

Detection of human faces and hands using colour and range data

Matjaž Divjak, Damjan Zazula

*University of Maribor, Faculty of Electrical Engineering and Computer Science,
Smetanova ulica 17, 2000 Maribor, Slovenia
E-mail: {matjaz.divjak, zazula}@uni-mb.si*

Abstract : The basis for all high-level human-computer interaction applications is a reliable tracking of human body. In this paper we present a method for non-intrusive, video-based detection and tracking of two of the most important body parts: hands and head. It combines colour and range information acquired by a stereocamera. Skin coloured regions are extracted by segmentation based on colour histogram comparison. Range data is partitioned into several depth intervals by observing its homogeneity. Partial results are fused by assuming that objects are present in both modalities. Additionally, 2D coordinates and motion trajectories of objects are determined. The method is evaluated on real-world video sequences and detailed performance estimation is carried out. With sensitivity at 80 %, specificity at 98 % and recognition rate at 95 %, the method shows great potential for home video-based tracking solutions.

Keywords : video tracking, stereocamera, bimodal segmentation, face and hand detection

Zaznavanje človeških obrazov in dlani s pomočjo barvnih in globinskih informacij

Povzetek. Sledenje gibanju človeškega telesa je podlaga za interakcijo med človekom in računalnikom. V članku predstavljamo postopek za optično sledenje obrazu in dlanem, najpomembnejšima telesnima deloma s stališča komunikacije. Temelji na združitvi barvnih in globinskih informacij, zajetih s stereokamero. Iz barvne slike s primerjavo histogramov izločimo regije kožne barve, globinsko sliko pa razdelimo na več intervalov s homogeno globino. Delne rezultate združimo pod predpostavko, da so iskani objekti prisotni v obeh množicah vhodnih podatkov. Določimo tudi 2D koordinate objektov in njihove trajektorije gibanja. Natančnost postopka ocenimo na realnih video posnetkih. Osemdesetodstotna občutljivost, 98-odstotna specifičnost in v povprečju 95-odstotna učinkovitost razpoznavne kažejo, da je primeren za uporabo v manj zahtevnih sledilnih sistemih.

Ključne besede: slikovno sledenje, stereokamera, bimodalna segmentacija, sledenje obrazu in dlanem

1 Introduction

Various 3D virtual environments are becoming more and more frequent in all aspects of modern computer science.

*Prejet 27. november, 2003
Odobren 24. december, 2004*

Such artificial environments can look and behave very realistically and naturally. But, as soon as one starts to use such a virtual reality (VR) system, a serious shortage becomes obvious: interaction with the virtual environment is usually anything but realistic. In simpler systems, it is generally based on the usage of a keyboard and mouse, two of the most reliable but the least evolved computer input devices. For interaction with more complex environments, specialized input devices that detect the movement of the user's body and transmit it into the virtual space are needed. Such tracking devices usually work on the principle of detecting the movement of a special sensor that is attached to the user's body. Examples of this approach are mechanical, magnetic, acoustic and inertial tracking devices [1]. Although they usually provide very accurate and fast measurements, they are intrusive, too expensive and obstruct the user's movement. For home applications, a simple, non-intrusive technology is needed. Optical tracking devices based on photo sensors and cameras seem to be the most appropriate. They can operate in a completely natural environment, without any additional equipment.

For interaction between a human and a computer, the head and hands are by far the most important body parts. They are involved in almost all our activities. We believe that the information about the position and orientation of head and hands is sufficient for effective navigation in VR worlds. As a result, our research will focus exclusively on fast and accurate detection of head and

hands. General overview of optical detection and tracking of human body can be found in [2] and [3].

To make the tracking system usable in non-professional environments and compatible with widely available home computers, simple and time-efficient data processing algorithms must be developed. It has been shown that multimodal approaches including range images serve the purpose in most cases. The most obvious solution is then a stereocamera [4], although multiple camera systems have also been introduced [5]. Usually, the additional range information is used only as an extra help for segmentation [6] and not for the purpose of estimating the real-world position of objects. In [7] the authors present a stereo system for detecting 3D position of face and hands, but they use infrared LEDs attached to the person's body.

Our intention, however, is to study how reliably and accurately can a fully non-intrusive stereocamera tracking system build trajectories of human hand and head movements. Head position is obtained by searching for a face in a scene. In Section 2, we describe the multimodal image segmentation procedure that uses colour and range information to locate human face and hands in a video frame. In Section 3, we present the fusion of partial results and the procedure for estimating the objects' movement trajectories. In Section 4, several experimental results on real video sequences are shown. Section 5 discusses the results and possible improvements. Final conclusions are given in Section 6.

