A Technique for Dynamic Background Segmentation using a Robotic Stereo Vision Head

Jose Prado* Luis Santos* Jorge Dias*

* *Institute of Systems and Robotics, Department of Electrical Engineering and Computers, University of Coimbra, Portugal*

Abstract— **Human-robot interaction approaches like face detection, face recognition, pedestrian detection are widely known in robotics field; however often they lead to performance problems. Additionally, false positive and false negative problems are commonly associated to bad illumination and strong featured images. Moreover background segmentation approaches are frequently used to solve this problem on static camera surveillance. Though all these approaches are unable to effectively deal with the constant background changes that certainly happens when the camera sensor is installed on a mobile robot. Hence, in this work we propose a *stereo vision dynamic background segmentation* solution to this problem.**

I. Introduction

The human society is becoming closer to the integration of robots into our daily environment. While looking to define the term robot, we found that robot often refers to the physical manifestation of a system in our physical and social space, and as such, virtual characters and/or avatar-based interfaces are not discussed in this work. Likewise while searching for a definition of social robot, the following has been proposed: "A physical entity embodied in a complex, dynamic, and social environment sufficiently empowered to behave in a manner conducive to its own goals and those of its community" [3]. Although we like to encourage the use of our approach to social robots, we prefer to call our robot an interactive-robot, since social robots might have several social aspects that our interactive-robot does not have (like facial expressions, arms and voice).

Several approaches like face detection and face recognition [11], [15], [16], [12], pedestrian detection [10], [9] often have to deal with issues associated to bad illumination and strong featured background. These problems also imply lack of performance because human detection algorithms will frequently analyze the whole image searching for features. Hence we propose a stereo vision dynamic background segmentation (DBS) to reduce the searching space to an *zone of interaction*[1]. In this work we show in section II-A the calibration method used for the stereo camera system, it is explained in section II-B how the horopter calculation proceeds. Further in section II-D we give an example of how face and hand recognition frequently used on gesture recognition algorithms could have better results with our approach. In section II-E we explain how we did implement our robotic head tracker in order to have better interaction with humans. Finally in section III-A, as a study case, we implemented this technique to improve the results of a gesture recognition algorithm based on Laban movement analysis proposed in [13].

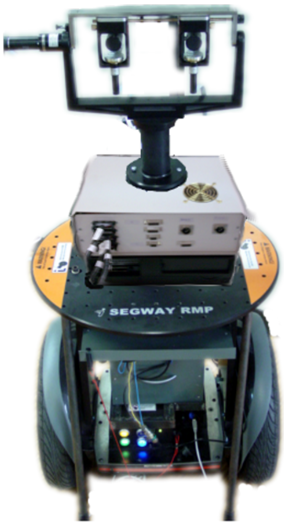


Fig. 1: Four degrees of freedom robotic head mounted on Segway robotic platform body

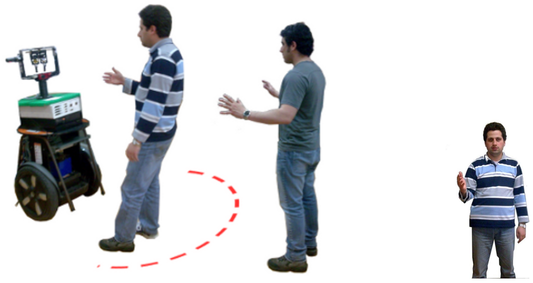
A. Related Work

By using a reference image, a video coding approach with Motion JPEG2000 has previously been developed in the context of road surveillance [17]. Moreover it was shown how the image reference was built during initialization phase. The classical background subtraction technique was used to perform the segmentation of mobile objects. Instead of updating the remote reference with a specific period, [17] presented a technique to update the remote background image by pieces. The updating of the remote reference is triggered when some specific conditions are met, depending on the amount of moving areas.

