International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 4 (2012) pp. 589-597

Interpretation of Meaningful Expressions by Integrating Gesture and Posture Modalities

Omer Rashid[1], **Ayoub Al-Hamadi**[1] **and Klaus Dietmayer**[2]

[1]Institute of Electronics, Signal Processing and Communications (IESK),
Otto von Guericke University Magdeburg, Germany
{*omer.ahmad,ayoub.al-hamadi*}*@ovgu.de*

[2]Institute of Measurement, Control, and Microtechnology,
University of Ulm, Germany
klaus.dietmayer@uni-ulm.de

Abstract: **Integration of different modalities aims to increase the robustness of a system and improves its performance in an unconstrained environment. With this motivation, this research is focused on integrating two modalities such as hand gesture and posture in a unified framework where the criterion is measured by employing Particle Filter system. The proposed framework has two main modules: 1) Gesture and posture recognition system, and 2) Particle filter based integration of these systems by incorporating CFG rules. In first module, the gesture and posture features are extracted and classified by HMM and SVM respectively. In the second module, the integration is carried by mapping the classification outcome on Particle filter system which acts as contribution-weight at decision level. Moreover, to infer and extract the "meaningful expressions" from the input sequence based on their contribution-weights, we have exploited regular grammar and developed the production rules using context free grammar (CFG) for our test scenario (i.e. a restaurant). Experiments are conducted on 500 different combinations of restaurant orders with the overall 98.3% inference accuracy whereas the classification accuracy of gesture and posture recognition is approximately 98.6% which proves the significance of proposed approach.**

Keywords: Posture Recognition, Gesture Recognition, Integration, Context Free Grammar, Particle Filter

I. Introduction

A lot of research has been done to provide the naturalness and intuitiveness while interacting with the computers. This naturalness and intuitiveness is achieved through HCI which bridges the gap between humans and computers by exploiting different modalities. These modalities are divided according to their input in three main categories namely visual, audio and sensor-based modality. First, the visual modality deals with solving the issues using human responses while taking the observations from a visual input device (i.e. camera). The main research applications in this modality includes gaze detection, face and facial expression, gesture and posture recognition. Second, the audio-based modality takes the audio signals as the input observations. Speech and speaker recognition, musical and emotion recognition are the application areas in this modality. The third type deals with the sensor-based modality and includes the pen-based interaction, mouse, keyboard and pressure sensors etc. However, the selection of these modalities is a crucial choice and it should provide the desirable level of applicability in terms of naturalness and intuitiveness for HCI systems.

In our proposed approach, first, the image acquisition is carried out by Bumblebee2 camera and objects of interest (i.e. hands and face) are extracted using color and depth information. Second, gesture and posture feature vectors are computed by exploiting different properties of hand. Further, fingertip categorization is performed to group the American Sign Language (ASL) signs. Thereafter, HMM is employed to recognize the gesture symbols from alphabets and numbers whereas SVM is used for finger-spelling ASL signs in posture recognition. For the integration of gesture and posture modalities, a particle filter system is proposed to define the integration-criteria (i.e. by computing the contribution-weights). Afterwards, the interpretation is performed by processing the Context Free Grammar (CFG) production rules which results in the inference of meaningful expression. The proposed framework is presented in Fig. 1.

The contributions of the paper are stated as under: Firstly, the extraction of invariant feature vectors in gesture and posture frameworks which results in robust recognition with lower number of training samples. Secondly, the categorization of ASL posture signs is computed by curvature analysis which results in better recognition rate. Thirdly, the integration of gesture and posture recognition system for which we have designed an effective interaction-interface for HCI where the integration criteria is computed by particle filter system at decision level. These contribution-weights sets a criteria for the combination of extracted symbols which are then mapped on CFG rules and results in the interpretation of new meaningful expressions based on the developed lexicon database. Lastly, the use of particle filter in gesture and posture recognition addresses the co-articulation issues by identifying the key frames and thus helps in resolving the ambiguities occurred due to low classification rate (i.e. posture signs are classified

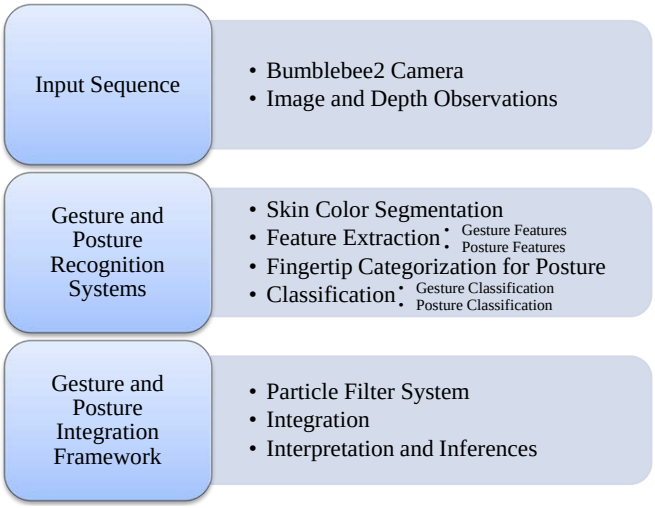


Figure. 1: (a) presents the process flow of the proposed framework.

by SVM even when classification percentage is 20%).
The main contents of this paper are described as follows: overview of literature is presented in Section II. Section III presents all the phases of gesture and posture recognition including image acquisition, object segmentation, feature extraction, categorization based on fingertip detection and classification. Particle filter system module is presented in Section IV. Section V describes the integration, interpretation and inference module along with CFG Grammar rules. The experimental results are demonstrated in Section VI whereas the concluding remarks are sketched in section VII.

II. Related Work

In the literature, two different approaches have been adopted to devise methods for vision based hand recognition namely model based approaches and appearance based approaches. In model based approaches, 3D model of the hand is constructed which contains 3D hand kinematics with certain Degrees Of Freedom (DOF). Moreover, hand parameters are extracted from this model and are matched with already observed images or image features. However, focus of our research is based on appearance based approaches in which the image features (i.e. contours, edges, image moments, eigenvectors etc.) are extracted and compared with the observed features set. There are two major issues which are to be considered in appearance based approaches namely feature selection and dataset training. Besides, the capability of operating in real-time environment motivates many researchers to explore the applicability of appearance based approaches to recognize gesture and posture for HCI. In the following, we have reviewed these in detail within the scope of this work.