2 Multimodal segmentation of stereodata

Multimodal segmentation technique combines data from multiple sources to achieve improved accuracy over the use of a single source alone [8]. To capture video, a Videre Design's STH-MD1-C digital stereocamera system [9] was selected among several similar systems. Camera calibration and range calculations were done with a supplied Small Vision System software library [4]. In this way, information on colour and range was available for further processing.

Although influenced by the varying lighting conditions, skin colour is the most useful feature for discriminating body parts from background. Range information is not affected by changes in illumination and can be used to measure the size of objects. However, it is easily confused by poorly textured surfaces [4]. Let us take a closer look at each of the two complementing data sources.

2.1 Colour segmentation

The colour of skin depends on two major factors: the quantity of haemoglobin and melanin [10]. Variations in

skin colour are caused mainly by different melanin concentrations, which affect the intensity of the skin colour, but not its hue. For efficient segmentation the intensity information (luminance) is usually discarded and only the hue information (chrominance) is used.

To separate luminance and chrominance, the RGB colour information must be transformed into a more suitable colour space. Some authors [11] argue that HSV

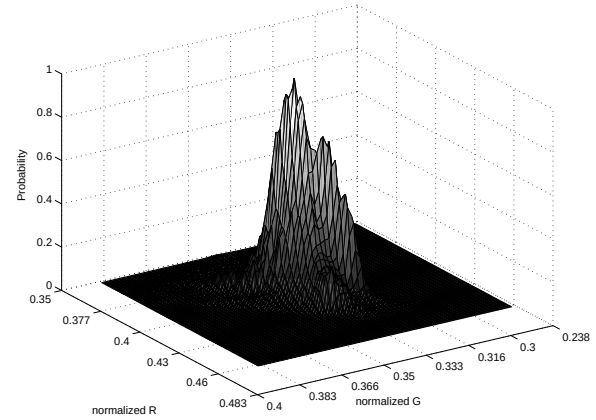


Figure 1. Location of skin colours in a normalized RGB space and their probability of appearance

colour space [12] provides better resistance to illumination variations. However, a comparison in [13] showed that using different colour spaces provided very little variation in skin segmentation results. Based on this, we decided to use the normalized RGB colour space [12]. Fig. 1 shows the region of skin colours in the 2D normalized RGB space, estimated on test images acquired in our experiments.

Our segmentation algorithm is based on a comparison of colour histograms. Prior to segmentation, a few test images are captured. Only the left stereocamera image is used for segmentation. Regions that represent face and hands are marked by hand. Images are transformed into normalized RGB space and then two 2D histograms of pixel colour values are constructed, one for skin regions and one for the background, respectively. Both histograms are normalized to match the bin values with the probability of input pixels having the selected colour.

During segmentation, each pixel of the k -th input image I_k is classified using Bayes' rule of maximum likelihood estimation. Image pixel $\mathbf{p} = [x \ y]^T$ is marked as "skin" if the ratio

$$\frac{P(\text{skin}|\mathbf{p})}{P(\text{background}|\mathbf{p})} = \frac{P(\mathbf{p}|\text{skin})}{P(\mathbf{p}|\text{background})} \quad (1)$$

is higher than 1, otherwise it is classified as “background”. Probabilities $P(\mathbf{p}|\text{skin})$ and $P(\mathbf{p}|\text{background})$ are determined from the skin histogram and the background histogram, respectively. This classification results in a binary image B with all skin pixels marked with ones. In a post-processing step, morphological erosion followed by dilation is used to fill small holes [14]. Additionally, all regions smaller than a user-specified threshold are removed and the remaining regions are uniquely enumerated. This algorithm, presented as Algorithm 1, produces results as shown in Fig. 2b.

