Visual Tracking of Objects Using Multiresolution

Raquel Aparecida de Freitas Mini, Mário Fernando Montenegro Campos

Departamento de Ciência da Computação
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627
31270-010 - Belo Horizonte - MG - Brazil
raquel,mario@dcc.ufmg.br

Abstract. This work presents a methodology for multiple object tracking in image sequences at high video rate using a multiresolution technique based on Haar basis of wavelet transform. The main objective of visual tracking is to closely follow objects in each frame of a video stream such that the object position as well as other geometric information are always known. The idea is to locate and accompany an object based on previously identified features such as salient geometric properties. The methodology was implemented and experiments were conducted where single and multiple objects were successfully tracked at high video rates. Important applications of this system include real time tracking of single or multiple targets.
Keywords: Visual tracking, multiscale analysis, wavelets, Haar basis.

1 Introduction

Tracking is one of the basic features accomplished by even the simplest biological systems, since it is fundamental to self-survival. Studies have shown that human visual systems are equipped with special cells that present high level response to fast variations in the visual field.

The main objective of visual tracking is to closely follow objects in each frame of a video stream such that the object position as well as other geometric information are always known. The idea is to locate and accompany an object based on previously identified features such as salient geometric properties. Important applications include real time tracking of single or multiple targets.

In most cases, tracking demands keeping a given object within the field of view such that a given task might be properly accomplished. This task usually involves the processing of several frames, which in many cases should happen at high rates. There is a large body of literature dealing with the subject, and in the next section we will be presenting a subset which is closely related to the work at hand.

The system proposed in this paper is based on a multiresolution technique to track multiples moving objects. The basic idea is to use the Haar basis of wavelet transform to reduce the resolution of each frame of the sequence. So, the processing is effected on low resolution frame, reducing the computational cost. A great advantage in the use of multiresolution methods is the possibility of determining the dimension of objects to be tracked. Thus, the wrong detection of small objects, possibly due to noise in the image capture process, can be ignored. The output of the system consists of the contours of existing objects in the scene.

1.1 Related Work

Several approaches for object tracking have been described in the literature. The key issue is to determine which objects in the scene need to be tracked and then locate each one of them in each consecutive frame. This is an inherently difficult task because the image of a given object may change from frame to frame due several factors such as differences in illumination and viewing angle. This can be further complicated if the object moves with respect to the camera in planes not parallel to the image plane, and with nonuniform acceleration [9]. Other difficulties include occlusion – where partially or completely covered objects are present in the scene – and the aperture problem. Clearly, some of the listed problems do not present a general solution, but may be overcome in some specific cases.

Several classifications of tracking algorithms have also been proposed, such as the global vs. local approach presented by Hager [6]. In the local type of algorithms, the first frame is entirely scanned in order to determine the position of an object. Then, all subsequent frames will be searched locally around the area where the object was found in the first frame. The range is determined based on object's velocity as-

sumptions and the elapsed time between two consecutive frames. Global approaches need the whole image in order to perform tracking. Those approaches are usually computationally intensive and demand custom hardware to achieve real time performance.

Elsewhere [13], other taxonomies for tracking algorithms has been proposed: low-level and high-level tracking. Low-level approaches use image features that would uniquely identify an object. They are usually very fast and parallel implementations are easily obtained. Problems with this methodology is that other objects in the image may have similar characteristics, and therefore be wrongly tracked. High-level based methods, on the other hand, are more robust, since specific object features are looked for. Algorithms using this approach are computationally intensive and usually require specialized hardware if they are to operate in real time.

Few approaches are concerned with estimating the parameters related to an object's movement in 3D, which is an inherently difficult problem. One of the techniques is based on views taken by more than one camera as projections, and from these projections estimate the objects movement. Three-dimensional tracking is described in the work by Krüger [15].

