Extraction of Mid-Level Semantics from Gesture Videos using a Bayesian Network

Dimitrios I. Kosmopoulos[a], Ilias Maglogiannis[b]

[a]National Centre for Scientific Research “Demokritos”,
Institute of Informatics and Telecommunications
15310 Aghia Paraskevi, Greece

[b]University of Aegean,
Department of Information and Communication Systems Engineering
83200 Karlovasi, Greece

Abstract. In this paper a method for extraction of mid-level semantics from sign language videos is proposed, by employing high level domain knowledge. The semantics concern labeling of the depicted objects of the head and the right/left hand as well as the occlusion events, which are essential for interpretation and therefore for subsequent higher level semantic indexing. Initially the low-level skin-segement descriptors are extracted after face detection and color modeling. Then the respective labels are assigned to the segments. The occlusions between hands, head and hands and body and hands, can easily confuse extractors and thus lead to wrong interpretation. Therefore, a Bayesian network is employed to bridge in a probabilistic fashion the gap between the high level knowledge about the valid spatiotemporal configurations of the human body and the extractor. The approach is applied here in sign-language videos, but it can be generalized to any other situation where semantically rich information can be derived from gesture.

Keywords: video processing, hand tracking, gesture semantics, Bayesian network, Zernike moments

1 Introduction

The extraction of mid- and high-level semantics from video content is important for tasks as video indexing and retrieval, video summarization and non-linear content organization. This applies also to videos depicting gestures, since they constitute a very useful source of semantic information for multimedia content analysis. The automated extraction of metadata, e.g., according to MPEG-7 or extensions of it, is a prerequisite for the above tasks. However, automated extraction regarding gesture is lagging behind processing of other modalities such as speech. Apart from the variability of spatiotemporal gesture patterns and coarticualtion effects (merging of gestures) that are responsible for this slow progress, occlusions introduce additional complexity in the extraction of gesture metadata. Failure to produce correct metadata as a result of using conventional extractors can lead to wrong semantics. Such metadata may concern association of color regions to the objects in this context, which are the head, the left and right hand or the visual *objects* that result from their mutual occlusions. They may also concern the *appearance* and *disappearance* or *occlusion* events for the *head* and *left*/*right hand* objects.

In this work the occlusion problem is handled through the analysis of temporally structured events by combining the two-dimensional visual features and the high – level knowledge about the human gestures and the related body configurations. A Bayesian network is employed for probabilistic modeling of this knowledge. Inferencing over this network serves the purpose of bridging the semantic gap in a top-down fashion.

The rest of the paper is organized as follows: in the next section the related work concerning semantics extraction from gesture videos is briefly discussed as well as the contribution of the proposed approach; in section 3 the skin segmentation approach is described; in section 4 the structure of the proposed semantic model through a Bayesian network is presented; in section 5 the experimental results are presented; finally section 6 summarizes the results and suggests future directions.

2 Related work

The extraction of semantics from gesture videos has attracted the interest of many researchers in the past. A big portion of them concerns sign language and gestures for human-computer interaction. Gesture recognition methods can be used for extraction of high-level semantics, with the Hidden Markov Models (HMMs) being the most remarkable (e.g., the works of Hyeon-Kyu and Kim (1999), Wilson and Bobick (1997)). The HMMs are probabilistic networks modeling processes. Tey have hidden and observable states. Usually each gesture is modeled by a single HMM. Another approach is based on neural networks, e.g., the work of Ming-Hsuan and Ahuja (1998), where time-delay neural networks are used; they are trained and tested using as input motion trajectories after multi-scale motion segmentation. The dynamic time warping approach (e.g., by Darel and Pentland (1993) and Yu et al (1998)) seeks to compare and fit feature trajectories to model trajectories considering variable motion duration. The motion history images were also used for this purpose (e.g., the works of Bobick and Davis (1996) and Kumar et al (2004)).

They create patterns using 2D images formed by the accumulation of the motion at every pixel during a time window. The comparative features of those approaches are extensively discussed in the surveys provided by Pavlovic et al (1997) and Wu and Huang (1999).

Trajectory-based techniques were presented recently by Rao et al (2002). They exploit the rapid changes in the hand trajectory to create view - invariant motion descriptors. One of the advantages of this approach compared with the ones above is the ability to learn new gestures. However it does not consider at all the hand shape during motion, which is critical especially for sign language and human-computer interaction.

Although the above techniques may provide significant results for gesture recognition tasks, they require accurate hand and provisionally head segmentation. Occlusions are not handled at all or become resolved through stereoscopic camera configurations, e.g., in the work of Vogler and Metaxas (2001), which is a rare exception and not the rule in the available content. Another approach regarding occlusion handling is the employment of a 3D hand model, as shown for example by Rehg and Kanade (1993), however optimization of models for articulated objects with so many degrees of freedom, such as hands, is a very challenging task. The fact that the exact hand size is unknown, poses an additional difficulty in solving the inverse kinematics problem as stated by Lien and Huang (1998).

In works similar to that of Tanibata et al (2002) the optical flow constraint is used, however it is assumed that the sampling rate is constantly high and that the movement is smooth in terms of shape and position.

All the above approaches use low-level features to infer higher-level semantics but they do not address the inverse information flow. Reasoning about low- and mid-level features using high-level knowledge (thus enabling a closed-loop semantic extraction process) would be a major step for bridging the semantic gap and could be complementary to the aforementioned methods. An approach that has motivated the present work is given in by Gong et al (2002), who employ a Bayesian network to model known spatio-temporal configurations of the human body. However, in order to extract semantic metadata concerning the occlusion types, all of them have to be modeled explicitly, an issue which is not addressed in that work. Furthermore, the shape representation concerns only size, aspect ratio and orientation, which is not adequate in the context of sign language videos, where a large variety of hand shapes is observed.

