Sign Language Recognition for Assisting the Deaf in Hospitals

Necati Cihan Camgöz[1](✉), Ahmet Alp Kındıroğlu[2], and Lale Akarun[2]

[1] University of Surrey, Guildford, Surrey GU2 7XH, UK
n.camgoz@surrey.ac.uk

[2] Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
{alp.kindiroglu,akarun}@boun.edu.tr

Abstract. In this study, a real-time, computer vision based sign language recognition system aimed at aiding hearing impaired users in a hospital setting has been developed. By directing them through a tree of questions, the system allows the user to state their purpose of visit by answering between four to six questions. The deaf user can use sign language to communicate with the system, which provides a written transcript of the exchange. A database collected from six users was used for the experiments. User independent tests without using the tree-based interaction scheme yield a 96.67 % accuracy among 1257 sign samples belonging to 33 sign classes. The experiments evaluated the effectiveness of the system in terms of feature selection and spatio-temporal modelling. The combination of hand position and movement features modelled by Temporal Templates and classified by Random Decision Forests yielded the best results. The tree-based interaction scheme further increased the recognition performance to more than 97.88 %.

Keywords: Sign language recognition · Assistive computer vision · Human computer interaction

1 Introduction

Sign Languages are the main communication medium of the hearing impaired. They are visual languages in which concepts are conveyed through the positioning, shape and movements of hands, arms and facial expressions. Similar to spoken languages, sign languages developed over time in local communities. For this reason, they show great variation from spoken languages and across other sign languages.

The education of the hearing impaired is a difficult task. Since they are socially isolated due to a communication barrier, they have difficulty learning the spoken language, even in its written form. Therefore literacy of spoken and written language is considerably lower for the hearing impaired. This greatly impedes their integration into society and causes difficulties in receiving education, finding jobs and using everyday public services such as healthcare and banking.

M. Chetouani et al. (Eds.): HBU 2016, LNCS 9997, pp. 89–101, 2016.
DOI: 10.1007/978-3-319-46843-3_6

However, laws mandate the provisioning of assistance to the hearing impaired by providing translators on demand. While it would have increased accessibility greatly, problems were present in its application as there simply were not that many Turkish Sign Language (TİD) translators available. A practical solution was found by making sign language call centers available. However, they had their drawbacks as the call centers had to employ large numbers of translators to service a large deaf population. The ideal solution to this problem is to have software that performs automatic sign language to spoken language translation, thus allowing the hearing impaired people to express themselves in public institutions to receive services.

With the development of machine learning and computer vision algorithms and the availability of different sign language databases, there has been an increasing number of studies in Sign Language Recognition (SLR). Since the work of Starner and Pentland [16] there have been many studies attempting to recognize sign language gestures using spatio-temporal modeling methods such as Hidden Markov Models (HMMs) [14] and Dynamic Time Warping (DTW) [1] based methods. Other approaches, such as Parallel Hidden Markov Models (PaHMMs) [19] and HMM-based threshold model [10], are also used in gesture and sign language recognition systems. Chai et al. [4] used DTW based classifiers to develop a translation system that interprets Chinese Sign Language to Spoken Language and vice versa. In more recent studies, Pitsikalis and Theodorakis et al. [13,18] used DTW to match subunits in Greek Sign Language for recognition purposes.

Prior to the release of consumer depth cameras, such as the Microsoft Kinect sensor [22], many computer vision researchers had to use color and data gloves, embedded accelerometers and video cameras to capture a users hand and body movements for sign language recognition [12]. However, the Microsoft Kinect sensor provides color image, depth map, and real-time human pose information [15], by which it diminishes the dependency to such variety of sensors.

Recently, there has been an increase in studies aimed at developing prototype applications with sign language based user interfaces. One of the earliest applications was the TESSA (Text and Sign Support Assistant) [5], that was developed for the UK Post Offices to assist a post office clerk in communicating with a Deaf person. The TESSA system translates a clerks speech into British Sign Language (BSL) and then displays the signs to the screen with an avatar to a Deaf customer at the post office. The authors used the entropic speech recognizer and performed semantic mapping on a "best match" basis to recognize the most phonetically close phrase. Lopez-Ludena et al. [11] have also designed an automatic translation system for bus information that translates speech to Spanish Sign Language (LSE) and sign language to speech.