INPUT: colour image I_k , skin histogram HIST_SKIN, background histogram HIST_BACK, minimum region area threshold THRESH_AREA

OUTPUT: image J_k with enumerated skin regions

1. Transform image I_k into normalized RGB colour space.
2. Create empty image B .
3. FOR each pixel $\mathbf{p} \in I_k$:

IF HIST_SKIN($\mathbf{p}$) / HIST_BACK($\mathbf{p}$) > 1 THEN
 $B(\mathbf{p}) = 1$
 ELSE
 $B(\mathbf{p}) = 0$
 END IF
4. $B = \text{dilate}(\text{erode}(B))$
5. Remove from B all regions with area smaller than THRESH_AREA.
6. $J_k = \text{enumerate_regions}(B)$

Algorithm 1. Segmentation of skin-coloured regions using colour histograms and Bayesian method

Range segmentation

Range images have the same structure as intensity images, except that each pixel value represents its distance from the camera. The main purpose of range image segmentation is to detect regions that match the actual position of face and hands as accurately as possible. The nature of range data usually makes this task a rather complicated one. If the variations of pixel depths are too small, pixels cannot be classified into separate sets and we have difficulties discriminating separate limbs [15]. This forces us to concentrate on detecting as much of the whole human body as possible and only then discard all the unnecessary image areas, such as the background. More thorough segmentation of the body region can be made later, with the help of colour information [6, 15].

With this in mind, we designed a range segmentation algorithm (Algorithm 2) that uses histogram analysis to find boundaries between the regions of homogenous depth. Firstly, an algorithm based on [16] is used to find the significant peaks of the range histogram. The

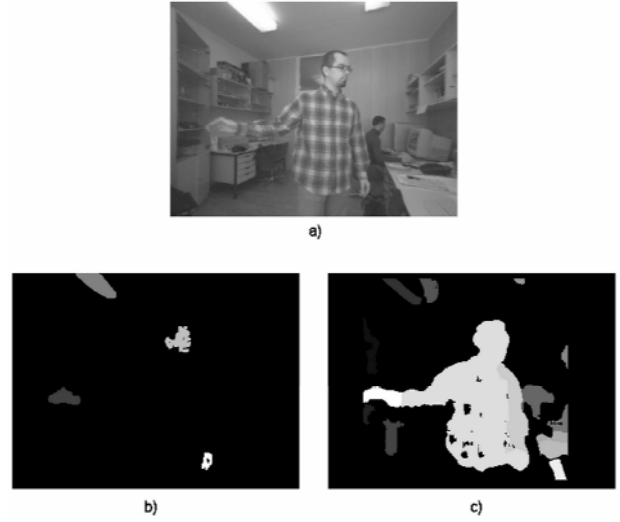


Figure 2. Example of segmentation results (different shades of grey mark different regions): a) input video frame, b) detected regions with skin colour, c) detected depth intervals

positions of the deepest valleys between the peaks are used as threshold values to partition the range image into several depth intervals. The most distant interval is considered the background and is removed. The remaining intervals are processed by Gaussian filter [14] to smooth the data and fill in some missing range values. At the end, regions from all depth intervals are uniquely enumerated and combined into a single image (Fig. 2c).

3 Fusion of multimodal information

Partial results from each modality, presented as images J_k and S_k with enumerated object regions, must be appropriately merged. The merging is based on a presumption that all objects of interest must be present in the images of both modalities. If this condition is not satisfied, the region in question is removed. The process of fusion consists of five main steps:

- search for overlapping colour and range regions
- merge the regions that lie very close together
- remove the redundant regions
- calculate the object positions
- track the positions of the detected objects.

Preliminary results [15] show that skin regions coincide with the actual position of face and hands much better than range regions. Therefore, the shape and size of final regions are determined by colour information, while the range information is primarily used to select only those skin-coloured regions that belong to the human in front of the camera.

The segmentation often results in homogenous input regions being detected as a set of separate, closely positioned regions. This effect is usually noticed on the hand objects, where fingers are often segmented separately from the palm. To rectify this, all colour

INPUT: range image R_k , minimum allowed peak size
 THRESH_SIZE, maximum distance between peaks
 THRESH_DIST, minimum valley depth between peaks
 THRESH_DEPTH, minimal region area threshold
 TRESH_AREA

