Automatic Recognition Fingerspelling Gestures in Multiple Languages for a Communication Interface for the Disabled[1]

Ahmet Alp Kindiroglu[a], Hulya Yalcin[a], Oya Aran[b], Marek Hrúz[c], Pavel Campr[c], Lale Akarun[a], and Alexey Karpov[d]

[a] *Bogazici University, Turkey*
[b] *Idiap Research Institute, Switzerland*
[c] *University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics, Czech Republic*
[d] *Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Russia*
e-mail: {alp.kindiroglu,hulya.yalcin,akarun}@boun.edu.tr, oya.aran@idiap.ch, {mhruz, campr}@kky.zcu.cz, karpov@iias.spb.su

Abstract—This paper presents the design and evaluation of a multi-lingual fingerspelling recognition module that is designed for an information terminal. Through the use of multimodal input and output methods, the information terminal acts as a communication medium between deaf and blind people. The system converts fingerspelled words to speech and vice versa using fingerspelling recognition, fingerspelling synthesis, speech recognition and speech synthesis in Czech, Russian and Turkish Languages. We describe an adaptive skin color based fingersign recognition system with a close to real-time performance and present recognition results on 88 different letters signed by five different signers, using above four hours of training and test videos.

Keywords: sign language, fingerspelling recognition, assistive technologies.

DOI: 10.1134/S1054661812040086

1. INTRODUCTION

Sign Languages are visual languages that make use of hand gestures, body movements and facial expressions to convey meaning. They are extensively used by hearing impaired communities as their primary means of communication. Although most sign languages are quite rich in terms of the number of signs in the language, there can be some foreign words, private names, or uncommon words, for which no sign is defined. In order to represent these words fingerspelling is used. Fingerspelling describes the letters of a written alphabet by using hand gestures. It is especially useful in a multilingual setting, where there are many foreign words. Fingersigns also differ from one community to another, adding one more challenge to the problem.

In this paper, we propose a multilingual fingersign recognition system to assist the communication between people who are speaking or hearing impaired, as well as people speaking different languages such as Czech, Russian and Turkish. We describe and present the performance of the fingerspelling recognition module on a database which consists of videos of people fingerspelling in Czech, Russian and Turkish languages.

The fingerspelling recognition system employs state of the art hand tracking, segmentation, feature extraction, and classification techniques. The performance of the system with a variety of features such as Hu moments, Fourier descriptors, local binary patterns and a modified radial distance function as well as their combined performance after fusion are evaluated. Without the underlying language model, the experimental results on multilingual fingersign videos of the aforementioned database yields 80 per cent user dependent identification performance and 42 per cent anonymous identification performance. In the real-time recognition system, using a word list of 10000 words, the most likely word in the list is chosen using a normalized Levenshtein distance [1].

2. MULTIMODAL INFOKIOSK SYSTEM OVERVIEW

The multimodal information terminal has been developed and integrated into the main system in Enterface2010 workshop, (Amsterdam, Netherlands) [2]. It consists of four standalone input and output modules that are united by a fifth server module. The input modules consist of speech [3] and fingerspelling recognition modules and the latter is the focus of this paper. The output modules consist of fingerspelling [4] and speech synthesizers. These modules operate on a client-server based architecture that makes use of sessions to enable simultaneous interaction between different input and output modules. The flowchart of the system is given in Fig. 1.

[1] The article is published in the original.

Received @@@

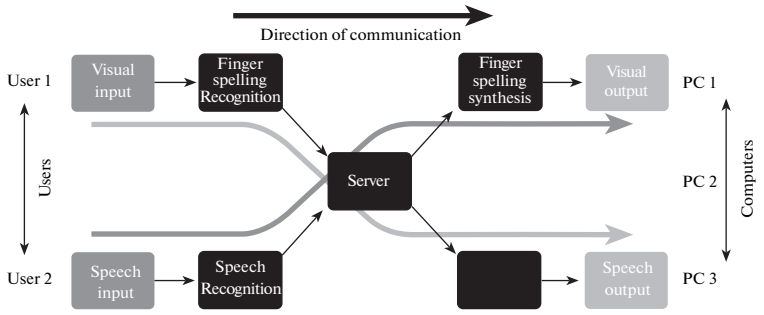

Fig. 1. FingerSign to Speech translator system flowchart.

The fingerspelling module receives input from a webcam and outputs the recognized letters. The letters are accumulated in the server until an end-of-word symbol marks the end of the given word, which is acknowledged when the user separates and lowers his hands for a relatively longer pause while signing.

3. FINGERSPELLING RECOGNITION

The fingerspelling recognition system has an offline and an online component. Implemented using MS Visual C++ and Intel OpenCV, the offline system is used for training the system and performing batch recognition operations on test sets, while the online system performs real-time letter and word level recognition using input from a single camera.

3.1. Preprocessing

Accurate tracking of hands and face of a natural signer is a challenging task that is crucial for a FSR applications. The accuracy of hand descriptor representation and modeling depends on hands being localized and perfectly segmented. As the signing human hand often tends to make fast and non-linear movements, modeling and tracking the hand becomes a greater challenge. Likewise, the occlusions of the hands with each other and with the face is very frequent. To handle this task of hand tracking we have implemented a Camshift [5] based joint hand and face tracking algorithm that is equipped with the necessary capabilities to handle the tracking of signing hands.