In [20], Weaver and Starner introduced SMARTSign, which aims to help the hearing parents of deaf children with learning and practicing ASL via a mobile phone application. The authors share the feedback they received from the parents on the usability and accessibility of the SMARTSign system. In [9], sign language tutoring is performed using a signing robot and interaction tests are used to asses system success. In [21], an avatar based sign language game is developed for teaching first grade curriculum in sign language to primary school children and assessing their knowledge.

When a deaf person arrives at a hospital, if he/she does not know how to read and write in spoken language, it is often a troublesome practice to communicate. To overcome this communication barrier, a sign language recognition platform called HospiSign was created. When deployed on a computer with a Microsoft Kinect v2 sensor, HospiSign works as a reception desk, welcoming deaf users and allowing them to express their purpose of visit.

The user interface of HospiSign was presented in [17]. In this paper, we focus on the sign language recognition aspects of the system. We proposed using several features, temporal modelling techniques and classification methods for sign language recognition. As features, we extracted upper body pose, hand shape, hand position and hand movement features from the data provided by the Microsoft Kinect v2 sensor to represent the spatial features of the signs. We model the temporal aspect of the signs by using Dynamic Time Warping (DTW) and Temporal Templates (TT). Finally, we classify spatio-temporal features extracted from the isolated sign phrases using k-Nearest Neighbours (k-NN) and Random Decision Forest (RDF) classifiers.

We evaluated the performance of the proposed recognition scheme on a subset of the BosphorusSign corpus [3], that contains a total of 1257 samples belonging to 33 signs, which were collected from six native TİD users. We investigated each features effect on the recognition performance and compared temporal modelling and classification approaches. In our experiments, combining hand position and hand movement features achieved the highest recognition performance while both of the temporal modelling and classification approaches yielded satisfactory recognition results. Moreover, we inspected the outcome of using the tree-based activity diagram interaction scheme and came to the conclusion that this approach increases the overall recognition performance.

In Sect. 2, we briefly explain the tree-based activity diagram interaction scheme in HospiSign. Section 3 describes our proposed sign language recognition method. Experimental results are given in Sect. 4 and finally, we conclude the paper in Sect. 5.

2 The Hospital Information System User Interface

The hospital information system user interface provides a communication medium for the hearing impaired in a hospital information desk setting. By asking questions in the form of sign videos and suggesting possible answers on a display, the system helps Deaf users to explain their problems. With the tree-based activity diagram interaction scheme, which can be seen in Fig. 1, the system only looks for the possible answers in each activity group, instead of trying to recognize from all the signs in the database. At the end of the interaction, the system prints out a summary of the interaction and the users are guided to take this print out with their ID to the information desk, where they can be assisted according to their needs.

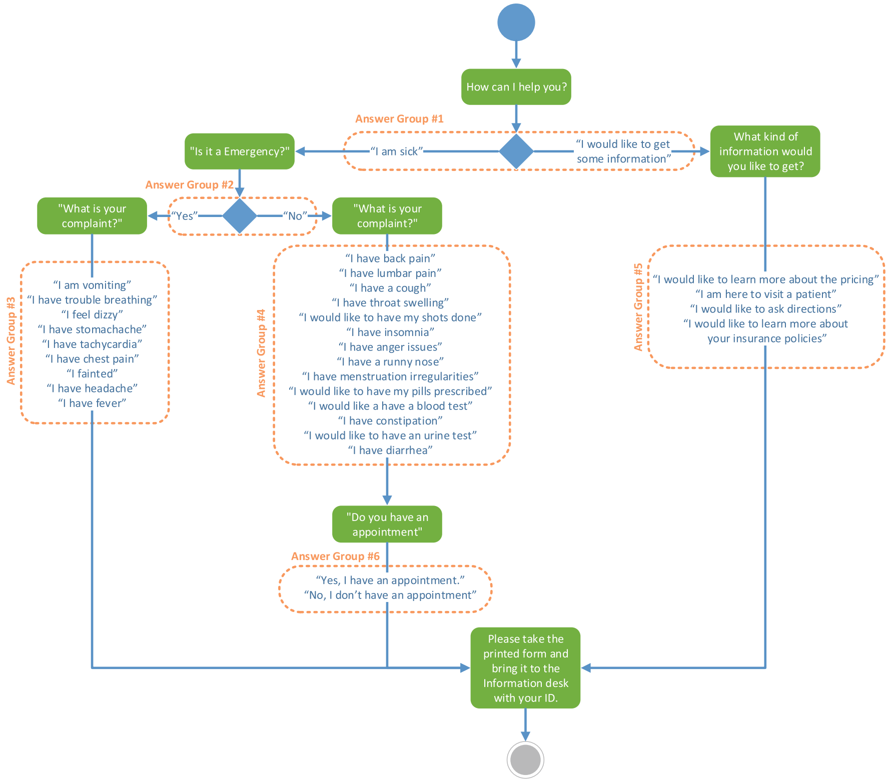
How can I help you?
Answer Group #1
"Is it a Emergency?"
"I am sick"
"I would like to get some information"
What kind of information would you like to get?
Answer Group #2
"What is your complaint?"
"Yes"
"No"
"What is your complaint?"
Answer Group #3
"I am vomiting"
"I have trouble breathing"
"I feel dizzy"
"I have stomachache"
"I have tachycardia"
"I have chest pain"
"I fainted"
"I have headache"
"I have fever"
Answer Group #4
"I have back pain"
"I have lumbar pain"
"I have a cough"
"I have throat swelling"
"I would like to have my shots done"
"I have insomnia"
"I have anger issues"
"I have a runny nose"
"I have menstruation irregularities"
"I would like to have my pills prescribed"
"I would like a have a blood test"
"I have constipation"
"I would like to have an urine test"
"I have diarrhea"
Answer Group #5
"I would like to learn more about the pricing"
"I am here to visit a patient"
"I would like to ask directions"
"I would like to learn more about your insurance policies"
"Do you have an appointment"
Answer Group #6
"Yes, I have an appointment."
"No, I don't have an appointment"
Please take the printed form and bring it to the Information desk with your ID.

