A New System for Automatic Recognition of Italian Sign Language

Marco Fagiani, Emanuele Principi, Stefano Squartini, and Francesco Piazza

Dipartimento di Ingegneria dell'Informazione
Università Politecnica delle Marche, Ancona, Italy
{m.fagiani,e.principi,s.squartini,f.piazza}@univpm.it

Abstract. This work proposes a preliminary study of an automatic recognition system for the Italian Sign Language (Lingua Italiana dei Segni - LIS). Several other attempts have been made in the literature, but they are typically oriented to international languages. The system is composed of a feature extraction stage, and a sign recognition stage. Each sign is represeted by a single Hidden Markov Model, with parameters estimated through the resubstitution method. Then, starting from a set of features related to the position and the shape of head and hands, the Sequential Forward Selection technique has been applied to obtain feature vectors with the minimum dimension and the best recognition performance. Experiments have been performed using the cross-validation method on the Italian Sign Language Database A3LIS-147, maintaining the orthogonality between training and test sets. The obtained recognition accuracy averaged across all signers is 47.24%, which represents an encouraging result and demonstrates the effectiveness of the idea.

1 Introduction

Sign languages [15] fulfil the same social and mental functions as spoken languages, allowing deaf and hearing impaired people to communicate in a comfortable way. These languages arise spontaneously, evolve rapidly, and are so geographical specific to present differences also in regions where the oral language is unique.

Interaction and communication between deaf and hearing people entails the same misunderstanding problem found in oral languages, given that is widely accepted that sign languages are independent on oral ones. The number of problems increase taking into account that the interpreters, with their indispensable work, cannot be always present and that less of the 0.1% of total population belong to hearing impaired community. Therefore, "signers" are faced with a considerable difficulty and a social inclusion problem. That is why the development of sign language recognition and synthesis tools have been gaining an increasing interest among the scientific community.

In [1], algorithms have been developed to robustly extract features from hands and face, with experiments conducted in uncontrolled environments. A multiple

B. Apolloni et al. (Eds.): *Neural Nets and Surroundings*, SIST 19, pp. 69–79.
DOI: 10.1007/978-3-642-35467-0_8 © Springer-Verlag Berlin Heidelberg 2013

hypotheses tracking approach is used for the extraction of hands features, while the extraction of face ones is based on the active appearance model. The recognition stage is based on Hidden Markov Models (HMM) and is designed for the detection of both isolated and continuous signs. Theodorakis *et al.* [17] propose a so-called product HMMs for efficient multi-stream fusion, while Maebatake *et al.* [14] discussed the optimal number of the states and mixtures to obtain the best accuracy. SignSpeak [9] is one of the first European funded projects that tackles the problem of automatic recognition and translation of continuous sign language. Among the works of the project, Dreuw and colleagues in [8,7,6] introduced important innovations: they used a tracking adaptation method to obtain an hand tracking path with optimized tracking position, and they trained robust models through visual speaker alignments and virtual training samples. A first framework for continuous sign language recognition applied to Italian has been proposed in [12]. The system uses features based on centroid coordinates of the hands and a self-organizing map neural network to classify the single sign and to provide a list of probable meanings. Experiments are performed on a corpus of 160 videos with a vocabulary of 40 signs (each sign is repeated four times).

This paper proposes an automatic sign language recognition system based on HMM. The structure and the parameters of the sign models are investigated by means of the resubstitution method, and a study on the feature set is presented using the Sequential Forward Selection approach. In addition, the system uses only features obtained from the evaluation of position and contours of face and hands. Facial features related to mouth, lips, nose and eyes are not used. Being a preliminary study, the simpler task of isolated sign recognition is addressed, while continuous sign recognition is reserved for future works. Experiments have been carried out on a recently proposed Italian Sign Language Database (A3LIS-147) [10], and the evaluation has been performed using the cross-validation approach, this way guaranteeing the independence between training and test sets. This represents a notable difference from [12], where the same signers appear in both sets.

The outline of the paper is as follows: Section 2 briefly illustrates the used video corpus. Section 3 proposes an overview of the developed system. Section 4 presents and discusses recognition results. Finally, Section 5 draws conclusions and proposes future developments.

