity:

$$\epsilon_{rh} = \frac{(m_{rh,20} - m_{rh,02})^2 + 4m_{rh,11}^2}{S_{rh}^4} \tag{5}$$

were P_{rh} is the surface area, S_{rh} the circumference, and $m_{rh,pq}$ the central moment of the order pq [59] determined for the binary object corresponding to the right hand in the image B_{rh}. The slope of the main axis was also calculated:

$$\phi_{rh} = 0.5 atan \frac{2m_{rh,11}}{m_{rh,20} - m_{rh,02}} \tag{6}$$

The binary image with the left hand B_{lh} and the values γ_{lh}, ϵ_{lh}, and ϕ_{lh} were determined analogously. When hands contacted or mutual occlusion was detected, $B_{rh} \cap B_{lh} \neq \emptyset$, the method used the values of compactness, eccentricity, and slope from the previous frame. In the case of contacts with another part of the body, the *maxSeedDist* parameter was used. Its value was determined during the calibration and was equal to the maximum distance between the hand's center of gravity and its contour. Thanks to this, other points of the body, located further from the hand, were not included, even though they met the first condition given in Line 11 of Algorithm 1. This approach did not ensure an accurate separation of the hand, but with a rough description of the shape, it seemed to be sufficient.

The following feature vector was proposed:

$$v = [x_{l,rh}, y_{l,rh}, z_{l,rh}, \gamma_{l,rh}, \epsilon_{l,rh}, \phi_{l,rh}, x_{l,lh}, y_{l,lh}, z_{l,lh}, \gamma_{l,lh}, \epsilon_{l,lh}, \phi_{l,lh}] \tag{7}$$

Linguistic studies on sign languages revealed that each sign has three distinctive phonological features: handshape, location, and movement [60]. In the proposed approach, γ, ϵ, and ϕ describe handshape and x, y, and z the location, time series, and movement. Data were standardized to have a mean of 0 and a standard deviation of 1. Time series corresponding to the expression "When the ID card is ready?" presented in PSL are shown in Figure 2.

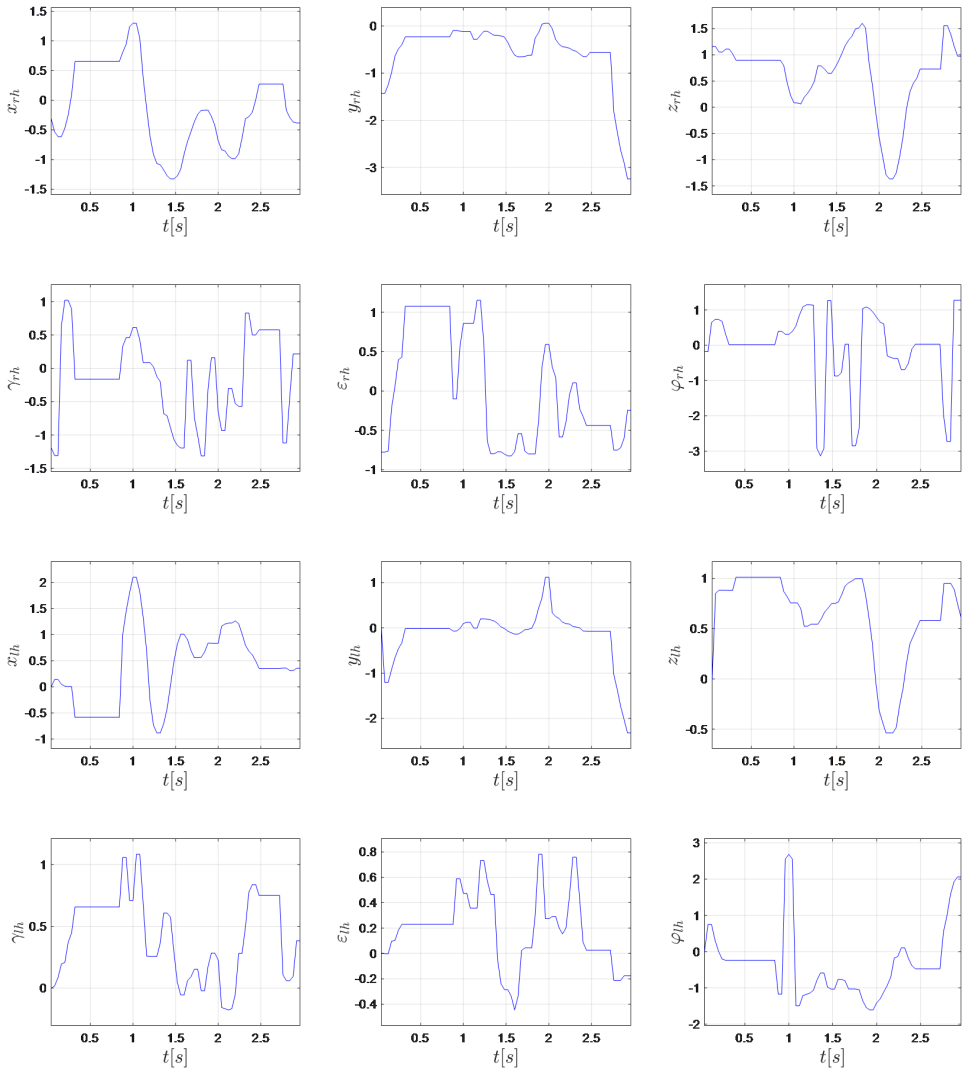

Figure 2. Time series corresponding to the expression "When the ID card is ready?" presented in Polish Sign Language (PSL).

5. The Classifiers

The classification was carried out using: (1) the k-nearest neighbors (k-NN) algorithm [61], (2) hidden Markov models (HMM) [62], and (3) bidirectional long short-term memory (BiLSTM) [63]. For k-NN, time series were compared using the dynamic time warping (DTW) method [64].