Fig. 1. Tree-based activity diagram interaction scheme of HospiSign.

The HospiSign platform consists of a personal computer, a touch display to visualize the sign questions and answers to the user, and a Microsoft Kinect v2 sensor. Since it is necessary to track the users' hand motions in order to recognize the performed signs, the Microsoft Kinect v2 sensor plays an essential role as it provides accurate real-time human body pose information.

The HospiSign system follows three stages to move from one question to the next in the tree-based activity diagram interaction scheme: (1) display of the question; (2) display of the possible answers to that question; and (3) the recognition of the answer (sign). The user first watches the question displayed on the top-center of the screen; then performs a sign from the list of possible answers displayed at the bottom of the screen, and then moves to the next question. This process is repeated until the system gathers all the necessary information from the user. After the user answers all the required questions, the system prints out a summary report to be given to the information desk or the doctor at the hospital. This summary contains the details of the user's interaction with HospiSign.

To make the classification task easier, the questions are placed into a tree-based activity diagram in such a way that each question will lead to another sub-question with respect to the answer selected by the user. With categorization of possible answers to each question, it is intended to help the users to easily describe their symptoms or intention of their visit.

One of the most important advantages of using such a tree-based scheme is that it makes the system more user-friendly and easy-to-interact. The tree-based activity diagram interaction scheme also increases the recognition speed and performance of the system as the task of recognizing a sign from possible answers to each question is much easier and faster than recognizing a sign from the all possible answers.

3 Proposed Sign Language Recognition Method

The proposed sign language recognition method consists of four modules: Human Pose Estimation, Feature Extraction, Feature Normalization and Selection, and Temporal Modeling and Classification, as visualized in Fig. 2. Taking this framework as a baseline, the usage of various features, their combinations, temporal modeling techniques and classification methods are proposed to represent, and to recognize isolated sign language phrases.

The first step of the recognition module, human pose estimation, is critical since illumination and background variations introduce great challenges. As it uses active projective light imaging, the Microsoft Kinect v2 sensor is able to overcome these challenges. By using its pose estimation library routines, we were able to extract world coordinates, pixel coordinates and orientations of the 25 body joints.

