ISPRS Journal of Photogrammetry and Remote Sensing, 45 (1990) 227–246
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

Pattern recognition in 3D automatic human motion analysis

G. Ferrigno[a], N.A. Borghese[a,b] and A. Pedotti[a*]
[a]*Centro di Bioingegneria, Politecnico of Milan, Via Gozzadini 7, 20148 Milan, Italy*
[b]*Istituto di Fisiologia dei Centri Nervosi, C.N.R., Via Mario Bianco 9, 20131 Milan, Italy*

(Received February 10, 1990)

ABSTRACT

Ferrigno, G., Borghese, N.A. and Pedotti, A., 1990. Pattern recognition in 3D automatic human motion analysis. In: P. Niederer (Editor), Medical Imaging and Stereometry. *ISPRS J. Photogramm. Remote Sensing.*, 45: 227–246.

The recent developments in the fields of pattern recognition and parellel computation open new possibilities to motion analysis. The quantitative analysis of movement is no longer restricted to the research, but is becoming an important tool for the assessment of patients and the therapeutic evaluation in clinical environment, as well as in industrial applications.

The main features of an automatic system for motion analysis using optoelectronic sensors (special nonmetric TV cameras) and passive lightweight markers (with no limits in number) will be described herein. The system is hierarchically organized on two levels: the first provides for marker recognition and is implemented by a dedicated hardware processor and the second is devoted to the software processing of the spatial coordinates of the markers. Special attention has been paid on the description of the algorithms for 3D spatial resection and intersection. The features of the system lead to high performance with respect to noise rejection, flexibility of set-up, accuracy and easyness of use, even in critical environmental conditions.

1. INTRODUCTION

The analysis of complex free movements has been revealed to be the most adequate tool to understand the general rules underlying human movement. Moreover, this kind of analysis has been recently introduced for an early diagnosis and therapy assessment in motor disorders.

Human motion analysis is a discipline that dates back to the most ancient times when it was a qualitative science, intended to a subjective assessment of the main features of the movement of men and animals. The first testimony can be found in a book written in 344 B.C. attributed to the Greek philosopher Aristotle. The title of the book is self-explaining: "De Motu Animalium".

The same title was adopted two thousand of years later in 1680 by Borelli,

*Correspondence should be addressed to A. Pedotti.

0924-2716/90/$03.50

a Renaissance scientist who observed for fifty years the way animals moved and tried to put his observations in a scientific form. Leonardo da Vinci too, in the same period, studied and drew pictures about bird flight and tried to understand motion by observing body structure. However, all these approaches were naive and could not be in any way quantitative until Luis Daguerre in 1839 invented the photographic process and, a few years later, cinematography became available. The first graphic representations of human movements date back to Muybridge (Muybridge, 1901) in United States and Marey (Marey, 1902) in France. Although movement analysis, as we conceive it today, had to wait for the first computers, we can date the first kinematic analysis to the turn of the century. These were conducted by Braune and Fisher in Leipzig and by Bernstein in Moscow (about fifty years later) (Bernstein, 1967; Whiting, 1984). Photography and cinematography possess, still nowadays, the paramount advantage of a very high resolution and of the fact that the whole image is recorded, allowing the recovery of additional information in the future.

However, the enormous amount of work required in manual digitization of the data pushes the scientific world toward an automated approach. Other reasons which support this appraoch are the relatively high inaccuracy of the digitization procedure, depending on the operator skill, and the practical impossibility to obtain a reliable 3D reconstruction.

1.1 Optoelectronic systems for kinematic analysis

Various physical principles have been used to gather the data about the trajectories of body landmarks. For example the angles between the long segments of the body (limbs) can be measured by using electrogoniometers, or accelerations can be measured by accelerometers and the data collected integrated in time. Both these approaches require devices, which are encumbering and constraining, to be attached to the subject. The least obtrusive way to approach this problem is to implement optical means, such as TV cameras or other types of optical sensors, which allow non contact measurements by their nature.

In order to simplify the data collection, the human body has been modelled as a set of rigid links connected by hinges (Fig. 1); this approach is a well established technique and simplifies the most complex steps of image analysis. The relevant body points are marked by suitable objects and the trajectories of these describe the movement of the body.

All the systems based on optoelectronic means are characterized by a common structure consisting in an interface to the environment, an optoelectronic sensor, a signal processor and a computer. These may have different importance within the system, but they are always present.

The first systems developed at the beginning of the seventies adopted active

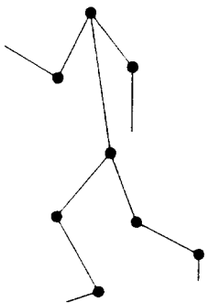

Fig. 1. Example of modelling of human body by a set of rigid links connected by hinges.

markers; LEDs attached to the subject at the selected anatomic points and sequentially lit. The lights are detected by a suitable recording system; in Watsmart and Selspot (Woltring, 1975, 1977; Woltring and Marsolais, 1980), the sensor is a lateral-effect photodiode where the current distribution depends on the position of the center of mass of the incident light. These systems have the advantage to require only a little software processing as the classification problem is overcome by the sequential marker activation; moreover, with few markers a high sampling rate can be achieved (for less than five markers more than 1 kHz). They have, however, some drawbacks in human motion analysis. The equipment which powers the LEDs and the relative connections is heavy and encumbering and may appreciably affect the natural movement. The accuracy of marker coordinates suffers both from sensor response unstability and from the background light and LED reflections, which cause the displacement of the center of mass of the impinging light.