A. Gesture Recognition

In gesture recognition, Yoon et al. [1] developed a hand gesture system by combining the location, angle and velocity for the recognition. Liu et al. [2] developed a system to recognize 26 alphabets by using different HMM topologies. In the similar context, Hunter et al. [3] used HMM for recognition where Zernike moments are used as image features for hand gesture sequences. With a different motivation, Chen et al. [4] presented a system for gesture recognition in which hands are recognized using Haar-like features. Moreover, the training algorithm is used based on AdaBoost which selects different Haar-like features for the classification. Bretzner et al. [5] presented an approach for static backgrounds to recognize the hand gestures using multi-scale color features. In a hierarchical model, the shape and color cues are combined with image features at different levels. Further, particle filter is used for the tracking and recognition of hand state.

B. Posture Recognition

Among many applications of posture recognition, sign language recognition is one of the application domains for HCI systems to interpret hand postures. Lamar and Bhuiyant [6] proposes a technique for feature extraction of hand postures based on Principal Component Analysis (PCA) and perform analysis on American Sign Language (ASL) and Japanese Sign Language (JSL) using multilayer Neural network classifier. However, drawback of the approach is the use of colored gloves and testing the approach on static background. Isaacs and Foo [7] proposed an approach using two layer feed-forward neural network to recognize fingerspelling ASL alphabets using still input images. Malassiotis and Strintzis [8] used Elliptic Fourier Descriptor for 3D hand posture recognition. In their system, orientation and silhouettes of the hand are used to recognize 3D hand postures. Similarly, Licsar and Sziranyi [9] used Fourier coefficients from modified Fourier descriptor approach to model hand shapes for hand gestures recognition. Altun et al. [10] proposed a method to increase the effect of fingers in fingerspelling Turkish Sign Language (TSL). In their method, hand shapes with strong edges are extracted and matched against the template.
It is observed that the co-articulation issues such as hand shape, position and orientation is a fundamental research objective in vision based hand gesture and posture recognition systems. Therefore, in our suggested approach, we have computed the robust and invariant features to make our system capable of operating under real-time conditions.

C. Data Fusion

Integration of different modalities have been used to improve the recognition (i.e. identification of a human by combining face and voice traits [11]) in the field of biometrics. In multi-modal biometric systems, fusion takes place at different levels which includes sample level, feature level, match score level and decision level fusion [12]. Chang et al. [13] proposed a face recognition system in which the fusion of 2D and 3D information of face images is done to improve the performance. Particular for the hand recognition, Kumar et al. [14] performed fusion at feature level and match score level to combine the palm prints and hand geometrical features. Similarly, Wu et al. [15] proposed a multi-model system to combine the gait recognition with face recognition system for the human recognition.
It is observed that the main motivation of exploiting different modalities is to achieve better performance and to cop the limitations of uni-modal approach. In the next sections, gesture and posture recognition systems along with their integration, interpretation and inferences modules are presented.

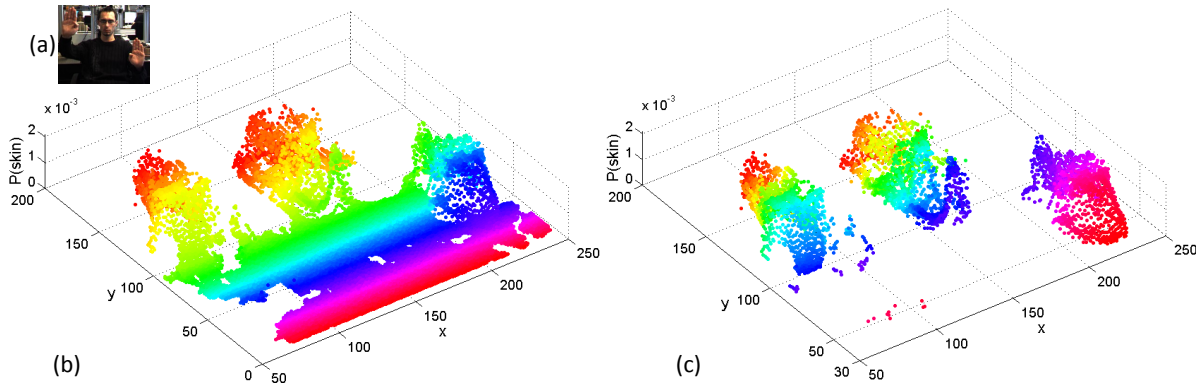


Figure. 2: (a) Input image (b) Skin color pixel probabilities of the depth region (i.e. 30cm to 200cm) using Normal Gaussian distribution (c) Selected skin color pixels for hand contour extraction and binarization.

III. Gesture and Posture Recognition Systems

In this section, components of the proposed gesture and posture recognition system are presented as follows:

A. Image Acquisition and Skin Color Segmentation

In image acquisition phase, Bumblebee2 camera is used which gives 2D image (i.e. Fig. 2(a)) and depth sequences. The depth sequences are exploited and processed to select region of interest (ROI) for segmentation of objects (hands and face) where the depth lies in range from 30 cm to 200 cm (i.e. in our experiments). Moreover, YC_bC_r color space is used because skin color lies in a small region of chrominance components whereas the effect of brightness variation is reduced by ignoring the luminance (Y) channel. In this region, hands and face are extracted from skin color distribution and are modeled by normal Gaussian distribution characterized by mean and variance as shown in Fig. 2(b). Normal Gaussian distribution probability for an observation $\mathbf{x}$ is calculated as:

$$\mathcal{P}(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-0.5\left((\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)} \quad (1)$$

where μ and Σ represents mean vector and covariance matrix respectively of training data. The computed probability $\mathcal{P}(\mathbf{x})$ derived from above model categorize the pixels as skin pixel or not (i.e. shown in Fig. 2(c)). After that, skin color pixels are binarized and contours are extracted by computing chain code representation for detection of hands and face.

