An Integrated Robot Vision System for Multiple Human Tracking and Silhouette Extraction

Jung-Ho Ahn, Sooyeong Kwak, Cheolmin Choi, Kilcheon Kim, and Hyeran Byun

Dept. of Computer Science, Yonsei University, Seoul, Korea, 120-749
{jungho, ksy2177, wxman, kimkch, hrbyun}@cs.yonsei.ac.kr

Abstract. In this paper, we propose a new integrated robot vision system designed for multiple human tracking and silhouette extraction using an active stereo camera. The proposed system focuses on robustness to camera movement. Human detection is performed by camera egomotion compensation and disparity segmentation. A fast histogram based tracking algorithm is presented by using the mean shift principle. Color and disparity values are combined by the weighted kernel for the tracking feature. The human silhouette extraction is based on graph cut segmentation. A trimap is estimated in advance and this is effectively incorporated into the graph cut framework.

Keywords: Object Detection, Tracking, Silhouette Extraction, Robot Vision.

1 Introduction

One of the main goals of virtual reality applications is to support natural, efficient, powerful interaction. If interaction is overly obtrusive, awkward, or constraining, the user's experience with synthetic environments can be severely degraded. To support natural communication, we want to not only track human movements, but interpret those movements in order to recognize semantically meaningful gestures. Gestures are used for control and navigation in CAVEs [11, 12] and in other virtual environments such as smart rooms [13] and virtual work environments [14].

This paper addresses a computer vision system which can detect and track multiple people and extract silhouettes with an active stereo camera in real time. This process assists in human gesture, gait and event recognition for virtual reality applications where accurate positions and silhouettes are needed to monitor and extract the features of body configuration. A prime application of this research is Human Robot Interaction (HRI).

Many computer vision systems with same purpose have been developed [4, 7, 8] but most of these were designed only for surveillance purpose, based on the background subtraction method which used fixed cameras. Unlike surveillance systems, the vision systems mounted on mobile robots are required to deal with camera movements so methods such as the background modeling. Therefore totally different approaches are necessary in these applications, and the efficiency requirements of live processing have restricted us to algorithms that are known to be capable of near frame-rate operation. Because of these problems, proper robot vision

Z. Pan et al. (Eds.): ICAT 2006, LNCS 4282, pp. 575 – 583, 2006.

systems for human behavior analysis have not been presented yet. To track people with a moving camera, Nursebot [1] used laser range sensors rather than a vision-based approach. However such hardware-based approaches are too limited to handle implicit communications in HRI such as gesture, gait and event recognition.

1.1 System Overview

The main contribution of the proposed system is the design of an integrated vision system for mobile robots that can perform multiple human tracking and segment silhouettes with an active stereo camera. Our proposed system is composed of three modules such as human detection, tracking and silhouette extraction. These modules work interactively. Their key features and contributions are as follows:

- The detection module detects when new people enter a scene. Their locations and sizes are determined by using egomotion compensation and disparity segmentation by flood fill algorithm. A detected person is called a *track* and the location of this person can be represented by a bounding box.
- The tracking module updates the locations of all tracks in every frame. In this paper we propose a new fast histogram-based tracking algorithm that works with color and disparity values simultaneously. For multiple tracking we present some strategies to handle the occlusion.
- The silhouette extraction module extracts the silhouette of a person of interest. The selected person is determined by the task manger that controls all operations of the robot. In this paper, we present a robust way of object segmentation method in which we estimate a trimap and effectively incorporated it into a graph cut framework.