The aim and the contribution of this work is the mid-level gesture metadata extraction using high level knowledge. It extends the work presented by Gong et al (2002) mainly in three ways: (a) A different probabilistic network structure is applied that allows explicit modeling of all possible states (occlusions) as well as their types, while not assuming known head position. (b) The modeled temporal relations are minimal in order to cope with motion discontinuities. (c) The observation variables provide a detailed skin region representation, including internal structure, based on Zernike moments. In the work presented by Gong et al (2002) the regions are represented only by the area the aspect ratio and orientation, which are insufficient for modeling the complex hand shapes appearing in sign language gestures.

3 Skin segmentation

The tracking of skin regions in the image presupposes skin detection in the image. Skin color modeling is a non-trivial problem, which includes selection of the color space, of the degree of accuracy concerning the color distribution and of the post-processing. In this work we handle all these problems not in the general case but under the assumption that the target faces the camera. This reduced problem is approached here by firstly locating the human face in the image and then using the corresponding region to train a color model.

3.1 Face detection

For face detection we use the Haar face detector, which is quite efficient. A brief description of the method is presented here due to self-containment of the paper, but the reader has to refer to the work Viola and Jones (2001) for more details.

Initially some features are defined as the difference of the sum of the pixel values in equally sized adjacent rectangular regions; the regions may be of various numbers, sizes and configurations, thus forming simple or more complex features. The features can be computed at any location or scale in constant time using the idea of summed area tables, known from the field of computer graphics. The calculated features are fed into a sequence of classifiers, which is used with the purpose of eliminating the largest number of negative inputs with little processing at the early stages (only positive results are further examined). The classifiers in the later stages are more accurate but combine more complex features. The training of each classifier proceeds according to the Adaboost algorithm which also selects the most proper features:

For each feature *j* a classifier h_j using this single feature is trained, which returns 0 or 1 for negative or positive examples respectively. The error is given by:

$$\varepsilon_j = \sum_i w_i \cdot | h_j(x_i) - y_i | \qquad \text{(eq 1)}$$

where x_i is the training sample (i=1..n), w_i the corresponding weight of the sample and y_i the ground truth. Then the classifier with the lowest error ε_j is selected and the weights are updated according to:

$$w_i' = w_i \cdot (\frac{\varepsilon_j}{1-\varepsilon_j})^{1-e_i} \qquad \text{(eq 2)}$$

where e_i is 0 if x_i is classified correctly or 1 otherwise. The procedure is repeated until we select the *T* features that satisfy the performance and accuracy requirements of the classifier. The final classifier is applied to many rectangular regions in the image and gives 1 (true) if

$$\sum_{t=1..T} \log\frac{1-\varepsilon_t}{\varepsilon_t} \cdot h_t(x) \geq \frac{1}{2} \sum_{t=1..T} \log\frac{1-\varepsilon_t}{\varepsilon_t} \qquad \text{(eq 3)}$$

or 0 (false) otherwise. The early stage classifiers seek to minimize false negatives using fewer features due to performance.

The rotated Haar-like features and the post optimization procedures may increase recognition rates as shown by Rainer and Jochen (2002). The sensitivity of the method to big rotations is known and can be enhanced by using methods such as inserting the Haar-like facial features into deformable graphs, but further elaboration in this subject is not within the scope of the presented work.

For the purposes of our research a fraction of the identified face bounding rectangle is used for training the skin-color classifier (training region-of-interest) as presented in the following subsection.

3.2 Color space selection

The color space for skin detection has been selected based on the criteria of calculation simplicity, low correlation between channels and skin color representation capability. A wide variety of color spaces has been presented in the past. Among them the most popular are the RGB, the HSV and the YCrCb. The RGB is one of the most widely used spaces, however the high correlation between channels and the mixing of chrominance and luminance make it improper for skin modeling as mentioned by Vezhnevets et all (2003). The HSV color space is more appealing, since it is closer to the human perception but it has its drawbacks too: the cyclic nature of hue-saturation makes it inconvenient for parametric color models that represent tight color clusters and the transformation from RGB is rather complex. The YCrCb color space is widely used for color compression and provides explicit separation of luminance and chrominance component. Furthermore, its simple calculation makes it appealing for color modeling and it has been applied successfully in many skin segmentation tasks, e.g., in Hsu et al (2002) and Phung et al (2002). The YCrCb model has been selected and has given higher robustness to noise than the other two color spaces, although the results for HSV and RGB were also acceptable.

3.3 Skin model

After the detection of the target's face, we use a part of the face region for skin color modeling. For solving this target-specific and thus reduced skin color modeling problem, we assume a single multivariate Gaussian model. The color probability density function for a pixel value x_i in the selected color space for each individual channel is given by:

$$\varphi_s(x_i;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \qquad \text{(eq 4)}$$

where μ, σ are the mean and standard deviation of the distribution. After calculating the individual channel distributions we will be able to calculate the rest covariance matrix coefficients σ_{YCr} , σ_{YCb}, , σ_{CrCb} . Then the skin regions are determined by selecting the pixels providing highest probability values based on the equation:

$$p(c \mid skin) = \frac{1}{\sqrt{(2\pi)^3 \mid \mathbf{\Sigma} \mid}} e^{-\frac{1}{2}(\mathbf{c}-\boldsymbol{\mu})^{\mathrm{T}} \mathbf{\Sigma}^{-1}(\mathbf{c}-\boldsymbol{\mu})} \qquad \text{(eq 5)}$$

where **c** refers to the color vector and **μ**, **Σ** are the mean vector and the covariance matrix respectively.