The great advantage introduced by systems adopting passive markers is the possibility to implement motion analysis without any encumbrance on the subject. These systems are based on the registration of the light reflected back by markers illuminated by flashing light sources whose direction of emission is coaxial to TV cameras. The markers are small plastic supports covered with reflective paper with dimensions varying accordingly with the field of view (Fig. 2) or small coloured prisms. These latter are used in the Coda III (Mitchelson, 1975) system. This system has the advantage over the other that markers classification is automatically accomplished by the markers colour coding. On the other hand, since it works with laser beams, its field of view is practically fixed.

The systems with uncoded passive markers allow the analysis of a great number of markers without reducing the temporal resolution. The use of these systems requires a good shuttering and flashing system and more software processing in order to classify the markers. this is the tradeoff associated with a more reliable and unconstrained motion analysis.

Fig. 2. Subject performing a flick. The markers on the body can be easily seen.

The first systems using TV cameras and passive markers, coated with reflective paper, were developed in the middle of seventies. The markers were illuminated by stroboscopes and their reflections detected by means of a simple threshold detection. The Vicon system (Jarrett et al., 1976) belongs to this category, together with the Motion Analyser and other noncommercial systems (Winter, 1972; Cheng et al., 1975). The threshold detection, although allowing to the subject to move freely, presents several drawbacks: first, everything having the same brightness of a marker is erroneously recognized and processed; second, the need of big markers (in order to increase their brightness) and the detection procedure severely limit the true accuracy of the analysis. Moreover, the criticality of recognition combined with the intrinsic inaccuracy (particularly on the time derivatives of the trajectories) make the automatic classification of markers an arduous task.

The recent developments in image processing techniques (pattern recognition, computer vision etc.) and in parallel computation by fast VLSI chips, allowed in the middle of the eighties the development of a new generation of fully automatic systems: for example, the Elite system (Ferrigno and Pedotti, 1985). These systems perform the data collection in several steps that are hierarchically organized and share some features of the visual system of hu-

man beings (Ullman, 19[illegible] Marr, 1982; Poggio and Poggio, 1984). First, the body segments, the movem[illegible] of which is to be analyzed, are recognized. This reduces the complexity of th[illegible] scenary into a few relevant primitives. Second, the two-dimensional coordinates of these primitives are computed and the distortions introduced by the surveying system are also corrected. In order to achieve a three-dimensional analysis more than one sensor is used and additional steps are required: these are the matching of the same primitives in different images and the computation of 3D coordinates.

2. ELITE: A HIERARCHICALLY ORGANIZED MOTION ANALYSER

The Elite system has been designed and realized at the Bioengeneering Centre of Milan during the first half of the 80s. Its innovative feature is the marker detection hardware which works on the shape and size of the markers rather than on their brightness. This characteristic allows a very easy use of the system even in the sunlight. The architecture of the system is hierarchically organized on two levels (Fig. 3).

The lower or first level is represented by the interface to the environment and by the Fast Processor for Shape Recognition (FPSR). The higher or second level is implemented on a commercial personal computer (IBM AT compatible with 80286 or 80386 processors).

2.1 First level: pattern recognition

The first block of the lower level is the Interface To the Environment (ITE); This includes markers of different dimensions according to the field of view. They range from 0.8 cm for a field of view of 2.5 m to 1 mm and less for 20 cm. This flexibility allows the system to be used for very different set-ups. The lighting system is realized by a circular ring of I.R. LEDs coaxial with the lenses, strobed in order to obtain sharp pictures without comet-tail effect. The TV cameras adopted are of the solid state CCD type which allow the best definition in the images at a sampling frequency of 50 or 100 Hz. In addition an electronic internal shutter system contributes to increase the signal to noise ratio and allows the recognition of the markers in the daylight.

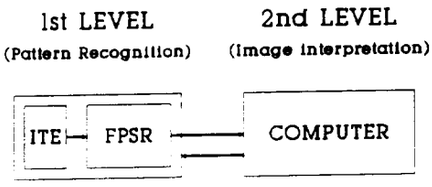

Fig. 3. Architecture of the Elite system.

The second block of the first level is the FPSR. This constitutes the core of the system and performs the recognition of the markers and the computation of their coordinates. The analog TV signal is digitized on a 256×256 useful pixel matrix. Sixteen gray levels are considered and coded in four bits by the A/D converter. The FPSR computes in real-time a two-dimensional cross-correlation between the incoming digitized signal and a reference kernel and drives the ITE with synchronization signals. The kernel is a 6×6 pixel matrix and is designed to achieve a high correlation with the marker shape and a low

−7	−7	−7	−7	−7	−7
−7	−1	0	0	−1	−7
−7	0	7	7	0	−7
−7	0	7	7	0	−7
−7	−1	0	0	−1	−7
−7	−7	−7	−7	−7	−7

Fig. 4. Kernel used for cross-correlation in the FPSR.