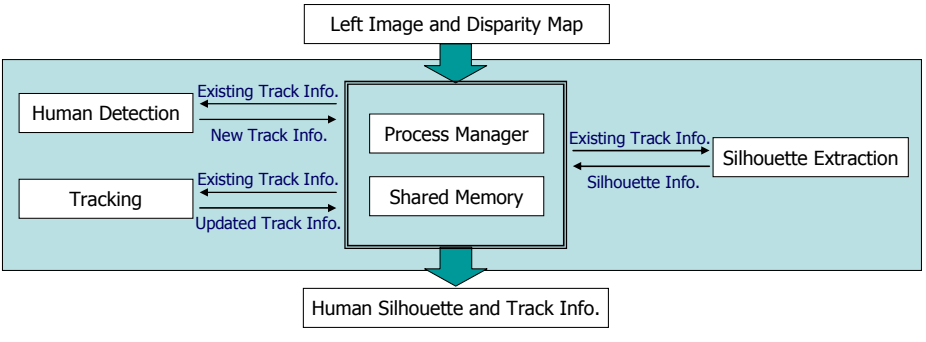


Fig. 1. Overall system flow chart of the proposed vision system

All three modules are robust to camera movements. The input of our system is the left image obtained by the stereo camera and the disparity map that is calculated by the Small Vision System(SVS)[10]. All modules share the track and silhouette information in the shared memory and are controlled by the process manager. Figure 1 shows a flow chart of the proposed vision system.

2 Human Detection and Tracking

The detection and tracking modules exchange the bounding box information obtained from all tracks. In every frame the tracking module updates the track information. Then the detection module observes the outside parts of the bounding boxes of existing tracks. If a new person is detected, a new track is initialized. The process manager associates its bounding box with the track and registers the results in the shared memory.

Human Detection: In general there are two motions that are determined with an active camera; object and background movements. To eliminate the background movements consecutive images can be calibrated by applying the egomotion compensation algorithm [9]. To accomplish this, feature points are generated by the Harris corner detector in every frame. Excluding the outliers, the correspondence between the feature points in the consecutive images can be estimated using the Kande-Lucas-Tomasi(KLT) feature tracking algorithm. Then, the camera egomotion can be estimated using the bilinear transformation model which is defined as:

$$\begin{pmatrix} f_x^t \\ f_y^t \end{pmatrix} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 & a_7 \end{pmatrix} \begin{pmatrix} f_x^{t-1} \\ f_y^{t-1} \\ 1 \\ f_x^{t-1} f_y^{t-1} \end{pmatrix} \tag{1}$$

where ($f_X{}^{t-1}$, $f_Y{}^{t-1}$) and ($f_X{}^t$, $f_Y{}^t$) are corresponding feature points in time *t*. After compensating for background motion, the frame difference between the compensated previous image and the current image can be obtained. The frame difference can detect only certain parts of the moving objects, for example, the legs and arms. To determine the accurate bounding box of the person, it is necessary to perform the disparity segmentation with the flood fill algorithm. The average of the disparity values in the detected moving area can be computed and the area can be expanded by adding the close pixels with similar disparity values.

Human Tracking: In order to track the detected people, a modified mean shift-based tracking algorithm is used. This algorithm works with color and disparity information. The kernel-based target and candidate models can be defined as in [6],

$$\hat{q}_u = C_q \sum_{i=1}^{n} k\left(\left\|x_i\right\|^2\right)\delta\left[b(x_i) - u\right], \qquad \hat{p}_u = C_p \sum_{i=1}^{n_h} k\left(\left\|\frac{y - x_i}{h}\right\|^2\right)\delta\left[b(x_i) - u\right] \tag{2}$$

When using mean shift-based color tracking it is difficult to separate the objects with similar color distribution. In order to solve this problem, we use a disparity-weighted kernel which combines a disparity value with a spatial kernel. We assume that the disparity of the body shows small variations compared with that of the background. Let M be the histogram bin with a maximum frequency of disparity values in the bounding box of a track. If the bounding box is tight enough, the pixels whose disparity values fall into bin *M* can be assumed to be target pixels. The weight value W_v of the disparity histogram bin *v* can be defined as:

$$W_v = 1 - \frac{|M - v|}{M} \tag{3}$$

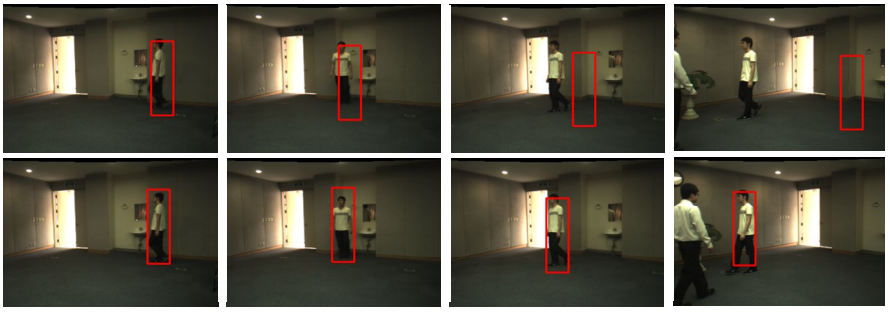
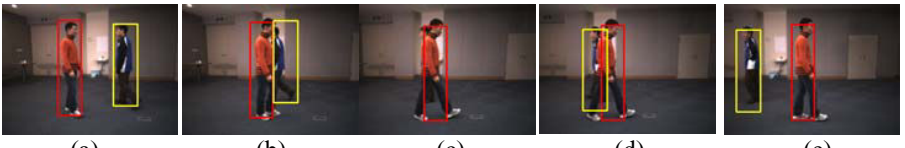


Fig. 2. Comparison of the tracking algorithms. The first row shows the results when using the conventional mean shift-based algorithm [6] and the second row shows the results when using the proposed tracking algorithm.



(a) (b) (c) (d) (e)

Fig. 3. The process of occlusion handling in human tracking. (a) No occlusion in the 150th frame of the test dataset *CM3*. (b) Before occlusion (checking similarity and distance) in the 163th frame. (c) Occlusion (checking similarity on the both sides of the occluding target) in the 166th frame. (d) Recovered target in the 170th frame (e) No occlusion in the 182nd frame.

for $\nu = 1, \cdots, k$. The new target and candidate representations with the disparity weighted kernel in (3) can be calculated by

$$\hat{Q}_u = C_{q,w} \sum_{i=1}^{n} W_{i,v}^{p} k\left(\left\| x_i \right\|^2\right) \delta[b(x_i) - u], \quad \hat{P}_u = C_{p,w} \sum_{i=1}^{n_h} W_{i,v}^{q} k\left(\left\| \frac{y - x_i}{h} \right\|^2\right) \delta[b(x_i) - u] \tag{4}$$

where $C_{q,w}$, $C_{p,w}$ are the normalization constants, and $W^{p}_{i,v}$, $W^{q}_{i,v}$ are the disparity weights for the target and candidate windows, respectively. Figure 2 shows the tracking procedures. For occlusion handling we detected the occluded tracks using the distance among the tracks as well as the histogram similarity after the mean shift tracking. Then, the occluded track was recovered by using the histogram similarities of the both sides of the occluding track. Figure 3 illustrates this procedure. The bounding boxes were accurately obtained by using disparity segmentation as the detection module.

3 Silhouette Extraction Using a Graph Cut with Estimated Trimap

For human silhouette extraction, we used a bounding box surrounding a person. For computational efficiency, this module used the set of pixels $\wp$ within the bounding

box only. We define $z = (z_1, z_2, \ldots, z_{|\wp|})$ as the image where z_n is the RGB color vector for the nth pixel, and $f = (f_1, f_2, \ldots, f_{|\wp|})$ as a binary vector whose component f_n's specifies the labels of either 1 or 0 where 1 represents an object and 0 represents a background. The vector f defines a segmentation. The segmentation of distant moving objects from the active stereo camera presented problems such as low resolution, shadows, poor stereo matching information and instability of color distributions. To overcome these problems we estimated a trimap in advance that assigned each pixel to one of three labels of 0, 1, and -1 where -1 represented the unknown. The estimated trimap was effectively incorporated in a new graph cut framework, and trimap estimation made the hard object segmentation problem easier.

3.1 Trimap Estimation

Our trimap was motivated by the user-interactive segmentation techniques proposed by Boykov and Jolly[2]. They assumed that a user imposes a hard constraints specifying some object and background pixels (*seeds*). For effective trimap estimation we first segment the input color image by mean shift [5]. The mean shift method segments images into homogeneous small regions and the set of regions is denoted by $\Re^t=\{R^t_i\}$ at the tth frame.

Background Seed Estimation: We assume that the person does not move further than d pixels away in each consecutive frame. Thus we dilate the human silhouette of the previous frame using a $d \times d$ square structuring element. The outside of the dilated silhouette is assumed to be the background area. Also, the regions R^t_i's in $\Re^t$ that are touching the bounding box are assumed to be the background area. The union of the two areas is estimated as the set of background seeds B^t.