OUTPUT: image S_k with enumerated range regions

1. Find histogram h of range values in image R_k .
2. Normalize histogram h .
3. Find local peaks of h and store their indices.
4. Treat histogram values with selected indices as a new histogram, find its local peaks and store the indices into list v .
5. Remove low peaks:
 - $a_{\max}$ = maximum amplitude in h
 - $i_{\max}$ = index of the highest peak in h , $h(i_{\max}) = a_{\max}$
 - IF $h(j) / h(i_{\max}) < \text{THRESH_SIZE}$, $\forall j \in v$
 - THEN remove peak index j from v
6. Remove peaks that are too close:
 - FOR every peak $h(i_1)$ and $h(i_2)$, $i_2 > i_1$, $\forall i_1, i_2 \in v$
 - IF $i_2 - i_1 < \text{THRESH_DIST}$
 - THEN remove the index of the lower peak from v
 - END FOR
7. Remove peaks with shallow valleys:
 - a. calculate mean histogram value between every two consecutive indices i_1, i_2 in v : $h_{\text{avg}} = \frac{\sum_{i=i_1}^{i_2} h(i)}{i_2 - i_1 + 1}$, $\forall i_1, i_2 \in v$,
 $i_1 = v(j)$, $i_2 = v(j+1)$, $j \in [1, \dots, \text{length}(v) - 1]$
 - b. IF $\frac{h_{\text{avg}}}{h(i_1) + h(i_2)} > \text{THRESH_DEPTH}$
 - THEN remove the index of the lower peak from v
8. Find the indices of valleys between consecutive peaks in v and store them into list t . Those values are the thresholds for segmentation.
9. Create empty image S_k .
10. FOR $m = 1$ to $\text{length}(t) - 1$
 - a. $B(\mathbf{p}) = \begin{cases} 1, & \text{if } t(m) < R_k(\mathbf{p}) < t(m+1) \\ 0, & \text{else} \end{cases}$, $\forall \mathbf{p} \in R_k$
 - b. Apply Gaussian filter to image B .
 - c. Remove all regions with their area smaller than THRESH_AREA from B .
 - d. Enumerate the regions in B and add them to S_k .

Algorithm 2. Segmentation of range image into several range intervals

regions that are positioned closer than a certain distance and share the same range region are treated as one compound object.

Only the three most promising objects are selected out of the candidate list, the others are discarded. As a selection criterion, the similarity between the mean object colour and typical skin colour is used. The mean colour of object O , $\mathbf{c}_{\text{avg}}(O)$, is calculated as the mean value of all

object's pixels. The most typical value of skin colour $\mathbf{c}_{\text{skin}}$ is defined by the position of the highest peak in the

normalized skin colour histogram, $\mathbf{c}_{\text{skin}} = [r_{\text{peak}} \quad g_{\text{peak}}]^T$.

The resemblance between object O and the typical skin region is evaluated by function ζ :

$$\zeta(O) = |\mathbf{c}_{\text{avg}}(O) - \mathbf{c}_{\text{skin}}|. \quad (2)$$

Three objects with the lowest ζ value are selected as the most probable body parts and constitute the final segmentation result.

2D position $\mathbf{s}$ of object O in the image plane is represented by its geometric centre of gravity. It is calculated as mean coordinate value of the object's skin region pixels.

Using the described procedure, the position of each detected object in a single video frame is acquired. In order to track the movement of objects, we have to find the correspondence between the objects in two successive images, I_k and I_{k+1} . Denote a set of objects in image I_k as O_k . This set contains up to three objects O_l^k , $l \in [1, 2, 3]$, which do not necessarily correspond to objects O_l^{k+1} from the next image I_{k+1} with the same indices. Note that subscript l denotes the index of a detected object in current image and superscript k denotes the image index. Assuming the video frame rate is sufficiently high to make an object displacement in two consecutive frames relatively small, correspondence can be determined by minimum object distance. Object O_l^k matches object O_m^{k+1} if distance

$$d(O_l^k, O_m^{k+1}) = |\mathbf{s}(O_l^k) - \mathbf{s}(O_m^{k+1})| \quad (3)$$

is minimum for each $O_l^k \in O_k$, $O_m^{k+1} \in O_{k+1}$. Once the correspondence of a specific object is uniquely determined, its position is added to the list of the object's past positions. This list constitutes the object's movement trajectory.

4 Experimental results

To test the proposed algorithm on real-world data, we captured several video sequences inside a typical office environment. Videos were recorded at different times of day, so scene illumination was significantly varied. The walls of the room were furnished with wooden panels of light brown colour, very similar to the colour of human skin. The test person was dressed in the casual clothes with no special markers attached. Each sequence consisted of 90 – 140 frames, captured at resolution of 320×240 pixels and approximately 7.5 frames per second. Seven video sequences were assessed for statistical purposes. The parameters of the algorithm were empirically set to the following values:

THRESH_AREA = 150, THRESH_SIZE = 0.05, THRESH_DIST = 15, THRESH_DEPTH = 0.75.