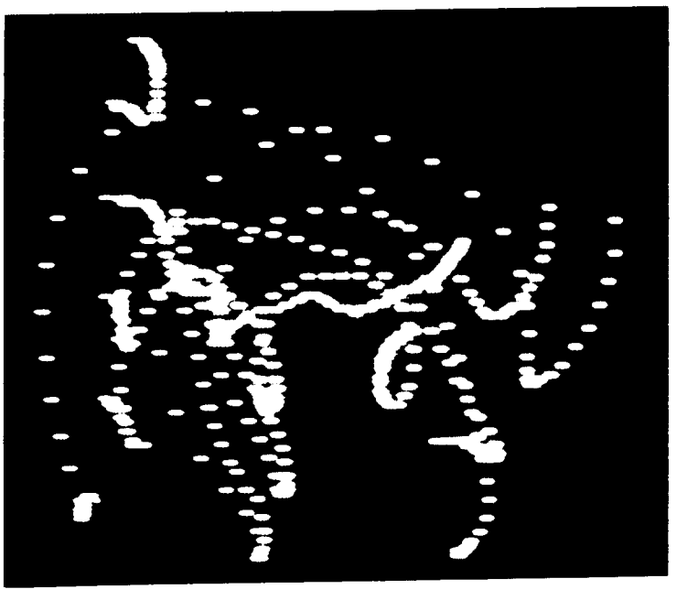

Fig. 5. Ensemble of 2D coordinates recorded by the first level of the Elite system. The movement is the same as in Figure 2.

one with the background. This is accomplished by subtracting from a repetition of the marker a model of the noise. This latter has been modeled as a whatever low spatial frequency shape exceeding the mask size, i.e. a constant value. The actual mask used following this principle is represented in Figure 4. The implementation has been done by using fast VLSI chips in a pipe-lined structure. This first level sends to the computer the 2D coordinates of the markers as registered during the movement (Fig. 5).

2.2 Second level: image interpretation

The second level is software implemented on a traditional serial personal computer and it performs high level processing (Borghese et al., 1988): image interpretation, dealing with marker classification, 2D calibration, 3D intersection and further processing such as filtering, computing derivatives, modelling, etc. Between the first and second level a further step is carried out. By using a coordinate enhancement algorithm, taking into account the cross correlation function, the 2D resolution is increased to 1/65, 000 of the field of view (Ferrigno and Pedotti, 1985)

3 MARKER CLASSIFICATION

The problem of marker classification is shared by all the motion analyzers using passive, non coded, markers. In fact, the computer receives the coordinates of the markers in an order (from top to bottom, from left to right of the image) which is not related to the arrangement of the landmarks of the body. For example, if the shoulder is marked during a flick movement (see Fig. 6), it is the 3rd marker at the beginning, but becomes marker 4 afterwards and then changes again. The purpose of the classification algorithm, is to assign a label to each pair of coordinates so that a landmark is always identifiable, disregarding the detection order. In the Elite system, this problem has been solved analytically by a software procedure of marker tracking. Other procedures can be found in the literature (Jarrett et al., 1976; Taylor et al., 1982) but these are less effective because the predictors they use to forecast the future markers positions are simpler and they do not use a priori information on the markers arrangement and the relationships between them. Our procedure takes into account the model compounded of segments and hinges used to schematize the subject. A model is defined for each different movement analyzed. The model definition procedure asks for the number of markers, of links and for the connections between links and markers. Not all the markers must be linked together and may also exist isolated points. Two other attributes must be set: priorities and weights. The former accounts for the probability that a marker can be hidden by another. A high priority must be associated to a marker that is never supposed to disappear and a low one to another

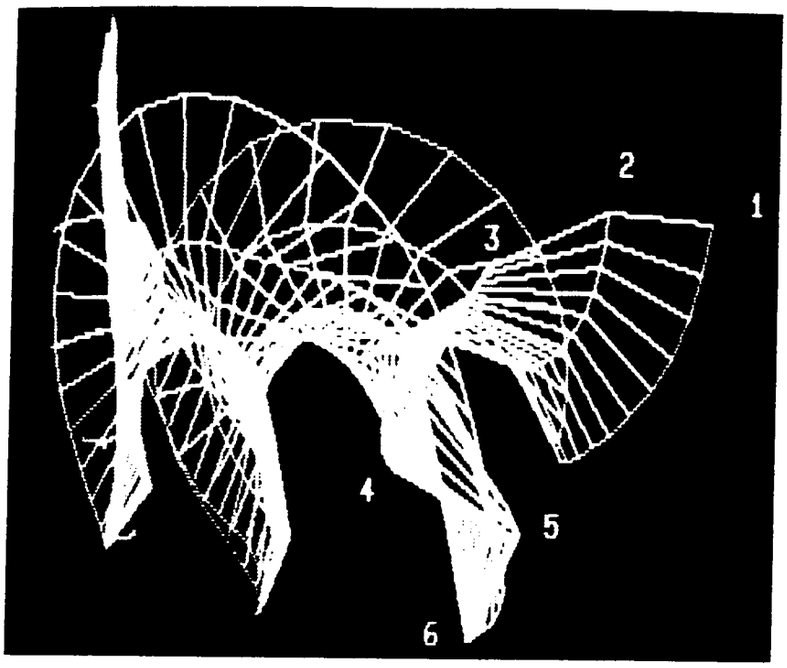

Fig. 6. Classification of the markers of the movement reported in Figure 2 and 5.

which can be masked during the movement. This information allows overcoming the uncertainty in classifying overlapped markers. Weights are associated to the links. These can be though as the mass of the links, although they are expressed as integer numbers ranging from 1 to 100. The weights are used to compute the confidence interval of the trajectory prediction of a marker during the tracking procedure. The interval depends on the reciprocal of the sum of the weights of all the links connected to the marker. This dependence expresses the property that the heavier the link, the smoother and thus more reliable the trajectory is. The prediction of the trajectories is performed by following the procedure outlined in (Ferrigno and Gussoni, 1988). A Taylor series of the trajectory is truncated at the third term and the derivatives are approximated by finite differences leading to the following formula:

$$x(n+1)=\frac{5}{2}[x(n)+2x(n-1)+\frac{1}{2}x(n-2)] \qquad (1)$$

Since the prediction assumes that the previous frames have been correctly classified, an initialization procedure is required. This is interactively performed by the operator who is requested to classify the markers (by graphic

pointers) of the first two frames. The third frame is classified with a linear two-point predictor and then equation (1) is used to proceed further.