2 Database A3LIS-147

This section describes the principal characteristics of the video corpus that has been used for the evaluation of the developed system. The database has been presented in [10], and is composed of 147 distinct isolated signs executed by 10 different signers: 7 males and 3 females. Each video presents a single sign which is preceded and succeeded by the occurrence of the "silence" sign.

Video sequences have been acquired by using a commercial camera located in front of the subject. Behind the signer, a green chroma-key background is placed and two diffuse lights (400 W each) are used to ensure an uniform lighting. The

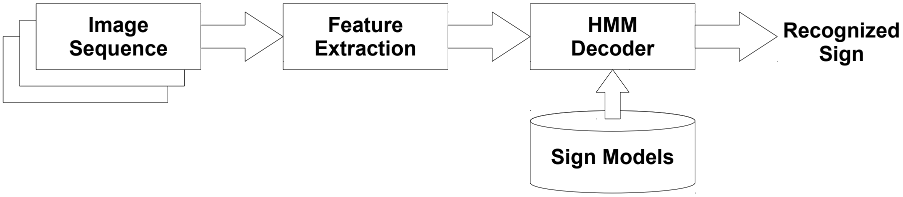

Fig. 1. Components of the sign language recognition system

video streaming has been acquired at 25 fps with a resolution of 720x576 pixel. The video sequences have been stored in Digital Video (DV) format in order to keep the maximum image quality.

3 System Overview

Given a sign language video input, the goal of the developed recognition system is to produce the corresponding sign as text. An overview of the whole system architecture is depicted in Fig. 1: in the first stage, the system processes the input video and extracts meaningful features. As a preliminary step, head and hands regions are located using a skin detection algorithm, then the actual feature matrix of size $N \times L$ is calculated. Here, N is the total feature vector dimension and L is equal to the video frames number.

The HMM decoder takes as input the feature matrix, calculates the output probabilities for each model, and selects the sign which scores best. The HMM models of each sign are not known and have to be estimated from the training data.

3.1 Feature Extraction

As aforementioned, the first step of the feature extraction stage is the discrimination of skin regions. In this work, skin recognition is based on the evaluation of each pixel colour in the YC_bC_r colour space, which produces better performance with respect to RGB since it is luminance independent. The corresponding skin cluster is given as [13]:

$$\begin{cases} Y > 80 \\ 85 < C_b < 135 \\ 135 < C_r < 180 \end{cases} \tag{1}$$

where $Y, C_b, C_r \in [0, 255]$.

The pixels satisfying the above conditions are included in the skin mask, with head and hands determined from the regions with largest areas (Fig. 2(b)). A morphological transformation [3,16], closing operation, has been introduced to merge the nearby regions on the skin mask that belong to the same zones of

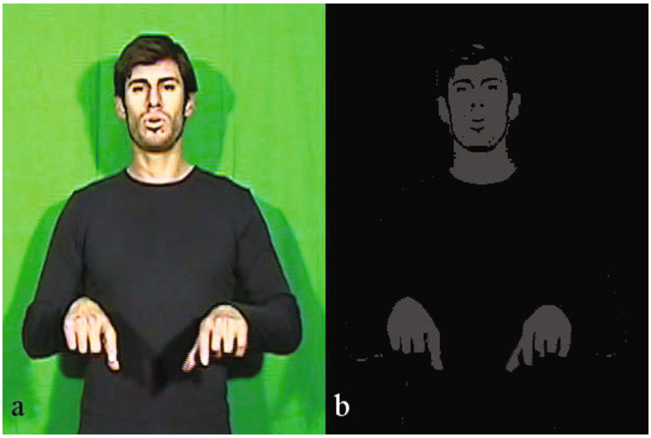

Fig. 2. Frame from the sign "abbonamento" (a). Skin regions detected (b).

interest.The closing is obtained applying first the dilation operator, then the erosion one. Fig. 3(a) shows the contours of the selected skin regions.

Table 1 shows the entire feature set calculated from the interest regions. The set can be divided into two different subsets: the first comprises features 1–9 and is obtained starting from the relative position of head and hands obtained by the mass center of the detected regions. It includes head and hands positions, position derivatives (movements velocity) and the hands distance. The second subset comprises features 10–18 and represents the general characteristics of head and hands, such as the area and area derivative, the shape type and the 2-D spatial orientation of the hands.