In [8] an integrated system for smart encoding in video surveillance was presented. Their system aims at defining an optimized code-stream organization directly based on the semantic content of the video surveillance analysis module. The proposed system produces a fully compliant motion stream that contains regions of interest (typically mobile objects) data in a separate layer than regions of less interest (e.g. static background). First the system performs a real-time unsupervised segmentation of mobiles in each frame of the video. The smart encoding module uses these regions of interest maps in order to construct a code-stream that allows an optimized rendering of the video surveillance stream in

[1] *zone of interaction* is the region inside the horopter 3D space (see theoretical horopter definition on section II-B)

978-1-4244-5081-7/09/$26.00 ©2009 IEEE



(a) (b)

Fig. 2: Horopter segmentation schema: a)Noisy scenario, another subject trying to interfere during the interaction; b) From the robot point of view, ignoring the interference

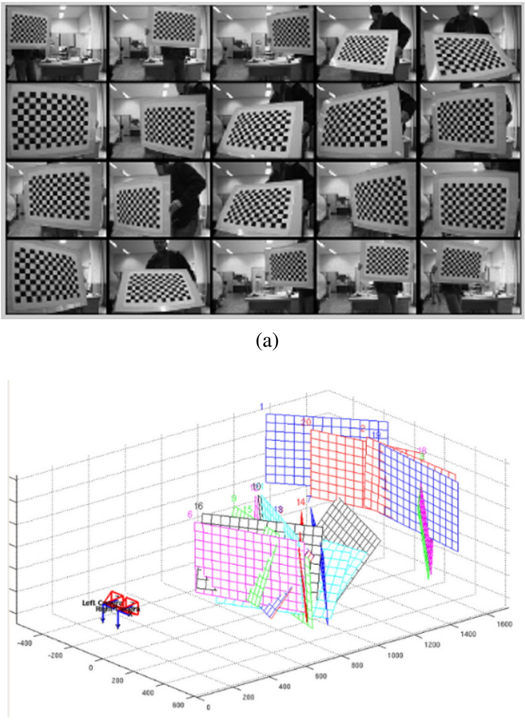
low bandwidth wireless applications, allocating more quality to mobiles than for the background. The integrated system of [8] improves the coding representation of the video content without data overhead. It can also be used in applications requiring selective scrambling of regions of interest as well as for any other application dealing with regions of interest.

On [14] horopter is calculated and vergence control is explained on a stereo-vision-system applied to *tracking by using optical flow*. The robotic stereo head presented on [14] is not mounted on a mobile robot. It is also noticeable by the result images of [14] that the resolution of the disparity map is 36x36 pixels. Our approach focus on applying the system to an *interactive mobile robot capable of recognize gestures*, in this case, the calculation of disparity need to be very fast (we have also the gesture to process) otherwise robot body rotation and translation will easily generate errors into the disparity map. Additionally computational power now allows us to use a better resolution for the disparity map, 80x60 in real time.

B. Motivation

Within the context of human-robot interaction, there is a pre-requisite, the need for the robot to recognize the person with whom it will interact. Usually it is done using a video sensing. Since the system is implemented in a mobile platform, to separate the person from the background demands more complex processing, due to dynamic characteristics of the background. This means that an approach based in static background, as in [17] and [8], is not possible. The challenge was thus to have a robust real time solution for dynamic background segmentation on mobile robotics.

Our approach is then based on the *Geometric Horopter* as will be shown in section II-B. The robot will consider visible objects only if they are inside the *zone of interaction* region (projected on 2D space of camera image plane). In figure 2a it is possible to observe the horopter represented by a semi-circle dashed line at the floor; the subject on the right of the image is purposely in a pose that would interfere on the analysis of several algorithms [11], [15], [16], [12], [10], [9]. Once applying our strategy of DBSH (Dynamic Background Segmentation based on Horopter) the robot will only see the person that is inside the horopter, according to figure 2b.



(a)

(b)

Fig. 3: Stereo camera calibration with Bouguet Matlab toolbox: a)Images with chessboard target used for calibration; b)Reconstructed target positions relative to the camera frame referential

II. Our Approach of Dynamic Background Segmentation

A. Camera Calibration

Camera calibration has been extensively studied, and standard techniques established. For this work, camera calibration was performed using the Camera Calibration Toolbox for Matlab [2]. The C implementation of this toolbox is included in the Intel Open Source Computer Vision Library [5].

The calibration uses images of a chessboard target in several positions and recovers the camera's intrinsic parameters, as well as the target positions relative to the camera, as shown in fig 3b.

The calibration algorithm is based on Zhang's work in estimation of planar homographies for camera calibration [19], but the closed-form estimation of the internal parameters from the homographies is slightly different, since the orthogonality of vanishing points is explicitly used and the distortion coefficients are not estimated at the initialization phase. The calibration toolbox will also be used to recover camera extrinsic parameters and holographic matrix between the two cameras of the stereo system.