As sign languages convey information through hand shape, upper body pose, facial expressions and hand trajectories, sign language recognition techniques extract features to represent each respective aspect of the signs. Kadir et al. [8] proposed specialized hand position and hand movement features in order to represent signs and capture their distinguishing properties. The features consist of hand positions and hand movements. Taking these features as baseline,

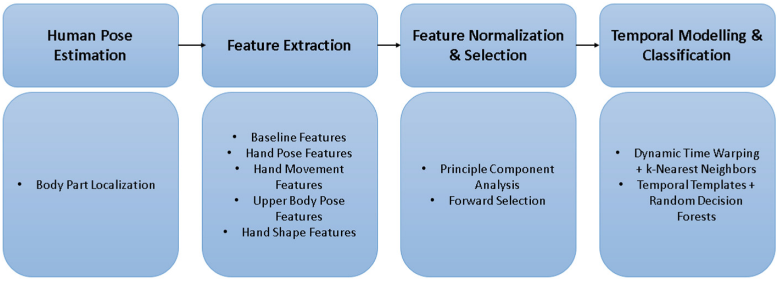
Human Pose Estimation
Feature Extraction
Feature Normalization & Selection
Temporal Modelling & Classification
• Body Part Localization
• Baseline Features
• Hand Pose Features
• Hand Movement Features
• Upper Body Pose Features
• Hand Shape Features
• Principle Component Analysis
• Forward Selection
• Dynamic Time Warping + k-Nearest Neighbors
• Temporal Templates + Random Decision Forests

Fig. 2. Four main modules of our sign language recognition framework.

hand position (Baseline Hand Position) and hand movement (Baseline Hand Movement) features were extracted from each video frame to represent sign samples.

In addition, we have extracted upper body pose (Normalized World Coordinates, Normalized Pixel Coordinates, Upper Body Joint Orientations), hand movement (Hand Joint Movement), and hand position (Hand Joint Distance) features using the body pose information provided by Microsoft Kinect v2 sensor.

Normalized World and Pixel Coordinates were extracted from the world and pixel coordinates that were provided by the Microsoft Kinect v2 sensor. The normalization was done by subtracting the Hip Center joint from the upper body joints, that are Head, Shoulder, Elbow, Wrist, Hand and Spine Joints, thus removing the location variance of the users. Then each joint coordinate is divided by the distance between the Shoulder Center and Hip Center joints in y axis, thus removing the scale (users' height) variance. We used Joint Orientation features as it is provided by the Microsoft Kinect v2 sensor.

Hand Joint Distance features were extracted by calculating the euclidean distance between the hand joints and the upper body joints, that were previously mentioned. The normalization of these features was done by dividing each distance by the sum of all Hand Joint Distances in its respective frame. Hand Movement Distance features represent the temporal dislocation of hands between subsequent frames and they were extracted by calculating the distance of each hands location from its location in the previous frame (in x, y, and z axis).

To represent hand shapes, we segmented the hand images using the hand joints' pixel coordinates and the signers' skin colors. We cropped a window of 80*80 pixel around both of the and joints and masked the hand using color based skin detection. Then we extracted Histogram of Oriented Gradients [6] with various Cell and Block Sizes from the segmented hand patches for each frame. A list of our features, the aspects they represent in a sign and their sizes can be seen in Table 1.

Table 1. Extracted features that are used to represent signs.

Feature name	Represented aspect	Feature size
Baseline Hand Positions	Hand Position	27
Baseline Hand Movements	Hand Movement	11
Normalized World Coordinates	Upper Body Pose	36
Normalized Pixel Coordinates	Upper Body Pose	24
Joint Orientations	Upper Body Pose	48
Hand Joint Distances	Hand Position	22
Hand Movement Distances	Hand Movement	6
HOG (L-M-H)	Hand Shape	18-108-432

As different features come from different distributions and have different scales we applied Principal Component Analysis (PCA) [7] to each feature separately before combining them in the temporal modeling and classification steps.

For temporal modeling and classification, we have proposed two approaches. The first approach is to model the temporal aspect of signs using Dynamic Time Warping (DTW) and classify samples using k-Nearest Neighbours (k-NN) algorithm. DTW is a popular tool for finding the optimal alignment between two time series. The DTW algorithm calculates the distance between each possible pair of points in terms of their spatial and temporal features. DTW uses these distances to calculate a cumulative distance matrix and finds the least expensive path through this matrix using dynamic programming. This path represents the ideal synchronization of the two series with the minimal feature distance. Usually, the samples are normalized to zero mean and smoothed with median filtering before distance calculation. The weighting of each feature inversely proportional to their feature size is applied to avoid features with larger sizes suppressing the effectiveness of features with smaller sizes. To classify a sign sample, its distance to the each training sample is calculated and the class of the sign is assigned using k-Nearest Neighbours algorithm.