In order to cope with the outliers in the training region of interest that are due to noise in feature measurements, which can corrupt the skin model leading the system to wrong segmentation, we employ robust statistics, which "is proper for recovering the structure that best fits the majority of the data while identifying and rejecting outliers or deviating substructures" (Hampel et al (1986)). The noise in the training region of interest is due to sensor noise, color of eyes, hair, or glasses. Using the color model extracted during initialization phase, we are able to segment the image into skin and non-skin regions resulting in a binary mask. Robust estimation in our case deals with the problem of fitting the value $\boldsymbol{\lambda} = [\mu, \sigma]^T$, in the presence of outliers, for the single channel model:

The estimation of the model parameters using robust statistics is performed by minimizing the quantity:

$$\mathrm{E} = \sum_{i=1..n} \rho(y(x_i) - \phi(x_i; \mu, \sigma), \omega) \qquad \text{(eq 6)}$$

where ρ is a robust estimation function and ω is a scale parameter of this function. We have used the robust estimator of Geman-McClure (1987) due to the gradual attenuation of the outliers that it provides. It is given by the equation:

$$\rho(s, \omega) = \frac{s^2}{s^2 + \omega} \qquad \text{(eq 7)}$$

Small values for scale parameter ω make the system more tolerant to outliers, since their effect is attenuated.

It can be proved (Memin and Perez, (1998)) that under concavity of $q(x) \equiv \rho(\sqrt{x})$, which is here the case, any multidimensional minimization problem of the form:

$$\arg\min_{\lambda} \sum_{i=1..n} \rho(g(x_i, \boldsymbol{\lambda})) \qquad \text{(eq 8)}$$

can be turned into a dual minimization problem:

$$\arg\min_{\boldsymbol{\lambda}, z_i} \sum_{i=1..k} [m z_i g(x_i, \boldsymbol{\lambda})^2 + \psi(z_i)] \qquad \text{(eq 9)}$$

involving weights z_i, which lie in (0,1]. Ψ is a continuous differentiable function and

$$m = \lim_{x->0^+} q'(x) \qquad \text{(eq 10)}$$

In our case $m=\sigma^{-1}$, $g(x_i, \boldsymbol{\lambda}) = y(x_i) - \phi(x_i; \mu, \sigma)$ and the dual minimization problem is a typical least square problem. It is solved in an iterative fashion, since the weights z_i are unknown. Initially the values for μ, σ are calculated as if there are no outliers, to give a first estimate. The initial value for all the weights z_i is 1. The non-linear least squares problem is solved as follows:

The deviation de_i from the model for a single pixel *(c)* is given by de_i=y_i-$\varphi(x_i;\mu,\sigma)$, which can be locally linearized to:

$$de_i = \frac{\partial \phi}{\partial \mu} d\mu \bigg|_{x_i} + \frac{\partial \phi}{\partial \sigma} d\sigma \bigg|_{x_i} \qquad \text{(eq 11)}$$

For *n* training pixels this relation is expressed as:

$$\begin{bmatrix} y(x_1) - \phi(x_1;\mu,\sigma) \\ \dots \\ y(x_n) - \phi(x_n;\mu,\sigma) \end{bmatrix} = \begin{bmatrix} \frac{\partial \phi}{\partial \mu}\Big|_{x_1} & \frac{\partial \phi}{\partial \sigma}\Big|_{x_1} \\ \dots & \dots \\ \frac{\partial \phi}{\partial \mu}\Big|_{x_n} & \frac{\partial \phi}{\partial \sigma}\Big|_{x_n} \end{bmatrix} \cdot \begin{bmatrix} d\mu \\ d\sigma \end{bmatrix} \qquad \text{(eq 12)}$$

Which can be written equivalently :

$$d\mathbf{e}=\mathbf{A}\cdot d\boldsymbol{\lambda} \Leftrightarrow \mathbf{A}^{T}d\mathbf{e}=(\mathbf{A}^{T}\mathbf{A})\, d\boldsymbol{\lambda} \Leftrightarrow d\boldsymbol{\lambda}=(\mathbf{A}^{T}\mathbf{A})^{-1}\mathbf{A}d\mathbf{e} \qquad \text{(eq 13)}$$

The calculated offset for μ, σ is then added to these values.

After solving the least squares problem we keep the μ,σ constant and we calculate the weights z_i as shown by Memin and Perez (1998), using a simple closed form given by the equation:

$$z_i = \frac{\rho'(g(x_i;\mu,\sigma))}{2mg(x_i;\mu,\sigma)} = \frac{\omega^2}{((y(x_i) - \phi(x_i;\mu,\sigma))^2 + \omega)^2} \qquad \text{(eq 14)}$$

The same calculations are repeated at this order until the calculated values for dμ, dς are smaller than a small positive value ε.

3.4 Post-processing

The final post-processing step includes noise elimination. The possible noise of the mask is removed by applying a morphological opening filter. The resulting regions are then processed with connected components, keeping a maximum of three regions (the bigger ones). The final mask is applied to the Y image channel to obtain a masked gray-level image including only these regions.

We assume here that the depicted person (target) is dressed and faces the camera with his/her upper body part captured in the image. The face rotation must not exceed thirty degrees at the initialization step. The hands may disappear from the image, they may be occluded by each other and they may occlude the head.

4 Semantic model

In this section we describe the semantic model used for the extraction of mid-level semantics from gestures. We present the employed Bayesian network and we describe how the skin regions are represented in the network using the complex Zernike moments.

4.1 Network structure

Using a fixed set of rules for this purpose creates many problems, because the results are dependent on fixed thresholds and can be easily corrupted by noise. Furthermore, rule-based approaches suffer from consistency problems because commitment to a single decision precludes feedback of higher level knowledge to refine lower-level uncertain observations as mentioned by Gong et al (2002).