Huwer *et al.* [10] present a methodology which is capable of tracking ROI (Region of Interest) in a sequence of 2D images based on the projection of the regions histogram. One of the main characteristics of this method lays on the use of special encoding arrangement for the pattern and the distance metric. The ROI is manually chosen, and will be tracked in each frame. A vector containing all line and column histograms projections of the ROI across all frames is then built, which can be easily manipulated. This technique, however, is only able to track a single object, whose image size may not vary too much.

KidRooms [1, 11] is a tracking system based on "closed-world regions". These are regions of space and time in which the specific context of what is in the regions is assumed to be known. This method is capable of simultaneously tracking multiple, non-rigid objects when erratic movement and object collisions are common.

There are instances, however, where objects characteristics are not known a priori. Surveillance and monitoring applications, where a camera may be pointing at a parking lot, or even at hallways of a building, are examples of such applications. Normally, those tasks are carried out by security people watching several TV monitors, when most of the time the scene is static. The work by Wren *et. al.* presents interesting results on the tracking of people [21], which differs from the W^4 system developed by Haritaoglu *et. al.* that performs tracking in $2\frac{1}{2}D$ [8]. For a unified approach to moving object detection in 2D and 3D scenes the reader is referred to a recent work by Irani and Anandan [12].

1.2 Overview of the paper

In the next section the methodology is presented preceded by a quick definition of the Haar basis of wavelets transform. It is also shown how Haar decomposition can be used to perform efficient tracking. Section 3 describes and discusses the results of the experiments performed both in the laboratory environment as well as in unstructured, real street scenes. Finally, Section 4 summarizes what we have done and outlines some areas needing further work.

2 Proposed Methodology

The goals of this methodology is to detect and accompany multiple objects in a frame sequence. It is based on a multiresolution technique using the Haar basis of wavelet transform. The application of Haar decomposition to frames of a sequence will provide images of smaller resolution which can be quickly processed by standard PCs. It is important to point out that the goal here is to simultaneously follow multiple objects in a scene and not necessarily to obtain motion parameters as performed in other works [14, 16].

Wavelet have been thoroughly discussed and used in recent years, however its roots dates back from seminal work by Karl Weierstrass [20]. Mallat [18] has shown how the multiresolution analysis developed in previous work [17] could be viewed as another form of the pyramid algorithms used in an earlier work by Burt [2].

The Haar basis is one of the simplest basis for wavelet transforms [19, 4, 5]. One of its special characteristics is the power of reducing the resolution by using the average of consecutive function values, which in the case of images are pixel (intensity) values.

The Haar wavelet can be defined by the function

$$\psi(x) = \begin{cases} 1 & \text{if } x \in [0, 1/2) \\ -1 & \text{if } x \in [1/2, 1) \\ 0 & \text{if } x < 0 \text{ or } x \geq 1 \end{cases} \quad (1)$$

and the set $\psi_{m,n}$ where

$$\psi_{m,n}(u) = 2^{-m/2}\psi(2^{-m}u - n), m, n \in \mathbb{Z}, \quad (2)$$

defines an orthornormal basis of $\mathbf{L}^2(\mathbb{R})$ that is known as the Haar basis.

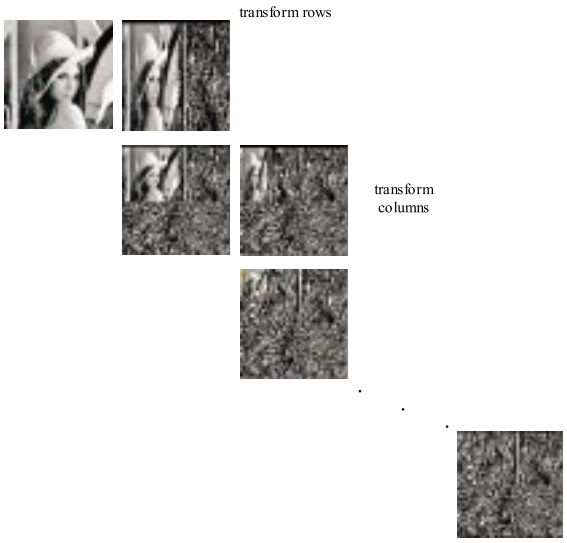


Figure 1: Haar basis decomposition.