The second approach is based on Temporal Templates (TT) and Random Decision Forest (RDF). Random Decision Forest is a supervised classification and regression technique that has become widely used due to its efficiency and simplicity. RDFs are an ensemble of random decision trees (RDT) [2]. However, RDFs do not inherently possess a temporal representation scheme. To incorporate the temporal aspect, Temporal Templates (TT), that represent each frame with the concatenated features of its neighbours are used in combination with Random Decision Forests. In template based temporal modelling, increasing template size enhances temporal representation. However, memory and computational power restrictions of development systems limit the feature vector size. To overcome this limitation, we downsample the data with various interval sizes to represent larger temporal windows while using the same number of frames. We classify the constructed temporal template of each frame by using Random Decision Forests (RDFs). Each tree is trained on a randomly sampled subset of the training data. This reduces over-fitting in comparison to training RDFs on the entire database; therefore increasing stability and accuracy. During training, a tree learns to split the original problem into smaller ones. At each non-leaf node, tests are generated through randomly selected subsets of features and thresholds. The tests are scored using the decrease in entropy, and best splits are chosen and used for each node [2]. Each tree ends with leaf nodes, that represent the probabilities of a given data to belong to the possible classes. Classification of a frame is performed by starting at the root node and assigning the frame either to the left or to the right child recursively until a leaf node is reached. Majority voting is used on the prediction of all decision trees to decide on the final class of the frame. Finally, signs are classified by taking the mode of its frames' classification results.

In the Dynamic Time Warping and K-nearest Neighbours based approach we choose the best combination of features by applying a greedy forward search algorithm, in which we iteratively added features by starting from the best performing feature until the recognition performance stopped increasing. There was no need for feature selection for the Temporal Template and Random Decision Forest based approach as the Random Decision Forests weight the features in their training.

4 Experiments and Results

All the experiments were conducted on a subset of the BosphorusSign database, which is used in the development of HospiSign. The subset contains 1257 sign phrase samples belonging to 33 phrase classes which were performed by six native TİD users in six to eight repetitions. In order to obtain user independent results we performed leave-one-user-out cross-validation and report the mean and standard deviation of recognition performance in all of our experiments.

The performance of the implemented methods were examined in terms of the features, temporal modeling techniques, and classification approaches. The first experiments were conducted to find the optimum parameters for Histogram of Oriented Gradients, which was used to represent hand shapes. Three HOG parameter setups were used that are Low Detailed (HOG-L, Cell Size: $[80 \times 80]$ Block Size: $[1 \times 1]$), Medium Detailed (HOG-M, Cell Size: $[40 \times 40]$ Block Size: $[2 \times 2]$), and High Detailed (HOG-H, Cell Size: $[20 \times 20]$ Block Size: $[4 \times 4]$). Examples of all the three parameter setups can be seen in Fig. 3. The parameter optimization results for different users demonstrate that while appearance based features worked well for some users, achieving up to 88 % accuracies, they did not produce reliable classifiers for others. The results can be observed in Table 2. Since HOG-M has the highest accuracy, we used it in the rest of our experiments.

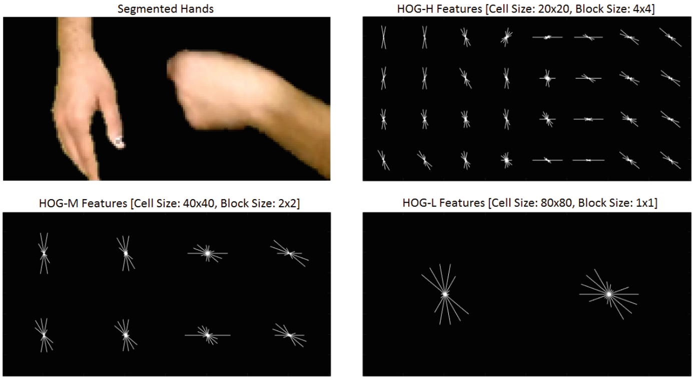
Segmented Hands
HOG-H Features [Cell Size: 20x20, Block Size: 4x4]
HOG-M Features [Cell Size: 40x40, Block Size: 2x2]
HOG-L Features [Cell Size: 80x80, Block Size: 1x1]

Fig. 3. Segmented hands and extracted Histogram of Oriented Gradients with different parameter setups. Top Left: Segmented Hands, Top Right: HOG-H, Bottom Left: HOG-M, Bottom Right: HOG-L.