B. Feature Extraction

Feature extraction is a fundamental step for the recognition process as the selection of good features always play a pertinent role. In the following, we have described the hand gesture and posture features as follows:

1) Gesture Features

In the proposed approach, hand orientation θ_t is used as a feature and is determined between two consecutive centroid points when drawing gesture path as follows:

$$\theta_t = arctan\left(\frac{y_{t+1}-y_t}{x_{t+1}-x_t}\right); t = 1, 2, ..., T-1 \quad (2)$$

where T represents length of gesture path, x_t and y_t are centroid points at frame t. The computed angle θ_t is quantized in range from 1 to 18 by dividing it by 20 degrees. These quantized values give us discrete vector which is used as feature in HMM to classify gesture symbols. The feature set $\mathcal{G}_{fv}$ is denoted as:

$$\mathcal{G}_{fv} = (\theta_t)^T \quad (3)$$

2) Posture Features

Two different types of feature vectors are computed for hand postures namely Hu-Moment (i.e. statistical feature vectors (i.e. FV)) $\mathcal{P}_{hu}$ and geometrical feature vectors $\mathcal{P}_{geo}$ [16]. These feature vectors are formulated as:

$$\mathcal{P}_{fv} = \{\mathcal{P}_{hu}, \mathcal{P}_{geo}\} \quad (4)$$

Statistical Feature Vector: In Eq. 4, $\mathcal{P}_{hu}$ represents Hu-Moments [17] which are derived from basic moments, and describes the properties of objects shape statistically (i.e. area, mean, variance, covariance and skewness etc). Hu [17] derived a set of seven moments which are translation, orientation and scale invariant, and are defined as follows:

$$\phi_1 = \eta_{20} + \eta_{02} \quad (5)$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \quad (6)$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (7)$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (8)$$

$$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (9)$$

$$\phi_6 = (\eta_{20} - \eta_{02})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \quad (10)$$

$$\phi_7 = (3\eta_{12} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (11)$$

These seven moments (i.e. $\phi_1, \ldots, \phi_7$) defines our statistical FV and are derived from second and third order moments. These moments are presented in the following set:

$$\mathcal{P}_{hu} = (\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7)^T \quad (12)$$

Geometrical Feature Vector: In Eq. 4, the second feature set is Geometrical FV ($\mathcal{P}_{geo}$) which is used to describe ASL sign shape. We have exploited the properties of circularity ($Cir = \frac{Perimeter^2}{4\pi \times Area}$) and rectangularity ($Rect = \frac{Area}{l \times w}$) as the measures of shape that defines how much object's shape is closer to circle and rectangle respectively. We show that these geometric attributes are capable of providing good representations for ASL signs as the shapes varies significantly for each corresponding ASL sign. Geometrical FV is stated as:

$$\mathcal{P}_{geo} = (Cir, Rect)^T \tag{13}$$

Both the statistical and geometrical FV set are combined together to form a feature vector set and is defined as:

$$\mathcal{P}_{fv} = \{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7, Cir, Rect\}^T \tag{14}$$

Normalization: Normalization is pertinent for posture feature vectors to keep them in a particular range for the classification (SVM) module. Our normalized FV is defined as:

$$\mathcal{NP}^i_{fv} = \frac{(\mathcal{P}^i_{fv} - c_{min})}{(c_{max} - c_{min})} \tag{15}$$

$$c_{min} = \mu - 2\sigma \ , \ c_{max} = \mu + 2\sigma \tag{16}$$

$\mathcal{NP}^i_{fv}$ is the normalized feature vectors for posture recognition where i refer to respective number of feature. c_{max} and c_{min} are the respective maximum and minimum values used for normalization of these features.

C. Fingertip Detection for Categorization

The fingertip categorization is performed based on number of detected fingertips from the hand's contour. The main idea in categorizing ASL alphabet signs based on fingertip detection is to reduce the mis-classifications observed among these symbols. In the proposed approach, four groups for ASL alphabets are formed according to number of detected fingertips (i.e. Group I with no fingertip, Group II with one fingertip, Group III with two fingertips and Group IV with three fingertips). This categorization is done only for alphabets because some numbers and alphabets share very similar shapes for instance 'D' and '1' are same with small change of thumb and it is hard to classify them together.

Given the contour of detected hand, curvature is estimated by considering the neighbor contour points to detect the fingertip [18]. Mathematically, curvature gives the ratio of length (i.e. sum of distances that a curve has) and displacement measures the distance from the first to last point if curve covers a straight line. Curvature $\mathcal{C}(k)$ is computed from the following equation:

$$\mathcal{C}(k) = \left\| \left(P_{k-n/2} - P_{k+n/2}\right) \right\|^{-1} \sum_{i=(k-n/2)}^{i=(k+n/2)} \left\| (P_i - P_{i+1}) \right\| \tag{17}$$

where k is the boundary point of object at which curvature $\mathcal{C}(k)$ is estimated, n is total number of pixels used for curvature estimation, and P_i and $P_{(i+1)}$ are the objects boundary points.

The main idea of fingertip categorization is to find high curvature values from contour points and results in detection of peaks from hands contour which provides a clue about the fingertip. We have adaptively determine the number of contour points (i.e. after conducting empirical experiments) by exploiting the depth information to find the distance between object and camera. In this way, a candidate for the fingertip is selected when curvature value is greater than $\sqrt{2}$.

In Fig. 3(a and b), experimental results show the contour pixels of left hand with curvature values on z-axis. The contour points with value greater than $\sqrt{2}$ are selected as the candidates for fingertip. Further, we have extracted two clusters named C_1 and C_2 with values above threshold and the maximum value from these individual clusters are selected using maximum local extreme value. The resulted points are marked as a fingertip. However, it is observed that both the peaks in the hand's contour (i.e. $\cap$) and valleys (i.e. $\cup$) can be inferred as a fingertip. Therefore, the next step is to remove valleys from being detected as a fingertip. For this purpose, selected contour points are taken and their distances are computed from the center point of the hand. Further, the normalization is done and these points are scaled ranging from 0 to 1. We pick the points whose values are greater than 0.5 for fingertip detection. In this way, fingertips are successfully detected for the categorization of ASL alphabet signs.