Figure 1 illustrates the process of image decomposition using the Haar basis. In this kind of decomposition the operations are alternated between rows and columns. First, it is performed one step of horizontal pairwise averaging and differencing on the pixel values in each row of the image. Next, it is applied the same operation on each columns of the result. After this, we have in the upper left corner a smaller resolution image. To continue this operation of reducing the image resolution we repeat this process recursively only on the quadrant containing averages in both directions. Considering that n represents the original image dimension (nxn), for one decomposition, we have a image of dimension $\frac{n}{2^1}$x$\frac{n}{2^1}$. For d decompositions, the image will have dimension equal $\frac{n}{2^d}$x$\frac{n}{2^d}$.

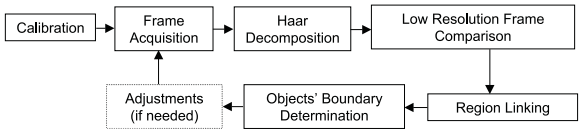


Figure 2: Proposed Methodology.

The tracking methodology is based on low resolution frame comparison to locate the differences that would represent existing objects in the scene. Figure 2 illustrates the schematical diagram of this methodology. It initiates with the *calibration* step which objective is to create the background model. This model represents the scene free of people or whatever movable object and consists of the pixel intensity range to account for background illumination variability.

After the frame acquisition, its resolution is reduced using the Haar basis decomposition. The Haar basis was used due to its implementation simplicity and speed efficiency. We can think as the original image being a function in a V_0 space (function peacewise constant in a interval of length $2^0 = 1$) and a low resolution image as a function in a V_d (function peacewise constant in an interval of length 2^d). Then, our goal is to project a function of the V_0 space to V_d space, where d is the number of decompositions. This number will depend on the dimension of the objects that will be tracked. Using a greater number of decompositions the frame dimension will be smaller and the object dimension will have to be greater in order to be tracked. If n represents the original image dimension, then the time complexity to perform this operation will be $n^2 = O(n^2)$. It is important to point out that we just find the projection of the function in V_d but we do not calculate the wavelet coefficients.

The second part of this methodology is responsible for comparing the low resolution image produced in the previous step with the background model. The result of this comparison is a binary image, called *difference matrix*, which bits are set to one only where current image and background model differs, and bits are set to zero otherwise. This matrix will indicate where to search for a moving object. In this step, the time complexity will depend on the number of decompositions performed in the image, and is equal to $\frac{n}{2^d} \times \frac{n}{2^d} = \frac{1}{2^{2d}} n^2 = O(n^2)$ where d is the number of decompositions.

The third part is responsible for grouping the pixels that belong to the same moving objects in the scene. In this step we use a variation of the *centroid linkage region growing* algorithm that is used in image segmentation [7]. The time complexity of this step is given by $\frac{n}{2^d} \times \frac{n}{2^d} \times 4 = \frac{4}{2^{2d}} n^2 = O(n^2)$.

In the last step, it is enough to find the contour of existing objects in the scene. It is used a kind of chain code to execute this task. The time complexity of this step is given by $\frac{n}{2^d} \times \frac{n}{2^d} \times 2 = \frac{2}{2^{2d}} n^2 = O(n^2)$.

The step called *adjustments* is executed just if is necessary and its objective is to update the background model in order to account for variations of illumination.

Clearly, the final complexity is still $O(n^2)$. However, it is important to note that the multiplying constant is given by $(1 + \frac{7}{2^{2d}})n^2$. This constant takes the small value of 1.11 for a number of decompositions $d = 3$.