An example of marker classification is reported in Figure 6. This figure shows a model with six markers and five links correctly classified including the reconstruction of the missing points (due to segment overlapping).

4. CALIBRATION

After having computed and classified the 2D coordinates of all the markers surveyed at least by two TV cameras, the system is ready to implement 3D reconstruction (space intersection). The obtained accuracy is very important as it influences the subsequent processing of the acquired data, particularly when derivatives must be computed (Lanshammar, 1981). Two ways may be followed in order to reach the goal. The black box approach (Fioretti et al., 1985) requires the computation of many coefficients and the use of many control points to reach an adequate accuracy. The other way to achieve a high 3D accuracy is to suitably tune all the factors which can influence it by using mathematical models. This latter way will be followed herein.

The geometrical transformation from the coordinates of a point in 3D space to the 2D target plane is described by the collinearity equations:

$$(x-x_o) = -c\,\frac{m_{11}(X-X_o)+m_{12}(Y-Y_o)+m_{13}(Z-Z_o)}{m_{31}(X-X_o)+m_{32}(Y-Y_o)+m_{33}(Z-Z_o)} \tag{2a}$$

$$(y-y_o) = -c\,\frac{m_{21}(X-X_o)+m_{22}(Y-Y_o)+m_{23}(Z-Z_o)}{m_{31}(X-X_o)+m_{32}(Y-Y_o)+m_{33}(Z-Z_o)} \tag{2b}$$

with:

X_o, Y_o, Z_o – TV camera location (3D coordinates of the perspective centre)

m_{ij} – nine director cosines, which are function of the three rotation angles: Ω, ϕ, k

x_o, y_o, c – TV cameras principal position (intersection of optical axis and image plane) and focal length (inner parameters)

X, Y, Z – 3D coordinates of the surveyed point

x, y – 2D target coordinates of its projection

$X_o, Y_o, Z_o \Omega, \phi$ and k are named external geometrical parameters.

This is an ideal model, but in the real situation, we must take into account the quantization error (stochastic) introduced by the measuring system and the optical distortions (systematic error); this latter error can be corrected, as it is consistent with time by a procedure of camera calibration.

In high-precision metric cameras, equipped with fixed lenses, laboratory procedures, like multicollimator and goniometer methods, are suitable to obtain internal parameters with a very high accuracy. A further refinement of these parameters is achieved by interpolating radial and tangential distor-

tions with an n-order polinomial function. This kind of calibration is not suitable neither for analogic TV cameras where target is not stable in the long run, nor for laboratory use, where focal length may be changed according to the type of the experiment and the required amplitude of the field of view. Moreover as set-up tuning is carried out on the field according to experimental needs, also the calibration procedure should be quickly performed on the field.

We have adopted a linear piece-wise approximation algorithm for correcting the distortions. This employes a small number of coefficients so that a quick assessment of the parameters is possible. A further calibration step is required in order to determine the geometrical parameters in equations (2); this procedure is called space resection. The project requirements are the same of camera calibration: operation speed and accuracy. Theoretically, self-calibration methods (Kenefick et al., 1972) provides the simplest way to accomplish it at the cost of using a great quantity of markers. In this approach the coordinates of the control points are themselves unknowns and only a few relations among them are known. The procedure, widely used in aerial photogrammetry and topography entails a high computational cost that is not here justified as, following the method presented herein, only few approximate measurements on the field are required.

Another widely used technique is the Direct Linear Transformation (DLT) (Abdel-Aziz and Karara, 1971; Marzan and Karara, 1975). It is based on the transformation of the collinearity equations in linear equations by rearranging the parameters. This approach, although leading to an exact solution, presents the drawbacks that one of the 11 obtained parameters is redundant causing some estimation problems and the coefficient matrix, 11×11 in dimension, introduces some error due to rounding problems. These may result in sizable reconstruction errors (Wood and Marshall, 1986; Hatze, 1988). Other analytical methods, like Church ones (Wolf, 1983) are, in some cases, error-sensitive.

Our approach uses the classical iterative least-squares estimation allowing a quick assessment of the correctness of the estimation itself by the physical meaning of the parameters and by the analysis of the residual. Also the maximum freedom in TV cameras positioning is preserved. Even if the parameters are initially estimated very coarsely (more than 50% error of final value), the algorithm converges to the correct value.

With regard to the problem of 3D intersection, it can be viewed as an error minimization problem. In the case of DLT, the 3D point is located in the position that minimizes the distance between the two sets of 3D points given by DLT equations for each TV camera. This solution has a geometrical explanation: if we look at the couple of DLT equations for one TV camera, we see that they have three unknowns: the X,Y,Z coordinates of the point. The system has ∞^1 solutions that are all the points on the straight line through the

perspective centre and the target projection. The least-squares solution of DLT equations for a couple of TV cameras minimizes the errors (i.e. the distance) between the two straight lines. Our algorithm performs this task by directly solving the intersection problem. The analytical solution has been described in (Borghese and Ferrigno, in press), a mixed (geometrical and analytical solution) is hereinafter described.

