Sequential Belief-Based Fusion of Manual and Non-Manual Information for Recognizing Isolated Signs

Oya Aran, Thomas Burger, Alice Caplier, Lale Akarun

Sequential Belief-Based Fusion of Manual and Non-Manual Information for Recognizing Isolated Signs

Oya Aran[1], Thomas Burger[2], Alice Caplier[3], Lale Akarun[1]

[1]Dep. of Computer Engineering, Bogazici University 34342 Istanbul, Turkey
aranoya@boun.edu.tr, akarun@boun.edu.tr
[2]France Telecom R&D, 28 ch. Vieux Chêne, Meylan, France
thomas.burger@orange-ftgroup.com
[3]GIPSA-lab, 46 avenue Félix Viallet, 38031 Grenoble cedex 1, France
alice.caplier@lis.inpg.fr

Abstract. This work aims to recognize signs which have both manual and non-manual components by providing a sequential belief-based fusion mechanism. We propose a methodology based on belief functions for fusing extracted manual and non-manual features in a sequential two-step approach. The belief functions based on the likelihoods of the hidden Markov models are used to decide whether there is an uncertainty in the decision of the first step and also to identify the uncertainty clusters. Then we proceed to the second step which utilizes only the non-manual features within the identified cluster, only if there is an uncertainty.

Keywords. Sign language recognition, hand gestures, head gestures, non-manual signals, hidden Markov models, belief functions

1 Introduction

Sign language (SL) is the natural communication medium of hearing impaired people. Similar to the evolution of spoken languages, many sign languages have evolved in different regions of the world. American Sign Language (ASL), British Sign Language, Turkish Sign Language, French Sign Language are different sign languages used by corresponding communities of hearing impaired people. These are visual languages and the whole message is contained not only in hand motion and shapes (manual signs - MS) but also in facial expressions, head/shoulder motion and body posture (non-manual signals - NMS).

Sign language recognition (SLR) is a very complex task: a task that uses hand shape recognition, gesture recognition, face and body parts detection, facial expression recognition as basic building blocks. For an extensive survey on SLR, interested readers may refer to [1]. Most of the SLR systems concentrate on MS and perform hand gesture analysis only [2]. As a state of the art, Hidden Markov models (HMM) and several variants are used successfully to model the signs [3]. In [4], a parallel HMM architecture is used to recognize ASL signs where each HMM models

the gesture of left and right hands respectively. A similar approach is applied to integrate the hand shape and movement [5].

However, without integrating NMS, it is not possible to extract the whole meaning of the sign. In almost all of the sign languages, the meaning of a sign can be changed drastically with the facial expression or body posture while the hand gesture remains the same. Moreover, the NMS can be used alone, for example to indicate negation in many SLs. Current multimodal SLR systems either integrate lip motion and hand gestures, or only classify either the facial expression or the head movement. There are only a couple of studies that integrate non-manual and manual cues for SLR [1]. NMS in sign language have only recently drawn attention for recognition purposes. Most of those studies attempt to recognize only non-manual information independently, discarding the manual information. Some works only use facial expressions [6], and some use only the head motion [7].

We propose a methodology for integrating MS and NMS in a sequential approach. The methodology is based on (1) identifying the level of uncertainty of a classification decision, (2) identifying sign clusters, and (3) identifying the correct sign based on MS and NMS. Our sequential belief-based fusion methodology is explained in Section 2 and our automatic sign cluster identification is explained in Section 3. In Section 4, we give the results of our experiments.

2 Sequential Belief Based Fusion

In a SLR problem, where a generative model such as an HMM is used to model the signs, the classification can be done via the maximum likelihood (ML) approach. In the ML approach, for a test sign, the sign class is selected as the class of the HMM that gives the maximum likelihood. The problem of the ML approach is that it does not consider the situations where the likelihoods of two or more HMMs are very close to each other. The decisions made in these kinds of cases are error-prone and further analysis must be made. We propose to use belief functions to consider such situations. Belief function formalism provides a way to represent hesitation and ignorance in different ways. This formalism is especially useful when the collected data is noisy or semi-reliable. Interested readers may refer to [8], [9] for more information on belief theories.