D. Classification

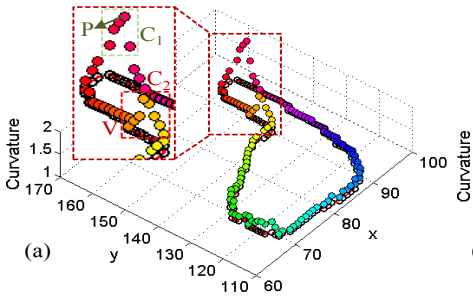
Two classification approaches are employed for gesture and posture recognition and are described as follows:

1) Gesture Classification

In the classification of gesture signs (i.e. alphabets and numbers), Baum-Welch (BW) algorithm is used to train the parameters of HMM by the discrete vector θ_t. We have used Left-Right banded model with 9 states for hand motion recognition of gesture path. Classification of hand gesture path is done by selecting the maximal observation probability of the gesture model by the Viterbi algorithm. In our case, the maximal gesture model is the classified symbol and has the largest observation probability among all the alphabets (i.e. A-Z) and numbers (i.e. 0-9).

2) Posture Classification

In the classification of posture signs, a set of thirteen ASL alphabets and ten ASL numbers are recognized using Support Vector Machines (SVM). SVM [19] is a supervised learning technique for optimal modeling of data. We have used normalized statistical and geometrical features vectors (i.e. $\mathcal{NP}^i_{fv}$) to train and classify signs using Radial Basis Function (RBF). RBF is a Gaussian kernel which works robustly with given number of features and provides optimum results when compared to other kernels. In the experiments, categorization is performed to make groups based on number of detected fingertips (see. confusion matrix in Table A). Confusion matrix (CM) in Table 1 and Table 2 presents classification probabilities of Group 2 (i.e. $\mathcal{A}, \mathcal{B}, \mathcal{D}, \mathcal{I}, \mathcal{H}/\mathcal{U}$) and Group 3 (i.e. $\mathcal{C}, \mathcal{L}, \mathcal{P}, \mathcal{Q}, \mathcal{V}, \mathcal{Y}$) respectively. Group 1 (i.e. $\mathcal{A}, \mathcal{B}$) has no mis-classifications and Group 3 has only one alphabet (i.e. $\mathcal{W}$). Table 3 presents confusion matrix of ASL numbers.



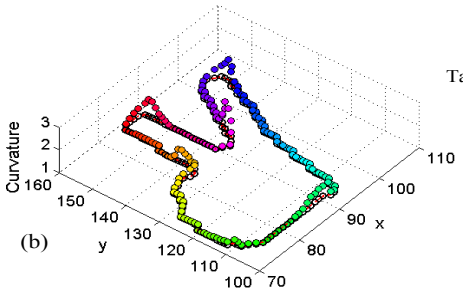


Table A: Confusion Matrix of Fingertips (i.e. all)

FT/Gr.	1	2	3	4
0	99.8	0.2	0.0	0.0
1	1.2	96.8	2.0	0.0
2	0.0	0.0	95.1	4.9
3	0.0	0.0	0.9	99.1

Figure. 3: (a) shows the posture signs "D" with contour pixels and curvature values. The clusters (C_1 and C_2) are presented with threshold above $\sqrt{2}$ and maximum local extreme values are selected from these clusters. These selected values are the interest points for fingertip detection which are then normalized to identify the peaks and valleys in the hand. The peak values represent the fingertip of the hand in the image. (b) presents the posture sign "V" (i.e. the cluster formation and normalization steps are same as in (a)). (c) Table A presents the confusion matrix of fingertips for categorization.

IV. Contribution-Weights using Particle Filter System

In this module, we have computed the contribution-weights for gesture and posture modalities through particle filter system which results in integration and inference of new symbols at any instance of time. Based on the computed contribution-weights, Integration $\mathcal{I}$ of gesture and posture systems fused at decision level is formulated as:

$$\mathcal{I} = \langle \alpha_{gstr} * \mathcal{R}_{hmm} \rangle \bigcap \langle \alpha_{pstr} * \mathcal{R}_{svm} \rangle \tag{18}$$

where $\mathcal{R}_{hmm}$ and $\mathcal{R}_{svm}$ are the classification outcome of gesture and posture system. α_{gstr} and α_{pstr} are the contribution-weights associated with gesture and posture system as integration criteria for fusion of these systems. We have employed Condensation algorithm [20] to approximate the probability density function which acts as contribution-weight using a collection of weighted samples of classification outcome and define the criterion for the fusion of gesture and posture systems [21]. In the proposed approach, two separate particle filters (i.e. for gesture and posture) are employed which are explained as follows.

Initialization. The classification outcome of both gesture and posture system is input to separate particle filter. A vector $\mathcal{S}(n)$ is constructed to initialize the Condensation process which is represented as follows:

$$\mathcal{S}(n) = \{s_k^{(gstr)}, s_k^{(pstr)}\} \tag{19}$$

A set of N random points (i.e. 100) called particles x_k^n with weights w_k^n denotes the initial distribution of particles at time k for both gesture and posture systems. These particles are denoted as:

$$s_k^{(gstr|pstr)} = \{x_k^n, w_k^n\}_N^{n=1} \tag{20}$$

(Note. From this point on, the same notation is used for both the particle filters (i.e. gesture and posture), except when stated otherwise).

Prediction. The a-priori probability $p\left(x_k|z_{k-1}\right)$ is computed from previous a-posteriori probability $p\left(x_{k-1}|z_{k-1}\right)$ and the dynamic model $p\left(x_k|x_{k-1}\right)$. The formulation is represented as:

$$p^{(n)}\left(x_k|z_{k-1}\right) = p^{(n)}\left(x_k|x_{k-1}\right) p^{(n)}\left(x_{k-1}|z_{k-1}\right) \tag{21}$$

Updation. The a-posteriori probability (i.e. contribution-weights) of the state is calculated from the a-priori probability $p(x_k|z_{k-1})$ and the likelihood function $p(z_k|x_k)$ by incorporating the new measurement data z_k. Likelihood function is formulated as follows:

$$p^{(n)}\left(z_k|x_k\right) = \pi_k^{(n)} = e^{-((z_k - x_k^n)^2)/(2\sigma^2)} \tag{22}$$

where σ is the standard deviation of particle weights. The contribution-weights $\alpha_{gstr|pstr}$ or a-posteriori probability $p(x_k|z_k)$ for gesture and posture system is computed as follows:

$$\alpha_{gstr|pstr} = p\left(x_k|z_k\right) = \frac{\sum_{n=1}^{N} p^{(n)}\left(z_k|x_k\right) p^{(n)}\left(x_k|z_{k-1}\right)}{\sum_{n=1}^{N} p^{(n)}\left(z_k|x_k\right)} \tag{23}$$