Table 2. Recognition performance of different HOG parameters

	User 1	User 2	User 3	User 4	User 5	User 6	Mean ± Std
HOG-H	51.01	47.17	66.83	20.00	20.20	20.60	37.64 ± 20.14
HOG-M	78.28	84.91	86.93	22.00	30.81	26.13	**54.84 ± 31.51**
HOG-L	70.20	82.64	88.44	25.00	23.23	37.19	54.45 ± 29.46

Table 3. Performance evaluation of features

	User 1	User 2	User 3	User 4	User 5	User 6	Mean ± Std
Hand Joint Distances	96.46	94.72	95.98	86.00	86.36	96.98	**92.75 ± 5.15**
Norm. Pixel Coordinates	94.95	93.96	98.49	86.00	83.84	94.97	92.04 ± 5.76
Norm. World Coordinates	94.95	95.85	97.49	85.00	80.30	91.46	90.84 ± 6.81
Hand Movement Distances	90.40	86.79	91.96	83.00	64.65	68.84	80.94 ± 11.50
Baseline Hand Movements	75.76	72.83	81.91	78.50	44.44	68.34	70.30 ± 13.49
Baseline Hand Positions	64.14	75.85	68.34	53.50	55.05	68.34	64.20 ± 8.58
HOG-M	78.28	84.91	86.93	22.00	30.81	26.13	54.84 ± 31.51
Joint Orientations	32.83	38.49	44.72	34.50	26.77	38.69	36.00 ± 6.12
All Features Combined	67,68	77,36	85,93	67,00	47,47	75,38	**70,14 ± 13,10**

By using the best performing HOG setup, we conducted experiments in order to find the combination of features that yield the highest recognition performance. In feature selection experiments Dynamic Time Warping (DTW) was used to measure the distance between isolated sign phrases. Using the distances provided by DTW, k-Nearest Neighbours (k-NN) algorithm was used to classify the isolated signs by taking the mode of its k nearest neighbours' class labels. Table 3 lists the recognition accuracies of individual features for each user. It is observed that Hand Joint Distances yield the highest performance. While some features show comparable performance with the Hand Joint Distances, the rest of the features such as Joint Orientations perform poorly. When all features are combined, average performance drops to 70.14 %.

Even though the performance of some features are inferior, they may have complementary value, and a combination of features may perform better. To see which combination performs better, we have employed forward search. Table 4 shows the first step of forward search: It is observed that the Hand Movement Distances, a dynamic feature, has complementary value and enhances performance. While appearance based features such as HOG contain complementary information, we see that their performance is not consistent across different users. We stop at two features because adding any third feature to the combination of Hand Joint Distances and Hand Movement Distances decreased the recognition performance. Tables 3 and 4 list the performance accuracies of different users separately. It is observed that the performance for User 5 is lower than other users. By inspection of sign videos, we have observed that User 5 performs signs differently: For example, that user performs signs repeatedly and much faster.

Table 4. Forward selection of features combined with Hand Joint Distances feature.

	User 1	User 2	User 3	User 4	User 5	User 6	Mean ± Std
Hand Joint Distances	96.46	94.72	95.98	86.00	86.36	96.98	**92.75 ± 5.15**
Norm. Pixel Coordinates	94.95	94.34	98.49	86.00	84.34	95.48	92.27 ± 5.70
Norm. World Coordinates	95.96	96.23	96.98	85.00	80.30	93.47	91.32 ± 6.98
Hand Movement Distances	96.46	95.09	98.99	91.00	82.32	98.99	**93.81 ± 6.36**
Baseline Hand Movements	87.88	85.28	89.95	89.00	62.12	80.90	82.52 ± 10.51
Baseline Hand Positions	87.88	87.17	94.97	84.50	78.28	91.46	87.38 ± 5.76
HOG-M	95.96	96.23	96.98	84.00	78.79	95.98	91.32 ± 7.87
Joint Orientations	37.88	44.53	51.76	38.50	29.80	42.21	40.78 ± 7.36

Table 5. Temporal Template Size and Interval Steps optimization results. TS: Template Size, IS: Interval Steps.