Our tracking method accomplishes tracking by using the color properties of tracked objects. Therefore the accuracy of the algorithm in distinguishing tracked objects solely depends on the accuracy of the histogram built during initialization. For accurate initialization, the users' hand needs to be flawlessly segmented from the background. In common implementations of the Camshift algorithm, the initialization process is implemented in a manual fashion, asking the user to draw a box on the object to track. As that method requires manual user input, we chose to implement a motion-based automatic hand detection module for initial detection of hands. The module requires the user to wave his hands for a few seconds before commencing signing.

The Camshift algorithm performs object tracking using a single tracking box and it is only capable of tracking at most one object of the same color any given time. As the task of fingerspelling tracking involves tracking a head and two hands, the increased number of object brings new challenges. Our joint tracking algorithm tracks three targets with similar color distributions. The algorithm assumes that a signer comes in front of the camera and starts signing with one or two of his hands. Cases such as hands leaving and entering the sign space, hands merging with each other, hands merging with the face and hands separating are handled using a heuristic algorithm based on human body information. A hierarchical redetection module is used with this algorithm to detect objects to be tracked in the sign space [6]. Motion and skincolor information are further utilized to predict the location and size of the tracking boxes during the segmentation stage; separating hand pixels from the background area.

3.2. Feature Extraction

We use a combination of different shape descriptors that can recognize one handed, two handed and overlapping hand shapes.

3.2.1. Hu Moments

These are derived from central image moments [7]. The first six Hu moments are rotation, scale and translation invariant, whereas the seventh moment is skew orthogonal invariant, a rather desired attribute to distinguish mirror images of identical objects.

3.2.2. Elliptic fourier descriptors. Calculated from the external shape contour of an object, these descriptors are used to represent an object in the frequency domain. Using a set of ellipses, the closed external outline of a shape can be transformed to yield a shape spectrum [8]. The lower frequency descriptors contain information about the general features of the shape, and the higher frequency descriptors contain information about finer details of the shape. Although the

number of coefficients generated from the transform is usually large, a subset of the coefficients is enough to capture the overall features of the shape.

For an n-harmonic elliptic Fourier descriptor representation of a 2-D closed shape is given in Equation 1. In the equation, the center of the curve is (a_0, c_0). (a_k, b_k, c_k, d_k) for $k = 1...n$ are elliptic Fourier coefficients of the curve up to n Fourier harmonics. T is the perimeter of the closed curve.

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} a_0 \\ c_0 \end{bmatrix} + \sum_{k=1}^{N} \begin{bmatrix} a_k \cos\left(\frac{2k\pi t}{T}\right) + b_k \sin\left(\frac{2k\pi t}{T}\right) \\ c_k \cos\left(\frac{2k\pi t}{T}\right) + d_k \sin\left(\frac{2k\pi t}{T}\right) \end{bmatrix} \quad (1)$$

The elliptic fourier descriptors (a_k, b_k, c_k, d_k) are calculated for each point k of the curve as seen in the Equations 2–5.

$$a_k = \frac{1}{2k^2\pi^2} \sum_{i=1}^{n} \frac{\Delta x_i}{\Delta t_i} \left[\cos\left(\frac{2k\pi t_i}{T}\right) - \cos\left(\frac{2k\pi t_{i-1}}{T}\right) \right] \quad (2)$$

$$b_k = \frac{1}{2k^2\pi^2} \sum_{i=1}^{n} \frac{\Delta x_i}{\Delta t_i} \left[\sin\left(\frac{2k\pi t_i}{T}\right) - \sin\left(\frac{2k\pi t_{i-1}}{T}\right) \right] \quad (3)$$

$$c_k = \frac{1}{2k^2\pi^2} \sum_{i=1}^{n} \frac{\Delta y_i}{\Delta t_i} \left[\cos\left(\frac{2k\pi t_i}{T}\right) - \cos\left(\frac{2k\pi t_{i-1}}{T}\right) \right] \quad (4)$$

$$d_k = \frac{1}{2k^2\pi^2} \sum_{i=1}^{n} \frac{\Delta y_i}{\Delta t_i} \left[\sin\left(\frac{2k\pi t_i}{T}\right) - \sin\left(\frac{2k\pi t_{i-1}}{T}\right) \right] \quad (5)$$

For each Fourier harmonic calculated, we obtain an ellipse that yields four invariant features [9]. The first two features are the major and minor axis lengths of the calculated ellipses. The latter two features can be derived from the first two features and vice versa. In our experiments with hand contours, we found ten harmonics sufficient to represent hand gestures, yielding 40 sized feature vectors. The formulas for the features are given in Equations 6–9

$$\mathbf{A}_k^2 = \frac{I_k + \sqrt{I_k^2 - 4J_k^2}}{2} \quad (6)$$

$$\mathbf{B}_k^2 = \frac{J_k^2}{A_k^2} \quad (7)$$

$$\mathbf{I}_k = a_k^2 + b_k^2 + c_k^2 + d_k^2 \quad (8)$$

$$\mathbf{J}_k = (a_k d_k) - (b_k c_k) \quad (9)$$

3.2.3. Radial distance function. RDF is a contour based method [10]. Using the distance of a seed-point in all directions to the closest background pixel, a representation of the object is obtained as illustrated in Fig. 2. The representation is invariant to translation, size and rotation. Rotation invariance is achieved by assigning the angle with the smallest radial distance as degree zero.