Using N values of $p(z_k|x_k)$, we have built a probability distribution for the whole space at any time instant. The conditional probability acts as a weighting factor for its corresponding state with successive iterations. The normalized weighting probabilities are calculated as follows:

$$\pi_k^{(n)} = \frac{p^{(n)}\left(z_k \mid x_k\right)}{\sum_{n=1}^{N} p^{(n)}\left(z_k \mid x_k\right)} \tag{24}$$

In this way, we obtain the contribution-weights which defines the integration-criteria for the fusion of these systems.

Table 1: Confusion Mat: 1-Detected FT

Sign	A	B	D	I	H/U
A	99.8	0.0	0.0	0.0	0.2
B	0.0	98.2	1.0	0.0	0.8
D	0.0	0.0	98.7	1.3	0.0
I	0.6	0.0	0.8	98.6	0.0
H/U	0.0	3.1	0.0	0.2	96.7

Table 2: Confusion Mat: 2-Detected FT

Sign	C	L	P	Q	V	Y
C	98.7	0.2	0.0	0.7	0.0	0.4
L	0.4	98.5	0.0	0.7	0.0	0.4
P	0.0	0.0	98.7	1.3	0.0	0.0
Q	0.0	0.0	3.8	96.2	0.0	0.0
V	0.2	0.0	0.0	0.0	99.3	0.5
Y	0.0	0.0	0.0	0.0	0.7	99.3

Table 3: Confusion Matrix of ASL Numbers

Nr.	0	1	2	3	4	5	6	7	8	9
0	99.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.3	99.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	98.3	0.4	0.0	0.0	1.3	0.0	0.0	0.0
3	0.0	0.0	0.4	98.2	0.9	0.0	0.5	0.0	0.0	0.0
4	0.0	0.0	0.0	0.2	98.2	1.6	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	2.4	97.6	0.0	0.0	0.0	0.0
6	0.0	0.0	0.8	0.6	0.0	0.0	98.6	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.7	98.3	0.6	0.4
8	0.0	0.0	0.0	0.4	0.0	0.0	0.2	0.4	98.4	0.6
9	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.4	0.5	98.6

V. Integration, Interpretation and Inference

After computing the contribution-weights for gesture and posture recognition, the integration of these modalities to generate interpretations and inferences is the main objective. To achieve this goal, we consider the integration as a problem of regular language and mapped the recognition outcome over the context free grammar (CFG) rules [22]. Before, we describe the concept of context specific interpretation and inference rules which we have employed in this research, it is essential to first describe the proposed structure of the language. The CFG grammar is defined in quadruple (i.e. 4-tuple) $Grammar = \langle \mathcal{V}, \mathcal{T}, \mathcal{S}, \mathcal{R} \rangle$ where $\mathcal{V}$ is the set of objects and contains non-terminals as well as terminals symbols, $\mathcal{T}$ is the set of terminals, $\mathcal{S}$ is start symbol and it is a subset of $\mathcal{V}$ (i.e. $\mathcal{S} \in \mathcal{V}$), and $\mathcal{R}$ is the set of production rules. The recognition outcomes are mapped on CFG rules(*) for the integration.

In the CFG production rules, $\langle Posture_{Alphabet} \rangle$ computes set of recognized posture alphabet signs, $\langle Gesture_{Alphabet} \rangle$ results in recognized gesture symbols and $\langle Posture_{Digit} \rangle$ is the set of recognized numbers. Different symbols can be devised in integration process depending upon the lexicon as shown in Table 4. The contribution-weights computed through particle filter system whose threshold (T) is above 70% are selected for the fusion process and is written as:

$$(\alpha_{gstr} | \alpha_{pstr}) \geq T \tag{25}$$

The inference for gesture recognition system takes place after processing some frames because of HMM classifier initialization. In contrast, posture system recognizes the symbol at every frame because a single frame is sufficient to recognize ASL symbols (i.e. Exceptions are "J" and"Z" and are not considered here). Integration is carried out when contribution-weights of gesture α_{gstr} and posture α_{pstr} signs satisfy the threshold (T) at any time frame. In this regard, different approaches are proposed for the fusion of different systems which includes AND/OR combination, majority voting, behavior knowledge method and weighted voting method [23]. However, we have used AND/OR combination for gesture and posture recognition symbols. Integration $\mathcal{I}$ is formulated as:

$$\mathcal{I} = \langle \alpha_{gstr} * \mathcal{R}_{hmm} \rangle \bigcap \langle \alpha_{pstr} * \mathcal{R}_{svm} \rangle \tag{26}$$

$$\mathcal{I} = \langle \alpha^i_{gstr} * \mathcal{R}^i_{hmm}; i = 1 \cdots m \rangle \bigcap \langle \alpha^j_{pstr} * \mathcal{R}^j_{svm}; j = 1 \cdots n \rangle \tag{27}$$

In our experiments, the combination of CFG rules yield us to the integration of gesture and posture recognition in which multiple posture symbols (i.e. described above as n) are combined with a gesture symbol (i.e. m). To make inferences of results from CFG (see. CFG Rules), the possible derivation of posture results is $\langle Posture_{Alphabet} \rangle$ followed by $\langle Posture_{Digit} \rangle$ whereas $\langle Gesture_{Alphabet} \rangle$ yields