In this work a Bayesian network is used as a semantic model due to its ability to capture uncertainty of the domain knowledge. The goal is the following:

> *Provided a set of variables* $\mathbf{x}$ *(represented as nodes in the network), we seek to find the instance vector* $\mathbf{x}_m$ *of those variables that maximizes their joint distribution, given some evidence* $\mathbf{e}$ *coming from image measurements (associated with some other network nodes). In other words:*

$$\mathbf{x}_m = \{\mathbf{x}_0 : \forall\, \mathbf{x},\ P(\mathbf{x}_0|\mathbf{e}) > P(\mathbf{x} \mid \mathbf{e})\} \qquad \text{(eq 15)}$$

The value of $\mathbf{x}_m$ provides the current gesture state, i.e., position of head and hands in the image as well as current occlusions.

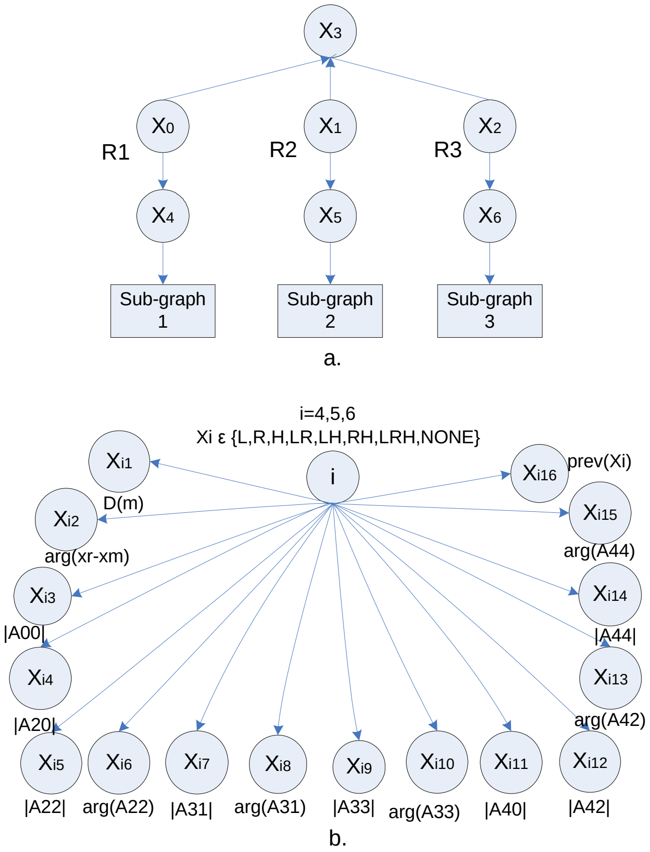


Fig. 1. (a) Bayesian belief network representing the semantic model and (b) the subgraphs.

The network that encodes the high level semantics of the gesture metadata extraction task is presented in Fig.1(a-b). The network variables and their dependencies are explained in the following:

- $X_{0,}$ $X_{1,}$ X_2: Variables corresponding to (maximum) three skin regions expressing the probability that one of them corresponds to left hand, right hand, head, mutual occlusion of hands, head – left hand occlusion, head-right hand occlusion, occlusion of head by both hands or noise. The corresponding values belong to the set **A**={*L, R, H, LR, LH, RH, LRH, N*}.

- X_3*:* Binary variable semantically related to $X_{0,}$ $X_{1,}$ X_2. It is used for excluding the non acceptable associations of regions to visual objects. It is only true when the following are the values for $X_{0,}$ $X_{1,}$ X_2: (*L*,*R*,*H*), (*L*,*H*,*N*), (*L*,*RH*,*N*), (*R*,*H*,*N*), (*R*,*LH*,*N*), (*H*,*N*,*N*), (*H*,*LR*,*N*), (*LH*,*N*,*N*), (*RH*,*N*,*N*), (*LRH*,*N*,*N*), at any possible order.

- X_4, X_5, X_6: Auxiliary variables used to decouple X_0, X_1, X_2, X_3 from the sub-graphs presented in detail in Fig.1b. Thus the network has a modular structure, i.e., during training and execution sub-graphs 1-3 are treated independently using the same learning and inference engine during training and execution. The factor matrices give values that coincide with the values of their parent nodes (X_0, X_1, X_2).

The aforementioned sub-graphs are identical, each of them is associated with a skin region and they are used to capture the semantics of visual evidence. Their root nodes are the X_i (i=7,8,9).

The child nodes in the sub-graphs are used to provide evidence for inferring the system state. More specifically:

- X_{i1}, X_{i2}: The distance and angle of the current region center of gravity from the common center of gravity of all regions.
- X_{i16}: The label of the region in the previous time instance. Models the motion continuity.
- X_{i3}, X_{i4}, X_{i5}, X_{i7}, X_{i9}, X_{i11}, X_{i12}, X_{i14}*:* They correspond to the Euclidean norm of the Zernike moments up to 4th order of the *i*-region around its center of gravity ($|A_{11}|=0$).
- X_{i6}, X_{i8}, X_{i10}, X_{i13}, X_{i15}*:* Correspond to the non-zero orientations of the above Zernike moments.

Some of the relations between the above semantic variables can be intuitively understood. The X_3 is true only if the extracted configurations are acceptable, e.g., identification of two right hands or

The complex Zernike moments have been selected here for shape representation due to their noise resiliency, the reduced information redundancy (orthogonality) and their reconstruction capability. Restating the goal expressed at the beginning of this section:

> *We aim to find the value* $\mathbf{x}_m$ *of the vector* ($X_{0,}$ $X_{1,}$ X_2) *that maximizes the joint distribution* P(**x**|**e**) *where* **e** *is the evidence provided by the measured variables and* **e={** X_3=1 U $X_{i,j}$: iє{4,5,6}, jє{1,2,3,..,15} }.