Object Seed Estimation: Using the mean m^{t-1}_D and the standard deviation s^{t-1}_D of the object disparity values of the previous $(t-1)$th frame, the set of candidate object pixels O^t_D is defined by

$$O^t_D = \{ p \in \wp \mid \; m^{t-1}_D - K_D s^{t-1}_D < d_p < m^{t-1}_D + K_D s^{t-1}_D \}$$

where d_p is the disparity value at pixel p. Using the set O^t_D and the previous object silhouette S^{t-1} we define the object region likelihood L_R of pixel p as 1 if $p \in O^t_D$, w_s if $p \in S^{t-1} - O^t_D$, and 0 otherwise. We select the regions $R \in \Re^t$ if

$$\sum_{p \in R} L_R(p) > w_s \tag{5}$$

where n_R is the number of pixels in R. We denote $\Re^t_s$ as the union of the selected regions. Then we remove the background pixels from $\Re^t_s$ by using background spatial-color probability. A five-dimensional histogram is built on the domain that concatenates the pixel locations and the RGB colors using the background seeds in B^t. We define P_B as the background probability induced by the histogram and the set of estimated object seeds $p \in O^t$ is finally determined if

(a) (b) (c) (d) (e) (f)

Fig. 4. Human silhouette extraction using the estimated trimap. (a) input left image, (b) cropped disparity map (c) object region likelihood L_R, (d) region selection, (e) trimap; white, black, and gray pixels indicate O^t, B^t, U^t, correspondingly, (f) segmented object by the proposed graph cut.

$$-\log P_E(p) > m_C^t + K_C \sigma_C^t \tag{6}$$

and $p \in \Re^t{}_S$ where $m^t{}_C$ and $\sigma^t{}_C$ are the mean and standard deviation of $-\log P_B(p)$'s of $p \in \Re^t{}_S$ and the parameter K_C is related to the level of confidence for the estimated object seeds. Figure 4(e) shows the estimated trimap with the input images in Figure 4(a) and Figure 4(b).

3.2 Graph Cut Framework with the Estimated Trimap

Given an image we try to find the labeling f that minimizes the energy[3]

$$E(f) = \Sigma_p D_p(f_p) + \Sigma_{\{p,q\} \in N} \delta(f_p, f_q) V_{p,q}(f_p, f_q), \tag{7}$$

where N is a standard 4-neighborhood system and $\delta(f_p, f_q)$ denotes the delta function defined by 1 if $f_p \neq f_q$ and 0 otherwise. The smooth term $V_{p,q}$ is defined by

$$V_{p,q} = \exp(-\| z_p - z_q \|^2 / \beta) \tag{8}$$

where β is chosen by the expectation of 2 $\exp(-\|z_p - z_q\|^2/)$ over all $\{p, q\} \in N$. The data term D_p measures how well label f_p fits pixel p given the observed data z_p. We model the object and background color likelihoods of $P(\cdot|1)$ and $P(\cdot|0)$ using Gaussian mixtures in the RGB color space, learned from image frames labeled from earlier in the sequence. In every frame we estimate a trimap that consists of O, B, and U. The set $U = \wp - (O \cup B)$. Then D_p is defined as

$$D_p(f_p) = \begin{cases} -\log P(z_p \mid f_p) & \text{if } p \in U^t \\ (K-c) f_p + c & \text{if } p \in O^t \\ (c-K) f_p + K & \text{if } p \in B^t \end{cases}, \tag{9}$$

where $K = \max_{\{p,q\} \in N} V_{p,q}(f_p, f_q)$ and c is a small number(usually accepted as 1 or 2). Minimization of the energy (3) is done by using a standard min-cut/max-flow algorithm[7]. The estimated trimap may have been wrong in some pixels. The role of c is to give the min-cut algorithm little chance to assign different labels from their pre-estimated labels of the trimap. This effect is shown in Figure 4. For the pixels in the estimated region, the data term D_p is determined by either of the constant K or c

no matter what the colors or disparities are. This is the reason why trimap estimation improves robustness to changes in illumination and camera movements. Figure 5 shows some human segmentation results for gesture recognition under the camera movements and illumination changes.