Defs. Rules 1 Context Free Grammar (CFG)*

$Definitions\ and\ Rules$:

$$\begin{aligned}
\mathcal{V} = \{&\mathcal{S}, Posture_{Alphabet}, X, Gesture_{Alphabet}, Alphabet, \\
&Posture_{Digit}, Y, Digit, 0_p|1_p, ..., 9_p, a_g|b_g, ..., z_g, \\
&a_p|b_p, ..., z_p\} \\
\mathcal{T} = \{&0_p|1_p, ..., 9_p, a_g|b_g, ..., z_g, a_p|b_p, ..., z_p\} \\
\mathcal{S} &\rightarrow \langle Posture_{Alphabet} \rangle \langle X \rangle \\
Posture_{Alphabet} &\rightarrow \langle Alphabet \rangle \langle Posture_{Alphabet} \rangle \mid \\
&\langle Alphabet \rangle \\
X &\rightarrow \langle Posture_{Digit} \rangle \langle Y \rangle \\
Posture_{Digit} &\rightarrow \langle Digit \rangle \langle Posture_{Digit} \rangle \mid \langle Digit \rangle \\
Y &\rightarrow \langle Gesture_{Alphabet} \rangle \langle Posture_{Digit} \rangle \\
Digit &\rightarrow 0_p|1_p|2_p, ..., 9_p \\
Alphabet &\rightarrow a_p|b_p|c_p, ..., z_p \\
Gesture_{Alphabet} &\rightarrow a_g|b_g|c_g, ..., z_g
\end{aligned}$$

Table 4: Lexicon of Symbols

Symbols ⇒ Fruits	Symbols ⇒ Fruits
A⇒Apple, Apricot	N⇒Nectarine
B⇒Blueberry, Banana	O⇒Orange, Oval Kumquat
C⇒Cherry, Cantaloupe	P⇒Pear, Peach
D⇒Date, Dewberry	Q⇒Quince
E⇒Elderberry, Eggfruit	R⇒Raspberry, Rambutan
F⇒Fig, Farkleberry	S⇒Star Fruit, Strawberry
G⇒Grapes, Gooseberry	T⇒Tangerine, Tart Cherry
H⇒Honeymelon, Hackberry	U⇒Ugli Fruit, Uniq Fruit
I⇒Imbe	V⇒Voavanga
J⇒Jackfruit, Jambolan	W⇒Watermelon, Wolfberry
K⇒Kaffir Lime, Kiwi	X⇒Xigua
L⇒Lemon, Lychee	Y⇒Yunnan Hackberry
M⇒Mango, Melon	Z⇒Zinfandel Grapes

to only one possible outcome in the integration. The inference derived from CFG rules is as follows:

$$S \rightarrow \langle Posture_{Alphabet}\rangle\langle Gesture_{Alphabet}\rangle \langle Posture_{Digit}\rangle$$

Different interpretations are devised for integration process which includes:

1. *Interpretation:*
   $\langle Gesture \Rightarrow Detected\rangle$; $\langle Posture \Rightarrow Detected\rangle$; $\langle Integration \Rightarrow Yes\rangle$
   Description: The ideal case of integration, both gesture and posture systems recognize the symbol at any time frame.

2. *Interpretation:*
   $\langle Gesture \Rightarrow NotDetected\rangle$; $\langle Posture \Rightarrow Detected\rangle$; $\langle Integration \Rightarrow No\rangle$
   Description : Gesture system does not classify any symbol because HMM is not activated when gesture drawing process starts. In contrast, the posture system classifies the sign with the contribution-weights α_{pstr} above the threshold.

3. *Interpretation:*
   $\langle Gesture \Rightarrow SemiDetected\rangle$; $\langle Posture \Rightarrow Detected\rangle$; $\langle Integration \Rightarrow Yes/No\rangle$
   Description: There can be some predictions about gesture symbols dependent upon the inference from HMM states. In this case, gesture symbol is still incomplete and it gives a clue about user's intention while drawing the gesture symbol. Intentions are predicted only when contribution-weight α_{gstr} of gesture sign pass the threshold criterion.

4. *Interpretation:*
   $\langle Gesture \Rightarrow NotDetected\rangle$; $\langle Posture \Rightarrow NotDetected\rangle$; $\langle Integration \Rightarrow No\rangle$
   Description: No match has occurred from gesture and posture systems. In this way, the symbols are not present in the lexicon.

VI. Experimental Results

In the proposed approach, the experimental setup involves the tasks of data acquisition, gesture and posture classification and particle filter system. Moreover, this system is directly linked to CFG rules which generates the "meaningful expressions" by the mechanism of interpretations and inferences. We have demonstrated the applicability of proposed approach on our real-time example scenario and show the description of meaningful expressions generated from the integration of these systems.

We have developed our own dataset in the laboratory which comprises of eight actors performing the gesture and posture signs where the image sequences are captured by Bumblebee2 stereo camera with 240*320 pixels image resolution. These developed signs are used for the training and testing purposes in which the gesture signs contain 30 video observations for each gestures sign whereas 3600 image observation are used for posture signs. Out of these video and image observations, 20 videos and 2400 image observations are used as training observations for gesture and posture respectively. It is worth to mention that, we have tested our examples case independent to classification process. So, the applicability can be extended by designing the lexicon and CFG rules according to the scenario under observation.

The proposed concept of integration is tested on a real-time example scenario, for instance we have designed restaurant lexicon which reflects the functionality of food and drink order placement at counter. For this purpose, we have studied type of food and drink item in a menu (e.g. name of fruit, drinks, fast food, etc.). In this work, we have chosen 45 different fruits for this choice as shown in Table. 4 and make different (i.e. currently our system supports 500 combinations) choices for the menu-order by combining recognized gestures and postures signs. For instance, an order is placed by integrating the first and second/third alphabet of the fruit name from gesture and ASL posture where the quantity of the desired fruit item is inferred by concatenating with another posture sign number as shown in Fig. 4.

Fig. 4(a and c) shows an interpretation based on fusion of gesture and posture recognition system. In Fig. 4(a), posture system firstly recognizes the alphabet "A". However, gesture recognition system did not recognize any symbol during the initial frames. The next posture symbol recognized is "2" which indicates the quantity of order. From frames 38 to 48, gesture recognition system computes the probability of possible signs which the user can draw depending on HMM states and most likely candidates for the gesture recognition. Moreover, it selects the highest probability element and mark it the "best" element for recognition. At frame 48, the gesture ends and the recognized symbol is "D", thus completing the order (i.e. $\langle Rec_{pstr} = \text{“}A\text{”}\rangle, \langle Rec_{pstr} = \text{“2”}\rangle, \langle Rec_{gstr} = \text{“}D\text{”}\rangle$). In Fig. 4(c), the next interpretation starts in which the user draws the posture symbol "L". The next posture recognized is the number which describes the quantity as "2" and finally the gesture symbol which has been recognized is the symbol "B" (i.e. $\langle Rec_{pstr} = \text{“}L\text{”}\rangle, \langle Rec_{pstr} = \text{“2”}\rangle, \langle Rec_{gstr} = \text{“}B\text{”}\rangle$). Gesture and posture recognition works optimally and recognizes the signs correctly.

