World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:3, No:11, 2009

A Robust Method for Hand Tracking Using Mean-shift Algorithm and Kalman Filter in Stereo Color Image Sequences

Mahmoud Elmezain, Ayoub Al-Hamadi, Robert Niese, and Bernd Michaelis
Institute for Electronics, Signal Processing and Communications (IESK)
Otto-von-Guericke-University Magdeburg
{Mahmoud.Elmezain, Ayoub.Al-Hamadi}@ovgu.de

***Abstract*—Real-time hand tracking is a challenging task in many computer vision applications such as gesture recognition. This paper proposes a robust method for hand tracking in a complex environment using Mean-shift analysis and Kalman filter in conjunction with 3D depth map. The depth information solve the overlapping problem between hands and face, which is obtained by passive stereo measuring based on cross correlation and the known calibration data of the cameras. Mean-shift analysis uses the gradient of Bhattacharyya coefficient as a similarity function to derive the candidate of the hand that is most similar to a given hand target model. And then, Kalman filter is used to estimate the position of the hand target. The results of hand tracking, tested on various video sequences, are robust to changes in shape as well as partial occlusion.

***Keywords*—Computer Vision and Image Analysis, Object Tracking, Gesture Recognition.

I. Introduction

The tracking of moving hands is an active area of research in the vision community, mainly Human-Computer Interaction (HCI). The goal of hand tracking is to push the advanced human-computer communication to bring the performance of HCI close to human-human interaction. This is due to the existing complexities in hand tracking such as hand appearance, illumination variation, and inter-hands occlusion. These issues undermine the performance and efficiency of tracking algorithms. In the last decade, there are several methods that attempt to develop robust techniques for varying video conditions such as partial or complete occlusions, noise, clutter, etc. Liu and Lovell [1] introduced a system for hand tracking in real-time based on the Camshift algorithm and the compound constant-acceleration Kalman filter algorithm. Nguyen *et al.* [2] proposed a system for hand gesture recognition, where the hand is tracked by Kalman filter and hand blobs analysis to obtain motion descriptors for hand region. This system is fairy robust to background cluster and uses skin color for hand gesture tracking and recognition. Whereas Nobuhiko *et al.* [3] used HSV color space to track hands and face in non complex background, where the overlapping problem between hands and face is solved by matching templates of the previous hands and face. Comaniciu *et al.* [4] proposed a technique to track the moving objects from a moving camera using Mean-shift algorithm and Kalman filter, where the implementation of this technique achieved real-time performance. Mostly, previous approaches have not been considered many points as the combination of accurate segmentation of both hands, robust tracking that containing overlap between hands and face, and the capability of the system to run in real-time on high resolution.

The main contribution of this paper is to introduce a robust method of efficient tracking for hands using Mean-shift algorithm [5] and Kalman filter [6] in conjunction with 3D depth map. The blob segmentation of the hands and face with complex background takes place using 3D depth map from a passive stereo camera, Gaussian Mixture Models (GMM) and color information, which is more robust to the disadvantageous lighting and partial occlusion. The depth information solve the overlapping problem between hands and face. After the target of the hand is localized, the Mean-shift analysis uses the gradient of Bhattacharyya coefficient as a similarity function to derive the target candidate of the hand that is the most similar to a given hand target model. The Bhattacharyya coefficient is obtained by masking the hand's color distributions with an Epanechnikov kernel. Moreover, the Mean-shift procedure is used to perform the optimization to determine the centroid point for each hand target. We considered a histograms as the representation of the hand's color probabilities density function (pdf's), as they can satisfy the low-cost requirement of real-time tracking. The measurement vector is determined based on mean shifts and then the next hand location is predicated by Kalman filter that is an optimal estimator to predicate and correct the states of linear processes. Above of all, we take in our consideration the mean depth value and the adaptive size of bounding box for each hand target in mean-shift tasks. This paper is organized as follows; Section II provides a detailed description of proposed tracking algorithms. The experimental results are described in the Section III. Finally, Section IV gives a few concluding remarks.