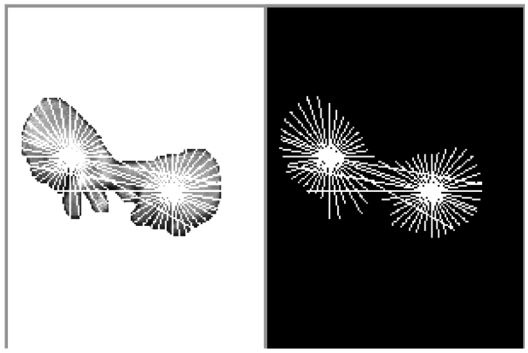

Fig. 2. Radial Distance Function Calculation For Two Hands

While describing convex shapes such as an isolated hand, the radial distance function is an efficient feature as it is possible to represent finger locations and notable extensions using their distance from the seed points. However, when describing blobs consisting of not completely overlapping, but touching hands, obtaining a point which has a straight line distance to both hands may not be possible. For this reason, we attempt to find the centers of gravity belonging to both hands. Using image moments, we calculate the parameters of the smallest ellipse that covers both hands using the formulas 10–13.

$$\mathbf{a} = \frac{M_{20}}{M_{00}} - x^2; \quad \mathbf{b} = 2\left(\frac{M_{11}}{M_{00}} - x_c y_c\right) \mathbf{c} = \frac{M_{02}}{M_{00}} - y^2 \quad (10)$$

$$l_1 = \frac{(a+c) + \sqrt{b^2 + (a-c)}}{2} \quad (11)$$

$$l_2 = \frac{(a+c) - \sqrt{b^2 + (a-c)}}{2} \quad (12)$$

$$\phi = \frac{1}{2}\tan_{-1}\left(\frac{b^2}{a-c}\right) \quad (13)$$

In the calculations, l_1 and l_2 are the principal axes of the bounding ellipse centered on the centroid of the image and ϕ is their rotation angle.

By dividing the image using the minor axis l_2 as a separator, we effectively divide the blob into two approximately equal parts belonging to different hands. After calculating the centroid of each part using image moments, we obtain seed locations for two radial distance functions that are sufficient to describe a hand blob consisting of two hands. The calculation of radial distance function for 2 hands can be seen in Fig. 2.

In our experiments, we have modeled our hand shapes using 72(360/5) distance calculations from two seed points. Thus, for each image, we obtained descriptors of size 144. The extracted feature sets are used in the subsequent classification stage.

3.2.4. Local binary patterns. Local Binary Patterns (LBP), first introduced by Ojala [11] for texture repre-

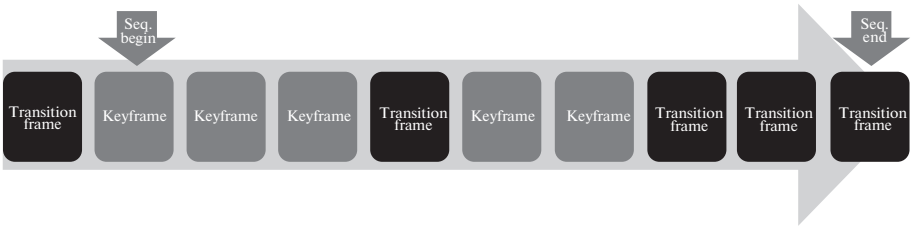

Fig. 3. Keyframe based gesture start and end point detection.

sentation, were later used for hand detection in cluttered images [12]. The algorithm moves a defined patch along all the pixels in an image. The LBP image changes depending on the size and shape of the patch. We use a circular 8-neighborhood patch with radius one and two pixels as shown in Fig 3. If the brightness of a pixel in the patch is greater or equal to the center pixels brightness, a binary label of 1 is assigned to the proper location in the patch. Otherwise, 0 is assigned. In each patch, we evaluate 8 locations (or combinations of locations) ending up with an 8 bit number. Next, we compute a histogram of the LBP image which is used as a feature vector. For an 8-neighborhood patch, one can have 256 histogram bins, each bin standing for one pattern. In practice though, it has been proven that not all patterns are necessary for recognition. The bulk of the information is enclosed within the patterns that have two or less changes between 0 and 1 in its binary representation. Such LBP patterns are called uniform LBPs. There exist 58 such patterns and the rest of the patterns are moved to the 59th bin of the histogram. Hence, the feature vector is of size 59 for uniform LBP. We implemented uniform as well as non-uniform LBP with the patch radius one and two.

3.3. Motion Modeling and Temporal Segmentation

An important challenge in continuous fingerspelling recognition is the problem of temporal segmentation. Using keyframe based temporal segmentation, we address two separate tasks. First, given a continuous sequence of images containing hand gestures, we try to designate when a sign starts and finishes. Secondly, having split sequences into chunks of images containing a single gesture, we try to decide on the letter, which is best suited to represent the entire sequence.