Matlab was used to calibrate the cameras, however, we implemented in C with QT, a Graphical interface where the calibration parameters can be manually inserted. So, we can save the Matlab calibration results in a file and then import the calibration using our graphical interface.

B. Horopter

Our approach is based on the *Geometric Horopter*. This technique used stereo vision to produce a *depth map*. It

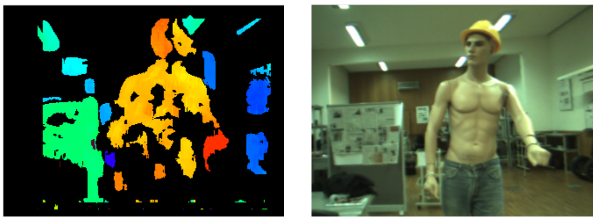

(a) (b)

Fig. 4: Results of depth-map calculation: a)Depth map (*'hot'* colors represent nearest areas, *'cold'* colors represent further ones; b) Dominant eye raw image

is presented in Fig.4a the *depth* map resulting from the application of this algorithm, while the right side shows the image from one of the stereo cameras.

The application of the *horopter* will introduce a definition, the *interaction zone*. As it will be explained in the next section, a circle will be defined, and the inside area of that region is the interaction zone. This means that only objects inside that area are possible of being detected, and thus to interact with the robot.

1) Geometrical Horopter :

a) Properties of ViethMuller Circle: The concept of *interaction zone* has been defined as dependent of a circle. That circle is called the Vieth-Muller Circle, the following properties can be defined (See Fig.4):

- In a pure version eye movement, the fixation point stays on the same ViethMuller Circle. Fig.4 a) illustrates this fact showing how P moves to P' along the Vieth-Muller Circle.
- It the fixation point remains static, the disparity for various points is studied. Disparity is defined as ϕLC ϕR.

b) Theorem 1: If a point Q lies on ViethMuller Circle, its disparity is zero.

As Q moves outside (e.g. point P moves to position Q in Fig.4 a)), ϕL decreases whilst ϕR will naturally increase. However ff point Q moves inside the circle, the opposite relation between ϕL and ϕR occurs.

c) Theorem 2: Disparity is nonzero outside the circumference line of the Vieth-Muller Circle (with opposite signals, depending on whether side of the circle it lies in, outside or inside).

For human vision system, when the disparity has high enough values, the object is seen in double (one from left eye and the other from right eye). This phenomenon is called *Diplopia*. The maximum disparity prior to the diplopia even is defined as *Panum's Fusional Limit*.

d) Calculating Disparity: The ϕL and ϕR are made by line of sight with the straight ahead direction. The $GazeAngle$ γ (see Fig.4 a) and $VergenceAngle$ μ (see fig. 6) are defined as

$$\gamma = \frac{1}{2}(\phi L + \phi R)$$

$$\mu = \frac{1}{2}(\phi L - \phi R)$$

CE represents the cyclopean eye and $(d + \delta)$ is the distance from CE to the target object (see fig. 6).

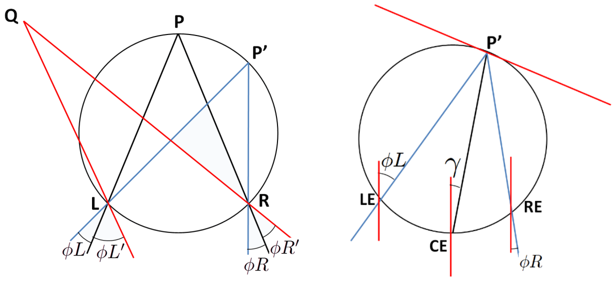


Fig. 5: a) Calculating the Disparity; b) Disparity Properties on Vieth-Muller Circle

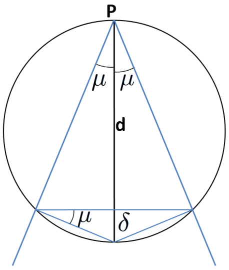


Fig. 6: Simples justification scheme for value γ

The Horizontal Disparity is

$$h = \frac{I\cos\gamma}{d}(\frac{\delta}{\delta + d} + \frac{d\tan\gamma}{\delta + d}x + x^2)$$

and Vertical Disparity

$$v = \frac{I\cos\gamma}{d}(\frac{d\tan\gamma}{\delta + d}y + xy)$$

where (x, y) are cyclopean image coordinates and I is the interocular distance.