In HMM based SLR, each HMM typically models a different class for the sign to be recognized [10]. Our purpose is to associate a belief function with these likelihoods. Then, it is possible to model these error-prone cases by associating high belief into the union of classes. By analyzing the proportion of belief which is associated with the union of classes, it is possible to decide whether the classification decision is certain or error-prone [11]: when the decision is certain, a single class is selected, whereas, when it is uncertain or error-prone, a subset of classes among which the good decision is likely to be found is selected. In this latter case, the decision is incomplete.

We propose the following SLR process: (1) Sign clusters are defined based on the similarity of the classes. (2) A first classification step is made. If the analysis of this classification indicates no uncertainty, then, a decision is made and the classification

process is over. On the contrary, if there is significant uncertainty, (3) a second classification step must be applied. This second step is applied to classes among which the uncertainty is detected.

The two-stage sequential belief based fusion technique is illustrated in Fig. 1. In this setup, the assumption is that the HMMs of the first bank are more general models which are capable of discriminating all the classes up to some degree. The HMMs of the second bank are specialized models and can only be used to discriminate between a subset of classes, among which there is an uncertainty.

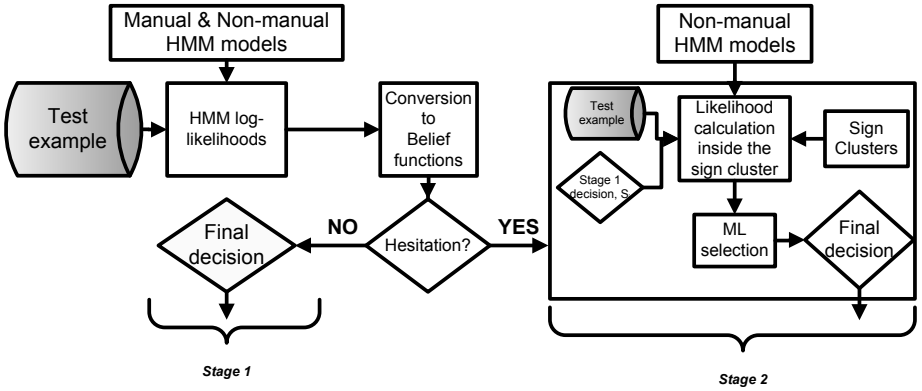


Fig. 1. Sequential belief-based fusion flowchart

3 Automatic Sign Cluster Identification

What we define as a sign cluster is a group of signs which are similar and the differences are either based on the non-manual component or variations of the manual component. In SLR point of view, a sign cluster indicates signs that are hard to discriminate. This can be a result of performance differences or systematic differences such as usage of NMS, or variations of MS. In linguistic point of view, a semantic interpretation of the signs may lead to totally different clusters. In a recognition task, although one can utilize prior knowledge such as the sign clusters based on semantic information, this has some disadvantages. First, it is not guaranteed that these semantic clusters are suitable for the recognition task, and second, the trained model will be database dependent and extending the database with new signs will require the re-definition of the cluster information. Thus, an automatic clustering method that depends on the data and considers the capabilities of the classifier would be preferable.

A first classical method is to use the confusion matrix of the HMM based classifier to automatically identify sign clusters. The confusion matrix is converted to a sign cluster matrix by considering the confusions for each sign. Signs that are confused form a cluster. For example, assume that sign i is confused with sign j half of the time. Then the sign cluster of class i is $\{i,j\}$. The sign cluster of class j is separately

calculated from its confusions in the estimation process. The disadvantage of this method is its sensitivity to odd mistakes which may result from the errors in the feature vector calculation as a result of bad segmentation or tracking.

We propose a more robust alternative which evaluates the decisions of the classifier and only consider the uncertainties of the classifier to form the sign clusters. For this purpose, we define a hesitation matrix. Its purpose is close to the classical confusion matrix, but it contains only the results of the uncertain decisions, regardless of their correctness. Then, when a decision is certain, either true or false, it is not taken into account in the calculation of the hesitation matrix. On the other hand, when a decision is uncertain between sign *i* and sign *j*, it is counted in the hesitation matrix regardless of the ground truth of the sign being, *i, j* or even *k*. As a matter of fact, the confusion between a decision (partial or not) and the ground truth can be due to any other mistake (segmentation, threshold effect, etc…) whereas, the hesitation on the classification process depends on the ambiguity at the feature level with respect to the class borders. Our method of determining clusters only based on the hesitation is more robust. In addition, it is not necessary to know the ground truth on the validation set on which the clusters are defined. This is a distinctive advantage in case of semi-supervised learning, to adapt the system to the signer's specificity.