We use a keyframe based system, where hand motion and motion blur are utilized to automatically separate keyframes from transitions. While signing fingerspelling gestures, the signer first moves his hands to a certain start position. After this initialization gesture, the signer moves his hands to perform the gesture and then either proceeds to the next gesture or back to the start position. While performing the gestures, he/she pauses at certain poses. Those hesitance points of time are actually candidates to serve as keyframes, since the hands contain meaningful shapes, move relatively slower and much less numbers of pixels are smeared into back-ground. We identify those pauses making use of motion and external hand contour change as features for keyframe selection.

The model displayed in Fig. 3 attempts to segment gesture start and end positions in a continuous image sequence. In a continuous sequence of images marked as keyframes and transition frames, the first encountered keyframe is designated as the sequence start frame. Then, each new frame is added to the sequence until a continuous transition frame of certain length arrives. In this method, we make use of two thresholds to optimize segmentation. First, by making sure that the recognized sequence is at least of length (0.30*seconds/systemfps*) frames, we attempt to prevent short noise from being recognized as a sequence. Secondly, we set the minimum length of continuous transition frames used to determine the end of a sequence to a fixes threshold of (0.20*seconds/systemfps*). This way, movements of dynamic gestures are segmented from transitions between gestures.

The thresholds setting minimum transition and minimum sequence length were chosen based on the fingerspelling performance of our native signer subjects performing approximately 3 signs per second. Choosing higher thresholds for slower, non-native signers prevents signer movement noise from being recognized as a gesture or a single gesture from being interpreted as two or three gestures.

To perform the classification of continuous fingerspelling gestures, we employ two different methods. In the first method (Fig. 4), we use our motion model with a a non-parametric k-nearest neighbor classifier. The classification results of each frame for Hu moments (Hu), Elliptic Fourier Descriptors

(EFD), Radial Distance Functions (RDF) and Local Binary Patterns (LBP) are fused by weighted majority voting. Then we fuse the decision results for each frame belonging to the same gesture using majority voting to obtain a gesture-wise result. In the second method, we use continuous HMM's with Gaussian Mixture Model's to model finger-spelling gestures as seen in Fig. 5.

In the real-time recognition system, due to the word level accuracy requirements of translation between different languages, the collected letters are checked for consistency at word level. Using a word list of 10000 words, the most likely word in the list is cho-

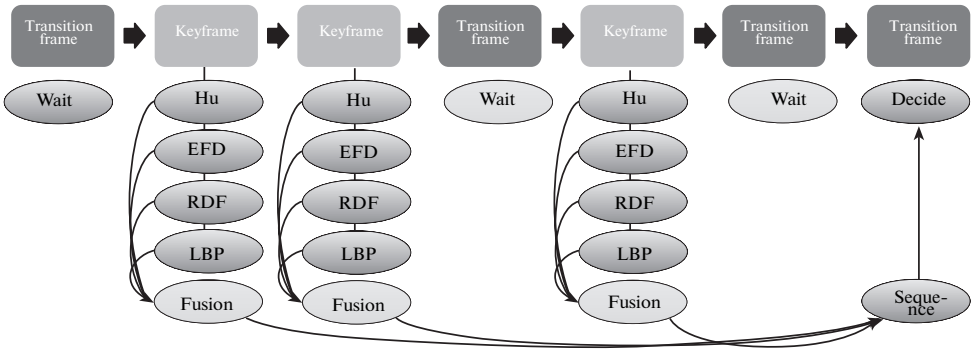

Fig. 4. Motion model for continuous FRR.

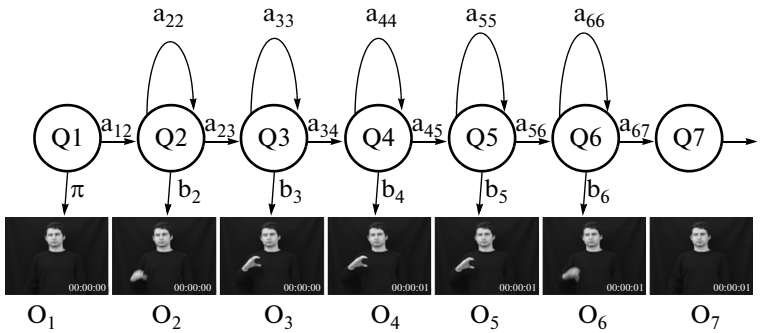

Fig. 5. FRR using continuous HMM's

sen using a normalized Levenshtein distance [9]. Translation between Czech, Russian and Turkish languages are then done, using the most likely word chosen by Levenshtein distance.

4. RESULTS

We created a video database consisting of signs of 3 the Turkish, Russian and Czech fingerspelling alphabets. Currently, the database which we have contains videos of five subjects performing the Turkish sign alphabet, three subjects performing the Czech sign alphabet and three subjects performing the Russian sign alphabet. Of these subjects, one is a natural signer and a signing instructor. The rest of the subjects are 3 university students who have performed fingerspelling gestures through visual guidance. The gestures of the Czech and Russian alphabets were taught to the performers by our study partners from those respective countries during the eNTERFACE'10 workshop [2]. Sample Images of the dataset can be seen in Fig. 6.

In total, there are 88 letters performed in the database, covering 29 Turkish, 26 Czech and 33 Russian sign alphabet letters. Some letters can be represented using only one hand, whereas some require the use of both hands. Moreover, one keyframe is enough to represent some letters whereas representation of some letters have a more dynamic nature since they can be performed with a movement rather than a static frame, for instance X. Each subject repeats each letter in the database five to nine times.