In the first subset of features, two different ways to compute the spatial moments for the mass center are used: features 1–3 use the Canny algorithm [4], while features 4–9 directly exploit the contours segmented from the skin mask (Fig. 3(a)). The Canny algorithm is applied within the bounding rectangle of the hands and obtains their contours as illustrated in Fig. 3(c). For the features 7–9 the head position is used as the center of the coordinate system. Features 10 and 11 are moments invariant to translation, changes in scale, and rotation [11]. Both for the head and hands regions, 7 different values are computed. As for the first features group, the Hu-moments are obtained in two different ways: using the Canny algorithm for 10 and using directly the contours for 11. The compactness 14 and eccentricity 15 of head and hands have been proposed in [17]. The first one is computed as the ratio of the minor and major axis lengths of the detected region. The second is obtained as the ratio of area and perimeter squared of the region. Features 16–18 have been introduced in [1], where $\mu_{p,q}$ indicates the corresponding central moments. The hands orientation $\alpha \in [-90^o, 90^o]$ is split into o_1 and o_2 to ensure stability at the interval borders. Note that all measurements are normalized respect the head dimension, so that they are independent of the distance from the camera and the signer height.

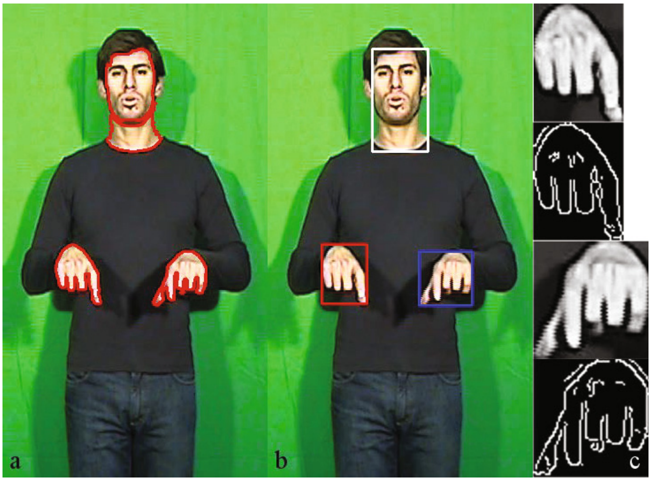

Fig. 3. Contours detection from the skin regions (a). Bounding rectangle of hands skin zones (b). The hands edge obtained by the Canny algorithm (c).

The overlap between head and hands, and between hands complicates the estimation of the mass centers of these regions. An overlap is detected when the skin mask contains two regions instead of three. When this occurs, the system evaluates whether the head area increased from the previous frame. An increase is considered as an head-hand overlap. Otherwise, if the area is unchanged, an hand-hand overlap is reported and the same region is assigned to both hands. With regard to the head, its position during the execution of a sign is considered as almost stationary. In order to limit incongruous deviations due to overlap with hands, a simple smoothing of the $(x, y)_t$ coordinates of the mass center at time t is computed:

$$(x, y)_t = \alpha(x, y)_t + (1 - \alpha)(x, y)_{t-1}, \tag{2}$$

where $\alpha \in [0, 1]$. With regard to the hands, the overlap with the head is managed assigning the position of the whole region at time t to the overlapped hand.

3.2 Sign Recognition

Hidden Markov Models are widely used in speech recognition applications, and they have been chosen for the sign recognition stage due to the similarity of the task. For each sign, a different left-to-right model with S states and G mixtures of Gaussians is created. The choice of the values of S and G is discussed in the following section. Recognition is performed calculating the likelihood for each sign model and selecting the sign whose model likelihood is highest.

Table 1. Feature set: A and P represent respectively the area of the region and its perimeter