5.1. DTW

To compare time series corresponding to expressions of different lengths, we used DTW [65]. It nonlinearly maps one sequence to another by minimizing the distance between them and compares similar series that are locally out of phase by time scale extension of compression. Users performed some parts of gestures representing the same expression with different velocities, so this advantage was especially important. For two time series $Q = \{q(1), q(2), \ldots, q(T_q)\}$ and $R = \{r(1), r(2), \ldots, r(T_r)\}$, a $T_q \times T_r$ matrix was considered, where the (i, j) element of the matrix contained the distance $d(q(i), r(j))$ between two points $q(i)$ and $r(j)$. A warping path, $W = w_1, w_2, \ldots, w_K$, where $max(T_q, T_r) \leq K \leq T_q + T_r - 1$, is a set of matrix elements' indexes $w_k = (i_k, j_k)$ that satisfies three constraints: boundary condition, continuity, and monotonicity. The boundary condition constraint requires $w_1 = (1, 1)$ and $w_K = (T_q, T_r)$. The continuity constraint limits the allowed steps to adjacent cells. The monotonicity forces the monotonic arrangement of points on the warping path. The warping path that has the minimum distance $d_{DTW} = \sum_{k=1}^{K} \frac{d(w_k)}{K}$ between the two series is of interest, where $d(w_k) = d(q(i_k), r(j_k))$. It is estimated using dynamic programming. To prevent mapping a relatively small section of one sequence to a much larger section of another one and to speed up the computation, the warping window constraint was applied [66]. It consists of defining a narrow strip around the diagonal connecting points w_1, w_K.

As a result of the experiments, the city block metric was chosen, and the warping window constraint was set to 20 (Figure 3) [67].

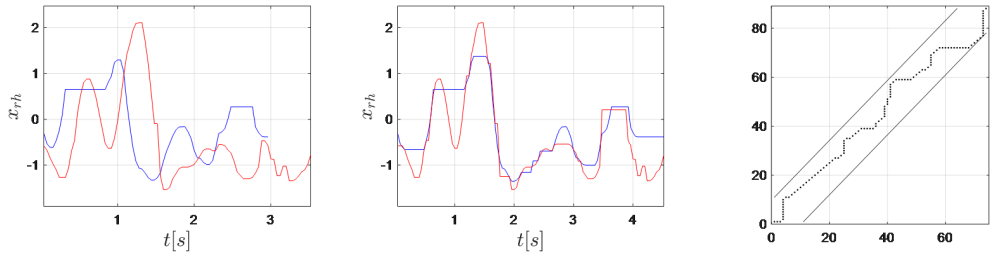

Figure 3. DTW of two time series corresponding to the expression "When the ID card is ready?" presented in PSL for the feature x_{rh}: (**a**) original time series, (**b**) time series after alignment, and (**c**) warping path with the warping window constraint.

The tests were carried out for the number of neighbors $k = 1, 3, 5$.

5.2. HMM

Gestures' executions are not perfect and may vary in speed and accuracy depending on mood or purpose. The human performance involves two distinct stochastic processes: immeasurable mental states and resultant actions that are measurable. Therefore to recognize dynamic gestures, we investigated hidden Markov models because they also consist of two stochastic processes. One of them was an unobservable Markov chain with a finite number of states, an initial state probability distribution, and a state transition probability matrix. The other one was a set of probability density functions associated with observations generated by each state. The model training consisted of an estimation of its parameters with the help of observation sequences. The expectation-maximization method (the Baum–Welch technique) can be used [62]. In the recognition step, the Viterbi algorithm

was used to identify the class represented by a model that gave the highest probability of generating a tested sequence.

The expression model was composed of a series of connected models corresponding to individual words. Bakis models were used. The number of emitting states per word and emission probability distributions were determined experimentally, separately for training on original and extended data (see Section 6.2). In the first case, the number of states per word was equal to 3, and the unimodal distribution of observations was chosen based on the results for the number of states from 1–5 and uni- and bi-modal observations (Figure 4).

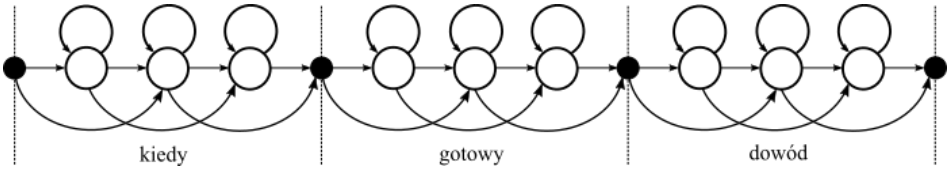

Figure 4. Bakis model for the expression "When the ID card is ready?" presented in PSL.

In the case of extended data, the number of states per word was equal to 9, and the unimodal distribution of observations was selected on the basis of experiments with the number of states being 3–13 and uni- and bi-modal observations.

5.3. BiLSTM

The LSTM classifier in the BiLSTM [68] version was used. The BiLSTM network is a modification of the long short-term memory (LSTM) network. The LSTM, first used by Hochreiter and Schmidhuber in 1997, is capable of learning long-term dependencies and is especially appropriate for the classification of time series. It has the chain structure shown in Figure 5 [69].

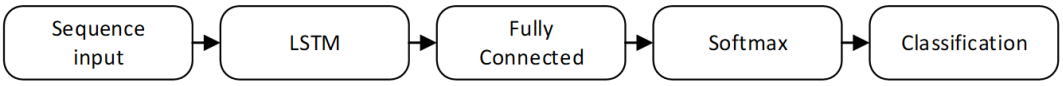

Figure 5. LSTM network architecture.

The sequence input layer introduces the data sequence or time series, and the LSTM layer learns the long-term relationships between the sequence time steps with its sophisticated structure, which consists of a set of recurrently connected memory blocks, each with one memory cell and three multiplicative gates: input, output, and forget gates. The gates control the long-term learning of sequence patterns. Each one is regulated by the sigmoid function, which learns during the training process, when to open and close, i.e., when to remember or forget information [69,70]. The network ends with a fully connected layer, a softmax layer, and a classification output layer to predict class labels. Unidirectional LSTM only preserves the information of the past because the only inputs it has seen are from the past. BiLSTM runs the inputs in two ways, one from past to future and one from future to past. This means that for every point in a given sequence, the BiLSTM has complete, sequential information about all points before and after it. The flow of data at time step t is shown in Figure 6.