One of the main advantages in using Haar decomposition for tracking moving objects is the possibility of determining the objects dimension to be followed. This determination is made by the number of decom-

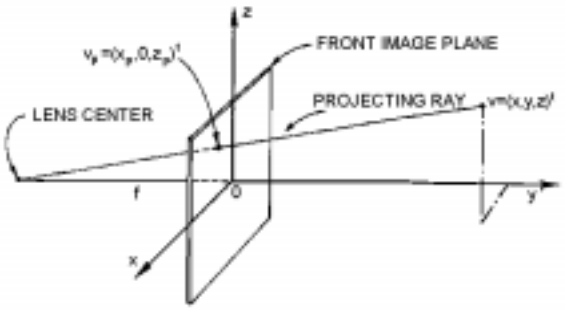


Figure 3: Camera Model [3].

positions used to generate the low resolution image. The bigger the number of decompositions, the lower will be the observed frequencies, therefore, the greater will have to be the object dimension so that this is tracked. More exactly, an object is followed if its dimension is bigger than the size of pixels of the low resolution sampling. For example, if d represents the number of Haar decompositions, each pixel in the low resolution sampling will occupy an area equivalent to 2^dx2^d pixels of the original image. Hence, for an object to be tracked, it will have to occupy an area bigger than 2^dx2^d pixels in the image.

If we are interested in finding the physical dimensions of the object so that it can be tracked, it is necessary to construct a frame acquisition model. Consider a perspective transformation model as illustrated in Figure 3 [3] and the following parameters: f = focal distance, y = object/camera distance, $v(x, y, z)$ = three-dimensional point, $v_p(x_p, 0, z_p)$ = projection of v onto the frame plane, w = CCD pixel width, h = CCD pixel height e d = number of decomposition. Then, the equations are

$$\begin{array}{l} \frac{x_p}{f} = \frac{x}{f+y} \Rightarrow x = \frac{x_p(f+y)}{f} \\ \frac{z_p}{f} = \frac{z}{f+y} \Rightarrow z = \frac{z_p(f+y)}{f}. \end{array} \quad (3)$$

Considering that in the low resolution image $x_p = 2^d w$ and $z_p = 2^d h$, we can find the threshold value that will determine which objects will be tracked in real world dimensions:

$$\begin{array}{l} x = \frac{2^d w(f+y)}{f} \\ z = \frac{2^d h(f+y)}{f} \end{array} \quad (4)$$

Therefore, for an object to be tracked, it will have to be greater than $x \times z$ in real world dimension.

3 Experimental Results

Two sets of experiments were conducted. In the first set, the system tracked the end effector of a PUMA 560 robot executing predefined trajectories in space. In the second set, the camera was pointed to a street on campus, where people and cars pass by. In both experiments the size of the original image is 256 by 256 pixels.

Figure 4: Setup for the experiments. The camera's image plane is parallel to the plane containing the robots displacements.

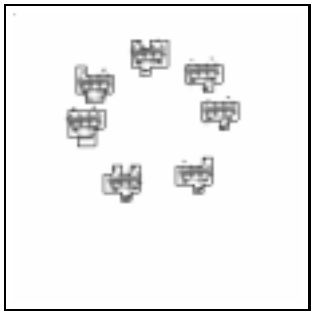
All experiments were performed on a 233MHz Pentium machine with 64Mbytes of RAM, running Windows 98 under interactive load only. Programs were developed using Borland C++ 5.0. Images were acquired with a Sony XC-77 video camera using a lens of f=25 mm. The video output from the camera was fed to a DT3155 very low jitter frame grabber at 30 frames per second rate.