In the videos, certain restrictions are imposed on the signers to make skin color based segmentation easy. The signers wear dark colored clothes with long sleeves that leaves only hands and faces bare. The signs are performed in front of a black background using a constant camera distance and angle. Before and after each performance, the signer brings his hands down to his sides, which is designated as the rest pose. The videos were recorded from a frontal pose approximately one and a half meters from the signers that captured the signers face and hands down to his/her waist. While the usage of certain lighting conditions allowed for easier segmentation, we have used normal room conditions in the recording of some videos to make the training set more robust.

The signs are recorded by a mini-dv camera at 25fps on resolutions varying from 1072 × 768 to 640 × 480. The total length of the training and test videos for each subject is between 25 and 30 minutes. Thus the length of the entire dataset borders four hours. In the videos, the interlacing effect caused by the camera interlacing effect is removed by applying a deinterlacing algorithm.

For our classification with kNN classifiers using motion models, we report two sets of results: signer independent and signer dependent. In the signer

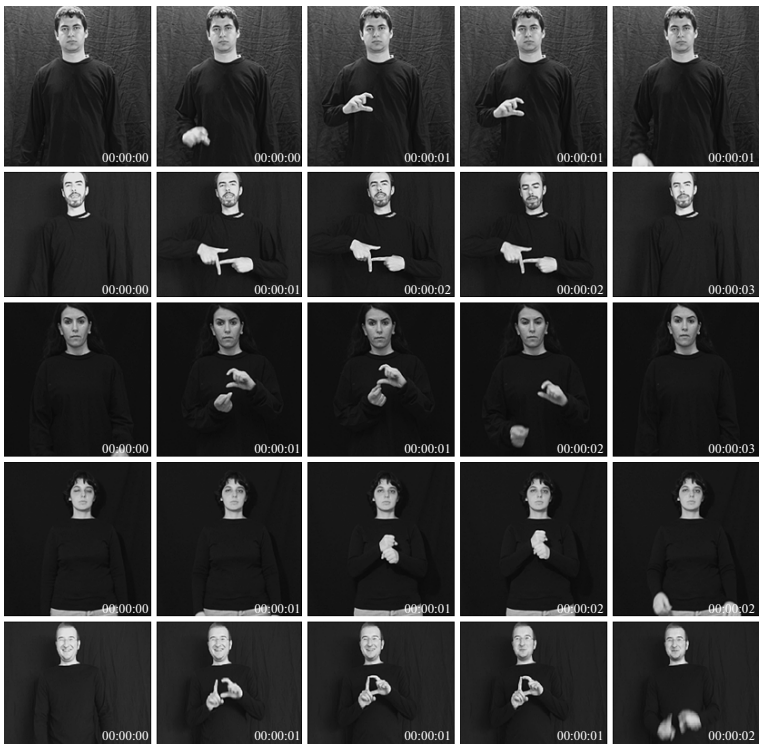

Fig. 6. Samples images from the multi-lingual fingerspelling dataset

dependent case, training and test examples are chosen from the same individuals. For each subject, half of the repetitions of each gesture are placed into the training set and the remaining half are placed into the test set.

For the signer independent recognition, a subject is chosen as test subject and the system is trained with sample gestures belonging to the rest of the users.

Using the signer dependent and independent test sets, recognition accuracy results (fingersign recognition rate—FRR) belonging to different subjects are obtained and shown in Tables 1 and 2. We observe that the fingersign recognition rate is 82 per cent for the signer dependent and 42 per cent for the signer independent test set. In online gesture recognition in the

Table 1. Multi-lingual user dependent fingerspelling recognition accuracy with manually selected keyframe sequences

	Language	HU	EFD	RDF	LBP	Fused
Sbj. 1	Turkish	0.53	0.63	0.67	0.85	0.88
Sbj. 2	Turkish	0.56	0.70	0.78	0.91	0.93
Sbj. 3	Turkish	0.54	0.75	0.77	0.87	0.87
Sbj. 4	Turkish	0.48	0.61	0.61	0.74	0.78
Sbj. 5	Turkish	0.59	0.79	0.76	0.94	0.96
Avg.	Turkish	0.54	0.70	0.72	0.86	0.88
Sbj. 1	Russian	0.64	0.61	0.72	0.69	0.75
Sbj. 2	Russian	0.38	0.54	0.75	0.8	0.8
Sbj. 3	Russian	0.63	0.61	0.64	0.69	0.73
Avg.	Russian	0.55	0.59	0.7	0.73	0.76
Sbj. 1	Czech	0.6	0.52	0.62	0.75	0.72
Sbj. 2	Czech	0.48	0.61	0.78	0.76	0.84
Sbj. 3	Czech	0.47	0.65	0.67	0.73	0.78
Avg.	Czech	0.52	0.59	0.69	0.75	0.78

3 **Table 2.** Multi-lingual user independent fingerspelling recognition accuracy with manually selected keyframe sequences