This vector provides the association of skin color regions in the image space to context objects, in the “world” space. The appearance or disappearance of those objects (whenever the vector elements change) signifies the visual events that are extracted in the form of metadata. For inference of the current state, Junction tree inference is employed, due to the absence of cyclic sub-graphs.

4.2 Skin area representation by Zernike moments

The final mask is applied to the Y channel of the image to obtain a masked gray-level image including the head and the two hands. The gray-level image provides a richer representation of the current gesture than the binary masks; the employment of the latter can lead to loss of information especially in the case of different gesture contours with similar silhouettes, e.g., head occlusion by the hand. We use the Zernike moments to represent the activity state as it is expressed by the relative position and shape of the head and the two hands. The Zernike moments have some very positive attributes that make them proper for our representation purposes and namely their noise resiliency, the reduced information redundancy and their reconstruction capability.

The complex Zernike moments of order *p* are defined as (see for example Mukundan and Ramarishnan (1998)):

$$\mathrm{A}_{pq} = \frac{p+1}{\pi} \int_{0}^{1} \int_{-\pi}^{\pi} R_{pq}(r) \cdot e^{-jq\theta} f(r,\theta) \cdot r \cdot dr d\theta \qquad \text{(eq 16)}$$

$$r = \sqrt{x^2 + y^2}\ ,\ \theta = \tan^{-1}(y/x),\ -1<x,y<1 \qquad \text{(eq 17)}$$

$$R_{pq}(r) = \sum_{s=0}^{\frac{p-q}{2}} (-1)^s \frac{(p-s)!}{s!(\frac{p+q}{2}-s)!(\frac{p-q}{2}-s)!} r^{p-2s} \qquad \text{(eq 18)}$$

p-*q* = even and 0≤*q*≤*p*.

The higher the order of moments that we employ, the more detailed will be the region reconstruction but also the more processing power will be required. Limiting the order of moments used is also justified by the fact that the details captured by higher order moments have much higher variability from person to person and are more sensitive to noise.

5 Experimental results

The system has been trained to associate regions to hands and head and to recognize occlusions, using a vocabulary of fifteen sign-language words. Typical examples of the tested vocabulary are displayed in Fig. 2, where fifteen signs are presented. They contain mutual occlusion of the hands, occlusions of the head by the left and right hand, mutual occlusions of the hands and occlusions of the head by both hands. Since we were not able to find a public database to satisfy our requirements in color, frame rate, occlusions, spatiotemporal variations, viewpoint and amount of content we have created our own database. The system has been trained for approximately 50 segmented custom made videos for each sign, trying to capture a variety of spatiotemporal variations. Each video lasts two to five seconds at a rate of 15 frames per second and each frame contains one to three regions that can be used for training. The types of the identified skin regions were manually identified and stored through a dedicated user interface.

As mentioned in previous section, to measure evidence we first locate the target's face in the image (Haar face detection), using a vision library (OpenCV), and we use a fraction of the face region for probabilistic skin color modeling. The resulting regions are then processed with connected components, keeping a maximum of three regions (the bigger ones). The final mask is applied to the Y image channel to obtain a masked gray-level image including only these regions. The image segmentation has always given reasonable results, provided that the face detector has been successful. Some typical results for the "add" gesture are displayed in figure 3.

The variation of the illumination conditions may affect the color segmentation because of the shifting that they cause to the curves modeling the color channels (mainly the intensity channel). When we omitted the intensity channel from the color model we experienced higher robustness in big illumination changes, however, the segmentation results appeared to be less accurate and therefore in the presented experiments we assumed constant illumination. Other options would be a periodic or event-based update of the model at a higher computational cost.

After skin segmentation the Zernike moment norms are normalized with regard to the initial face area to decrease influence of non-uniform body dimensions or distance from camera. For simplicity all continuous variables are discretized (angle variables at 8 values, norms and lengths at 10 values). The state space of the variable set with this discretization is 1.237×10^{54}. However, using the proposed network structure we have only to populate through training probability tables containing only 1540 probabilities. They concern the sub-graph presented in figure 1, while the rest tables are easily defined manually through considering the acceptable configurations mentioned in sub-section 4.1.

The confusion matrix for 2430 test frames including one to three skin segments (for all signs) is displayed in Table 1. It is obvious that the most difficult semantics to recognize regard the occlusion type. This is particularly difficult in cases where the hands occlude the head without altering its contour, when the respective regions are falsely recognized as "head". Less frequent (on a percentage basis) is the error that regards assignment of the labels "left" or "right" to regions corresponding to "left-right", which is reasonable especially when we have almost total occlusion of one hand by the other. The system is more accurate when assigning labels to left and right hands (true positive rate over 93 percent). The system runs in real time for each frame, while it requires about two seconds for initialization on a Pentium 2.5 GHz PC (face detection and color modeling).

The approach has been tested when the spatial resolution changes. The results for the half resolution are given in Table 2. The basic success rates do not change dramatically, which is reasonable considering the scale-invariant measurement variables. However, the reduced resolution preserves less area details and some results are consequently less accurate.

The results for reduced time resolution are given in Table 3, after measuring using every second video frame (7.5 fps instead of 15). Although some accuracy is lost, the system is still able to label successfully the regions. This is expected, since the time constraint posed (by variable X_{i16}) is probabilistic and not a hard rule-based constraint, allowing some deviations. Preserving high accuracy even when the movement is fast appears to be discontinuous, is a significant advantage compared to algorithms based on optical flow.

Although we have tried to include a variety of spatiotemporal variations in our training procedure it is still possible that in real world applications the skin segments may not appear as expected due to occlusion by clothes or incorrect segmentation. It can also be that some personal traits or increased emotionality can lead to different spatiotemporal patterns during gesturing. Currently, when the system finds a measurement vector that was not met during training then it marks the respective segment as noise. In the future we plan to handle the issue by modeling each measurement variable as continuously distributed, thus allowing higher tolerance in the observed values.