3.1 Tracking known targets

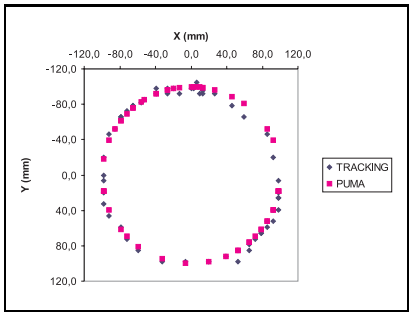
In order to test the methodology and verify the overall accuracy of the system, two experiments using a robotic manipulator were conducted. The PUMA manipulator has 6 degrees of freedom, repeatability = 0.1 mm and maximum linear velocity of 468 mm/sec. The experimental setup can be seen in Figure 4. In each experiment, the end effector of the manipulator was programmed to perform predefined spatial patterns, while the system tracked the end effector's movements. Several runs were executed, and the results were compared to the movement actually performed.

In the first experiment, the robot was programmed to describe a circular pattern with a radius of 100mm, and the distance between the camera and the robot's end effector was 3.17m. Several runs were made with different speeds, from 10% to 35% of maximum speed (46.8mm/s to 163.8mm/s). Figure 5(a) depicts several instances of the PUMA end effector as seen by the

camera and tracked by the system. The illumination was held constant, and no other objects were moving in the scene (the robot's links were covered with neutral color material so that they were not visible during the experiment).



(a)



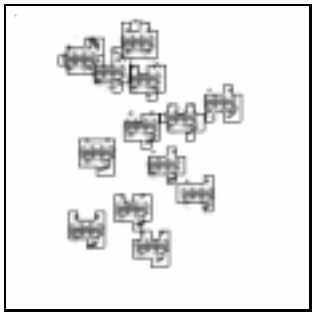
(b)

Figure 5: End effector as seen by the camera: (a) Actual robot end effector displacement as tracked by the system for circle of radius of 100 mm and (b) Points tracked by the system superimposed by the actual points described by the robot's end effector.

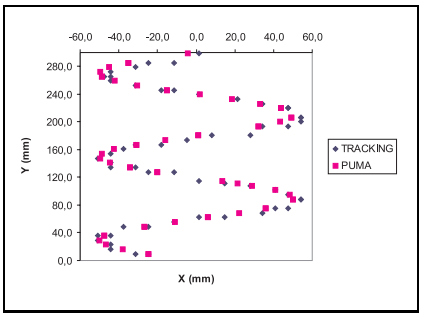
Figure 5(b) shows the points determined by the tracking system during the robot's movement superimposing the actual points. Later the measured errors will be discussed.

For the second pattern the robot was programmed to perform a sinusoidal movement of amplitude 50mm and total spacial displacement of 120mm. In this experiment, the distance between the camera and the robot was 3.25m. Figure 6(a) depicts the robot's end effector being tracked by the system. As in the case of the circle, the illumination was held constant, and no other objects were moving in the scene.

The points determined by the tracking system during the robot's movements can be seen in Figure 5(b)



(a)



(b)

Figure 6: End effector as seen by the camera: (a) Actual robot end effector displacement as tracked by the system for sinusoid of 50mm amplitude and overall span of 120mm and (b) Points tracked by the system superimposed by the actual points described by the robot's end effector.

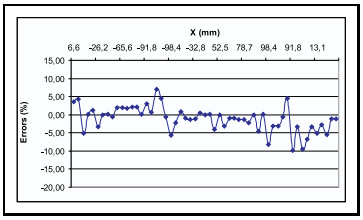
superimposing the actual points. Once again, one can see that the system provided good results.

The tracking system presented very good results in terms of accuracy, with average error values around 3.5%. The errors shown in Figure 7 are from one run of the circle and sinusoid experiments, both at 10% of maximum robot speed.

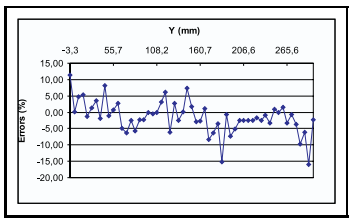
3.2 Tracking real world objects

After the performance of the system has been evaluated in controlled lab conditions, the camera was positioned pointing down at one of the sides of the building, and several sequences were acquired and processed in real time. Figure 8 shows some frames of the taken sequences. The system was able to track the moving objects that appeared on the scene.