	Features	**Parameters**
1	Hands centroid center Canny-Filter normalized respect head position	$(x,y)_{right}$ $(x,y)_{left}$
2	Hands centroid center derivatives Canny-Filter normalized respect head position	$(\dot{x},\dot{y})_{right}$ $(\dot{x},\dot{y})_{left}$
3	Hands distance Canny-Filter normalized respect head dimension	d_{norm}
4	Hands centroid center contours (segmented by skin detection)	$(x,y)_{right}$ $(x,y)_{left}$
5	Hands centroid center derivatives contours	$(\dot{x},\dot{y})_{right}$ $(\dot{x},\dot{y})_{left}$
6	Hands distance contours	d_{norm}
7	Hands centroid center contours referring at head coordinates	$(x,y)_{right}$ $(x,y)_{left}$
8	Hands centroid center derivatives contours referring at head coordinates	$(\dot{x},\dot{y})_{right}$ $(\dot{x},\dot{y})_{left}$
9	Hands distance contours referring at head coordinates	d_{norm}
10	Head and hands Hu-moments Canny-Filter	I_1 I_2 I_3 I_4 I_5 I_6 I_7
11	Head and hands Hu-moments contours	I_1 I_2 I_3 I_4 I_5 I_6 I_7
12	Hands area normalized with head area	A_{right} A_{left}
13	Hands area derivative	$\dot{A}_{right}$ $\dot{A}_{left}$
14	Head and hands compactness 1	$AXIS_{min}/AXIS_{max}$
15	Head and hands eccentricity 1	A_{norm}/P^2
16	Head and hands compactness 2	$(4\pi A_{norm})/P^2$
17	Head and hands eccentricity 2	$((\mu_{2,0}-\mu_{0,2})^2+4\mu_{1,1})/A_{norm}$
18	Hands orientation	$o_1=\sin(2\alpha)$ $o_2=\cos(\alpha)$

4 Experimental Results

All the video elaboration have been executed with the OpenCV, an open source computer vision library [2]. The aforementioned A3LIS-147 database has been used for all the recognition test and training, and the sign recognition stage has been implemented using the Hidden Markov Model toolkit [19].

The first experiment studies the number of states and mixtures in left-to-right HMM representing each sign. The experiments has been performed evaluating the recognition results for different number of states (from 1 to 10) and mixtures (from 1 to 8). The second experiments analyses the performance of the system for different combinations of features.

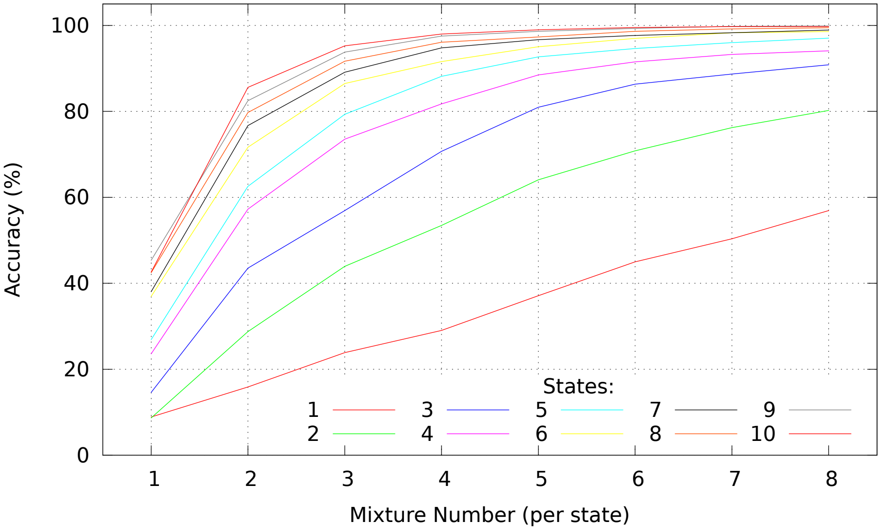

Fig. 4. Resubstitution accuracy for different states and mixtures number

The performance have been evaluated using the sign recognition accuracy. It is defined on the basis of the number of substitution (S), deletion (D) and insertion errors (I) as in speech recognition tasks [19]:

$$\text{Accuracy} = \frac{(N - D - S - I)}{N}. \tag{3}$$

The term N is the total number of labels in the reference transcriptions. It is worth pointing out that in the isolated sign recognition task addressed in this work only the substitution errors are present.

4.1 Hidden Markov Models Parameters Estimation

In order to find an estimate of the number of states and of the number of mixtures per state of the sign models, the resubstitution method [18] has been used. The method entails the utilization of the same data set both for training and for testing with the aim of providing an optimistic estimation of the true error probability. The method has been applied recursively with different states and mixtures number using features 1, 3, 10, 12, 14 and 15.

Fig. 4 shows the obtained results: as expected, increasing the number of states and the number of mixtures per state produces better recognition accuracies. A *plateau* is reached when the number of states is above 7 and the number of mixtures is more than 5, suggesting that a good trade-off between computational cost and performance can be obtained using such values. Using 7 states and 5 mixtures gives a recognition accuracy of 96.69%.