4 Experimental Results

The proposed system was implemented in C/C++ and was run on Pentium IV-3.0 GHz PC with 1G RAM. The test videos were acquired with a the pan-tilt stereo camera and multiple people entered an indoor scene. Each video contained 700 frames of 320×240 images. The test video $\mathcal{KC}1$ and $\mathcal{SS}1$ were taken by a fixed stereo camera and one person entered the scene. The other five test videos were taken by the pan-tilt stereo cameras. In theses videos two or three people entered the scene, moved around, occluded with each other, and assumed natural poses. Excessive illumination changes occurred in the test video $\mathcal{SY}3$. The average running time of the proposed system was about 9 fps.

Performance Evaluation of the Detection and Tracking Algorithm: Performance was evaluated by using the ground-truth bounding boxes. We segmented the silhouettes of moving people manually at every 10th frame and determined their ground-truth bounding boxes. Table 1 shows the results. In the table *Entrance* refers to the total number of people entering the scene. *Detected* refers to the total number of detected objects, and *True+D* and *False+D* represent the number of correct detections and incorrect detections, respectively. If the detected bounding box overlapped by more than 90 percent with the ground-truth bounding box, we counted it as a correct detection. *Tracking rate* shows the average of the ratio of the number of correct tracking instances to that of the correctly detected tracks. If the center point of the tracked bounding box fell in the middle half area of the ground-truth bounding box we counted it as correct tracking. The average detection rate of the seven test video sets ws 98.31 percent.

Table 1. Performance of the detection/tracking and silhouette extraction algorithms

		CM2	*CM3*	*JH3*	*KC1*	*KC3*	*SS1*	*SY3*
Detection/ Tracking	*Entrance*	4	6	9	1	10	1	7
	Detected	4	8	10	1	10	1	7
	True +D	4	6	9	1	10	1	7
	False +D	0	2	1	0	0	0	0
	Tracking Rate(%)	98.96	99.63	98.83	99.71	99.88	99.91	91.72
Silhouette extraction	*Error rate(%)*	1.06	2.55	1.25	1.36	1.26	0.61	0.42

Performance Evaluation of the Silhouette Extraction Algorithm: Performance was evaluated with respect to the ground-truth segmentation of every 10th frame in each of seven test stereo sequences. Each pixel of each ground truth image was labeled as object, background or unknown. The unknown label occurred when there

were one plus pixel and one minus pixel along the object boundaries to mark the mixed pixels. Ignoring the mixed pixels, the errors were counted. Table 1 shows the averages of the *errors rates* within the bounding boxes. Some results obtained by the proposed vision system are shown in Figure 5.