	IS: 1	IS: 2	IS: 3	IS: 5
TS: 9	84,56 ± 5,65	90,89 ± 4,17	92,94 ± 3,16	95,54 ± 1,87
TS: 11	87,37 ± 6,32	91,77 ± 3,43	94,41 ± 2,13	95,96 ± 1,57
TS: 13	87,91 ± 5,35	93,07 ± 3,41	95,23 ± 2,13	**96,67 ± 1,80**
TS: 15	89,86 ± 4,39	94,26 ± 2,66	95,27 ± 1,95	96,65 ± 2,04
TS: 17	89,86 ± 4,39	94,26 ± 2,66	95,90 ± 2,03	96,38 ± 1,67
TS: 19	90,83 ± 3,57	94,91 ± 1,82	96,19 ± 1,69	96,08 ± 2,28
TS: 21	91,45 ± 3,82	94,91 ± 1,68	96,40 ± 2,03	95,96 ± 3,03
TS: 23	92,69 ± 2,95	95,35 ± 1,73	96,48 ± 2,02	95,22 ± 3,97

One other observation is that in Table 4, while the recognition performance of Normalized Pixel and World Coordinates and HOG-M increases performance for Users 2 and 3 who are expert level signers, they decrease significantly for Users 4 and 5 who show variations in their performance with respect to speed and sign positions.

Then we conduct experiments in order to find the optimum window size and interval steps (down-sampling rate) for the Temporal Templates (TT). We classify the Temporal Templates using Random Decision Forest (RDF) that contains 100 trees.

As it can be seen in Table 5, as the template size and interval steps increase, the recognition performance also gets better until an optimum size of represented temporal window. While lower template sizes benefit from higher interval steps, this trend is lost with higher template sizes. We choose a template size of 13 and down-sampling rate(interval step) of 5 since that yields the best performance.

In the light of our experiments, we have seen that DTW and RDF reach 93.81 % and 96, 67 % average recognition accuracies respectively on 33 classes of signs of six different users in leave-one-user-out cross-validation tests. However, in HospiSign, the tree-based activity diagram interaction scheme, that is displayed in Fig. 1, guides its users to perform signs from a limited subset at each step.

Table 6. Mean and standard deviation of recognition results of DTW and RDF based methods with and without the Activity Diagram based recognition scheme

Setup	nClasses	DTW+k-NN	TT+RDF	Combined
All Signs	33	93.81 ± 6.36	96,67 ± 1,80	N/A
Activity Group 1	2	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Activity Group 2	2	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Activity Group 3	9	100.00 ± 0.00	98,78 ± 1.53	100.00 ± 0.00
Activity Group 4	14	95.86 ± 3.05	97,88 ± 1,67	97.88 ± 1.67
Activity Group 5	4	97.92 ± 5.1	98,09 ± 3,39	98,09 ± 3,39
Activity Group 6	2	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

We have conducted experiments using the best performing parameters for both the DTW+k-NN and TT+RDF based approaches and reported the results in Table 6.

As the number of classes that the systems requires to classify from decreases, the recognition performance improves drastically. Moreover, as each activity group in the tree-based activity diagram interaction scheme is a different recognition task, we can combine the best performing temporal modeling and classification approaches, thus further increasing the recognition performance. By choosing the best performing approach for each activity group, we have achieved 100 % recognition performance for four activity groups and more than 97.88 % recognition performance for the renaming two activity groups, suppressing the best recognition performance of recognizing signs from 33 classes (96.67 % using TT+RDF approach). The reason that the two activity groups that hand lower recognition performance then the rest is the similarity of signs in Activity Group 4 (All of phrases are ending in the same way) and the larger number of classes that system is required to be classified from in Activity Group 5 (14 sign phrase classes).

5 Conclusion

In this study, a real time sign language recognition system was designed with the aim of working as a communication platform for a hospital information desk. The system was developed using a Microsoft Kinect v2 sensor to aid with the human pose estimation. The recognition system, trained with a subset of the BosphorusSign database [3], extracts hand shape, hand position, hand movement and upper body pose features and performs temporal modelling using Dynamic Time Warping and Temporal Templates. The spatio-temporally represented signs are then classified using k-Nearest Neighbours algorithm and Random Decision Forests.

The experiments demonstrate that the highest recognition (93.81 %) was achieved by using the Hand Joint Distance and Hand Movement Distance features while using the Dynamic Time Warping and k-Nearest Neighbours based recognition approach. These features were selected using a greedy forward selection scheme. Forward selection demonstrated that the presence of any other feature reduced overall recognition performance for all users. However, it is interesting to note that while appearance and coordinate based features performed well with recognition from three users, they were not effective with other users who performed the signs with more variation in location and speed. This can be explained by the fact that while these features do posses complementary information that may be helpful in recognition, their variation among different users makes them user and recording environment dependent. This is especially important when designing an online recognition system such as HospiSign, as the system becomes more robust the less it is over-trained on users who perform the signs perfectly with little room for variations.