4.2 Features Selection

The best feature combination has been obtained by means of the Sequential Forward Selection (SFS) [18] technique. SFS is sub-optimal, and consists of the following steps:

1. compute the criterion value for each of the features;
2. select the feature with the best value;
3. form all possible vectors that contain the winner from the previous step;
4. compute the criterion value for each of the new vectors;
5. select the best one;
6. repeat until the end of the features.

The results for each combination of features have been evaluated by means of the cross-validation method [18], a variation of the leave-one-out method, where $K > 1$ samples are used for testing and the remaining $N - K$ samples are used for training. The procedure is repeated for N iterations, each time excluding a different set of K samples. This way, training is performed using all samples and the independence between the training and test sets is maintained, differently from the approach followed in [12]. In the experiments, $N = 1470$ and $K = 147$, where the K samples have been chosen so that they all belong to the same signer.

Preliminary recognition tests have been executed and, at the fourth step, the selection reached a maximum accuracy of 38.76% with the combination of features 5, 13, 7 and 8: hands centroid center derivatives contours, hands area derivative, hands centroid center contours and derivatives referring at head coordinates. The analysis of these results has led to the introduction of crucial improvements (e.g., smoothing of the head position), already discussed in Section 3.1.

Table 2. Sequential Forward Selection experimental results (accuracy of first four best selection/selected step). The "Total" row values are obtained as an average of the recognition rate performed for each cross-validation step, that is for each signer.

Signer	1 SFS	2 SFS	3 SFS	4 SFS
	Feature 5	Feat. 5-13	Feat. 5-13-8	Feat. 5-13-8-7
fal	51.03%	55.86%	57.24%	47.59%
fef	17.93%	20.69%	15.86%	22.07%
fsf	28.28%	37.24%	30.34%	35.86%
mdp	39.31%	42.76%	46.21%	44.83%
mdq	47.59%	49.66%	46.21%	53.10%
mic	48.97%	51.03%	53.10%	45.52%
mmr	49.66%	57.93%	52.41%	55.17%
mrla	49.66%	53.10%	51.72%	**61.38%**
mrlb	50.34%	55.17%	56.55%	60.00%
msf	48.97%	48.97%	46.90%	46.21%
Total	43.17%	**47.24%**	45.65%	47.17%

The best results of the subsequent feature selection are reported in Table 2. Feature 5 obtains the best performance (43.17%) for the first evaluation of the feature selection. In the second step the combination with feature 13 achieves an average recognition rate of 47.24%. The combinations with the feature 8, at the third step, and then with the feature 7, fourth step, produce respectively an overall accuracy of 45.65% and 47.17%. At the end of the sequential forward feature selection, an overall best accuracy of 47.24% is obtained at the second step of the selection by the combination of the hands centroid center derivatives and hands area derivative.

The main causes of errors have been analyzed by means of the confusion matrix among all signers. The analysis gives better insights of the system weak points, and indicates which signs limit the overall recognition accuracy. One of the most frequent causes of errors is the overlap between hands, and between head and hands, which causes the wrong revelation of their position and boundaries. Another source of errors, is the inability to discriminate in a proper way the 2-D projection respect the 3-D true form of a hand. This problem is emphasized if there are signs that have the same movements for the arms. In addition, some movements, such as those parallel to the line of sight of the camera, are not detectable in a 2-D field.

5 Conclusions

In this paper, a preliminary study of an automatic recognition system of Italian Sign Language based on Hidden Markov Models has been proposed. Using features obtained from the position of hands and face, a study on the parameters of the model representing each sign has been proposed, and the feature set has been selected by means of the Sequential Forward Selection approach. Experiments have been performed on the A3LIS-147 database of isolated Italian signs through the cross-validation approach, this way guaranteeing the orthogonality between training and test sets. In the final system, each sign is represented with an HMM composed of 7 states and 5 mixtures and 4 of the original 18 features are used. The obtained recognition accuracy averaged across all signers is 47.24%, with a maximum of 57.93%. Noteworthy the recognition accuracy of 47.17% reached by the fourth step of the selection with a remarkable maximum signer accuracy of 61.38%.