Fig. 4(b and d) presents the classification and weight-contribution results of gesture and posture recognition for the whole sequence. The recognition of gesture and posture system after applying the threshold is presented in Fig. 4(b) along with the integration of these systems. In this sequence, the recognized gesture elements for the first order is $\langle Date = \text{“}D\text{”}\rangle$, $\langle Posture_{Alphabet} = \text{“}A\text{”}\rangle$ which means $Date$ and from the posture recognized symbol, it is "2". It means $\langle Two\ Date\ Juices\rangle$. The second order is $\langle Two\ Blueberry\ Juices\rangle$. By changing the lexicon, the proposed approach can be used in other scenarios.

We have tested our proposed approach on the restaurant lexicon database with the overall 98.3% inference accuracy. It is observed that the classification inaccuracies do not affect the performance due to particle filter based weight computation technique. One of the potential reasons is, the particle filter works on the principle of prediction and updation mechanism, therefore, the inference of meaningful expression is achieved successfully.

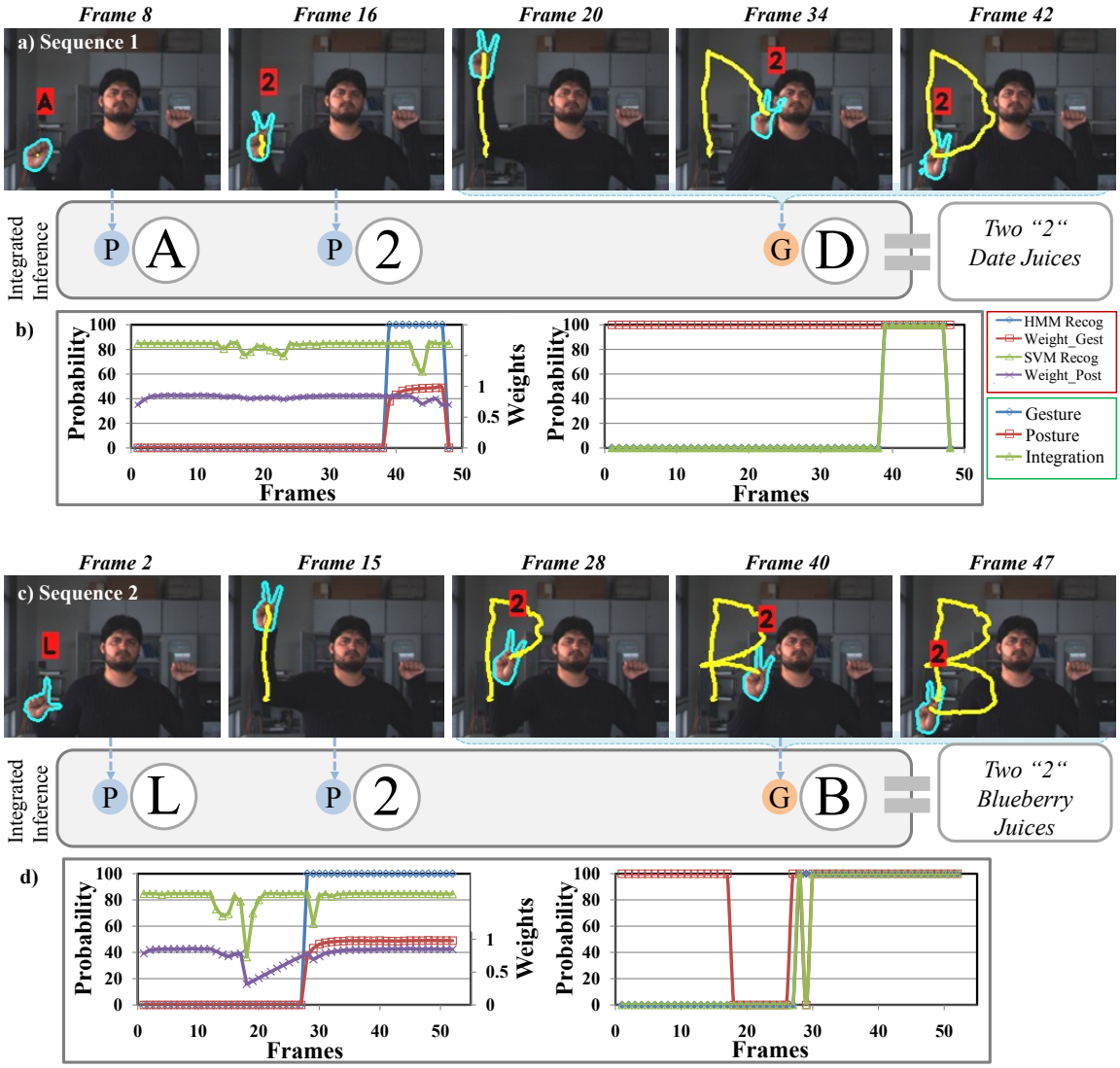


Figure. 4: a) shows the recognized gesture symbol "D" whereas the classified postures are alphabet "A" and Number "2". The meaningful expression "Two Date Juices" is inferred from this sequence. c) Second meaningful expression is "Two Blueberries Juices" which results from recognized gesture symbol "B" and classified posture alphabet "L" and Number "2". b and d) Graphs on left show the recognition rate along with contribution-weights from particle filter system whereas graphs on right present gesture and posture results with their integration. Legends of left graph are bordered with red whereas green bordered legends are for the graphs on right.