II. Suggested Method

Our method is proposed to introduce an efficient tracking for hands from stereo color image sequences using Mean-shift algorithm, Kalman filter in conjunction with 3D depth map. This requires technique for skin segmentation and handling occlusion between hands and face to overcome the difficulties of overlapping regions and partial occlusion. In particular, the proposed method consists of three main modules (Fig.1).

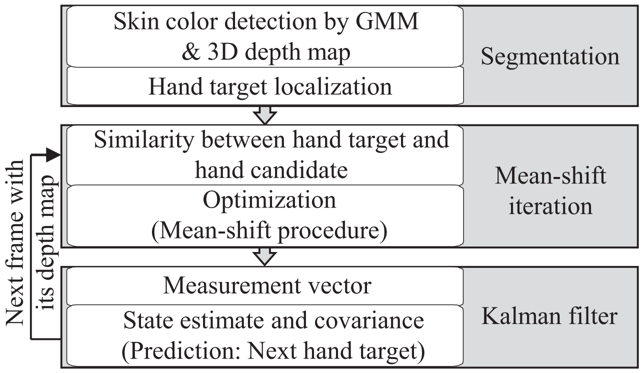

Skin color detection by GMM & 3D depth map
Hand target localization
Segmentation
Similarity between hand target and hand candidate
Optimization (Mean-shift procedure)
Mean-shift iteration
Measurement vector
State estimate and covariance (Prediction: Next hand target)
Kalman filter
Next frame with its depth map

Fig. 1. Simplified structure showing the main computational modules for the proposed tracking in real image sequences.

- **Segmentation**; the segmentation of hands & face targets in the first frame takes place using 3D depth map, GMM over YC_bC_r color space.
- **Mean-shift analysis**; measure the similarity between hand target and hand candidate that is localized from the next image frame with its depth map, perform the optimization to determine the correct mean shift point.
- **Kalman filter**; determine the measurement vector based on mean shifts and predicate the next hand target location.

A. Automatic Hand Segmentation

In this paper, a method for detection and segmentation of a hand in stereo color images with complex background is described where the hand segmentation takes place using 3D depth map and color information. Segmentation of skin colored regions becomes robust if only the chrominance is used in analysis. Therefore, YC_bC_r color space is used in our method where Y channel represents brightness and (C_b, C_r) channels refer to chrominance. We ignore Y channel to reduce the effect of brightness variation and use only the chrominance channels which fully represent the color information. A large database of skin and non-skin pixels is used to train the Gaussian model. In the training set, 18972 skin pixels from 36 different races persons and 88320 non-skin pixels from 84 different images are used. The GMM technique begins with modeling of skin using skin database where a variant of k-means clustering algorithm performs the model training to determine the initial configuration of GMM parameters. For more details, the reader can refer to [7], [8], [9], [10]. For the skin segmentation of hands and face in stereo color image sequences an algorithm is used, which calculates the depth value Z in addition to skin color information according to Eq.1 (Fig. 2(d)).

$$Z = \frac{f.b}{x_L - x_R} \tag{1}$$

where f is the identical effective focal length. b is base line, which represents the distance between two optical centers (left: O_L, right: O_R). The angle subtended by the two optical axes be 2θ. A point $P(X,Y,Z)$ in 3D space is projected onto the points (x_L, y_L) and (x_R, y_R) in the image plane of left and right camera. For more details, the reader can refer to [11], [12]. By the given depth information (Fig.2(c)) from camera set-up system, the overlapping problem between hands and face is solved since the hand regions are closer to the camera rather than the face region. Also, blob analysis is used to derive the hand boundary area, bounding box and centroid point.