4.1 2D calibration

The image of a square placed in front of a TV camera, is distorted and it appears as a curvilinear quadrilateral. Decreasing the side of the square, the quadrilaterals sides tend to approximate rectilinear segments and we can reasonably suppose that the distortions are uniformly distributed inside each square. Therefore, we can divide the camera calibration problem into a set of subproblems, restricted to each square: the transformation of each quadrilateral in a virtual square; all the virtual squares having the same side length. We introduce here the word "virtual", meaning that the measure of this side of the so obtained square can differ from the real one; but this is not a problem for the accuracy in 3D coordinates as it results in a scaling of the 2D image. The coefficients of a couple of quadratic functions for each square, that transform the sides of the distorted squares in the real ones, are computed as reported in appendix A.

The algorithm for distortions correction consists of two steps: marker assignment in order to decide which square the marker belongs to, and location correction with the aid of the parameters of that square.

In order to determine the square which a marker P_i, with coordinates (x_i,y_i), belongs to, we compute the integer part of the expressions:

$$h=\text{Int}((x_i-x_{m11})/L)+1 \tag{3a}$$

$$k=\text{Int}((y_{m11}-y_i)/L)+1 \tag{3b}$$

where h,k indicate the square assigned to point P_i, L is the measure of the virtual square grid side, x_{m11},y_{m11} are the measured coordinates of the upper left corner of the grid. Then, transforming the marker coordinates by functions (A8a) and (A8b) of appendix A, we obtain the point P_{ir}, with corrected coordinates (x_{ir},y_{ir}). We check if the corrected point belongs to the square of the virtual grid having the same indexes h,k (position) of the square of the surveyed grid including the point P_i. If it does, the square results to be the right one and so are the associated parameters; otherwise, we presume that the point P_{im} belongs to an adjacent square: it is chosen the square having the same position in the surveyed grid that it results to have in the virtual grid. The procedure is repeated until the virtual square corresponds to the surveyed one. Iterations may become necessary only for those points very close

to squares borders. In no case more than four iterations are required. In order to get the parameters we operate five sequential transformations on each of the surveyed grid squares (appendix A) which apply to all the points belonging to the related square: points on the border are transformed in border points and internal points in internal points. Operatively, the determination of calibration parameters is performed by the acquisition of a set of markers arranged in a square grid on a plane parallel to the TV camera target; small misalignments of TV camera target and grid result in second-order errors. However, they are automatically taken into account when space resection is performed and geometrical parameters are computed so that this inaccuracy does not appreciably affect system measurement. The criterion for the choice of the number of squares depends on the accuracy to be achieved and on the basic distortion of the transducing system. In our case, grids of 5×6 have given good results.

4.2 3D calibration

As mentioned before, in order to perform space resection, the nine geometrical parameters of the collinearity equations are estimated from the real coordinates of a set of known location control points and their measured coordinates using standard least-squares techniques. The main problem of this algorithm is that a good computation of the initial values is necessary; they can be computed by direct analytical methods like DLT, Church methods and so on, otherwise an approximate initial estimation of the six external parameters value should be introduced by the operator. The centre of the virtual 2D calibration grid is assumed as the initial position of the principal point and the initial value of focal length is computed from these eight parameters. Algorithm convergence is not guaranteed a priori: the function norma [function (B2) in appendix B] can also decrease towards a local minimum, giving a solution different from the correct one; however, with a good correction of systematic errors and a wide distribution of control points, the local optima tend either to be absent or not to affect parameters estimation (Woltring, 1980). In order to obtain the great number of control points required without having to spend much time to compute their 3D coordinates by theodolites or other precise measurement systems, we confined the control points on a plane, which is shifted according to reference locations on the floor, or on a supporting board (depending on the field of view). Very precise measurements are required only once, when the markers are put on the grid and the references on the floor are located, so that every time that the system set-up is changed, little time is required to calibrate it.

5. 3D RECONSTRUCTION

After the parameters in the geometrical transformation are computed, the

system can compute the 3D coordinates of the markers. This is accomplished by finding the minimum distance segment between two straight lines, r and s, conveyed through the perspective centre and 2D image projection. For this purpose, let us write the equations of two lines in a parametric form:

$$[P]_i = [T] + h[L] \tag{4}$$

where:
$[P]_i$ is the 3D coordinates vector
$[T]$ is the perspective centre point
$[L]$ is the vector of the direction cosines
h is a real parameter equivalent to the distance $|P_i - T|$
The expression of the direction cosines is:

$$[L] = \{[M]^{\mathrm{T}}([p]_i - [t])\} \{|[M]^{\mathrm{T}}([p]_i - [t])|\}^{-1} \tag{5}$$

where:
$[M]$ is the rotation matrix between the laboratory and camera reference frames
$[M]^{\mathrm{T}} = [M]^{-1}$ due to the properties of the orthonormal matrix
$[p]_i$ are the coordinates of the projection of P_i (x_i, y_i, c) in the camera reference frame
$[t]$ are the coordinates of the perspective centre (x_o, y_o, c) in the camera reference frame

Intersection can be expressed as:

$$[T]_r + h_r[L]_r = [T]_s + h_s[L]_s \tag{6}$$

which is a linear system of three equations in the two unknown h_r and h_s. If there is a perfect intersection or an approximation of 3D coordinates is required, it is sufficient to solve the system using only two equations; otherwise, the minimum distance segment between the lines must be found. The analytical solution has been described in (Borghese and Ferrigno, in press). We will report in this paper a mixed geometrical and analytical solution. For this purpose, the equation of the sheaf of planes with generatrix the straight line r is computed, each plane is identified by its perpendicular line. We extract from this sheaf, the plane with the perpendicular line also perpendicular to line s and compute the equation of this straight line; this will contain the minimum distance segment and the 3D point is located in the middle of it (appendix C). The mean length of this segment is on the order of 1/2500 of the field of view, thus allowing the use of two of equations (6) for an approximate assessment of the 3D coordinates.