It is important to point out one difficulty encountered by this methodology when two objects are too



(a)



(b)

Figure 7: Typical errors presented by the system when tracking (a) a circle and (b) a sinusoid.

(a) t=0 (b) t=0,55sec. (c) t=1,10sec.

(d) t=1,59sec. (e) t=2,14sec. (f) t=3,35sec.

Figure 8: Image Sequence.

close. In this case they tend to be considered as one object, but as the sequence continues, they are naturally separated. This can be seen in frames (f) to (h) of Figure 9.

In all experiments the system took about 18ms to process each frame.

4 Conclusions

We have presented a simple and fast methodology for tracking multiple objects in cluttered scenes based on multiresolution analysis. Previous knowledge of the objects to be tracked is not used. The system detects and follows any object which dimensions are greater than a threshold value previously determined. The Haar basis was chosen due to its simplicity and because it produces fast implementations. Its use to reduce the resolution of each frame was interesting for resulting in a reduction of the computational cost.

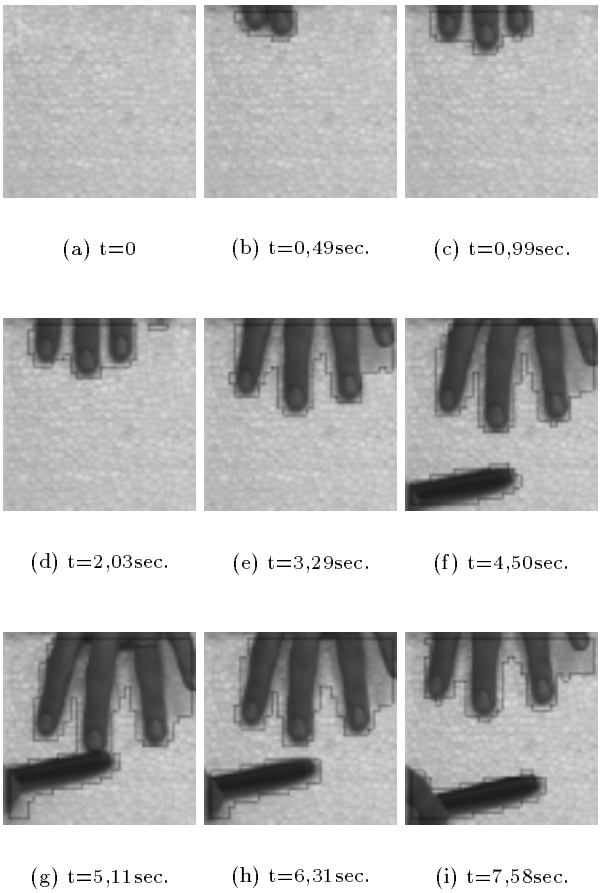
The use of a multiresolution technique for tracking has several advantages when compared to working with the image itself. Many details (such as high spatial frequency components) contained in an image may be of little interest when all that is needed is to follow one or a group of "moving" pixels in a sequence of frames. Furthermore, the scale can be chosen accordingly, either by manually pre-setting or by automatic means. There is no need to do any further processing in the image such as edge detection or binarization, since the algorithm works with the raw image. However, the method suffers from many of the problems that also occur in other methodologies such as occlusion.

The technique was validated with real data both from the lab (controlled environment) and from outdoor scenes (unknown environment). Even though error estimation was not feasible in the second experiment, the first two sets presented average error of about 5%. Both experiments took about 18ms to process each frame.

Surveillance and supervision of remote areas are among the most common applications for the proposed methodology. The detection of an object in the scene can be used to trigger an alarm indicating the existence of intruders in the monitored environment. This is a good example of a task that is better executed by computers because a monitoring work for a long time causes the observer to be tired and loose his concentration.