	Language	HU	EFD	RDF	LBP	Fused
Sbj. 1	Turkish	0.31	0.15	0.39	0.28	0.43
Sbj. 2	Turkish	0.23	0.06	0.24	0.21	0.28
Sbj. 3	Turkish	0.36	0.30	0.48	0.29	0.45
Sbj. 4	Turkish	0.07	0.28	0.61	0.48	0.54
Sbj. 5	Turkish	0.32	0.20	0.45	0.23	0.40
Avg.	Turkish	0.26	0.20	0.43	0.30	0.42
Sbj. 1	Russian	0.36	0.16	0.44	0.44	0.47
Sbj. 2	Russian	0.36	0.25	0.39	0.42	0.4
Sbj. 3	Russian	0.27	0.28	0.43	0.31	0.33
Avg.	Russian	0.33	0.23	0.42	0.39	0.4
Sbj. 1	Czech	0.39	0.34	0.46	0.41	0.48
Sbj. 2	Czech	0.27	0.18	0.49	0.44	0.46
Sbj. 3	Czech	0.22	0.18	0.37	0.37	0.38
Avg.	Czech	0.29	0.23	0.44	0.4	0.44

handicapped kiosk system, performance is substantially improved by the language model.

In Tables 3 and 4 we present the results of our continuous HMM based classification. In this method, feature vectors belonging to each descriptor are modeled with Gaussian Mixture models and combined in Hidden Markov Models. Each feature vector of an individual gesture is modeled using a single HMM with five Gaussian mixture components, which was determined through cross-validation.

While modeling gestures using HMMs, an effective parameter of the model is the number of states used to generate a gesture. We have used two different methods to estimate the optimal state count of our models. In the first approach we have trained several HMMs with different number of states. Using fivefold cross validation, we have compared system recognition accuracies of these HMMs to determine the state number of the best fitting model for each feature (Fig. 7). In Table 3, we present the recognition of modeling with a fixed HMM with five states.

Table 4 shows a four percent increase in overall system FRR from 64 to 68 per cent. Further improvement of recognition results requires a breakdown of recognition accuracy for different gestures. Figure 8 shows the confusion matrix for the fused Turkish FRR from the experiment in Table 4. By closely examining the confusion matrix, we can detect letter pairs that are most likely to be confused with one another.

It is observed that the letter pairs "A-H", "E-H" and "M-N" are often confused with each other. The most likely reason of this confusion is the similarity in the posture of hands which only slightly differs by the position of middle and ring fingers of the right hand. Likewise, the letter pairs "S-Ş" and "K-Y" also often appear to be confused. The interesting fact to note about these letter pairs is that, while they have similar hand postures, "S" and "K" are static gestures and "Ş" and "Y" are dynamic. Therefore, while recognizing these gestures, the system fails to distinguish between letters that may contain similar gestures but contain different movements over time. Another highly confused pair of letters is "I-İ" which is highly unlikely and unexpected. While "I" is performed with one hand, the letter "II" is performed using two hands. Through experimentation, this confusion was found to be a failure in tracking. During tracking, a redetection failure of the right hand while performing the letter "İ" caused the system to create both one handed and two handed models for that letter. Likewise, due to failures in the tracking of the loosely merged two

Table 3. Fingerspelling recognition accuracy on keyframe 3 sequences using continuous HMM's

	HU	EFD	RDF	LBP	Fused
Turkish	0.2714	0.6156	0.3291	0.6457	0.6407
Russian	0.2371	0.6983	0.1905	0.5086	0.6147
Czech	0.2768	0.7288	0.2373	0.6271	0.7054

Table 4. Turkish FRR with dynamic and static numbers of state continuous HMM's

	HU	EFD	RDF	LBP	Fused
Static	0.2714	0.6156	0.3291	0.6457	0.6407
Dynamic	0.2764	0.6256	0.3510	0.6608	0.6884

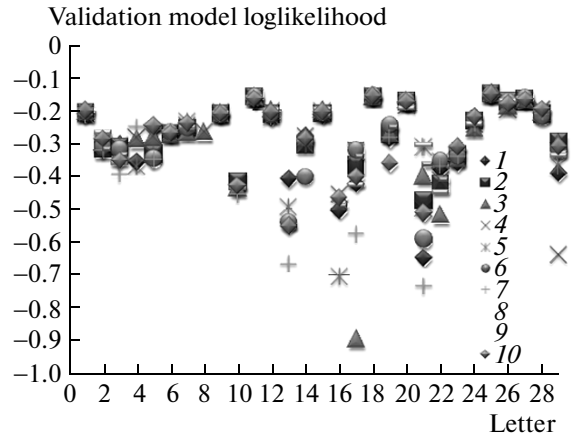

Fig. 7. Cross-validation results of continuous HMM log-likelihood values with different number of states for Turkish fingerspelling gestures.

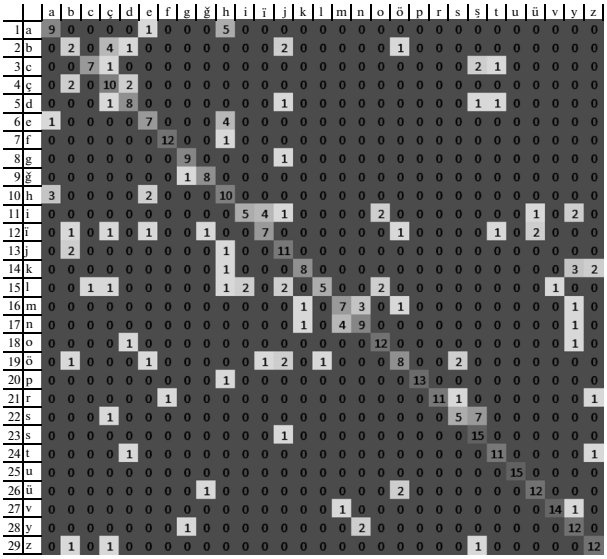

Fig. 8. Confusion matrix for Turkish FRR using continuous HMM's.

handed letter "B", the letter is often confused to separated two handed letters such as "Ç" and "J".