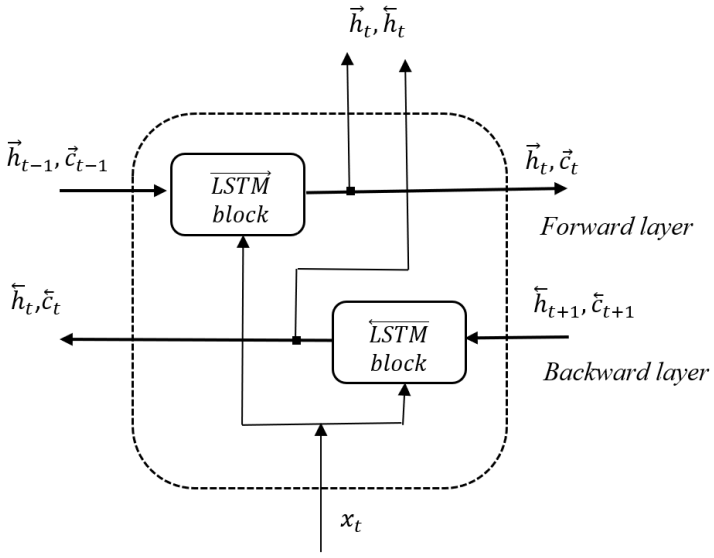

Figure 6. BiLSTM flow of data at time step t.

The hidden state $(\vec{h}_t, \overleftarrow{h}_t)$ is the output of the BiLSTM layer at time step t. The memory cell state $\vec{c}_{t-1}$ ($\overleftarrow{c}_{t+1}$) contains information learned from the previous (subsequent) time steps. At each time step t, the forward layer and the backward layer add information to or remove information from the respective cell state, based on the actual step of the sequence x_t. The layers control these updates using gates, as mentioned earlier. The BiLSTM network architecture is like that in Figure 5 with LSTM replaced by BiLSTM. It is especially useful when there is a need to teach the entire time series at every step. The BiLSTM usually learns faster than a one-directional approach, although it depends on the task. In our research, using BiLSTM was justified, because sign language expressions are sequences of interrelated elements. The hyperparameters were selected experimentally (see Section 6.4).

6. Experimental Results

6.1. Dataset and Tools

One-hundred-twenty-one Polish Sign Language expressions used to submit an application for an ID card or collect an ID card were recognized. Time series were obtained from recordings prepared in the SyKoMi system, as described in Section 4. Each expression was shown five times by four people, A, B, C, and D. Among them were two women and two men. Their age ranged from 20 to 63 years, height from 162–189 cm, and weight from 62 to 108 kg. The minimum number of words was 1, maximum 8, and average 3. The data acquisition frequency was 25 frames per second. The shortest expression lasted 0.960 and the longest 10.240 seconds. The average length was 3.327 seconds.

Expressions were divided into seven groups corresponding to thematic threads, which could be distinguished in typical conversation schemes. Group numbers, their descriptions, and sizes are given in Table 1.

Table 1. Division of expressions into groups.

Group Number	Description	Number of Expressions
I	Welcome and explanation of the reason for the visit	29
II	The reason for applying for an ID card	33
III	Application form	19
IV	Identity document	23
V	Photo	10
VI	Receiving ID card	14
VII	Fee	8

The following tools were used: MS Visual C++ 2015 (data acquisition and processing), MATLAB R2019b (DTW and BiLSTM), and the Hidden Markov Model Toolkit 3.4.1 (HMM) [71]. The experiments were carried out on a computer with an i7-6820HQ 2.7 GHz processor and 32 GB RAM.

6.2. Data Augmentation

The training set (set of patterns) was augmented with artificially generated data. Each pair of original time series related to the same expression was aligned using DTW and used to create three new ones according to the formula:

$$c = \tau a\left(a_s\right) + \left(1-\tau\right) b\left(b_t\right) \tag{8}$$

where a, b are the original time series, c the result series, a_s, b_t the indexes determining the "warping path" for a and b, and $\tau = -0.1, 0.5, 1.1$.

During the experiments, the leave-one-subject-out (l-o-s-o) protocol was adopted, i.e., sign expressions performed by a specific person were recognized by the classifier trained on data from remaining people. Thus, one expression had $5 \times 3 = 15$ original utterances in the training set. From 15 original time series, two could be selected in 105 ways, so $105 \times 3 = 315$ artificial time series for one expression were obtained. One expression in the training set was represented by original data and artificial data, i.e., a total of $15 + 315 = 330$ utterances.

6.3. Results

Three classifiers were used: (1) k-NN, (2) HMM, and (3) BiLSTM. Four variants were tested: ACOD, ACED, GOD, GED. OD and ED mean original and extended data, respectively. AC indicates that classifiers were trained using all classes. G means that classifiers were trained separately for each group. For G, the recognition rate of expressions from a given group was taken into account while assessing the classifier, and for the AC recognition rate, all classes were tested. To compare both variants, the results for all classes were decomposed into appropriate groups. The obtained results are presented in Table 2 and Figure 7.

Table 2. Recognition rates (%). AC, all classes; G, group; OD, original data; ED, extended data.