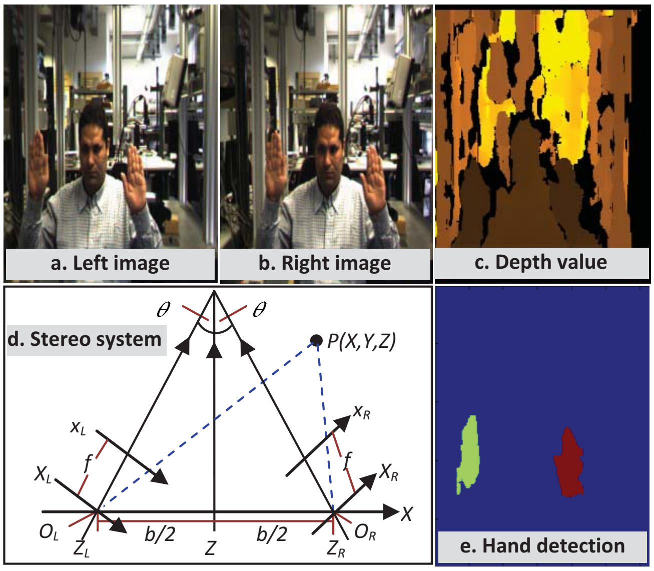


Fig. 2. (a) Left image frame of video stream. (b) Right image frame. (c) Depth value of left and right image from the Bumblebee stereo camera system. (d) Geometry of a stereo camera system. (c) Labeled skin color segmentation of two hands.

B. Mean-shift Analysis

Mean-shift algorithm is a kernel (i.e. non-parametric) density estimator that optimizes a smooth similarity function to find the direction of the hand target's movement. We consider m-bin histograms as the representation of the hand's color probabilities density function (pdf's), as they can satisfy the low-cost requirement of real-time tracking. After localization of the hand's target from the segmentation step, we find its color histogram with Epanechnikov kernel (monotonic decreasing kernel profile $k(x)$) [4], [5] (Fig.3). Epanechnikov kernel assigns smaller weights to pixels father from the center. Using these weights increases the robustness of the density estimation since the peripheral pixels are the least reliable, being often affected by occlusions.

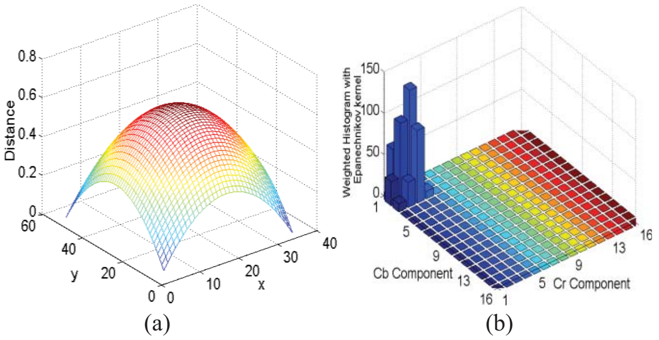


Fig. 3. (a) The Epanechnikov kernel for the hand target model of first image frame. (b) 2D weighted histogram by Epanechnikov kernel for (C_b, C_r) components with 16×16 bins.

Let $x_i^*, i = 1...n$ be the normalized pixel locations in the **hand target model**. The probability of the feature $u = 1...m$ in the hand target model histogram is computed as;

$$q_u = F \sum_{i=1}^{n} k(\|x_i^*\|^2)\delta[b(x_i^*) - u] \tag{2}$$

where δ is the Kronecker delta function, equal to 1 only at u and 0 otherwise. The normalization constant F is determined by imposing the condition $\sum_{u=1}^{m} q_u = 1$, where

$$F = \frac{1}{\sum_{i=1}^{n} k(\|x_i^*\|^2)} \tag{3}$$

For the **hand candidate model** in the next frame, Let $x_i, i = 1...n_h$ be the normalized pixel locations of the hand candidate, centered at y. Using the same kernel profile $k(x)$, but with bandwidth h. The probability of the feature $u = 1...m$ in hand candidate histogram is calculated as;