5. CONCLUSION

In this study, we investigated the extraction and recognition of fingerspelling gestures from sign language videos. We worked with Czech, Russian and Turkish sign alphabets and dealt with the segmentation, tracking, representation, temporal modeling and classification of hand gestures.

The experiments performed with isolated fingerspelling gestures show that the system is capable of recognizing isolated gestures. The KNN classification with majority voting of sequences achieved approximately 82 percent recognition accuracy with all languages. While the results with other approaches seem comparatively lower, they can be improved through the availability of more training data and examples for each class. By modeling gestures with Hidden Markov Models using Gaussian Mixture Models, we obtain a fingerspelling recognition accuracy of 64 per cent. By optimizing the number of HMM states to model each gesture using cross validation, we obtain an accuracy increase of approximately four per cent. However, a disadvantage of using cross validation to determine the number of states for an HMM is that training multiple HMMs from scratch is computationally expensive. In addition, since the success of training a model depends on initial parameters, it is possible to compare ideal models with models that have been poorly initialized. To overcome this, a multiple number of models with the same state number may be trained to choose the best among them. However, this also causes an increase in computational complexity. Therefore in order to train the system for new gestures, the usage of methods such as State Splitting HMM's may be preferred over cross validation for determining the number of states [13].

Our tests with signer independent gestures shadow the acquired signer dependent accuracy ratings. We observe that the recognition rate of fingersign keyframes drops from 82 per cent for the signer dependent test set to 42 per cent for the signer independent test set. This result clearly shows that as the signers articulation derivate from the training data, the recognition accuracy decreases. We attribute this performance decrease in signer independent recognition to several factors such as variances in signer performances, differences in interpretation of signs, signing speed and hand shape or sizes. Better results are obtained through the usage of language models at word level recognition. Further improvement of these results is possible through the application of dedicated adaptation methods successfully applied in automatic speech and sign language recognition.

Of our different feature representation methods, we have found Local Binary Patterns and Radial Distance Function to provide the highest recognition accuracies. While LBP yields the highest recognition accuracy in signer dependent case, the modified RDF performs best in signer independent recognition. Elliptic Fourier Descriptors also show a good performance with GMM's modeled using HMMs. A problem in recognizing dynamic letters becomes apparent from comparing the recognition results of all three languages. The experiments showed that the applied features are effective in distinguishing hand gestures, but improvements in accuracy for dynamic hand gestures are possible.

A possible weakness of our fingerspelling recognition method is that we do not make use of spatial position of hands while describing hand gestures. Although we make use of the location of hands and face while initializing tracking and keeping track of keyframes, the spatial coordinates of hand gestures are not represented in

any of our features. Although this provides us with an advantage in dealing with performance variations of different signers, it also causes the system to neglect valuable information while dealing with two handed gestures. Therefore, a possible feature level improvement of the system can be achieved by incorporating normalized spatial information to our feature vectors.

3 Among three fingerspelling alphabets, the Turkish one yields the highest accuracy rates. We can attribute this to several factors, the most likely being the presence of a higher percentage of two handed gestures and static gestures. In addition, the tracking, feature extraction and recognition modules were all implemented and calibrated using the videos

3 from Turkish fingerspelling dataset. The Russian and Czech languages were incorporated into the system at a later point in the development schedule, which may be a reason of their lower performance.

6. ACKNOWLEDGMENTS

This research was supported by the Ministry of Education of the Czech Republic, project no. ME08106; by the EU and the Ministry of Education of the Czech Republic, project no. CZ.1.07/2.3.00/09.0050; by the Grant Agency of Academy of Sciences of the Czech Republic, project no. 1ET101470416; by the Ministry of Education and Science of Russia, contract no. 11.519.11.4025; by the grant of the President of Russia, project no. MK-1880.2012.8; by the Bogazici University Research Fund, grant agreement no. 6061.