4 Methodology & Experiments

In order to assess the appropriateness of our belief-based method, we have performed experiments on a sign language database which has been collected during the eNTERFACE'06 workshop. In the following section, we give details about this database.

Table 1. Signs in eNTERFACE'06 Database

Base Sign	Variant	Variation on hand motion	Head Motion (NMS)
Clean	Clean		
	Very clean		✓
Afraid	Afraid		
	Very afraid	✓	✓
Fast	Fast		
	Very fast		✓
drink	To drink		✓
	Drink (noun)	✓	
open (door)	To open		
	door (noun)	✓	
Here	[smbdy] is here		✓
	Is [smbdy] here?		✓
	[smbdy] is not here		✓
Study	Study		
	Study continuously	✓	✓
	Study regularly	✓	✓
Look at	Look at		
	Look at continuously	✓	✓
	Look at regularly	✓	✓

4.1 eNTERFACE'06 ASL Database

The signs in the eNTERFACE'06 American Sign Language (ASL) Database [12] are selected such that they include both manual and non-manual components. There are eight base signs that represent words and a total of 19 variants which include the systematic variations of the base signs in the form of NMS, or inflections in the signing of the same MS. A base sign and its variants will be called as a "*base sign cluster*" for the rest of this paper. Table 1 lists the signs in the database. As seen from Table 1, some signs are differentiated only by the head motion; some only by hand motion variation and some by both.

A single web camera with 640x480 resolution and 25 frames per second rate is used for the recordings. The camera is placed in front of the subject. The database is collected from eight subjects, each performing five repetitions of each sign. Fig. 2 shows example signs form the database.

The dataset is divided to training and test set pairs where 532 examples are used for training (28 examples per sign) and 228 examples for reporting the test results (12 examples per sign). The distributions of sign classes are equal both in training and test sets. For the cases where a validation set is needed, we apply a stratified 7-fold cross validation (CV) on the training set.

Fig. 2. Sign CLEAN and VERY CLEAN. The main difference between these two signs is the existence of NMS, motion of the head.

Since we concentrate on the fusion step in this paper, we have directly used the processed data from [13] where the features of hand shape, hand motion and head motion are extracted. In the following sections, we summarize the detection and feature extraction methodology. Further details can be found in [13].

4.2 Hand and Face Detection

To ease the hand and face detection, subjects in the eNTERFACE'06 ASL database wear gloves with different colors when performing the signs. We use the motion cue and the color cue of the gloves for hand segmentation. Although skin color detection can be applied in restricted illumination and lighting conditions, segmentation becomes problematic when two skin regions, such as hands and face, overlap and occlude each other. For the signs in the eNTERFACE ASL database, the hand position is often near the face and sometimes, in front of the face.

Hands are segmented by using trained histograms for each glove color using HSV color space [14]. The connectivity within the hands is ensured by double thresholding and the largest connected component over the detected pixels is considered as the

hand. A bounding box around the face is found by applying the Viola and Jones face detection algorithm [15], using the MPI toolbox [16].

4.3 Feature Extraction

Sign features are extracted both for MS (hand motion, hand shape, hand position with respect to face) and NMS (head motion). The resulting feature vector, for two hands and the head, is composed of 61 features per frame.

The system tracks the center of mass of the hand and calculates the coordinates and velocity of each segmented hand at each frame. Two independent Kalman filters, one for each hand, are used to obtain smoothed trajectories. This is required since the original calculations are corrupted by segmentation noise and occlusion. The motion of each hand is approximated by a constant velocity motion model, in which the acceleration is neglected. We calculate the hand motion features for each hand from the posterior states of each Kalman filter: x, y coordinates of the hand center of mass and velocity. The hand motion features in each trajectory are further normalized to the range [0,1] by min-max normalization.

Hand shape features are appearance-based shape features calculated on the binary hand images. These features include the parameters of an ellipse fitted to the binary hand and statistics from a rectangular mask placed on top of the binary hand [13]. Most of the features are scale invariant. The recordings are with a single camera and the features do not have depth information; except for the foreshortening due to perspective. In order to keep this depth information, some features were not normalized. Prior to the calculation of the hand shape features, we take the mirror reflection of the right hand so that we analyze both hands in the same geometry; with thumb to the right.

Hand position is calculated with respect to the face center of mass. The distance between the hand and the face is calculated by the x and y coordinates and normalized by the face width and height respectively.