In the experiments, in which the signs were temporally modeled using Temporal Templates and classified using Random Decision Forests, the best recognition performance (96.67 %) was achieved using a template size of 13 with interval steps of 5. As the Random Decision Forests does the feature selection in its training, no additional feature selection scheme was applied in these experiments.

Our experiments demonstrate that while using the 33 class classification scheme the highest recognition performance (96.67 %) was achieved by using the Temporal Template and Random Decision Forest based classification approach. However, by using the tree-based activity diagram interaction scheme, we were able to improve the recognition performance for all of the activity groups, as the systems has to recognize signs from a lower number of classes in each step of the interaction. One of the main benefits of using the tree-based activity diagram interaction scheme is that the best performing approaches can be used for the classification of each activity group. By combining the best performing classification approach for each activity group, we were able to reach 100 % recognition performance in four activity groups and more then 97.88 % recognition performance for the remaining two activity groups, thus suppressing the recognition performance of 96.67 %.

References

1. Berndt, D., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Workshop on Knowledge Knowledge Discovery in Databases, vol. 398, pp. 359–370 (1994)
2. Breiman, L.: Random forests. Mach. Learn. **45**(5), 1–35 (1999)
3. Camgöz, N.C., Kindiroglu, A.A., Karabüklü, S., Kelepir, M., Akarun, L., Ozsoy, S.: BosphorusSign: a Turkish sign language recognition corpus in health and finance domains. In: LREC (2016)
4. Chai, X., Li, G., Chen, X., Zhou, M., Wu, G., Li, H.: VisualComm: a tool to support communication between deaf and hearing persons with the Kinect. In: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (2013)
5. Cox, S., Lincoln, M., Tryggvason, J.: TESSA, a system to aid communication with deaf people. In: Proceedings of the Fifth International ACM Conference on Assistive Technologies. ACM (2002)

6. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (2005)
7. Jolliffe, I.: Principal Component Analysis. Wiley Online Library, Chichester (2002)
8. Kadir, T., Bowden, R., Ong, E., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: British Machine Vision Conference (2004)
9. Kose, H., Yorganci, R., Algan, E.H., Syrdal, D.S.: Evaluation of the robot assisted sign language tutoring using video-based studies. Int. J. Social Robot. **4**(3), 273–283 (2012)
10. Lee, H., Kim, J.: An HMM-based threshold model approach for gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. **21**(10), 961–973 (1999)
11. Lopez-Ludena, V., Gonzalez-Morcillo, C., Lopez, J.C., Barra-Chicote, R., Cordoba, R., San-Segundo, R.: Translating bus information into sign language for deaf people. Eng. Appl. Artif. Intell. **32**, 258–269 (2014)
12. Ong, S.C.W., Ranganath, S.: Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6), 873–91 (2005)
13. Pitsikalis, V., Theodorakis, S., Vogler, C., Maragos, P.: Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2011)
14. Rabiner, L., Juang, B.: An introduction to hidden Markov models. ASSP Magazine, IEEE (1986)
15. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR, vol. 2 (2011)
16. Starner, T., Pentland, A.: Real-time American sign language recognition from video using hidden Markov models. In: 1995 Proceedings of the Computer Vision (1995)
17. Süzgün, M.M., Özdemir, H., Camgöz, N.C., Kindiroglu, A.A., Başaran, D., Togay, C., Akarun, L.: HospiSign: an interactive sign language platform for hearing impaired. In: Proceedings - Eurasia Graphics 2015, Istanbul (2015)
18. Theodorakis, S., Pitsikalis, V., Maragos, P.: Dynamic-static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image Vis. Comput. **32**, 533–549 (2014)
19. Vogler, C., Metaxas, D.: Parallel hidden Markov models for American sign language recognition. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, pp. 116–122 (1999)
20. Weaver, K.a., Starner, T.: We Need to Communicate! Helping Hearing Parents of Deaf Children Learn American Sign Language. Assets (Xiii), p. 91 (2011)
21. Yorganci, R., Akalin, N., Kose, H.: Avatar Tabanlı Etkileşimli İşaret Dili Oyunları. In: Uluslararası Engelsiz Bilişim 2015 Kongresi. Manisa (2015)
22. Zhang, Z.: Microsoft Kinect sensor and its effect. IEEE Multimedia **19**(2), 4–10 (2012)