e) Theorem 3: $d = I\cos\gamma/\sin 2\mu$

A simple justification can be presented for the value of $\gamma = 0$, as it can bee seen in Fig.6.

$$I/2 = d \times \sin u \times \cos u \Rightarrow d = I\cos r/\sin 2u$$

Having disparity calculated, the resulting depth image (Fig.3 a)) is correlated with the CE image. Pixels that present negative values for disparity, will be assigned zero value (black color pixels). The result is a segmented image where the pixels calculated to be inside the *Vieth-Muller* circle define the visible objects within the circle (the interaction zone). The segmented image (right column of figure 7) results in a region of interest and this region will define the true input pixels for the *face/hand* detector. Consequently the robot will interact only with subjects inside *Vieth-Muller* circle, *i.e.* inside its current horopter.

f) Noise: at the segmented images, the noisy areas exists usually due to homogeneous areas in the original image. Homogeneous areas and also very similar neighbor features of the image can add noise to our depth map and consequently to the final horopter segmented image. Although we have this noise the result is still better for hand and face detection than if you have no segmentation.

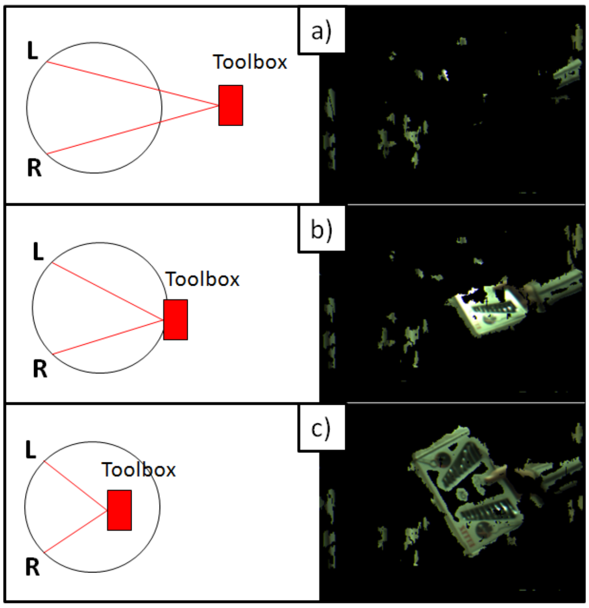


Fig. 7: a) The toolbox is yet outside the Vieth-Muller Circle; b) Toolbox starting to enter the horopter zone; c) The object is fully inside the Vieth-Muller circle, and thus, visible.

C. Hardware Platform

Our robotic head (fig.1) is a common platform consisted by many sensors. In this work we are using two monocular cameras (which compose our stereo vision system) and the four degrees of freedom of our robotic head fig.1 (head pan, head tilt, vergence in eye left, vergence in eye right). The cameras are two AVT Guppy Fire-Wire Cameras. Thus, we calculate stereo imaging for real-time depth-map using the triangulation principle. Camera images are transferred to a PC using the IEEE 1394 (Fire-Wire) bus. The PC is a laptop computer that can be attached in a tray inside the robot body which is an adapted Segway RMP (Robotic Mobility Platform). The Segway RMP adaptions made by us consists basically on four suspended legs to fall avoidance, a strong box for sensor batteries and a tray for the robotic head hardware controllers and laptop attachment.

D. Face and Hand Detection

1) Featured base face detection: A multi-stage classification procedure has been proposed by [18], that reduces the processing time substantially while achieving almost the same accuracy as compared to a much slower and more complex single stage classifier. Later [6] extends their rapid object detection framework in two important ways: Firstly, their basic and over-complete set of haar-like feature is extended by an efficient set of 45° rotated features, which adds additional domain-knowledge to the learning framework and which is otherwise hard to learn. These novel features can be computed rapidly at all scales in constant time. Secondly, [6] derive a new post-optimization procedure for a given boosted classifier that improves its performance significantly.

More recently Bau-Cheng Shen and Chu-Song Chen proposed a new method to retrieve similar face images from large face databases. The proposed method extracts a set of Haar-like features, and integrates these features with supervised manifold learning. Haar-like features are intensity-based features. The values of various Haar-like features comprise the *rectangle feature vector* (RFV) (detailed on [15]), to describe faces. Compared with several popular unsupervised dimension reduction methods, RFV is more effective in retrieving similar faces. To further improve the performance, [15] combine RFV and a supervised manifold learning method and obtain satisfactory retrieval results.