Classifier	Problem	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
k-NN	ACOD	86.90	88.48	87.37	86.30	83.50	88.21	98.13
	ACED	90.52	91.82	90.79	89.57	87.50	90.00	100.00
	GOD	96.21	91.36	96.05	91.96	92.00	97.14	100.00
	GED	98.10	93.03	98.42	93.70	94.50	98.57	100.00
HMM	ACOD	70.52	81.36	66.32	68.91	72.00	70.36	58.13
	ACED	75.52	88.94	72.63	75.00	75.00	77.50	61.25
	GOD	87.07	88.49	82.37	80.44	83.00	90.00	100.00
	GED	89.83	93.64	85.79	81.96	84.00	91.07	100.00
BiLSTM	ACOD	87.41	85.00	84.21	75.65	88.00	81.43	79.37
	ACED	90.34	86.82	88.68	78.04	89.00	78.21	86.88
	GOD	95.17	90.15	92.11	86.96	90.50	90.71	100.00
	GED	94.65	91.52	90.53	90.22	92.50	92.86	100.00

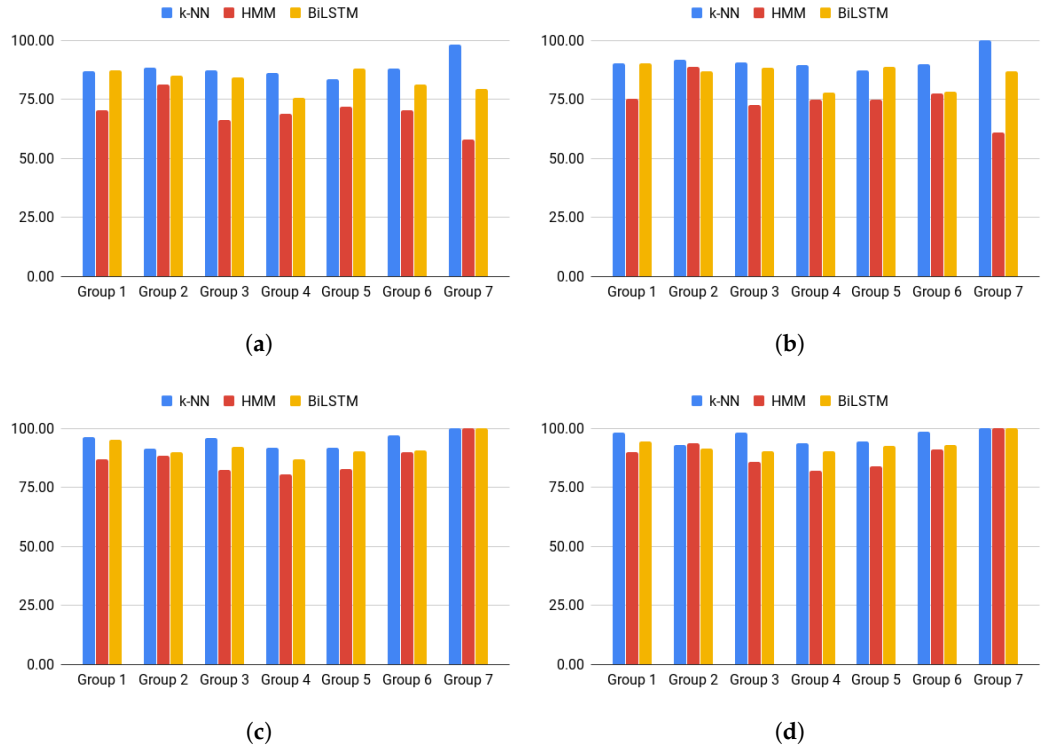

Figure 7. Recognition rates: (**a**) ACOD, (**b**) GOD, (**c**) ACED, (**d**) GED.

Classifier response times are given in Table 3.

Table 3. Classifier response times (s).

Problem	k-NN	HMM	BiLSTM
ACOD	0.200	0.120	0.001
ACED	3.500	0.260	0.001
GOD	0.055	0.080	0.002
GED	0.860	0.110	0.002

For the ACOD problem, the best results were obtained for k-NN (five of seven groups). BiLSTM was the best for the other two groups. In all seven cases, the worst results were obtained for HMM. It turned out that models with three states per word were insufficient to precisely model the recognized expressions.

After considering extended data (ACED problem), k-NN turned out to be the best classifier in six out of seven cases. Only for Group 5, the best result was obtained for BiLSTM. However, it is worth noting the significant increase in the response time of k-NN, which must compare a recognized sequence with each pattern from the extended dataset. Considering this, BiLSTM may be the best option. In most cases, an increase in recognition rates was observed compared to the variant based only on original data (ACOD). Depending on the group, kNN results improved from around 1.97 to 4.00 percentage points. For HMM, the improvement was much more significant, ranging from 3.00 to 7.58 points. It was possible to train larger models thanks to the extended data available. In the case of BiLSTM, results were better by 1.00 to 7.51 points, except for Group 6, for which the result was 3.22 points worse.

After division into groups (GOD problem), a significant increase in recognition efficiency was observed compared to ACOD. Depending on the group, it ranged from 1.87 to 9.31 points (k-NN), 7.13–41.87 points (HMM), and 2.50–20.63 points (BiLSTM). In six out of seven cases, the k-NN classifier was the best. For Group 7, 100% recognition rates were obtained for all classifiers.

After considering extended data, when training took place only on utterances belonging to the recognized group (GED), an increase in recognition rates was observed by 1.43 to 2.50 points (k-NN) and 1.00 to 5.15 points (HMM). For BiLSTM, recognition rates were better by 1.37 to 3.26 points for four of the seven groups. For Groups 1 and 3, there was a decrease in recognition rates by 0.52 and 1.58 points, respectively. In the GED problem, in five out of six cases, the k-NN classifier turned out to be the best. For Group 2, the best result was obtained for HMM, while for Group 7, all classifiers reached 100%.

6.4. BiLSTM Hyperparameters' Selection

BiLSTM network parameters were selected experimentally (Tables 4–6).

Table 4. Hyperparameter values in BiLSTM network training.