VII. Conclusion

In this paper, we have proposed a framework for recognition of gesture signs , ASL posture signs and their integration. In gesture and posture recognition systems, features are extracted which are invariant to translation, orientation and scaling. Besides, fingertips are detected for ASL alphabets and used as a measure to categorize thus avoids the misclassifications between posture signs. SVM is applied for recognition of ASL signs whereas HMM is used for the classification of gesture symbols. Moreover, a novel approach is proposed for the integration of gesture and posture recognition in which contribution-weights are computed using Particle filter system by incorporating CFG rules. The proposed approach is tested on restaurant lexicon which successfully integrates both systems and enables to interpret multiple inferences at the same instance of time. Future research is focused on words recognition for gesture and posture systems along with their integration.

Acknowledgments

This work is supported by Transregional Collaborative Research Center SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by German Foundation and LSA (Graduiertenfoerderung OvG University).

References

[1] Yoon, H., Soh, J., Bae, Y., Yang, H.: Hand gesture recognition using combined features of location, angle and velocity. Pattern Recognition **34** (2001) 1491–1501

[2] Liu, N., Lovel, B., Kootsookos, P.: Evaluation of hmm training algorithms for letter hand gesture recognition. In: IEEE Int. Sym. on SPIT. (2003) 648–651

[3] Hunter, E., Schlenzig, J., Jain, R.: Posture estimation in reduced-model gesture input systems. In: International Workshop on Automatic Face-and Gesture-Recognition. (1995) 290–295

[4] Chen, Q., Georganas, N., Petriu, E.: Real-time vision based hand gesture recognition using haar-like features. In: Instrumentation and Measurement Technology Conference. (2007)

[5] Bretzner, L., Laptev, I., Lindeberg, T.: Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In: Automatic Face and Gesture Recognition. (2002) 423–428

[6] Lamar, M., Bhuiyant, M.: Hand alphabet recognition using morphological pca and neural networks. In: International Joint Conference on Neural Networks. (1999) 2839–2844

[7] Isaacs, J., Foo, S.: Hand pose estimation for american sign language recognition. In: Thirty-Sixth Southeastern Symposium on IEEE System Theory. (2004) 132–136

[8] Malassiotis, S., Strintzis, M.: Real-time hand posture recognition using range data. Image and Vision Computing **26** (2008) 1027–1037

[9] Licsar, A., Sziranyi, T.: Supervised training based hand gesture recognition system. In: ICPR. (2002) 999–1002

[10] Altun, O., Albayrak, S.: Increasing the effect of fingers in fingerspelling hand shapes by thick edge detection and correlation with penalization. In: Pacific-Rim Symp. on Image and Video Tech. (2006) 1133–1141

[11] Brunelli, R., Falavigna, D.: Person identification using multiple cues. IEEE Trans. on PAMI **17** (1995) 955–966

[12] Ross, A., Jain, A.: Multimodal biometrics: An overview. In: 12th Signal Processing Conference. (2004) 1221–1224

[13] Chang, K., Bowyer, K.W., Flynn, P.J.: Face recognition using 2d and 3d facial data. In: ACM Workshop on Multimodal User Authentication. (2003) 25–32

[14] Kumar, A., Wong, D., Shen, H., Jain, A.: Personal verification using palmprint and hand geometry biometric. In: 4th Int. Conf. on Audio and Video-based Biometric Person Authentication. (2003) 668–678

[15] Wu, Q., Wang, L., Geng, X., Li, M., He, X.: Dynamic biometrics fusion at feature level for video-based human recognition. (2007) 152–157

[16] Rashid, O., Al-Hamadi, A., Michaelis, B.: Utilizing invariant descriptors for finger spelling american sign language using svm. In: ISVC. (2010)

[17] Hu, M.: Visual pattern recognition by moment invariants. IRE Trans. on Information Theory **8** (1962) 179–187

[18] Kim, D.H., Kim, M.J.: A curvature estimation for pen input segmentation in sketch-based modeling. Computer-Aided Design **38** (2006) 238 – 248

[19] Lin, C., Weng, R.: Simple probabilistic predictions for support vector regression. Technical report, National Taiwan University (2004)

[20] Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. Int. Jour. of Computer Vision **29** (1998) 5–28

[21] Rashid, O., Al-Hamadi, A., Michaelis, B.: Interpreting dynamic meanings by integrating gesture and posture recognition system. In: VECTaR 2010, ACCV. (2010)

[22] Hopcroft, J.E., Motwani, R., Ullman, J.D.: Introduction to Automata Theory, Languages, and Computation. 2. edn. Pearson Addison-Wesley (2001)

[23] Monwar, M., Gavrilova, M.: A robust authentication system using multiple biometrics. In: Comp. and Information Science. (2008) 189–201

A. Biography

Omer Rashid is a PhD student in Otto-von-Guericke University, Magdeburg, Germany. He received his Bachelors degree in Computer Engineering from University of Engineering and Technology, Lahore, Pakistan. Further, he has completed his Master degree from Otto-von-Guericke University, Magdeburg, Germany. His current research is focused on human computer interaction, image processing and pattern recognition.

Ayoub K. Al-Hamadi was born in Yemen in 1970. He received his Masters Degree in Electrical Engineering in 1997 and his PhD in Technical Computer Science at the Otto von Guericke University, Magdeburg, Germany in 2001. Since 2002 he has been Junior-Research-Group-Leader at the Institute for Electronics, Signal Processing and Communications at University of Magdeburg. In 2008 he became Professor of Neuro-Information Technology. His research work concentrates on the field of image processing, pattern recognition and artificial neural networks. Professor Al-Hamadi is the author of more than 170 articles in peer-reviewed international journals and conferences.

Klaus Dietmayer received the Dipl.-Ing. degree in electrical enginieering from the Technical University of Braunschweig, Germany, in 1989 and PhD degree from the University of Armed Forces. Hamburg, Germany, in 1994. In 1994, he joined the Philips Semiconductors Systems Laboratory, Hamburg, as a Research Engineer and became a Manager in the field of networks and sensors for automotive applications in 1996. In 2000, he was appointed as a Professor in the field of measurement and control with the University of Ulm, Germany, where he is now Director of the Institute of Measurement, Control and Microtechnology. His research interests include information fusion, multi-object tracking, environment perception and situation assessment for driver assistance and safety systems. Prof. Dietmayer is member of the IEEE and of the German Association of Electrical Engineers.