5.1 Accuracy tests

Two accuracy tests both in static and dynamic conditions have been performed. The first test has been performed by using a fixed length bar with two markers applied at the extremities. The bar has been positioned in different orientations (11) in the field of view and an acquisition of 100 samples has been performed in each attitude (static test). The second test, (dynamic test) is based on the same object used in the static test. In this case it has been moved in the calibrated volume seven times and 100 samples have been taken each time. A summary of the results is reported in Table 1, for additional results see Ferrigno (1990).

These results are comparable with the dynamic 2D standard deviation of the error of the Elite system (Ferrigno and Pedotti, 1985). The camera positioning also plays an important role. The ratio between the intercamera distance (L) and the distance between the cameras and the surveyed object (D), L/D shows the optimum value of 2 (Borghese and Ferrigno, in press).

TABLE 1

Summary of accuracy tests

Static test	$\sigma_{xyz}=1/4040$	$N=1100$
Dynamic test	$\sigma_{xyz}=1/2835$	$N=700$

σ_D = standard deviation of the errors on the distance.
$\sigma_x, \sigma_y, \sigma_z, \sigma_{xyz}$ = standard deviation of the errors along coordinate axes.

6. CONCLUSIONS

Thanks to progress in electronics, a great step towards automatic image processing and motion measurement has been achieved since the early 1980s. These systems have been conceived as having a two-level structure; the first, hardware implemented, performs in real-time the recognition of repere points on the moving subject compressing the information and reducing it to a small set of primitives; the second level performs a flexible processing of these in order to recover the 3D coordinates.

The Elite system has some innovative features that increase its reliability and the easiness of use by nonskilled persons. The image processing, performed as a shape recognition, guarantees a higher signal to noise ratio which allows the elimination of unwanted reflexes and other light objects from the surveyed images. The software level performs the stereo-matching between images by means of a model-based system which is able to reconstruct the position of markers that disappeared during the movement. Calibration is divided into two steps resulting in a very simple operative procedure and an easy setting-up of the system.

The capability to acquire and to process a large set of data, opens new possibilities to these systems. The marking of the surfaces allows to analyze the movement of complex body features; moreover, transputers and parallel processors will allow to perform primitives processing in real-time, obtaining a whole real-time system which could be used not only as a motion analyzer but also in those fields where monitoring and controlling movement is required.

APPENDIX A

The geometrical operations required to determine the coefficients used for distortion correction are outlined here. We will indicate with (x_{rihk}, y_{rihk}) $1<i<4$ the virtual coordinates of the four vertexes of the square h,k and with (x_{sihk}, y_{sihk}) the coordinates of the same vertexes as measured by the surveying system. The first operation is the approximate estimation of the position of the centre (x_o, y_o) of the grid and of the real side of the squares. By means of these parameters, a "virtual" grid can be built up. The centre of the grid is simply estimated computing the mean point of its external sides as:

$$x_o = (x_{s111} + x_{s41m} + x_{s2n1} + x_{s3nm})/4 \tag{A1a}$$

$$y_o = (y_{s111} + y_{s41m} + y_{s2n1} + y_{s3nm})/4 \tag{A1b}$$

and the side of each square as:

$$L = (x_{s41m} - x_{s111} + y_{s41m} - y_{s3nm})/(n+m) \tag{A2}$$

This estimation is not very accurate, but a great accuracy is not required at this step. The virtual grid must have the same shape (squared) as the original one; this operation builds up a virtual target of the TV camera parallel to the real one (scaled) and it does not change ratios between the various parts of the images; we shall automatically take into account the difference between real and virtual target position in the 3D calibration when geometrical parameters are going to be estimated. Let's go through the sequential geometrical transformations of the surveyed squares; in the following we omit the squares indexes. The first transformation (Fig. A1) is a rigid translation in order to overlap vertex P_{1s} to vertex P_{1r}:

$$x_i^a = x_{is} - x_{1s} + x_{1r} \tag{A3a}$$

$$y_i^a = y_{is} - y_{1s} + y_{1r} \tag{A3b}$$

The second transformation is the orthogonal projection of the upper side (P_1P_4) on the virtual square side $P_{1v}P_{4v}$ (Fig. A1):

$$x_i^b = x_i^a \tag{A4a}$$

$$y_i^b = y_i^a + (y_{4s} - y_{1s})(x_{2s} - x_{1m})/(x_{1s} - x_{4s}) \tag{A4b}$$