Parameter	Value
Size of the mini-batch (miniBatchSize)	15 (OD), 330 (ED)
Number of hidden units (numHiddenUnits)	150
Maximum number of epochs (maxEpochs)	100
Regularization factor (l2Regularization)	0.0001
Probability to drop out input elements (dropoutProbability)	Table 5
Initial learning rate (initialLearnRates)	Table 5
Learning rate schedule (learnRateSchedule)	"none"
Number of epochs for dropping the learning rate (learnRateDropPeriod)	10
Factor for dropping the learning rate (learnRateDropFactor)	0.1
Solver for training network (solverName)	"Adam"

The dropoutProbability and initialLearnRate were selected using the grid search approach through a specified discrete subset of parameters. The l-o-s-o results were used as a performance metric. Values obtained for the ACOD problem are shown in Table 5. In the same way, the selection of parameters for other problems was carried out (Table 6).

Table 5. An example of dropoutProbability and initialLearnRate selection in the ACOD problem.

DropoutProbability	InitialLearnRates	Recognition Rate (%)
0.1	0.0005	82.40
0.1	0.001	82.52
0.1	0.005	81.49
0.2	0.0005	81.24
0.2	0.001	81.57
0.2	0.005	80.04
0.3	0.0005	80.00
0.3	0.001	83.39
0.3	0.005	82.40

Table 6. DropoutProbability and initialLearnRate hyperparameters.

	DropoutProbability		InitialLearnRates	
	OD	ED	OD	ED
AC	0.3	0.3	0.001	0.001
G1	0.3	0.1	0.001	0.005
G2	0.2	0.1	0.001	0.005
G3	0.3	0.2	0.005	0.001
G4	0.2	0.3	0.001	0.005
G5	0.3	0.2	0.005	0.001
G6	0.2	0.3	0.001	0.005
G7	0.2	0.3	0.001	0.001

6.5. Comparison with Other Works

The comparison of our work with recent papers on dynamic sign language recognition using RGB-D data is given in Table 7.

Table 7. Comparison with other works. l-o-s-o, leave-one-subject-out.

Work	Year	Dataset	Deaf Involvement	Vocabulary Domain	Modality	Gesture Spotting	Classifier	Recognition Rate	Protocol l-o-s-o
[72]	2020	Chinese SL 500 words 50 users 5 repetitions	n/a	daily life	skeleton	manual	Bi-LSTM	82.55%	no
[73]	2019	Indian SL 200 words 10 users 10 repetitions	n/a	words describing human actions	RGB depth	manual	CNN	89.69%	yes
[74]	2018	Indian SL 30 words 10 users 9 repetitions	n/a	different unrelated words	skeleton	manual	HMM	83.77%	yes
[75]	2017	Chinese SL 370 words 5 users 5 repetitions	n/a	n/a	skeleton RGB	manual	Adaptive HMM	67.34%	yes
[76]	2017	Mexican SL 20 words 35 users 1 repetition	yes	greetings (4) questions (2) family (5) pronouns (2) places (2) others (2)	skeleton	manual	DTW	98.57%	no
[77]	2016	Chinese SL 500 words 1 user 5 repetitions	yes	n/a	skeleton RGB	manual	Adaptive HMM	98.80%	no
[78]	2016	Chinese SL 100 words 50 users 5 repetitions	n/a	daily life	skeleton	manual	HMM	82.70%	no
[79]	2016	Chinese SL 21 words 8 users 20 expressions 2 users	n/a	daily communication	skeleton	manual	HMM	88.00%	no
[80]	2015	American SL 73 words 63 expressions 10 users 3 repetitions	n/a	selected signs used by beginning signers	skeleton RGB	manual	Latent SVM	82.90%	n/a
[81]	2015	Indian SL 37 words 15 users 5 repetitions	yes	different unrelated words	skeleton	automatic	SVM	86.16%	no
[82]	2015	Arabic SL 16 words 4 users 3 repetitions	yes	words that can be used in hospital	skeleton RGB	manual	HMM	64.61%	yes
our	**2020**	**Polish SL** **121 expressions** **4 users** **5 repetitions**	**yes**	**submitting application for ID card**	**skeleton** **depth**	**automatic**	**k-NN** **HMM** **BiLSTM**	**98.57%** **91.07%** **92.86%**	**yes**

The proposed solution was practically oriented; therefore, it had some advantages over other works.

1. Only four papers mentioned the involvement of the deaf in a dataset's preparation. In our case, deaf people participated not only in the data collection, but also during the design of the user interface and verification of the system in a laboratory and real conditions. It is necessary to create a useful application for people with disabilities.
2. The vocabulary considered in this work concerned a specific application from a very narrow field. With the current state of knowledge about sign language recognition, only this approach guaranteed reliability that was acceptable in a real system. Other works concerned broader domains (e.g., daily activity) requiring vocabularies larger than the considered one to ensure consistent communication, or they focused on unrelated words.
3. Only two of the compared works concerned the recognition of signed expressions. However, the number of considered classes was smaller than in our case.
4. In all compared works except one, recognized words or expressions were extracted manually from the video stream. In our solution, a simple but effective algorithm for automatic expression spotting was employed.
5. More than half of the work used only skeletons. Therefore, these methods cannot be applied for gestures, for which the shape of the hand is the only distinctive feature.
6. Most works, especially those with better results, did not use the leave-one-subject-out (l-o-s-o) protocol, which allowed for the more accurate evaluation of the method.

7. Conclusions

The paper discussed the recognition of dynamic PSL expressions in an experimental system supporting deaf people in an office when applying for an ID card. A method of processing a continuous stream of RGB-D data recorded by the Kinect sensor and feature vector inspired by linguistic research was proposed. The classification was tested using three methods most commonly used to recognize dynamic gestures: k-NN, HMM, and BiLSTM. The l-o-s-o protocol was used for the dataset containing 121 PSL expressions performed five times by four people. Since the data collection process was a burdensome task and required the involvement of deaf people, the data augmentation method was also tested. Because the identified conversation patterns between an office clerk and a deaf person could be divided into compact thematic threads, the considered expressions were divided into inseparable groups, and the effectiveness of recognition in individual groups was examined. This allowed for higher recognition rates and shorter response times, which was important in the implemented system, especially for the k-NN classifier. In the current version of the system, switching between these groups was done by the office clerk, who was a conversation moderator.