Actual		Recognised as							
		L	**R**	**H**	**LR**	**LH**	**RH**	**LRH**	**NONE**
	L	1224	33	9	0	0	0	0	45
	R	71	1168	0	0	0	0	0	3
	H	9	0	2415	0	0	0	0	36
	LR	3	0	33	414	0	0	0	45
	LH	3	0	6	0	105	0	0	12
	RH	0	0	22	0	0	134	0	9
	LRH	0	0	9	0	0	0	55	2
	NONE	1	7	4	0	0	0	0	1425

Table 1: Classification results for 2460 frames (full resolution)

Actual		Recognised as							
		L	**R**	**H**	**LR**	**LH**	**RH**	**LRH**	**NONE**
	L	1190	72	42	0	0	0	0	7
	R	96	1096	20	0	0	0	0	30
	H	22	63	2321	21	0	0	0	33
	LR	23	19	42	378	0	0	0	60
	LH	0	0	13	0	101	0	0	12
	RH	6	0	36	0	0	117	0	6
	LRH	0	0	9	0	0	0	47	10
	NONE	6	7	6	3	0	0	0	1415

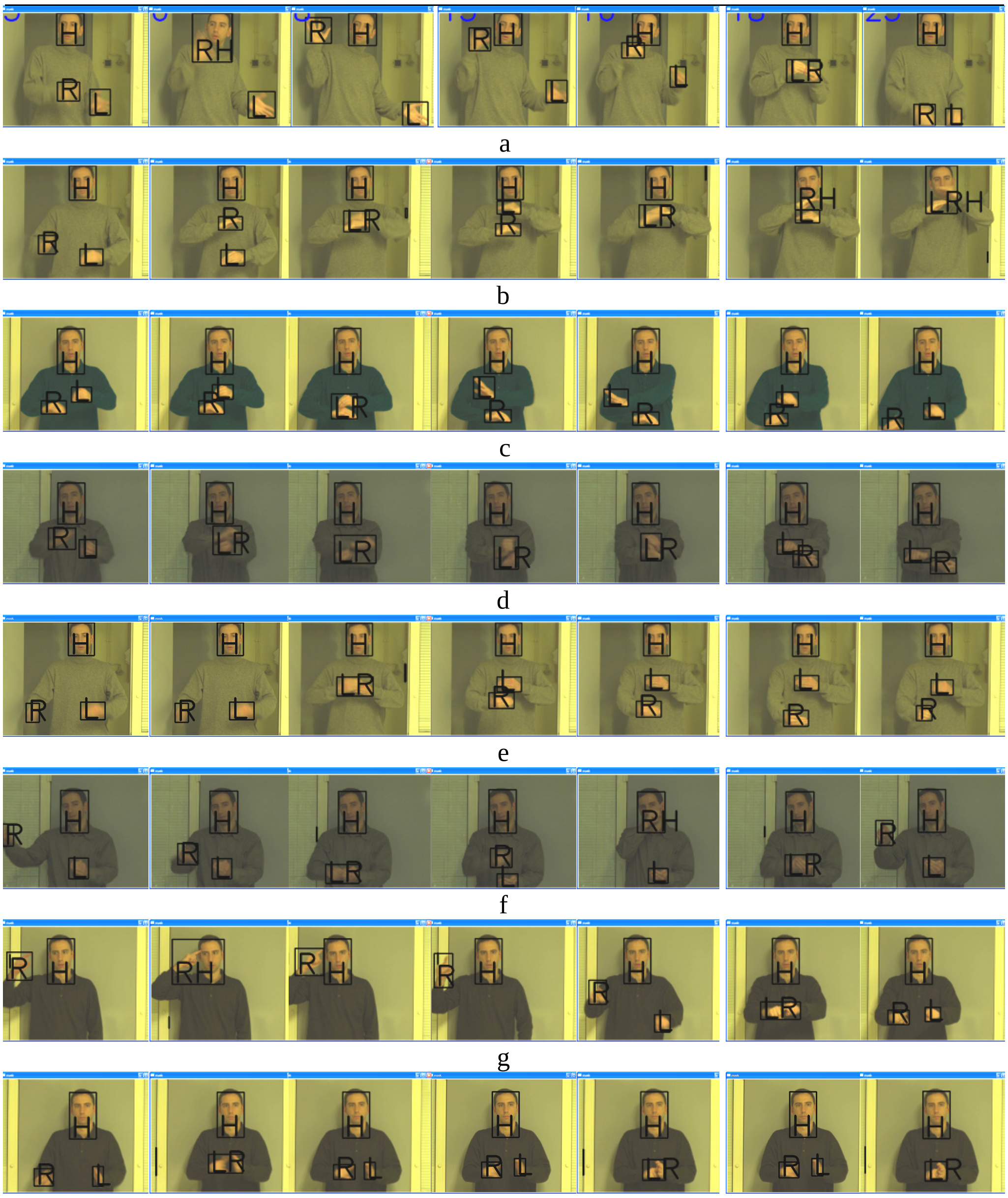
Table 2: Classification results for 2460 frames (half resolution)

		Recognised as							
		L	**R**	**H**	**LR**	**LH**	**RH**	**LRH**	**NONE**
Actual	**L**	633	11	10	0	0	0	0	3
	R	20	601	7	0	0	0	0	0
	H	6	0	1206	0	0	0	0	18
	LR	12	9	11	180	0	0	0	33
	LH	3	1	5	0	49	0	0	15
	RH	0	0	15	0	0	45	0	0
	LRH	0	0	6	0	0	0	27	0
	NONE	24	11	20	0	0	0	0	615

Table 3: Classification results for time resolution reduced to half (1230 frames)

Regarding comparison to other existing systems, it has been already reported that probabilistic approaches like the one presented in this work have been proved superior compared to particle filtering algorithms and namely the Condensation as regards tolerance to occlusions and performance, provided that proper training has been previously performed (see for example Gong et al (2002)). This applies also to the Kalman filter.