REFERENCES

1. V. Levenshtein, Soviet Phys. Dokl. **10** (1966).
2. P. Campr, E. Dikici, A. Kindiroglu, M. Hruz, Z. Krnoul, A. Ronzhin, H. Sak, D. Schorno, L. Akarun, O. Aran, et al., in *Proc. eNTERFACE 2010, The Summer Workshop on Multimodal Interfaces* (Amsterdam, 2010).
3. E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, IEEE Trans. Audio, Speech, Language Processing **17**, 874 (2009).
4. Z. Krvnoul, J. Kanis, M. Železný, and L. Muller, in *Proc. 5th Int. Workshop on Machine Learning for Multimodal Interaction* (Utrecht, 2008), pp. 180–191.
5. J. Allen, R. Xu, and J. Jin, in *Proc. Pan-Sydney Area Workshop on Visual Information Processing* (Australian Computer Soc. Inc., 2004), Vol. 36, pp. 3–7.
6. D. Exner, E. Bruns, D. Kurz, and A. Grundh, Fast and Reliable CamShift Tracking (2005).
7. M. Hu, IRE Trans. Inf. Theory **8**, 179 (1962).
8. C. Lin and C. Hwang, Pattern Recogn. **20**, 535 (1987).
9. A. Tort, Math. Geol. **35**, 873 (2003).
10. E. Yoruk, E. Konukoglu, B. Sankur, and J. Darbon, IEEE Trans. Image Processing: Publ. IEEE Signal Processing Soc. **15**, 1803 (2006).
11. T. Ojala, M. Pietikainen, and T. Maenpaa, IEEE Trans. Pattern Anal. Machine Intell. **24**, 971 (2002).
12. T. Nguyen, Nguyen, and H. Bischof, in *Proc. 8th IEEE Int. Conf. on Automatic Face & Gesture Recognition* (Amsterdam, 2008), pp. 1–6.
13. S. Siddiqi, G. Gordon, and A. Moore, in *Proc. AIST-ATS* (Citeseer, 2007).

Ahmet A. Kindiroglu graduated from Sabanci University in 2008 and obtained his Sc degree in computer science from Bogazici University in 2011. His main research areas are gesture recognition, sign language recognition, transfer learning and human computer interactions. He is a research and teaching assistant at Bogazici University where he is working towards attaining a PhD degree on Computer Engineering.

Hulya Yalcin is an assistant professor at the Mechanical Engineering department of Istanbul Technical University. Prior to that, she was a postdoctoral fellow in the Robotics Institute at Carnegie Mellon University working with Prof. Takeo Kanade and Prof. Martial Hebert on tracking moving vehicles from airborne video imagery. She received her PhD at Brown University in Division of Engineering under the guidance of Prof. Michael Black and William Wolovich in 2004. She received her BSc and MSc degrees in Electrical & Electronics Engineering from Bogazici University in 1996 and 1999 respectively. Her current research interests include ambient intelligence, human computer interaction for the elderly and disabled, motion estimation, tracking, and surveillance applications. For more details of her research, see her webpage: http://www.cs.cmu.edu/hulya/Research.html

Oya Aran received the BS, MS and PhD degrees in Computer Engineering from Bogazici University, Istanbul, Turkey in 2000, 2002 and 2008, respectively. During her PhD study she focused on hand gesture recognition and sign language recognition. She has published papers in leading computer vision and pattern recognition journals and conferences. She was awarded a Marie Curie Intra-European fellowship in 2009 with NOVICOM (Automatic Analysis of Group Conversations via Visual Cues in Non-Verbal Communication) project. She joined the Idiap Research Institute, Martigny, Switzerland in June 2009 as a postdoctoral researcher and a Marie Curie fellow. Her research interests include pattern recognition, machine learning, computer vision, human-computer interaction and social computing.

Marek Hrúz was born in 1983 in Slovakia. He received his MS degree in cybernetics from the University of West Bohemia (UWB) in 2006. As a PhD candidate at the Department of Cybernetics, UWB, his research interests focus on hand gesture and sign language recognition, particularly tracking and image parametrization, computer vision, machine learning and multimodal human-computer interaction. He is participating in the research projects MUSSLAP and POJABR and is a teaching assistant at the Department of Cybernetics.

Pavel Campr was born in 1981 in the Czech Republic. He graduated in cybernetics from the University of West Bohemia (UWB) in 2005. As a PhD candidate at the Department of Cybernetics, UWB, his research interests focus on gesture and sign language recognition, computer vision, machine learning and multimodal human-computer interaction. He is participating in the research projects MUSS-LAP (Multimodal Human Speech and Sign Language Processing for Human-Machine Communication), ARET (Automatic Reading of Educational Texts for Vision Impaired Students) and POJABR (Language handicap elimination for hearing-impaired students via automatic language processing). He is also teaching assistant, enthusiastic about web technologies, maintainer of the departmental website and coorganizer of the eNTERFACE'11 workshop.

Lale Akarun received the BS and MS degrees in electrical engineering from Bogazici University, Istanbul, Turkey, in 1984 and 1986, respectively, and the PhD degree from Polytechnic University, Brooklyn, NY, in 1992. From 1993 to 1995, she was Assistant Professor of electrical engineering at Bogazici University, where she is now Professor of computer engineering. Her current research interests are in image processing, computer vision, and computer graphics.

Alexey A. Karpov was born in 1978. Graduated from St. Petersburg State University of Aerospace Instrumentation (SUAI) in 2002 and obtained PhD degree in computer science from SPIIRAS in 2007. His main research interests are in ASR, TTS, HCI, multimodal user interfaces. He is the senior researcher of the Speech and Multimodal Interfaces Laboratory of SPIIRAS. Senior Researcher of the Speech and Multimodal Interfaces Laboratory of SPIIRAS. Scientific interests: multimodal user interfaces, audiovisual speech processing, speech recognition, assistive information tech- 4
nologies. Author of more than 130 publications in journals and proceedings of International conferences. Member of ISCA, EURASIP (Local Liaison Officer in Russia), IAPR associations. Awarded with the medal of RAS for the best research work of young Russian scientists in 2011.

SPELL: 1. loglikelihood, 2. FingerSign, 3. fingerspelling, 4. assistive