The installation of the system in the office was experimental. However, it allowed gathering valuable experience. Preliminary tests of the system in real conditions showed that further work was needed, not only in the area of gesture recognition. Thanks to the use of the RGB-D sensor, the number of restrictions imposed on the deaf person significantly decreased, but the hands still should be kept on the knees in the intervals between utterances. It turned out that this was not an easy task for people for whom hands are the primary means of expression. In the future, therefore, it will be necessary to develop a better gesture spotting system, as well as a method of distinguishing intentional utterances from involuntary gestures and expressions from classes not belonging to the dictionary under consideration. With the current state of knowledge, solving this problem in a way that can be used in a practical system is still a big challenge [1].

Deaf people, especially the elderly, are afraid of modern technology. Therefore, it is necessary to increase their involvement not only in the process of collecting training datasets, but also in other stages of the software development and maintenance cycle. Their participation in the phase of user

interface design and system evaluation is particularly important. Training on the use of the system by deaf people is also necessary.

The system was dedicated, and its transfer or extension to another domain is a time-consuming and costly process. Recording extensive datasets and training classifiers are necessary. There are different sign language dialects in the country, which is an additional difficulty. Many words have their local equivalents or several interchangeable forms. Therefore, deaf people from a given region must participate in the dataset recording process.

It is also necessary to enrich the training set constantly, taking into account its diversity in terms of gender, age, clothing, lighting, and background. Moreover, further research in sign language linguistics should contribute to the development of better classifiers that increase system flexibility. However, the undeniable positive effect that has already been observed is a better understanding of deaf people's needs by office clerks operating the system.

Author Contributions: Idea of the study and methodology, T.K. and M.W.; software, T.K.; datasets, T.K.; experiments' design and discussion of the results, T.K. and M.W.; writing, review and editing, T.K. and M.W.; supervision and project administration, T.K. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This project is financed by the Minister of Science and Higher Education of the Republic of Poland within the "Regional Initiative of Excellence" program for the years 2019–2022, Project Number 027/RID/2018/19; amount granted 11,999,900 PLN.

Acknowledgments: The authors would like to express their deepest gratitude to the employees in the Office of Civil Affairs of Rzeszow and the deaf from the Subcarpathian Association of the Deaf for their kind assistance and support. This work is a continuation of the research carried out under The National Centre for Research and Development Grant TANGO1/270034/NCBR/2015.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bragg, D.; Koller, O.; Bellard, M.; Berke, L.; Boudreault, P.; Braffort, A.; Caselli, N.; Huenerfauth, M.; Kacorri, H.; Verhoef, T.; et al. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, Pittsburgh, PA, USA, 28–31 October 2019.
2. Cheok, M.J.; Omar, Z.; Jaward, M.H. A review of hand gesture and sign language recognition techniques. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 131–153. [CrossRef]
3. Lun, R.; Zhao, W. A Survey of Applications and Human Motion Recognition with Microsoft Kinect. *Int. Pattern Recognit. Artif. Intell.* **2015**, *29*, 1555008. [CrossRef]
4. Pisharady, P.K.; Saerbeck, M. Recent methods and databases in vision-based hand gesture recognition: A review. *Comput. Vis. Image Underst.* **2015**, *141*, 152–165. [CrossRef]
5. Rautaray, S.S.; Agrawal, A. Vision based hand gesture recognition for human computer interaction: A survey. *Artif. Intell. Rev.* **2015**, *43*, 1–54. [CrossRef]
6. Wadhawan, A.; Kumar, P. Sign Language Recognition Systems: A Decade Systematic Literature Review. *Arch. Comput. Meth. Eng.* **2019**, 1–29. [CrossRef]
7. Grimes, G.J. Digital Data Entry Glove Interface Device. U.S. Patent 4,414,537, 8 November 1983.
8. Fels, S.S.; Hinton, G.E. Glove-Talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Trans. Neural Netw.* **1993**, *4*, 2–8. [CrossRef]
9. Liang, R.-H.; Ouyoung, M. A real-time continuous gesture recognition system for sign language. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 558–567.
10. Oz, C.; Leu, M.C. American Sign Language word recognition with a sensory glove using artificial neural networks. *Eng. Appl. Artif. Intell.* **2011**, *24*, 1204–1213. [CrossRef]
11. Cooper, H.; Bowden, R. Sign language recognition using linguistically derived sub-units. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Valetta, Malta, 22–23 May 2010; pp. 57–61.