To evaluate the accuracy of segmentation, the detected object regions were compared with reference images that contained hand-marked regions of face and hands. First, a statistical significance of the obtained results was calculated (Table 1). Sensitivity and specificity reveal the percentage of correctly detected reference object positions and the percentage of correctly recognized background area. Positive predictive value describes the percentage of the detected objects' area that truly belongs to the objects of interest. Similarly, negative

Statistical index	Mean value (%)	σ (%)
Sensitivity	79.5	4.7
Specificity	98.5	0.9
Positive predictive value	88.3	4.7
Negative predictive value	97.7	0.9
Efficiency	96.6	1.3

Table 1. Overall statistical significance of the results

predictive value describes the percentage of correctly detected background. Efficiency presents the overall success rate of pixel classification.

Additionally, the area of detected and reference regions was compared using the r_1 and r_2 ratios (Table 2) [17]. The r_1 ratio compares the intersection of regions and the area of reference region, while r_2 is the ratio between the area of intersection and the area of detected region. The region is considered recognized if both r_1 and r_2 are greater than some threshold value. The recognition rate (RR) depends on this threshold. The threshold value of 0.5 corresponds to variations of the object centre by ± 20 % of the object diameter. Higher threshold values force more strict RR estimation. For example, the threshold of 0.6 corresponds to ± 16 % variation and the threshold of 0.7 to ± 12 %. The actual centre positions of objects in test sequences varied on average by ± 3.5 pixels (11.3 %). If correct recognition is considered at the object-center displacement of 10 %, the RR equals 86 %. For contour-based quality estimation, the mean absolute distance d_{MAD} and the Hausdorff distance d_{HAUS} between

Index	Mean value	σ
RR in %, $r_1 = 0.5, r_2 = 0.5$	94.8	10.3
RR in %, $r_1 = 0.6, r_2 = 0.6$	85.5	13.2
RR in %, $r_1 = 0.7, r_2 = 0.7$	73.0	17.9
d_{MAD} in pixels	3.7	8.9
d_{HAUS} in pixels	8.5	9.7
κ	24.1	8.7

Table 2. Average recognition rate and contour-based performance indices, estimated on the test set of seven video sequences

the outer boundaries of detected and reference regions were calculated [17]. The average compactness κ [14] of detected regions was also determined (Table 2).

An example of final face and hand detection is depicted in Fig. 3. Video frames from the same sequence as in Fig. 2 are overlaid with the detected object boundaries, shown in white. In Fig. 4 an example of 2D motion trajectories of the face and hands is shown – a swing with the right hand.

5 Discussion

The presented results show that approximately 80 % of the hand and face area is correctly detected on average. This results in a recognition rate of 95 % with $r_1 = r_2 = 0.5$. The contours of detected objects lie on average 3.7 pixels away from their reference positions, but the maximum distance between the contours is always smaller than 8.5 pixels, while the mean diameter of objects measures 36 pixels. These results are assumed sufficient for a fairly reliable face and hand detection that can be used for simple navigation in virtual environments. As Fig. 4 shows, the obtained data can be used to follow the movement of objects across the image plane. Motion trajectories are very useful for recognition of the user's actions and behaviour. Of course, the threshold selection affects the performance of the segmentation, but the empirically established set of values proved to be suitable for most situations and all our test video sequences were processed with the same parameter set.

Bimodal segmentation proved greatly advantageous. With relatively simple segmentation algorithms and fusion of their results, we obtained significantly better recognition than by using a single source of information. If static background with no skin-like colours is available, face and hands can easily be segmented by background subtraction and skin detection. However, this is very uncommon in real-world scenes. With range data providing supplementary information to colour, robustness of the algorithm is greatly increased. This idea could be exploited further for solving the problem of occlusion, where the range information could help in resolving the overlapping of objects.

As the next step, the range information could be employed to infer about the size of objects, their 3D position and similar. Unfortunately, the range images provided by STH-MD1-C and Small Vision System are very sparse. On average, only 35 % of image pixels contain valid depth information, mainly because of low intensity variance and inconsistent left-right matches [4]. This problem must be addressed in order to enhance the range data information. Another shortcoming is a very poor colour resolution of STH-MD1-C. Our test images contain on average only about 24,000 unique colours, making the variance between certain colours very small and thus limiting the detection process.

6 Conclusions

We proposed a procedure for tracking the position of faces and hands with stereocamera. Experimental results suggest the method is reliable enough to be used for navigation. Although the tracking is adapted to face and hand properties, it can quickly be modified to suit other objects. We are planning to verify and validate our algorithm's ability to detect 3D object coordinates and follow the behaviour of objects in space, thereby expanding its area of applicability.



Figure 3. Example of final segmentation results. The detected region contours are marked in white.