$$p_u(y) = F_n \sum_{i=1}^{n_h} k(\|\frac{y - x_i}{h}\|^2)\delta[b(x_i) - u] \tag{4}$$

where

$$F_h = \frac{1}{\sum_{i=1}^{n_h} k(\|\frac{y-x_i}{h}\|^2)} \tag{5}$$

Moreover, the Bhattacharyya coefficient [13] is more suitable to measure the similarity between the hand target model and the chosen candidate. To find the best match of our hand target in the sequential frames, the Bhattacharyya coefficient is maximized for the Bayes error that arising from the comparison of the target and candidate pdf's. The maximization of the Bhattacharyya coefficient between the unit vectors $\sqrt{q}$ and $\sqrt{p(y)}$ that representing the hand target histogram and hand candidate histogram respectively takes the following form;

$$\rho[p(y_0), q] = \sum_{u=1}^{m}\sum_{u=1}^{m} \sqrt{p_u(y_0)q_u} \tag{6}$$

That means, we need to maximize the term;

$$\sum_{i=1}^{n_h} w_i k(\|\frac{y - x_i}{h}\|^2) \tag{7}$$

where h is the kernel's smoothing parameter or bandwidth and the weights w_i is derived according to Eq. 8.

$$w_i = \sum_{i=1}^{n_h} \sqrt{\frac{q_u}{p_u(y_0)}}\delta[b(x_i) - u] \tag{8}$$

The mean-shift procedure is defined recursively and performs the optimization to compute the mean shift vector [14]. In short, mean-shift iteration uses the gradient of similarity function as an indicator of the direction of hand's movement (Eq.9).

$$y = \frac{\sum_{i=1}^{n_h} x_i w_i}{\sum_{i=1}^{n_h} w_i} \tag{9}$$

C. *Kalman Predication*

Kalman filter is an optimal estimator that predicts and corrects the states of linear processes [6]. After each mean-shift optimization that gives the measured location of the hand target, the uncertainty of the estimate can also be computed and then followed by the Kalman iteration, which drives the predicated position of the hand target. The Uncertainty is determined by image noise, the similarity between hand target colors, clutter colors and the percentage of occlusion. As such, the equations for the Kalman filter fall into two groups: time update equations and measurement update equations. The measurement vector z_k consists of the location of centroid point of hand region. First, we measure hand location and velocity in each image frame. So, we define the state vector as x_t;

$$x_t = (x(t), y(t), v_x(t), v_y(t))^T \tag{10}$$

where $(x(t), y(t))$ refers to hand location and $(v_x(t), v_y(t))$ represents the velocity of the hand in the t^{th} image frame.

- **Time Update Equations**

As x_k is not measured directly therefore the information provided by measurement is used to update the unknown state x_k. A priori estimate of state $\hat{x}_k^-$ and covariance error estimate P_k^- is obtained for the next time step k.

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1} \tag{11}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{12}$$

where A is the transition matrix with associated noise Bu and Q that is the Gaussian process noise with zero mean.

- **Measurement Update Equations**

The objective is to estimate a posteriori $\hat{x}_k$ which is a linear combination of the a priori estimate $\hat{x}_k^-$ and the new measurement z_k.