2) Skin color hand detection: According to [7], skin color can provide a useful and robust cue for human-related image analysis, such as face detection, hand detection and tracking, people retrieval in databases and Internet, etc. The major problem of such kinds of skin color detection algorithms is that it is time consuming and hence cannot be applied to a real time system. To overcome this problem, Tarek [7] introduced a fast technique for skin detection which can be applied in a real time system. In this technique, instead of testing each image pixel to label it as skin or non-skin (as in classic techniques), Tarek [7] suggested to skip a set of pixels to improve performance. The reason of their skipping process is the high probability that neighbors of the skin color pixels are also skin pixels.

3) Our solution: In this work we combined face and hand detection algorithms with the horopter dynamic segmentation. We firstly do the dynamic background segmentation, hence it is only necessary to slide on the remaining pixels; this significantly increases the detection performance. Thus we have very fast (10 fps) results on the segmentation plus detection. The gesture recognition algorithm proposed in [13] assumes always the same default initial position for face and hands, later on the process it tracks the real position; this approach implies on performance lost during godfather[2] localization. Thus, in order to save start up time, our choice was to firstly detect the face and the hands position with the algorithms previously mentioned and give this as input to the gesture recognition algorithm.

The red oval on *fig. 11 c)* is an approximation of the search region. It is observable on the *right c)* image that there are areas with skin color on the wall and floor, so if the full image was passed to the hand algorithm hand false positives would certainly occurs. Furthermore similar errors could happen for the face algorithm if the background was strongly and randomly featured.

E. Tracking

If a subject be inside horopter for some seconds, the robot will elect this subject to be it's godfather. Let's call godfather the human elected for interaction with the robot. Hence the robot locate his face and hands as explained in section II-D. As our robot is an interactive robot, we want it to track the godfather while he moves also. In order to have an intuitive interaction it is necessary that the man see the robot looking to him; or, in other words, the robotic head needs to move targeting at the center of the subject head.

In homogeneous coordinates consider an image point $P(x, y, z, 1)$, after normalization $P(u, v, 1)$; knowing focal length f from camera intrinsic calibration, d from horopter calculation (see fig. 6) and a empirically found multiplier

[2] godfather is defined on [13] as the person whom the robot is supposed to interact

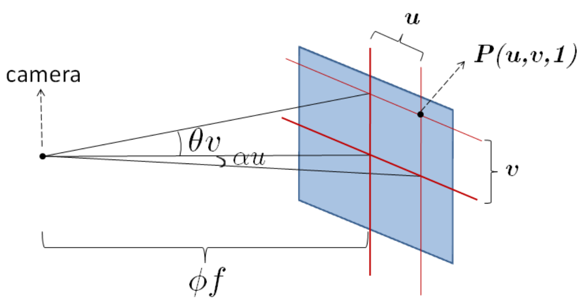


Fig. 8: αu= pan ,θv = tilt — Tracking angles to the robotic head

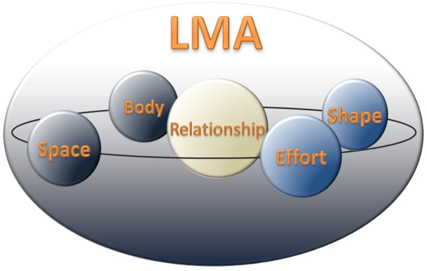


Fig. 9: The five LMA components

λ. We have: $\phi = d * \lambda$. Due to the fact that u and v are initially in pixels while d and f are initially in centimeters, the multiplier $\lambda$is necessary.

Then, as it is visible on fig. 8, $tan\theta v = \frac{v}{\phi f}$ and $tan\theta u = \frac{u}{\phi f}$.

III. Applications of Dynamic Background Segmentation to Interactive Robotics

As mentioned on previous sections one of the principles we are focused in is *interaction*. The interaction scheme can be simplified and thus divided in two stages:*Whom* to interact with; *How* to interact. The whom question as been described throughout sections II-B to II-E. This section will give a general overview on the how.

A. Laban Movement Analysis

Rett J. in his work [13], investigated the possibility of using Laban Movement Analysis (LMA) to classify human movements. Laban Movement Analysis, is a descriptive language of dancing movements. It was developed by Rudolf Laban (1879 to 1958), considered by many a pioneer of European modern dance and theorist of movement education. There are some studies related to LMA, but this is particularly interesting, because an interactive robot was developed to serve as a demonstrator of the usability of this technique.