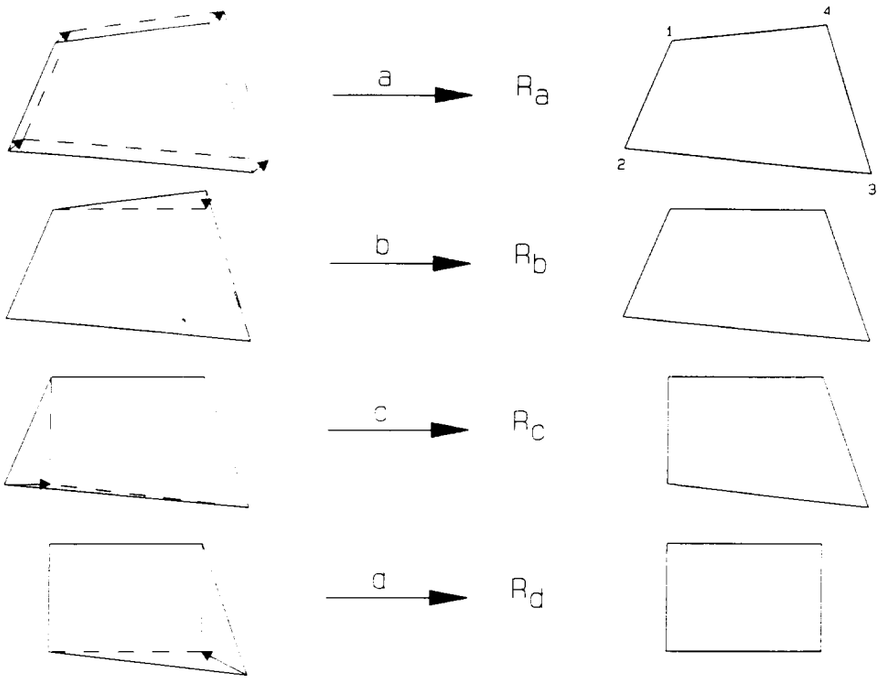

Fig. A1. Sequential transformations for distortions correction.

The third transformation is the orthogonal projection of the vertical left side ($P_1^b P_2^b$) on the virtual square side ($P_{1r}P_{2r}$) (Fig. A1):

$$x_i^c = x_i^b + (x_{2s} - x_{1s})\,(y_i^b y_{1r})/(y_{1r} - y_{4s}^b) \tag{A5a}$$

$$y_i^c = y_i^b \tag{A5b}$$

The fourth transformation is the contraction of vertex 3 toward the centre of the square as long as opposite sides are parallel to each other; the quadrilateral has been now transformed in a rectangle (Fig. A1):

$$x_i^d = (x_3 - x_4)^c (y_i - y_4)^c (x_i - x_1)/(y_3 - y_1)\,(x_3 - x_1)^c \tag{A6a}$$

$$y_i^d = (y_2 - y_3)^c (x_i - x_1)^c (y_i - y_1)/(x_3 - x_1)\,(y_3 - y_1)^c \tag{A6b}$$

Scaling is the last operation that allows to obtain a square with the estimated length from the rectangle (Fig. A1):

$$x_{ir} = (x_i^d - x_{1r})\,(x_{4r} - x_{1r})/(x_{4r} - x_{1r}) + x_{1r} \tag{A7a}$$

$$y_{ir} = (y_i^d - y_{1r})\,(y_{2r} - y_{1r})/(y_{1r} - y_2^b) + y_{1r} \tag{A7b}$$

from the expressions above, rearranging the various terms, we obtain the second-order polinomial expression for x_{ir} and y_{ir}:

$$x_{ir} = A_1 x_{is} + B_1 y_{is} + C_1 x_{is}^2 + D_1 x_{is} y_{is} + E_1 y_{is}^2 + F_1 \quad \text{(A8a)}$$

$$y_{ir} = A_2 x_{is} + B_2 y_{is} + C_2 x_{is}^2 + D_2 x_{is} y_{is} + E_2 y_{is}^2 + F_2 \quad \text{(A8b)}$$

APPENDIX B

In order to use the least-squares method, the collinearity equations must be linearized leading to:

$$x_i = a_{i1}\,d\Omega + a_{i2}\,d\phi + a_{i3}\,dK + a_{i4}\,dX_o + a_{i5}\,dY_o + a_{i6}\,dZ_o + a_{i7}\,dc + dx_o + n_i \quad \text{(B1a)}$$

$$y_i = b_{i1}\,d\Omega + b_{i2}\,d\phi + b_{i3}\,dK + b_{i4}\,dX_o + b_{i5}\,dY_o + b_{i6}\,dZ_o + b_{i7}\,dc + dy_o + m_i \quad \text{(B1b)}$$

where a_{ij}, b_{ij} are the first derivatives of collinearity equations with respect to the nine parameters and n and m are the errors due to linearization and quantization. Least-squares estimation is obtained by minimizing the norma of the residuals D:

$$\min D = [\textstyle\sum_i (n_i)^2 + \sum_i (m_i)^2]^{1/2} \quad \text{(B2)}$$

Rewriting equations B1 in a matrix form: $V + L = A\,X$, condition (B2) is satisfied if: $A^T L - A^T A\,X = 0$, which leads to the solution:

$$X = (A^T A)^{-1} AL \quad \text{(B3)}$$

Equations (B1) can be rewritten as follows:

$$P_i = f(P) \text{ where } P = P(\Omega, \phi, K, X_o, Y_o, Z_o, x_o y_o, c) \quad \text{(B4)}$$

The solution of the normal equations (B3) gives us the incremental value to give to point P in the direction in which the normal is minimized ($\min \|L\|$). The new point will be the starting point for a new iteration until $\|L\| < k$ small enough. The point to which the algorithm converges is the same starting from different initial points.