$$K_k = P_k^- H^T(HP_k^- H^T + R)^{-1} \tag{13}$$

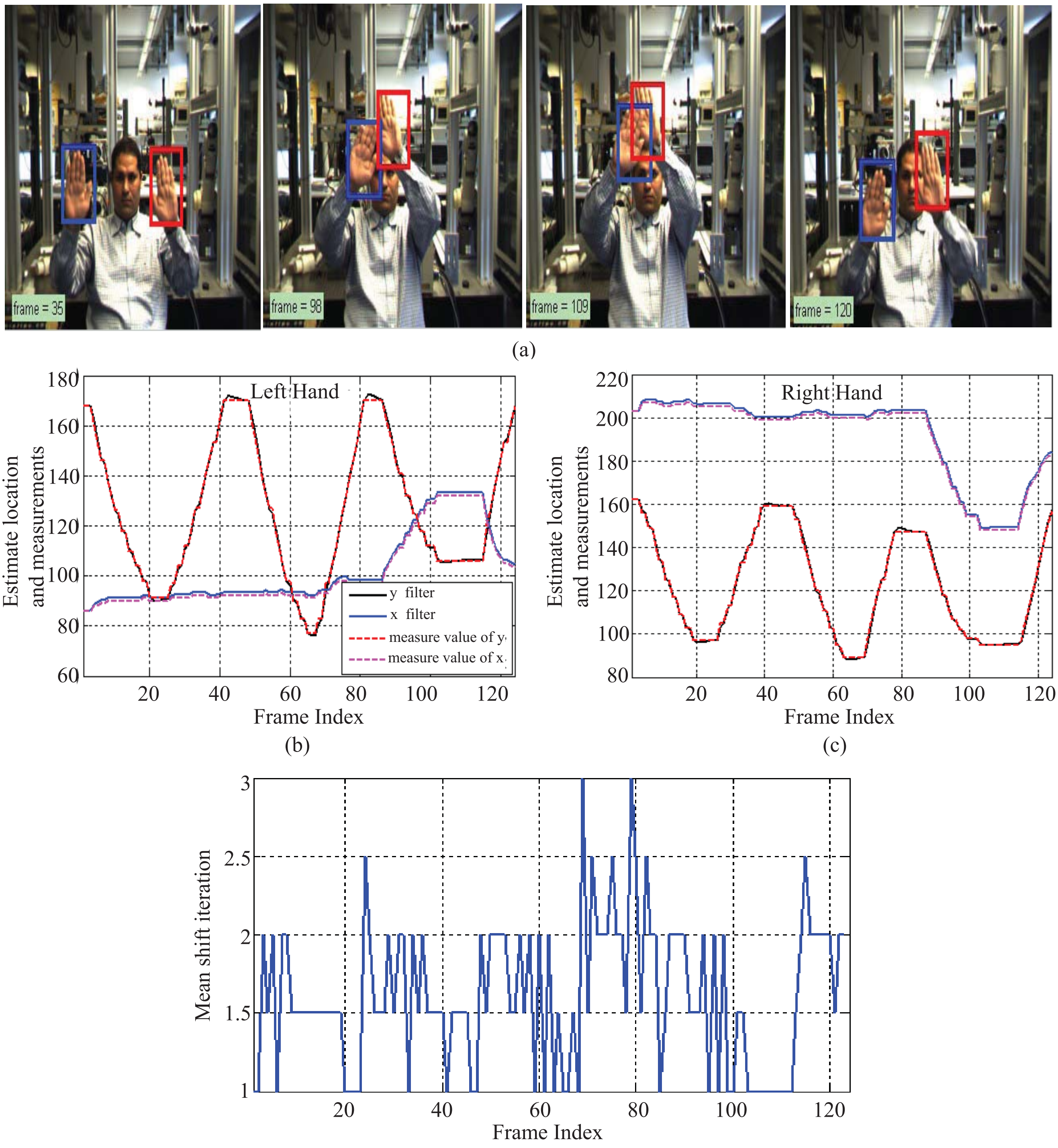
$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \tag{14}$$

$$P_k = (I - K_k H)P_k^- \tag{15}$$

where H represents a measurement matrix with associated noise R that is the error between real and detected location of the hand, K_k is Kalman gain. Process noise represents the accuracy of the model and is determined empirically and the measurement noise is derived directly from the off-line calibration test where an estimate of $\hat{x}_{k-1}$ and P_{k-1} is initialized. We obtain hand trajectory by taking the correspondences of detected hand between successive image frames.