Literature is not in consensus about the number of LMA components. Most notably, the work of Norman Badler's group [21], [1], [20], divides LMA into five components (Fig. 9) that are: *relationship, space, body, shape, effort*. Each of this latter four components deals with a specific aspect of movements. **Non-kinematic** components: *Body* specifies which body parts are moving, their relation to the body center; *Space* deals directly with the trajectory executed by the body parts while performing a movement. Within the **Kinematic** ones there are: *Effort* which deals with the dynamic qualities of the movement, and the inner attitude towards the use of energy; *Shape* (emerging from *Body* and *Space*) is focused on the body itself. Then we have *Relantionship* that appears as the less explored component, and describes the interaction with oneself, others and the environment. Some literature only considers the first four mentioned components [4].

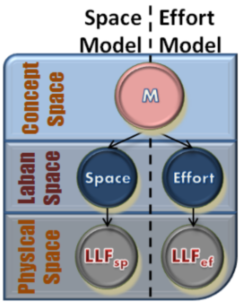


Fig. 10: LMA Global Model

Movement	Interpretation	Action
Circle	turn 360º	Rotation
Pointing	Acknowledgment	Perform Action
Wave Left	Step aside (left)	Move Left
Wave Right	Step aside (right)	Move right
Sagittal Wave	Come closer	Move Forward
Bye-bye	Ignore Gesture, Stop interaction	Switch system off

TABLE I: Movement and correspondent robot actions

B. Interaction

In [13], a Bayesian framework is used as support to the implementation of LMA. The Bayes net implementation is out of the scope of this work, however, Fig. 10 presents the global model for contextualization purposes. Since LMA is composed of four main components, Bayesian approach gives us the flexibility of component integration, i.e. each component can be modeled separately and integrated in a final global model. Also probabilistic approaches allow us to deal with uncertainty and incomplete data, which may also occur, in case tracking fails at some point. As input to the Bayesian network, features are provided as evidences. While movements are being performed, the tracking of body parts generate 2-D trajectories. The features (e.g. angle displacements, vectorial displacements, acceleration, etc.) emerge directly from these tracked trajectories.

A set of movements was learned, and a set of actions was assigned in response, i.e. the robot, through the probabilistic approach, estimated a determined movement through inference of the features, and consequently would react to its assumption. Table I shows the movements and the action responses.

As it can be seen, the actions of the robot are a direct consequence of the movement identification, and this identification relies on the robustness of the tracking algorithm.

IV. Dynamic Results

As already previously stated, when using color tracking schemes, the tracker sometimes loses the target by means of generating false positives for body part identification. This is due to multi-colored backgrounds, which are very common within dynamic scenarios. Thus, by applying the geometric horopter technique to the system used in [13]

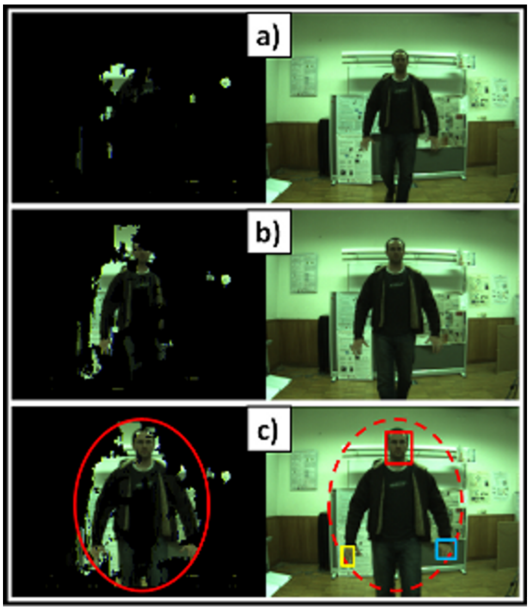


Fig. 11: a) and b) Subject entering in horopter, consequently entering in the field of view of the robot. c) subject is inside the horopter and thus have his face and hands localized.

was able to reduce the search area within the image. The perfect scenario occurs when a perfect bounding box around the human silhouette is generated, as it was theoretically represented on Fig. 2b. There is no shadow effect on this kind of segmentation, since shadows are 2D and will be considered according to the distance of the plane they are projected, the errors noticeable on Fig. 11 takes place due to the small number of correlated points we set up to guaranty a real time application runnable in a ordinary computer. Moreover, homogeneous regions like the illuminated white wall might also generate some erroneous correlated pixels (II-B.1.f). The algorithm slowed its tracking computational time, from deploying 15 frames/second to 10 frames/second, which is not considered critical, as 10 frames is still a good rate. This happened because the old version used one camera only, and after the application of this method, most processing time is dedicated to the computation of the depth image. However tracking results increased dramatically, by reducing the tracking false positives in 87%. To strengthen our tracking rate, geometric constraints were also applied. The results of movement classification are out of the scope of this work and hence, will not be discussed.