12. Pezzuoli, F.; Corona, D.; Corradini, M.L.; Cristofaro, A. Development of a Wearable Device for Sign Language Translation. In *Human Friendly Robotics*; Ficuciello, F., Ruggiero, F., Finzi, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 115–126.
13. Starner, T.E. Visual Recognition of American Sign Language Using Hidden Markov Models. Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1995.
14. Rybach, D.; Ney, I.H.; Borchers, J.; Deselaers, D.I.T. Appearance-Based Features for Automatic Continuous Sign Language Recognition. Master's Thesis, Diplomarbeit im Fach Informatik Rheinisch-Westf Alische Technische Hochschule Aachen, RWTH University, Aachen, Germany, 2006.
15. Zaki, M.M.; Shaheen, S.I. Sign language recognition using a combination of new vision based features. *Pattern Recognit. Lett.* **2011**, *32*, 572–577. [CrossRef]
16. Imagawa, K.; Lu, S.; Igi, S. Color-based hands tracking system for sign language recognition. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 462–467.
17. Tanibata, N.; Shimada, N.; Shirai, Y. Extraction of hand features for recognition of sign language words. In Proceedings of the 15th International Conference on Vision Interface, Calgary, AB, Canada, 27–29 May 2002; pp. 391–398.
18. Sako, S.; Kitamura, T. Subunit Modeling for Japanese Sign Language Recognition Based on Phonetically Depend Multi-stream Hidden Markov Models. In *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*; Stephanidis, C., Antona, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 548–555.
19. Bauer, B.; Kraiss, K. Video-based sign recognition using self-organizing subunits. In Proceedings of the Object Recognition Supported by User Interaction for Service Robots, Quebec City, QC, Canada, 11–15 August 2002; Volume 2, pp. 434–437.
20. Wang, J.; Gao, W. A Fast Sign Word Recognition Method for Chinese Sign Language. In *Advances in Multimodal Interfaces—ICMI 2000*; Tan, T., Shi, Y., Gao, W., Eds.; Springer: Berlin/Heidelberg, Germany, 2000; pp. 599–606.
21. Fang, G.; Gao, W.; Chen, X.; Wang, C.; Ma, J. Signer-Independent Continuous Sign Language Recognition Based on SRN/HMM. In *Gesture and Sign Language in Human-Computer Interaction*; Wachsmuth, I., Sowa, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2002; pp. 76–85.
22. Wang, C.; Gao, W.; Shan, S. An approach based on phonemes to large vocabulary Chinese sign language recognition. In Proceedings of the Fifth IEEE International Conference on Automatic Face Gesture Recognition, Washington, DC, USA, 21 May 2002; pp. 411–416.
23. Huang, C.L.; Huang, W.Y. Sign language recognition using model-based tracking and a 3D Hopfield neural network. *Mach. Vis. Appl.* **1998**, *10*, 292–307. [CrossRef]
24. Su, M.-C. A fuzzy rule-based approach to spatio-temporal hand gesture recognition. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2000**, *30*, 276–281.
25. Grobel, K.; Assan, M. Isolated sign language recognition using hidden Markov models. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, Orlando, FL, USA, 12–15 October 1997; Volume 1, pp. 162–167.
26. Lichtenauer, J.F.; Hendriks, E.A.; Reinders, M.J.T. Sign Language Recognition by Combining Statistical DTW and Independent Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 2040–2046. [CrossRef] [PubMed]
27. Holden, E.J.; Owens, R. Visual Sign Language Recognition. In *Multi-Image Analysis*; Klette, R., Gimel'farb, G., Huang, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 270–287.
28. Kapuscinski, T.; Wysocki, M. Using Hierarchical Temporal Memory for Recognition of Signed Polish Words. In *Computer Recognition Systems 3*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 355–362.
29. Terrillon, J.; Shirazi, M.N.; Fukamachi, H.; Akamatsu, S. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 28–30 March 2000; pp. 54–61.
30. Bose, S.; Jain, R.; Wu, Y.; Iyengar, S. A New Generalized Computational Framework for Finding Object Orientation Using Perspective Trihedral Angle Constraint. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *21*, 961–975.

31. Marcel, S.; Bernier, O.; Viallet, J.; Collobert, D. Hand gesture recognition using input-output hidden Markov models. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 28–30 March 2000; pp. 456–461.
32. Just, A.; Marcel, S. A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition. *Comput. Vis. Image Underst.* **2009**, *113*, 532–543. [CrossRef]
33. Yang, M.-H.; Ahuja, N. Extraction and classification of visual motion patterns for hand gesture recognition. In Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, USA, 25 June 1998.
34. Yang, M.-H.; Ahuja, N.; Tabb, M. Extraction of 2D motion trajectories and its application to hand gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1061–1074. [CrossRef]
35. Ng, C.W.; Ranganath, S. Real-time gesture recognition system and application. *Image Vis. Comput.* **2002**, *20*, 993–1007. [CrossRef]
36. Corradini, A. Dynamic time warping for off-line recognition of a small gesture vocabulary. In Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Vancouver, BC, Canada, 13 July 2001.
37. Caridakis, G.; Asteriadis, S.; Karpouzis, K. Non-manual Cues in Automatic Sign Language Recognition. *Pers. Ubiquitous Comput.* **2014**, *18*, 37–46. [CrossRef]
38. Yang, H.D.; Lee, S.W. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognit. Lett.* **2013**, *34*, 2051–2056. [CrossRef]
39. Rusu, R.B.; Cousins, S. 3D is here: Point Cloud Library (PCL). In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011.
40. Li, L. Time-of-Flight Camera—An Introduction. Tech. White Paper. 2014. Available online: http://www.ti.com/lit/wp/sloa190b/sloa190b.pdf (accessed on 30 March 2020).
41. Wang, Q.; Chen, X.; Zhang, L.G.; Wang, C.; Gao, W. Viewpoint invariant sign language recognition. *Comput. Vis. Image Underst.* **2007**, *108*, 87–97. [CrossRef]
42. Dreuw, P.; Steingrube, P.; Deselaers, T.; Ney, H. Smoothed Disparity Maps for Continuous American Sign Language Recognition. In *Pattern Recognition and Image Analysis*; Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I., Eds.; Springer Berlin Heidelberg: Berlin/Heidelberg, Germany, 2009; pp. 24–31.
43. Laskar, M.A.; Das, A.J.; Talukdar, A.K.; Sarma, K.K. Stereo Vision-based Hand Gesture Recognition under 3D Environment. *Procedia Comput. Sci.* **2015**, *58*, 194–201. [CrossRef]
44. Uebersax, D.; Gall, J.; Van den Bergh, M.; Van Gool, L. Real-time sign language letter and word recognition from depth data. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 383–390.
45. Zafrulla, Z.; Brashear, H.; Starner, T.; Hamilton, H.; Presti, P. American Sign Language Recognition with the Kinect. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; ACM: New York, NY, USA, 2011; pp. 279–286.
46. Oszust, M.; Wysocki, M. Polish sign language words recognition with Kinect. In Proceedings of the 6th International Conference on Human System Interactions (HSI), Gdansk, Poland, 6–8 June 2013; pp. 219–226.
47. Kapuscinski, T.; Oszust, M.; Wysocki, M.; Warchol, D. Recognition of Hand Gestures Observed by Depth Cameras. *Int. J. Adv. Robot. Syst.* **2015**, *12*, 36. [CrossRef]
48. Koller, O.; Zargaran, S.; Ney, H.; Bowden, R. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *Int. J. Comput. Vis.* **2018**, *126*, 1311–1325. [CrossRef]
49. Koller, O.; Camgoz, C.; Ney, H.; Bowden, R. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef] [PubMed]
50. Cui, R.; Liu, H.; Zhang, C. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Trans. Multimedia* **2019**, *21*, 1880–1891. [CrossRef]
51. Five App—Sign Language Messenger. Available online: https://fiveapp.mobi/ (accessed on 12 November 2019).
52. SSMS—Sign Short Message Service. Available online: http://www.ssmsapp.com/ (accessed on 12 November 2019).