For head motion analysis, the system detects rigid head motions such as head rotations and head nods working in a way close to the human visual system [17]. By analyzing the head motion, two features are extracted: the quantity of motion and motion event alerts. Only on the motion events alerts, with an optic flow algorithm, both the orientation and velocity information are provided. As head motion features, we use the quantity of motion and the vertical, horizontal velocity of the head at each frame.

4.4 Clustering for Sequential Fusion

As explained in Section 4, we propose a belief based method for automatic identification of the clusters via the hesitation matrix: The clusters are defined by transforming the hesitation matrix so that it is closed, transitive and reflexive. The cluster identification is done by applying 7-fold CV on the training data. The hesitation matrices of each fold are combined to create a joint matrix, which is used to identify the clusters.

Fig. 3 shows the sign clusters identified by the uncertainties provided by the belief functions. The automatically identified sign clusters are the same as the base sign clusters except for the base signs LOOK AT and STUDY.

For the LOOK AT sign, the differentiation is provided by both non-manual information and variations in signing. However, the hands can be in front of the head for many of the frames. For those frames, the face detector may fail to detect the face and may provide wrong feature values which can mislead the recognizer.

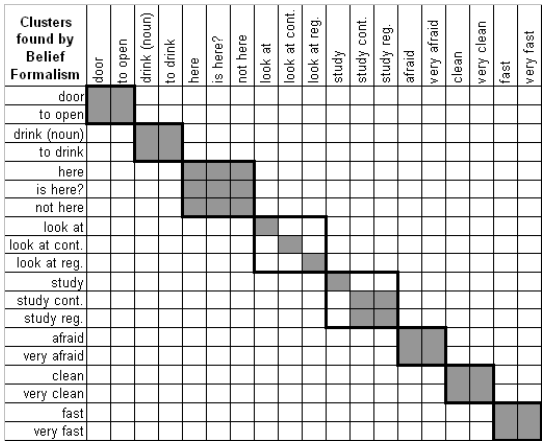


Fig. 3. Sign clusters identified by the uncertainties between the classes in 7-fold cross validation. Clusters are shown row-wise, where for each sign row, the shaded blocks show the signs in its cluster.

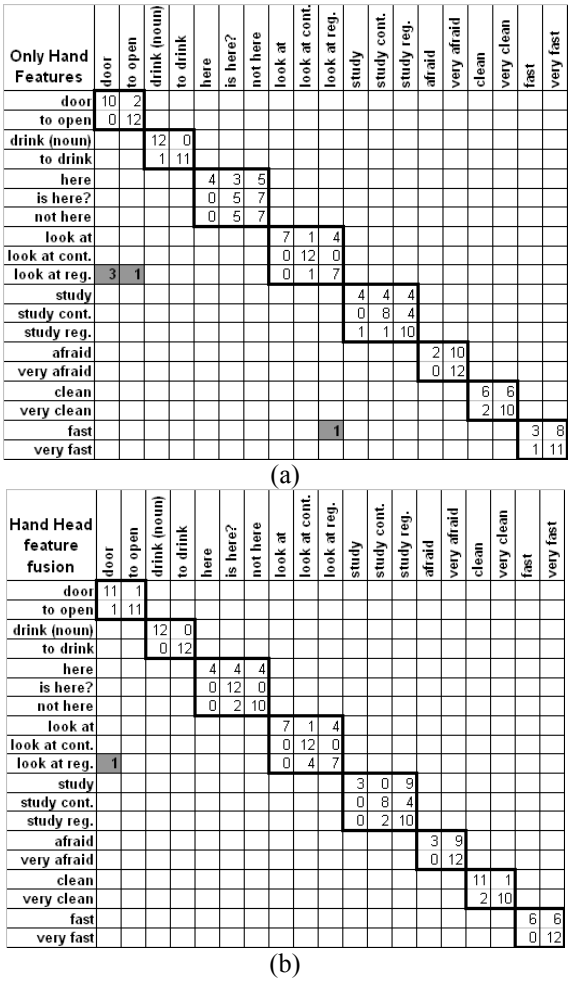
It is interesting to observe that the base STUDY sign is clustered into two sub-clusters. This separation agrees with the nature of these signs: In the sign STUDY, the hand is stationary and this property directly differentiates this sign from the other variations. The confusion between STUDY REGULARLY and STUDY CONTINUOUSLY can stem from a deficiency of the 2D capture system. These two signs differ mainly in the third dimension. However a detailed analysis of the non-manual components can be used at the second stage to resolve the confusion.