III. Experimental Results

Our proposed method has been tested on various video sequences with various hand shape as well as overlapping and partial occlusion. The hand target and hand candidate histograms has been derived in (C_b, C_r) channels with 16×16 bins. Since the scale of the hand candidate often changes in time, the bandwidth h of the kernel profile in Eq. 4 has been adapted accordingly. The bandwidth is measured in the current



(a)

(b)

(c)

(d)

Fig. 4. (a) Tracking result where at frame 109, each hand is determined correctly notably in case of overlapping and partial occlusion. (b) & (c) Measurement value and location of estimated state for left and right hand sequences, respectively. (d) The number of mean-shift iteration is 1.61 per frame for both left and right hand (suitable for real-time implementation).

frame by running the hand candidate localization three time with small fractions ± 0.1. The best yielding of hand candidate localization is obtained according to the largest Bhattacharyya coefficient. The input images were captured by Bumblebee stereo camera system that has 6 mm focal length at 25FPS with 240×320 pixels image resolution, Matlab implementation. Our experiment focused on the hand tracking only rather than the work presented in the reference [4], because our focus is to sign language recognition and hand gesture spotting in the field of HCI. The examples of tracking are very convincing, because in many video samples, the hands are posed in a natural pose that maximizes visible surface area: the hand closed or semi-closed or completely flat and also the fingers of the hand close to each other and sometimes divergent. Moreover, both hands palm at different orientations of the camera. Some video samples also contain the stop movement during the process of tracking and then again re-movement. Fig. 4 shows the successful tracking in the presence of a partial occlusion and overlapping between hands and face (frame 109). The sequence has 122 frames and the mean-shift iterations were computed for both two hands, which proved to be robust to real-time implementation and online tracking. The measurement value and location of estimated state for left and right hand sequences receptively is shown in Fig. 4(b)&(c).

IV. Conclusion

This paper proposes an automatic method to track hand with superior performance and low computational complexity using Mean-shift analysis, Kalman predication and 3D depth map. The proposed method has shown good performance when applied on several video samples containing confusing situations such as partial occlusion and overlapping. The future research will address the full occlusion of hands and hand gesture spotting for sign recognition via multi-camera system.

Acknowledgment

This work is supported by BMBF Bernstein-Group (FKZ: 01GQ0702), Forschungspraemie (BMBF-Förderung, FKZ 03FPB00213) and Transregional Collaborative Research Centre SFB/TRR 62 ”Companion-Technology for Cognitive Technical Systems” funded by DFG.

References

[1] N. Liu, and B. C. Lovell, *MMX-accelerated Real-Time Hand Tracking System*, Proceedings of Image and Vision Computing, pp. 381-385, 2001.

[2] D. B. Nguyen, S. Enokida, and E. Toshiaki, *Real-Time Hand Tracking and Gesture Recognition System*, International Conference on Graphics, Vision and Image Processing, CICC, pp. 362-368, 2005.

[3] T. Nobuhiko, S. Nobutaka, and S. Yoshiaki, *Extraction of Hand Features for Recognition of Sign Language Words*, International Conference on Vision Interface, pp. 391-398, 2002.

[4] D. Comaniciu, V. Ramesh, and P. Meer, *Kernel-Based Object Tracking*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, pp. 564-577, 2003.

[5] D. Comaniciu, V. Ramesh, and P. Meer, *Real-Time Tracking of Non-Rigid Objects Using Mean Shift*, Conference on CVPR, Vol. 2, pp. 1-8, 2000.

[6] G. Welch, and G. Bishop, *An Introduction to the Kalman Filter*, In Technical Report, University of North Carolina at Chapel Hill, pp. 95-041, 1995.

[7] M. Elmezain, A. Al-Hamadi, and B. Michaelis, *Real-Time Capable System for Hand Gesture Recognition Using Hidden Markov Models in Stereo Color Image Sequences*, Journal of WSCG, Vol. 16, No. 1, pp. 65-72, 2008.