V. Conclusion

Dynamic background segmentation is a good strategy to reduce the false positives of several algorithms that are based rather on pixel color or features. By reducing the scope of the searching image to an *zone of interaction* area, the applications of the DBS we proposed here are wide open on the field of Social Robots. In all the cases (haar like features face detection, skin color hand detection, gesture recognition with LMA), our DBS approach shown to improve the performance **and** the results.

References

[1] D. Bouchard and N. Badler. Semantic segmentation of motion capture using laban movement analysis. *In Proc. of Intelligent Virtual Agents*, 1:37–44, 2007.

[2] Jean-Yves Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calibdoc/index.html, 2006.

[3] B. Duffy. The social robot. *Ph.D. Thesis, Department of Computer Science, University College Dublin*, 2000.

[4] Afra Foroud and Ian Q. Whishaw. Changes in the kinematic structure and non-kinematic features of movements during skilled reaching after stroke: A laban movement analysis in two case studies. *Journal of Neuroscience Methods*, 158:137–149, 2006.

[5] Intel. Intel open source computer vision library. http://www.intel.com/technology/computing/opencv/index.htm, 2006.

[6] Rainer Lienhart and Jochen Maydt. An extendeed set of haar-like features for rapid object detection. *ICIP*, 2002.

[7] Tarek M Mahmoud. A new fast skin color detection technique. In *World Academy of Science*, 2008.

[8] J. Meessen, C. Parisot, C. Lebarz, D. Nicholson, and J.F. Delaigle. Smart encoding for wireless video surveillance. In *In SPIE Proc. Image and Video Communications and Processing*, volume 1, 2005.

[9] L R Oliveira and Urbano Nunes. On integration of features and classifiers for robust vehicle detection. *IEEE Conference on Intelligent Transportation Systems (ITSC08)*, 2008.

[10] L R Oliveira and Urbano Nunes. On using cell broadband engine for object detection in its. *IEEE International Conference on Intelligent RObots Systems (IROS)*, pages 54–58, 2008.

[11] Minh Tri Pham and Tat Jen Cham. Fast training and selection of haar features using statistics in boosting-based face detection. *In Proc. 11th IEEE International Conference on Computer Vision (ICCV'07)*, 2007.

[12] Iain Matthews Ralph Gross and Simon Baker. Active appearance models with occlusion. *Image and Vision Computing*, 24:593–604, 2006.

[13] Joerg Rett. *Robot-Human Interface Using Laban Movement Analysis Inside a Bayesian Framework*. PhD thesis, University of Coimbra, 2009.

[14] S. Rougeaux and Y. Kuniyoshi. Velocity and disparity cues for robust real-time binocular tracking. *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 1997.

[15] Bau-Cheng Shen, Chu-Song Chen, and Hui-Huang Hsu. Face image retrieval by using haar features. *Pattern Recognition ICPR*, pages 1–4, 2008.

[16] Barry-John Theobald, Iain Matthews, and Simon Baker. Evaluating error functions for robust active appearance models. *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 1:149–154, 2006.

[17] Theodore Totozafiny, Olivier Patrouix, Franck Luthon, and Jean-Marc Coutellier. Dynamic background segmentation for remote reference image updating within motion detection jpeg2000. In *ICIP International Conference on Image Processing*, volume 1, 2008.

[18] Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE CVPR*, 2001.

[19] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. *In Proceedings of the Seventh International Conference on Computer Vision (ICCV' 99)*, 1:666–673, 1999.

[20] Liwei Zhao. Synthesis and acquisition of laban movement analysis qualitative parameters for communicative gestures. *PhD Thesis, University of Pennsylvania*, 2002.

[21] Liwei Zhao and Norman I Badler. Acquiring and validating motion qualities from live limb gestures. *Graphical Models*, 67:1–16, 2005.