Another interesting characteristic is that with small modifications this methodology can be used for video data compression. The knowledge of an object motion allows the prediction of its position in successive frames, removing the need of retransmiting identical frame data and leading to a reduction in the bit rate required to transmit the video.

Improvements of this methodology are happening in two main directions. In one of them, it is necessary to improve the *adjustment* step in order to make the algorithm more adaptable to the background il-



(a) t=0 (b) t=0,49sec. (c) t=0,99sec.

(d) t=2,03sec. (e) t=3,29sec. (f) t=4,50sec.

(g) t=5,11sec. (h) t=6,31sec. (i) t=7,58sec.

Figure 9: Image Sequence.

lumination variability. On the other direction, some techniques are being studied in order to minimize the problem of jointing two or more objects that move too closely to each other.

Acknowledgments

The authors wish to thank Prof. Paulo Cupertino de Lima from the Department of Mathematics for his help in the mathematical formulations. This work is partially funded by FAPEMIG grant, and CNPq.

References

[1] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, A. Ivanov, A. Schutte, and A. Wilson. The KidsRoom: A perceptually-based interactive and immersive story environment. Technical Report 398, MIT - Media Lab Perceptual Computing, November 1996.

[2] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

[3] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, 1973.

[4] Jonas Gomes and Luiz Velho. *From Fourier Analysis to Wavelets*. SIGGRAPH-ACM, Orlando Florida, July 1998.

[5] Amara Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2), 1995.

[6] D. Gregory Hager and Kentaro Toyama. The XVision system: A general-purpose substrate for portable real-time vision applications. *Computer Vision and Image Understanding*, 69(1):23–27, 1998.

[7] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision, Volume I*. Addison Wesley, 1992.

[8] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4s: A real-time system for detecting and tracking people in 2 1/2d. In Hans Burkhardt and Bernd Neumann, editors, *5th European Conference on Computer Vision - ECCV'98*, volume 1, pages 877–892, Freiburg/Germany, June 1998. Springer.

[9] B.K.P. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1987.

[10] S. Huwer and Niemann H. 2d-object tracking based on projection-histograms. In Hans Burkhardt and Bernd Neumann, editors, *5th European Conference on Computer Vision - ECCV'98*, volume 1, pages 861–876, Freiburg/Germany, June 1998. Springer.

[11] Stephen S. Intille, James W. Davis, and Aaron F. Bobick. Real-time closed-world tracking. In *CVPR'97*, June 1997.

[12] Michal Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):577–589, June 1998.

[13] Michael Isard and Andrew Blake. Icondesation: Unifying low-level and high-level tracking in a stochastic framework. In Hans Burkhardt and Bernd Neumann, editors, *5th European Conference on Computer Vision - ECCV'98*, volume 1, pages 893–908, Freiburg/Germany, June 1998. Springer.

[14] N. Kingsbury and J. Magarey. Wavelet transforms in image processing. In *European Conference on Signal Analysis and Processing*, June 1997.

[15] Stefan Krüger and Andrew Calway. Multiresolution motion estimation using an affine model. CSTR 96-002, University of Bristol - Departament of Computer Science, February 1996.

[16] J. Magarey and N. Kingsbury. Motion estimation using complex wavelets. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2371–2374, Atlanta USA, May 1996.

[17] Stephane Mallat. *Multiresolution representation and wavelets.* PhD thesis, University of Pennsylvania, 1988.

[18] Stephane Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7((11)):674–693, July 1989.

[19] Eric J. Stollnitz, Tony D. Derose, and David H. Salesin. *Wavelets for Computer Graphics.* Morgan Kaufmann Publishers Inc., 1996.

[20] Karl Weierstrass. *Mathematische Werke*, volume II. Mayer & Muller, Berlin, 1895.

[21] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.