[8] M. Elmezain, A. Al-Hamadi, and B. Michaelis, *A Novel System for Automatic Hand Gesture Spotting and Recognition in Stereo Color Image Sequences*, Journal of WSCG, Vol.17, No. 1, pp. 89-96, 2009.

[9] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, *A Hidden Markov Model-Based Continuous Gesture Recognition System for Hand Motion Trajectory*, International Conference on Pattern Recognition (ICPR), pp. 519-522, 2008.

[10] M. Elmezain, A. Al-Hamadi, and B. Michaelis, *Spatio-Temporal Feature Extraction-Based Hand Gesture Recognition for Isolated American Sign Language and Arabic Numbers*, IEEE Symposium on Image and Signal Processing and Analysis (ISPA), pp. 254-259, 2009.

[11] R. Klette, K. SChlüns, and A. Koschan, *Computer Vision: Three-Dimensional Data from Images*, Springer, Singapore, ISBN 981-3083-71-9, 1998.

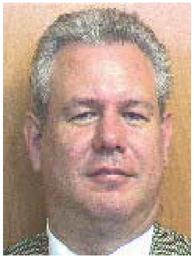
[12] R. Niese, A. Al-Hamadi, and B. Michaelis, *A Novel Method for 3D Face Detection and Normalization*, Journal of Multimedia, Vol. 2, pp. 1-12, 2007.

[13] S. Khalid, U. Ilyas, S. Sarfaraz, and A. Ajaz, *ABhattacharyya Coefficient in Correlation of Gary-Scale Objects*, Journal of Multimedia, Vol. 1, pp. 56-61, 2006.

[14] D. Comaniciu, and P. Meer, *Mean Shift: A Robust Approach Toward Feature Space Analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, pp. 603-619, 2002.

Mahmoud Elmezain was born in Egypt. He received his Masters Degree in Computer Science in 2004. Between 1997 and 2004 he worked as Demonstrator in Dept. of Statistic and Computer Science. Since 2004 he is Assistant lecturer in Dept. of Computer Science, Faculty of Science, Tanta University, Egypt. His current work on a Ph.D. thesis focuses on image processing, pattern recognition and human-computer interaction, at the Institute for Electronics, Signal Processing and Communications at Otto-von-Guericke University of Magdeburg, Germany.

Ayoub K. Al-Hamadi was born in Yemen in 1970. He received his Masters Degree in Electrical Engineering & Information Technology in 1997 and his Ph.D. in Technical Computer Science at the Otto-von-Guericke University of Magdeburg, Germany in 2001. Since 2002 he has been Assistant Professor and 2005 Post-Doc in KFST in Magdeburg. 2004 until 2005 he graduated Professional Training for Industrial Project Management and Start-Up of Business Establishment at University Magdeburg, Germany. Between 2006 and 2008 he has been Junior-Research-Group-Leader at the Institute for Electronics, Signal Processing and Communications at the Otto-von-Guericke University Magdeburg. In August 2008 he became Professor of Neuro-Information Technology at the Otto-von-Guericke University Magdeburg. His research work concentrates on the field of image processing, computer vision, pattern recognition, human-computer interaction, artificial intelligence and information technology. Prof. Dr.-Ing. Al-Hamadi is the author of more than 100 articles in peer-reviewed international journals and conferences.



Bernd Michaelis was born in Magdeburg, Germany in 1947. He received a Masters Degree in Electronic Engineering from the TH Magdeburg in 1971 and his first Ph.D. in 1974. Between 1974 and 1980 he worked at the TH Magdeburg and was granted a second doctoral degree in 1980. In 1993 he became Professor of Technical Computer Science at the Otto-von-Guericke University Magdeburg. His research work concentrates on the field of image processing, artificial neural networks, pattern recognition, processor architectures, and microcomputers. Professor Michaelis is the author of more than 200 articles.