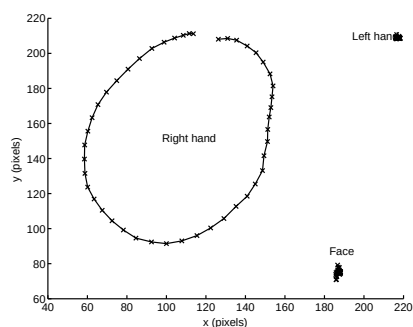


Figure 4. Motion trajectories of three detected objects. The right hand moves in a circular motion while the face and the left hand remain mostly stationary.

7 References

- [1] G. Welch, E. Foxlin, Motion Tracking: No Silver Bullet, but a Respectable Arsenal, *IEEE Computer Graphics and Applications*, Vol. 22, No. 6, pp. 24-38, 2002.
- [2] J. J. L. Wang, S. Singh, Video analysis of human dynamics: A Survey, *Real Time Imaging*, Vol. 9, No. 5, pp. 320-345, October 2003.
- [3] E. Hjelmås, B.K. Low, Face Detection: A Survey, *Comp. Vision and Image Understanding*, Vol. 83, No. 3, pp. 236-274, 2001.
- [4] K. Konolige, D. Beymer, SRI Small Vision System, User's manual, ver. 2.3i, 2002.
- [5] S.L. Dockstader, A.M. Tekalp, Multiple Camera Tracking of Interacting and Occluding Human Motion, *Proc. of the IEEE*, Vol. 89, No. 10, pp. 1441-1455, 2001.
- [6] T. Darrell, G. Gordon, M. Harville, J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, *Int. Journal of Computer Vision*, Vol. 37, pp. 175-185, 2000.
- [7] K. Dorfmueller, H. Wirth, Real-Time Hand and Head Tracking for Virtual Environments Using Infrared Beacons, N. M.-Thalmann, D. Thalmann, *CAPTECH'98*, pp. 113-127, Heidelberg, Springer-Verlag, 1998.
- [8] N. Strobel, Fusion of multisensor data, B. Girod, G. Greiner, H. Niemann, *Principles of 3D Image Analysis and Synthesis*, pp. 309-322, Boston, Kluwer Academic Publishers, 2000.
- [9] Videre Design, *STH-MD1/-C Stereo Head - User's Manual*, www.videredesign.com, 2001.
- [10] M.J. Jones, J.M. Rehg, Statistical Color Models with Application to Skin Detection, *Int. Journal of Computer Vision*, Vol. 46, No. 1, pp. 81-96, 2002.
- [11] K. Sobottka, I. Pitas, Looking for Faces and Facial Features in Color Images, *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*, Vol. 7, No. 1, Russian Academy of Sciences, 1997.
- [12] H. Palus, Representation of colour images in different colour spaces, S.J. Sangwine, R.E.N. Horne, *The Colour Image Processing Handbook*, pp. 67-90, Chapman & Hall, 1998.
- [13] B.D. Zarit, B.J. Super, F.K.H. Quek, Comparison of Five Color Models in Skin Pixel Classification, *Int. Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, pp. 58-63, Greece, 1999.
- [14] M. Sonka, V. Hlavac, R. Boyle, *Image Processing, Analysis and Machine Vision*, London, Chapman & Hall, 1994.
- [15] M. Divjak, *3D motion tracking using a stereocamera*, Masters Thesis, University of Maribor, Slovenia, 2003.
- [16] H.-D. Cheng, Y. Sun, A Hierarchical Approach to Color Image Segmentation Using Homogeneity, *IEEE Trans. on Image Proc.*, Vol. 9, No. 12, pp. 2071-2082, 2000.
- [17] B. Potočnik, D. Zazula, Assessing the Efficiency of the Image Segmentation Algorithms, *Electrotechnical Review*, Vol. 68, No. 2-3, pp. 97-104, 2001.

Matjaž Divjak received his M.Sc. degree in Computer Science from the University of Maribor, Slovenia in 2003. Currently he is a teaching assistant and a postgraduate student at the Faculty of Electrical Engineering and Computer Science in Maribor and is working on his Ph.D. His main research interests include image processing, machine vision and virtual reality.

Damjan Zazula graduated from the University of Ljubljana, Slovenia, where he also received his M.Sc. and a Ph.D. degrees in signal processing. He has worked in industrial research for eleven years and in 1986 joined the Faculty of Technical Sciences, University of Maribor. He is currently holding a position of a Full Professor lecturing on system software, computer system performance, operating systems, and digital signal processing. His main research interests are digital signal and image processing, especially in biomedical applications.