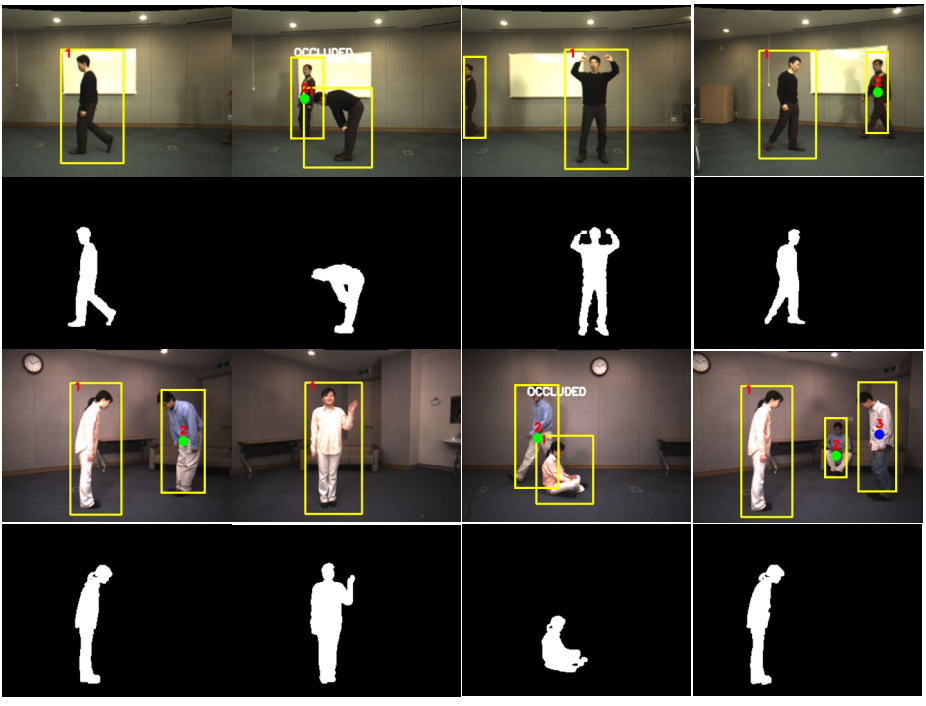


Fig. 5. The results of human tracking and silhouette extraction under an active camera. The images are taken from *CM*2 and *SY*3.

5 Conclusions

In this paper we presented a new integrated robot vision system for multiple human detection, tracking and silhouette extraction when using an active stereo camera. A prime application of this system in gesture, gait and event recognition when working with HRI systems. The background subtraction methods presented easy ways to detect and track people and extract their silhouettes but they were not applicable to mobile robot environments. The proposed system focused on robustness to camera movements and presented several novel vision techniques, such as a fast histogram-based tracking and graph cut segmentation with an estimated trimap, etc. The proposed system showed very good results when working with real sequences at near real-time speed. Further research will involve human identification since the detection module presently only detects moving objects. Mean shift segmentation takes about 60% of the processing time of the human silhouette extraction module so we are also developing faster image segmentation techniques.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019).

References

1. M. Bennewitz, W. Burgard, and S. Thrun: Using EM to learn motion behaviors of persons with mobile robots: Int. Conf. on Intelligent Robots and Systems (2002) 502-507.
2. Y. Boykov and M. Jolly.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. Int. Conf. on Computer Vision, CD-ROM, 2001.
3. Y. Boykov and V. Kolmogorov.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision: IEEE PAMI.: 26. No. 9. (2004) 1124-1137.
4. R.T. Collins, A.J.Lipton, T.Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt and L. Wixson: A system for video surveillance and monitoring: CMU-RI-TR-00-12 (2000)
5. D. Comaniciu and P. Meer. Mean Shift: a robust approach toward feature space analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence: Vol. 24. No. 5. (2002) 603-619.
6. D. Comaniciu, and V. Ramesh.: Kernel-Based Object Tracking. IEEE Trans. Pattern Anal. Mach. Intell. Vol. 25. (2003) 564-577
7. A. Hampapur, L.M. Brown, J. Connell, M. Lu, H. Merkl, S. Pankanti, A.W. Senior, C.F. Shu, and Y.L. Tian.: Multi-scale tracking for smart video surveillance: IEEE Tran. on Signal Processing: Vol. 22. No. 2. (2005) 38-51.
8. I. Haritaoglu, D. Harwood, and L. S. Davis.: W4 :Real-time surveillance of people and their activities: IEEE Trans. Pattern Anal. Mach. Intell..: Vol. 22. No. 8. (2000) 809-830.
9. B.Jung and G.S.Sukhatme.: Detecting Moving Objects using a Single Camera on a Mobile Robot in an Outdoor Environment: Int. Conf. on Intell. Auto. Sys. (2004) 980–987.
10. K. Konolige.: Small Vision Systems: Hardware and Implementation: Eighth International Symposium on Robotics Research, (1997) 111-116.
11. C. Cruz-Neira, D.J. Sandin, and T.A. DeFanti: Surround-screen projection-based virtual Reality: The Design and Implementation of the CAVE: Computer Graphics ACM SIGGRAPH, (1993) 135-142
12. V. Pavlovic, R. Sharma, and T. Huang: Gestural interface to a visual computing environment for molecular biologists: Proc. Int'l Conf. Automatic Face and Gesture Recognition, (1996)
13. S. Shafer, J. Krumm, B. Brumitt, B. Meyers, M. Czewinski and D. Robbins: The new easyLiving project at Microsoft Research: Proc. Joint DARPA/NIST Smart Space Workshop (1998).
14. W. Kruger, C. A. Bohn, B. Frohlich, H. Schuth, W. Strauss and G. Weche: The responsive workbench: A virtual work environment: IEEE Computer, (1995) 28(7) 42- 48.