We have tried to compare our work with the one of Gong et al (2002) by using the network structure proposed there and a similar training procedure with our system, although we know that a direct comparison is not possible due to many implementation details that may affect the provided results. We were able to achieve approximately 3% less classification errors for full and half spatial resolution and 25% less classification errors in half time resolution test. This could be explained by the fact that only one network variable (per skin region) models the temporal relations (X_{i16}) in comparison to 12 such variables used by Gong et al (2002), which require better training to model motion discontinuities.



a

b

c

d

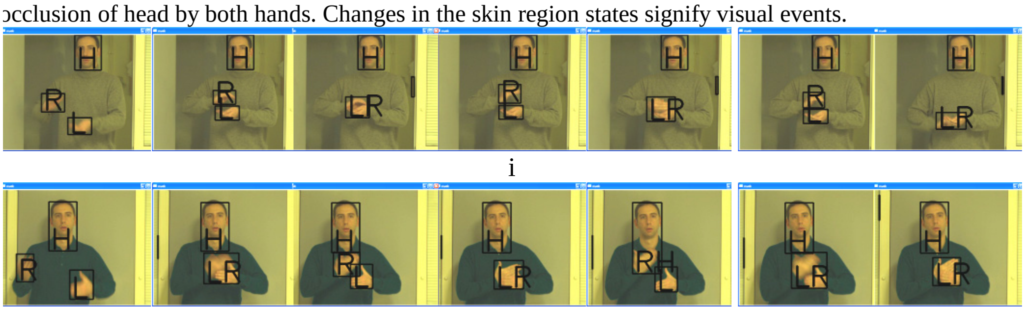
e

f

g

h

Fig.2 (a)-(h) The extraction of mid-level semantics from skin regions in sign language videos for the words: (a) add (b) achieve (c) advice (d) change (e) after (f) correct (g) agree (h) alarm. The identified skin regions are enclosed in rectangles, annotated with the region type L, R, H for left hand, right hand and head, LR, LH, RH for mutual occlusion of hands, head occlusion by left and right hand, LRH occlusion of head by both hands. Changes in the skin region states signify visual events.



i

j

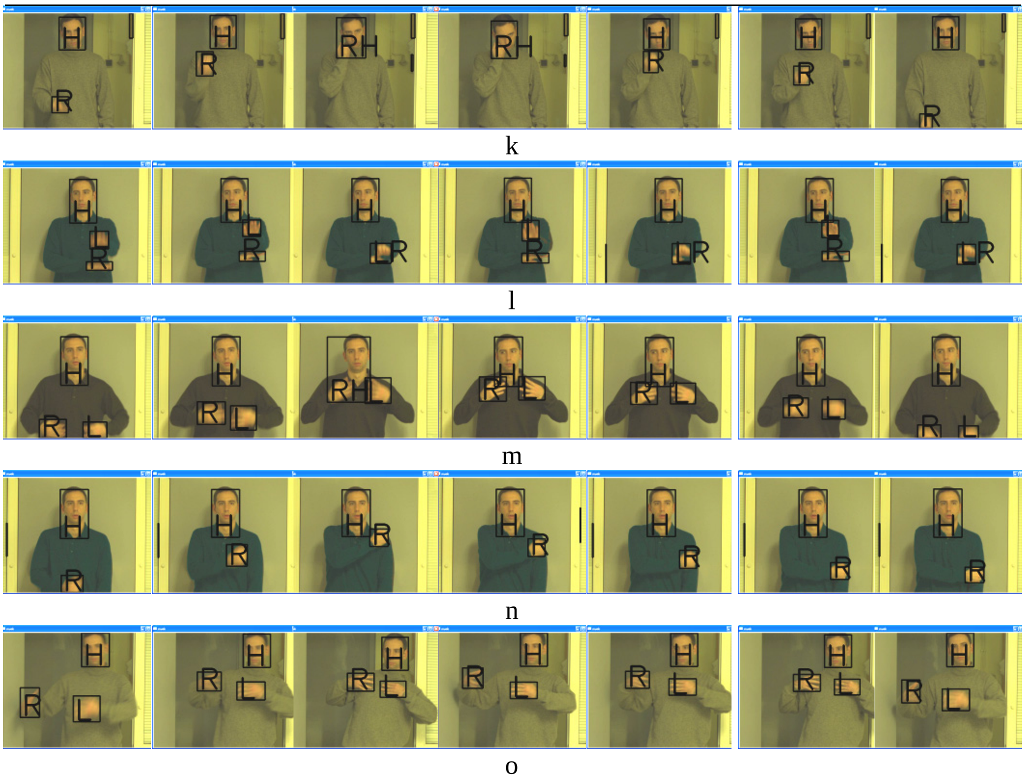


Fig.2 (i)-(o) The extraction of mid-level semantics from skin regions in sign language videos for the words: (i) amount (j) annoy (k) ashamed (l) afternoon (m) anger (n) anger, (o) assume

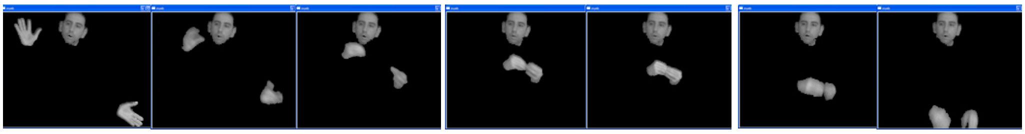


Fig.3 Typical skin segmentation results for gesture “add”

6 Conclusions

A method for extraction of gesture-related mid-level semantics from video based on a Bayesian network has been presented. It has been demonstrated that the method is able to overcome difficulties posed by occlusion events, which is difficult to achieve with conventional methods. Furthermore, it is not affected by image discontinuities as the optical-flow based methods.