4.5 Results

To model the MS and NMS and perform classification, we trained three different HMMs. The first one is trained for comparison purposes and the last two are for the first and second steps of our fusion method:
(1) HMM_M uses only *manual* features;
(2) $HMM_{M\&N}$ uses both *manual and non-manual* features
(3) HMM_N uses only *non-manual* features

The classification of a sign is performed by the maximum likelihood approach. We train HMMs for each sign and classify a test example by selecting the sign class whose HMM has the maximum log-likelihood. The HMM models are selected as left-

to-right 4-state HMMs with continuous observations where Gaussian distributions with full covariance are used to model the observations at each state. Baum-Welch algorithm is used for HMM training. Initial parameters of transition and prior probabilities and initial parameters of Gaussians are randomly selected.



Only Hand Features	door	to open	drink (noun)	to drink	here	is here?	not here	look at	look at cont.	look at reg.	study	study cont.	study reg.	afraid	very afraid	clean	very clean	fast	very fast
door	10	2																	
to open	0	12																	
drink (noun)			12	0															
to drink			1	11															
here					4	3	5												
is here?					0	5	7												
not here					0	5	7												
look at								7	1	4									
look at cont.								0	12	0									
look at reg.	3	1						0	1	7									
study											4	4	4						
study cont.											0	8	4						
study reg.											1	1	10						
afraid														2	10				
very afraid														0	12				
clean																6	8		
very clean																2	10		
fast										1								3	8
very fast																		1	11

(a)

Hand Head feature fusion	door	to open	drink (noun)	to drink	here	is here?	not here	look at	look at cont.	look at reg.	study	study cont.	study reg.	afraid	very afraid	clean	very clean	fast	very fast
door	11	1																	
to open	1	11																	
drink (noun)			12	0															
to drink			0	12															
here					4	4	4												
is here?					0	12	0												
not here					0	2	10												
look at								7	1	4									
look at cont.								0	12	0									
look at reg.	1							0	4	7									
study											3	0	9						
study cont.											0	8	4						
study reg.											0	2	10						
afraid														3	9				
very afraid														0	12				
clean																11	1		
very clean																2	10		
fast																		6	6
very fast																		0	12

(b)

Fig. 4. (a) Confusion matrix of HMM_M, 97.8% base sign accuracy, 67.1% total accuracy (b) Confusion matrix of $HMM_{M\&N}$, 99.5% base sign accuracy, 75.9% total accuracy. Rows indicate the true class and columns indicate the estimated class. Base sign and its variations are shown in bold squares. The classification errors are mainly between variations of a base sign.

We compared the classification performance of HMM_M and $HMM_{M\&N}$ to see the information added by the non-manual features via feature level fusion. The classification results of these two models should show us the degree of effective utilization of the non-manual features when combined into a single feature vector with manual features. Although there is no direct synchronization between the manual

and non-manual components, the second model, $HMM_{M\&N}$, models the dependency of the two components for sign identification.

The classification results and confusion matrices for the two techniques are shown in Fig. 4. Although the classification accuracy of $HMM_{M\&N}$ is slightly better than HMM_M, total accuracy is still low. However, it is worth noting that the classification errors in both of the models are mainly between variants of a base sign and out of cluster errors are very few.

From these confusion matrices, it appears that some mistakes occur between signs which are completely different. It illustrates that the use of such matrices to define the clusters is less robust than the method we propose.

Table 2. Classification performance

Models Used	**Fusion method**	**Cluster identification**	**Test Accuracy**
HMM_M	No fusion	-	67.1 %
$HMM_{M\&N}$	Feature fusion	-	75.9 %
$HMM_{M\&N} \dashrightarrow HMM_N$	Sequential *belief-based* fusion	Hesitation matrix	**81.6 %**

The accuracies of the techniques are summarized in Table 2. Although the time dependency and synchronization of MS and NMS are not that high, feature fusion ($HMM_{M\&N}$) still improves the classification performance (13% improvement over HMM_M) by providing extra features of NMS. However, NMS are not effectively utilized by $HMM_{M\&N}$. Nevertheless, it is important that the classification errors are mainly between variants of a base sign and out of cluster errors are very few, with 99.5% base sign accuracy (Fig. 4). We further improve the accuracy by sequential-belief based fusion: up to 81.6%. The improvement is mainly based on (1) the possibility of accepting the first stage classification decision thanks to belief formalism and the robustness of the belief-based cluster identification, and (2) the robustness of the method to define the clusters.