53. Lee, S.; Henderson, V.; Hamilton, H.; Starner, T.; Brashear, H.; Hamilton, S. A Gesture-based American Sign Language Game for Deaf Children. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2005; pp. 1589–1592.
54. Brashear, H.; Henderson, V.; Park, K.H.; Hamilton, H.; Lee, S.; Starner, T. American Sign Language Recognition in Game Development for Deaf Children. In Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, Portland, OR, USA, 23–25 October 2006; pp. 79–86.
55. Reis, J.; Solovey, E.T.; Henner, J.; Johnson, K.; Hoffmeister, R. ASL CLeaR: STEM Education Tools for Deaf Students. In Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility, Lisbon, Portugal, 26–28 October 2015; pp. 441–442.
56. AlShammari, A.; Alsumait, A.; Faisal, M. Building an Interactive E-Learning Tool for Deaf Children:Interaction Design Process Framework. In Proceedings of the IEEE Conference on e-Learning, e-Management and e-Services, Langkawi Island, Malaysia, 21–22 November 2018; pp. 85–90.
57. Kinect for Windows v2 Windows Runtime API Reference. Available online: https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn758675(v=ieb.10) (accessed on 28 December 2019).
58. Camera Calibration Toolbox for Matlab. Available online: http://www.vision.caltech.edu/bouguetj/calib_doc/ (accessed on 28 December 2019).
59. Hu, M.-K. Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theor.* **1962**, *8*, 179–187.
60. Stokoe, W.C., Jr. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *J. Deaf Stud. Deaf Educ.* **2005**, *10*, 3–37. [CrossRef]
61. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.* **1967**, *13*, 21–27. [CrossRef]
62. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
63. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
64. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 43–49. [CrossRef]
65. Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*; Elsevier: New York, NY, USA, 2003.
66. Ratanamahatana, C.A.; Keogh, E. Three myths about dynamic time warping data mining. In Proceedings of the SIAM International Conference on Data Mining, Newport Beach, CA, USA, 21–23 April 2005; pp. 506–510.
67. Paliwal, K.; Agarwal, A.; Sinha, S. A modification over Sakoe and Chiba's dynamic time warping algorithm for isolated word recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France, 3–5 May 1982; pp. 1259–1261.
68. Schuster, M.; Paliwal, K. Bidirectional recurrent neural networks. *Signal Process. IEEE Trans.* **1997**, *45*, 2673–2681. [CrossRef]
69. Long Short-Term Memory Networks. Available online: https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html (accessed on 1 January 2020).
70. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]
71. HTK Speech Recognition Toolkit. Available online: http://htk.eng.cam.ac.uk/ (accessed on 18 March 2020).
72. Xiao, Q.; Qin, M.; Yin, Y. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Netw.* **2020**, *125*, 41–55. [CrossRef] [PubMed]
73. Ravi, S.; Suman, M.; Kishore, P.; Kumar, K.; Kumar, A. Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition. *J. Comput. Lang.* **2019**, *52*, 88–102. [CrossRef]
74. Kumar, P.; Saini, R.; Roy, P.P.; Dogra, D.P. A position and rotation invariant framework for sign language recognition (SLR) using Kinect. *Multimedia Tools Appl.* **2018**, *77*, 8823–8846. [CrossRef]
75. Guo, D.; Zhou, W.; Li, H.; Wang, M. Online Early-Late Fusion Based on Adaptive HMM for Sign Language Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* **2017**, *14*. [CrossRef]
76. García-Bautista, G.; Trujillo-Romero, F.; Caballero-Morales, S.O. Mexican sign language recognition using kinect and data time warping algorithm. In Proceedings of the International Conference on Electronics, Communications and Computers (CONIELECOMP), Cholula, Mexico, 22–24 February 2017; pp. 1–5.

77. Zhang, J.; Zhou, W.; Xie, C.; Pu, J.; Li, H. Chinese sign language recognition with adaptive HMM. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
78. Pu, J.; Zhou, W.; Zhang, J.; Li, H. Sign Language Recognition Based on Trajectory Modeling with HMMs. In *International Conference on Multimedia Modeling*; Springer International Publishing: Cham, Switzerland, 2016; pp. 686–697.
79. Yang, W.; Tao, J.; Ye, Z. Continuous sign language recognition using level building based on fast hidden Markov model. *Pattern Recognit. Lett.* **2016**, *78*, 28–35. [CrossRef]
80. Sun, C.; Zhang, T.; Xu, C. Latent Support Vector Machine Modeling for Sign Language Recognition with Kinect. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*. [CrossRef]
81. Mehrotra, K.; Godbole, A.; Belhe, S. Indian Sign Language Recognition Using Kinect Sensor. In *International Conference Image Analysis and Recognition*; Springer International Publishing: Cham, Switzerland, 2015; pp. 528–535.
82. Sarhan, N.A.; El-Sonbaty, Y.; Youssef, S.M. HMM-based arabic sign language recognition using kinect. In Proceedings of the Tenth International Conference on Digital Information Management (ICDIM), Gyeongju, Korea, 15–17 October 2015; pp. 169–174.