Higher order Zernike moments can be easily included for more detailed area representation, as well as more temporal variables too, but the application context has to be considered in that case for a performance-effectiveness trade-off. Tests with bigger vocabularies will verify the scalability of the approach.

Future work includes the integration of the network in a closed-loop gesture recognition scheme, e.g., in combination to a HMM, for more focused low level feature extraction. Furthermore, within the scope of research is the development of invariant measures, to minimize the effect of different camera viewpoints as well as variations in human body. These measures will be integrated as evidence nodes in the network. A preparation of a release version of the gesture database for public use is also among our future goals.

References

Bobick, A.F., Davis, J.W., 1996, “Real time recognition of activity using temporal templates”, *International Conference on Automatic Face and Gesture Recognition*, VT, USA.

Darrell, T., Pentland, A., 1993, “Space-time gestures”, *IEEE Conference on Computer Vision and Pattern Recognition*, 335 – 340

Geman, S., McClure, D. E., 1987, “Statistical methods for tomographic image reconstruction”, *Bull. Int. Stat. Inst.*, LII-4, 5-21

Gong, S., Ng, J., Sherrah, J., 2002, “On the semantics of visual behavior, structured events and trajectories of human action”, *Image and Vision Computing*, 20, 873-888

Hampel, F. R., Ronchetti, E. M., Rousseau, P.J., Stahel, W. A., 1986, “Robust Statistics: The Approach Based on Influence Functions”, Wiley, New York

Hyeon-Kyu Lee, Kim, J.H., 1999, “An HMM-based threshold model approach for gesture recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10), 961 – 973

Isard, M., Blake, A., 1998, “CONDENSATION conditional density propagation for visual tracking”, *International Journal of Computer Vision,* 29 (1), 5-28

Jensen, F. V., 2001, “Bayesian Networks and Decision Graphs”, Springer Verlag, New York

Kumar, S.; Kumar, D.K.; Sharma, A.; McLachlan, N., 2004, “Classification of visual hand movements using multiresolution wavelet images”, *Proceedings of International Conference on Intelligent Sensing and Information Processing*, 373 – 378

Lien, CC., Huang, CL, 1998, :Model-based articualed hand motion tracking for gesture recognition, *Image and Vision Computing*, 16, 121-134

Memin, E., Perez, P., 1998, "Dense Estimation and Object-based Segmentation of the Optical Flow with Robust Techniques", *IEEE Transactions on Image Processing*, 7(5), 703-719

Ming-Hsuan Yang; Ahuja, N., 1998, “Extraction and classification of visual motion patterns for hand gesture recognition”, *IEEE Conference on Computer Vision and Pattern Recognition*, 892 – 897

Mukundan, R., Ramakrishnan, K. R., 1998, "Moment Functions in Image Analysis: Theory and Applications", World Scientific, Singapore

Pavlovic, V., Sharma, R., Huang, T. S., 1997, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: a Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 677 – 695

Phung, S. L., Bouzerdoum, A., Chai, D., 2002, "A novel skin color model in YCrCb color space and its application to human face detection", *IEEE International Conference on Image Processing*, Rochester, New York, USA, 1, 289-292

Rainer L. and Jochen M., 2002, "An Extended Set of Haar-like Features for Rapid Object Detection", *IEEE International Conference on Image Processing*, New York, USA, 1, 900-903

Rao, C., Yilmaz, A., Shah, M., 2002, View-Invariant representation and recognition of actions, *International Journal of Computer Vision*, 50(2), 203-226

Rehg, J., Kanade, T., 1993, "DigitEyes: Vision-Based Human Hand Tracking", Tech. Rep. CMU-CS -93-220, School of Comp. Science, Carnegie Mellon University, Pittsburgh

Tanibata, N., Shimada, N., Shirai, Y., 2002, "Extraction of Hand Features for Recognition of Sign Lan-guage Words", *International Conference on Vision Interface*, 391-398

The OpenCV Computer Vision Library http://www.intel.com/research/mrl/research/opencv/

Vezhnevets V., Sazonov V., Andreeva A., 2003, "A Survey on Pixel-Based Skin Color Detection Techniques", *Proc. Graphicon-2003,* Moscow, Russia, pp. 85-92

Viola P., Jones M. J., 2001, "Rapid Object Detection using a Boosted Cascade of Simple Features", *International Conference on Computer Vision and Pattern Recognition*, I511-I518, 2001

Vogler, C., Metaxas, D., 2001, "A Framework for Recognizing the Simultaneous Aspects of American Sign Language", *Computer Vision and Image Understanding*, 81, 358-384

Wilson, A. D., Bobick, A., 1997, "Parametric Hidden Markov Models Approach for Gesture Recogni-tion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12) 1325 – 1337

Wu, Y., Huang, T., 1999, "Vision-based gesture recognition: a Review", Gesture-Based Communication in Human-Computer Interaction (A. Braffort, R. Gherbi, S. Gibet, J. Richardson, D. Teil, Eds.), 1739, *Lecture Notes in Artificial Intelligence*, 103-115

Yu Huang, Yuanxin Zhu, Guangyou Xu, Hui Zhang, 1998, “Spatial-temporal features by image registration and warping for dynamic gesture recognition”, *IEEE International Conference on Systems, Man, and Cybernetics*, 5, 4498 - 4503