5 Conclusions

We have proposed a technique for integrating MS and NMS in an isolated sign recognition system. A dedicated fusion methodology is needed to accommodate the nature of MS and NMS in sign languages. Although NMS can also be used alone, we concentrated on the signs in which MS and NMS used in parallel, to complement or emphasize the meaning of the MS. Our fusion technique makes use of the fact that the manual information gives the main meaning of the sign and the non-manual information complements or modifies the meaning. The sequential fusion method processes the signs accordingly. The key novelties of our fusion approach are two-fold. The first novelty is the two stage decision mechanism which ensures that if the decision at the first step is without hesitation, the decision is made immediately. This would speed up the system, since the system can understand that there is no need for further analysis. Even in the case of a hesitation, the decision of the first step

identifies the cluster which the test sign belongs to, if not the exact sign class. The second novelty is the clustering mechanism: the sign clusters are identified automatically at the training phase and this makes the system flexible for adding new signs to the database by just providing new training data. Our results show that automatic belief based clustering outperforms the feature fusion and increases the accuracy of the classifier.

Acknowledgments. This work is a result of a cooperation supported by SIMILAR 6FP European Network of Excellence (www.similar.cc). This work has also been supported by TUBITAK project 107E021 and Bogazici University project BAP-03S106.

References

1. Ong, S.C.W. and Ranganath, S.: Automatic Sign Language Analysis: A survey and the Future beyond Lexical Meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27, 6, 873-891 (2005)
2. Wu, Y. and Huang, T.S.: Hand modeling, analysis, and recognition for vision based human computer interaction. IEEE Signal Processing Magazine. 21, 51-60 (2001)
3. Vogler C. and Metaxas, D.: Adapting Hidden Markov models for ASL recognition by using three-dimensional computer vision methods. In: IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 156—161 (1997)
4. Vogler C. and Metaxas, D.: Parallel Hidden Markov Models for American Sign Language Recognition. In: International Conference on Computer Vision, Kerkyra, Greece, pp. 116--122, (1999)
5. Vogler C. and Metaxas, D.: Handshapes and Movements: Multiple-Channel American Sign Language Recognition. In: Gesture Workshop, pp. 247-258 (2003)
6. Ming, K.W., Ranganath, S.: Representations for Facial Expressions. In: Proceedings of International Conference on Control Automation, Robotics and Vision, vol. 2, pp. 716-721 (2002)
7. Erdem U.M., S. Sclaroff, S.: Automatic Detection of Relevant Head Gestures in American Sign Language Communication. In: International Conference on Pattern Recognition. vol. 1, pp. 460-463, 2002.
8. Shafer, G.: A Mathematical Theory of Evidence, Princeton University Press (1976)
9. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence. 66, 2, 191-234 (1994)
10. Rabiner, L.R.: A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of IEEE, vol.77, pp.257-285 (1989)
11. Burger , T., Aran, O., and Caplier, A.: Modeling hesitation and conflict: A belief-based approach. In: International Conference of Machine Learning and Applications (ICMLA), pp. 95-100 (2006)
12. eNTERFACE06 ASL Database, http://www.enterface.net/enterface06/docs/results/databases/eNTERFACE06_ASL.zip
13. Aran, O., Ari, I., Benoit, Campr, P., A., Carrillo, A.H., Fanard, F., Akarun, L., Caplier, A. & Sankur, B.: SignTutor: An Interactive System for Sign Language Tutoring. IEEE Multimedia, 4 (2008)
14. Jayaram S., Schmugge, S., Shin, M.C. and Tsap L. V.: Effect of Color space Transformation, the Illuminance Component, and Color Modeling on Skin Detection.

In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04), vol 2, pp. 813-818 (2004)
15. Viola, P. and Jones, J.: Robust real time face detection. International Journal of Computer Vision, 57, 2, 137-154 (2004)
16. Machine Perception Toolbox (MPT), http://mplab.ucsd.edu/grants/project1/free-software/MPTWebSite/API/.
17. Benoit, A. and Caplier, A.: Head Nods Analysis: Interpretation of Non Verbal Communication Gestures. In: IEEE International Conference of Image Processing (2005)