In order to have better insights on the performance of the system, the confusion matrix among all signers has been analysed: the main causes of errors identified are the overlap between head and hands and between the two hands, and the inability to discriminate the 3-D position of the subject. These problems are more evident for the signs related to the days of the week, they have similar arm movements and the information is entirely contained in the hands shapes. Therefore, an overlap impedes to detect the correct shape of the hands resulting in loss of valuable information. The computational burden of the recognition process has been evaluated in terms of real-time factor, defined as the ratio between the processing tine and the length of the video sequence. On average, the whole process, requires about 8 seconds, and video segments are 3 seconds long. This results in a real-time factor of about 2.6, without taking into account solutions to lower the computational cost.

As future works, the overlap problem will be primarily addressed with the introduction of tracking techniques, as Mean-Shift or CAM-Shift. In addition, the inclusion of novel features containing lip movements and Histogram of Oriented Optical Flow (HOOF) [5] will be investigated. HOOF features, in particular, are suitable for real-time computation, and could be obtained directly from the flow information encoded in a compressed video. Finally, the authors are considering the development of a new sign database including audio and spatial information (e.g., using multiple cameras, or Microsoft Kinect$^{\text{TM}}$) to augment the feature set and reduce errors due to overlapping and wrong spatial detection of the subject.

References

1. von Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent Developments in Visual Sign Language Recognition. Univers. Access Inf. Soc. 6(4), 323–362 (2008)
2. Bradski, G.: The OpenCV library. j-DDJ 25(11), 122–125 (2000)
3. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Inc., Sebastopol (2008)
4. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
5. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of Oriented Optical Flow and Binet-Cauchy Kernels on Nonlinear Dynamical Systems for the Recognition of Human Actions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1932–1939. IEEE Computer Society, Los Alamitos (2009)
6. Dreuw, P., Forster, J., Deselaers, T., Ney, H.: Efficient Approximations to Model-Based Joint Tracking and Recognition of Continuous Sign Language. In: 8th IEEE International Conference on Automatic Face Gesture Recognition, pp. 1–6. IEEE Computer Society, Los Alamitos (2008)
7. Dreuw, P., Neidle, C., Sclaroff, V.A.S., Ney, H.: Benchmark Databases for Video-Based Automatic Sign Language Recognition. In: Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). European Language Resources Association (ELRA), Marrakech (2008)
8. Dreuw, P., Ney, H.: Visual Modeling and Feature Adaptation in Sign Language Recognition. In: ITG Conference on Voice Communication (SprachKommunikation), pp. 1–4 (2008)
9. Dreuw, P., Ney, H., Martinez, G., Crasborn, O., Piater, J., Moya, J.M., Wheatley, M.: The SignSpeak Project - Bridging the Gap Between Signers and Speakers. In: Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA), Valletta (2010)
10. Fagiani, M., Principi, E., Squartini, S., Piazza, F.: A New Italian Sign Language Database. In: Zhang, H., Hussain, A., Liu, D., Wang, Z. (eds.) BICS 2012. LNCS, vol. 7366, pp. 164–173. Springer, Heidelberg (2012)
11. Hu, M.-K.: Visual Pattern Recognition by Moment Invariants. IRE Trans. on Information Theory 8(2), 179–187 (1962)

12. Infantino, I., Rizzo, R., Gaglio, S.: A Framework for Sign Language Sentence Recognition by Commonsense Context. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(5), 1034–1039 (2007)
13. Kukharev, G., Nowosielski, A.: Visitor Identification - Elaborating Real Time Face Recognition System. In: Proc. of WSCG (Short Papers), pp. 157–164. UNION Agency, Plzen (2004)
14. Maebatake, M., Suzuki, I., Nishida, M., Horiuchi, Y., Kuroiwa, S.: Sign Language Recognition Based on Position and Movement Using Multi-Stream HMM. In: Proc. of the 2nd Int. Symposium on Universal Communication, pp. 478–481. IEEE Computer Society, Los Alamitos (2008)
15. Sandler, W., Lillo-Martin, D.: Sign Language and Linguistic Universals. Cambridge University Press, Cambridge (2006)
16. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, Inc., Orlando (1983)
17. Theodorakis, S., Katsamanis, A., Maragos, P.: Product-HMMs for Automatic Sign Language Recognition. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1601–1604. IEEE Computer Society, Washington (2009)
18. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, Fourth Edition. Academic Press, Burlington (2008)
19. Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book, version 3.4. Cambridge University Press, Cambridge (2006)