APPENDIX C

In order to recover the 3D coordinates, the first step is to write the equation of the straight line r as an intersection of two planes:

$$(X - X_{or})/\cos\alpha_r = (Y - Y_{or})/\cos\beta_r \quad \text{(C1a)}$$

$$(Y - Y_{or})/\cos\beta_r = (Z - Z_{or})/\cos\tau_r \quad \text{(C1b)}$$

where $\cos\alpha$, $\cos\beta$ and $\cos\tau$ are the director cosines of r and (X_{or}, Y_{or}, Z_{or}) are the coordinates of C_r (Fig. C1). From these, we can write the equation of the sheaf of planes with generatrix r:

$$\theta[(X-X_{or})/\cos\alpha_r-(Y-Y_{or})/\cos\beta_r] + \mu[(Y-Y_{or})/\cos\beta_r-(Z-Z_{or})/\cos\tau_r]=0 \quad \text{(C2)}$$

The plane parallel to s is the one that satisfies the relation:

$$\theta\cos\alpha_s/\cos\alpha_r+(\theta+\mu)\cos\beta_s/\cos\beta_r+\mu\cos\tau_s/\cos\tau_r=0 \quad \text{(C3)}$$

and posing $\theta=1$, the value of μ can be computed. The equation of the plane π_r is:

$$(X-X_{or})/\cos\alpha'_r+(Y-Y_{or})/\cos\beta'_r+(Z-Z_{or})/\cos\tau'_r=0 \quad \text{(C4)}$$

where:

$\cos\alpha'_r=\cos\alpha_r/p$

$\cos\beta'_r=\cos\beta_r(1+\mu)/p$

$\cos\tau'_r=\cos\tau_r\mu/p$

$p=[1/\cos^2\alpha_r+(\mu+1)^2\cos^2\beta_r+\mu^2/\cos^2\tau_r]^{1/2}$

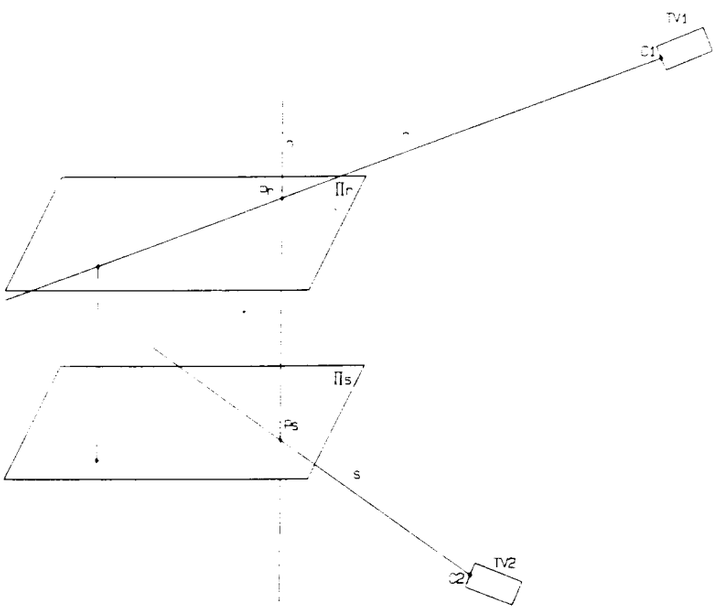

Fig. C1. 3D geometrical intersection.

The distance d of the line s from π_r is:

$$d=(X_{os}-X_{or})/\cos\alpha'_r+(Y_{os}-Y_{or})/(\cos\beta'_r)+(Z_{os}-Z_{or})/(\cos\tau'_r) \quad (C5)$$

and this is also the minimum distance (d^) between the two straight lines r and s. This minimum distance is measured on the line n connecting the two points Pr and Ps. The following relations holds:

$$X_s=X_r+d^{\wedge}\cos\alpha' \quad (C6a)$$

$$Y_s=Y_r+d^{\wedge}\cos\beta' \quad (C6b)$$

$$Z_s=Z_r+d^{\wedge}\cos\tau' \quad (C6c)$$

because the segment of minimum distance is contained into line n which is directed perpendicular to both s and r, and as line r is perpendicular to plane π_r, the director cosines of n are the same as those of plane π_r. We can rewrite equations (C6) as:

$$X_{os}+h_s\cos\alpha_s=X_{or}+h_r\cos\alpha_r-d^{\wedge}\cos\alpha' \quad (C7a)$$

$$Y_{os}+h_s\cos\beta_s=Y_{or}+h_r\cos\beta_r-d^{\wedge}\cos\beta' \quad (C7b)$$

$$Z_{os}+h_s\cos\tau_s=Z_{or}+h_r\cos\tau_r-d^{\wedge}\cos\tau' \quad (C7c)$$

that are three equations in 2 unknowns: h_r and h_s. We can solve the system by a least-squares best estimation, but as we need only one of the two unknowns, we can solve it directly and obtain:

$$h_r=(\cos\alpha_r-f\cos\alpha_s)b_1+(\cos\beta_r-f\cos\beta_s)b_2+(\cos\tau_r-f\cos\tau_s)b_3)/(1-f^2) \quad (C8a)$$

$$h_s=(f\cos\alpha_r-\cos\alpha_s)b_1+(f\cos\beta_r-\cos\beta_s)b_2+(f\cos\tau_r-\cos\tau_s)b_3)/(1-f^2) \quad (C8b)$$

with $f=[L]_r^T[L]_s$ and $b_1=X_{os}-X_{or}+d^{\wedge}\cos\alpha$, $b_2=Y_{os}-Y_{or}+d^{\wedge}\cos\beta$ and $b_3=Z_{os}-Z_{or}+d^{\wedge}\cos\tau$.

REFERENCES

Abdel-Aziz, Y.I. and Karara, H.M., 1971. Direct Linear Transformation from Comparator Coordinates into object space coordinates in close-range photogrammetry. Proc. ASP/UI Symp. on Close Range Photogrammetry, Urbana, IL, pp. 1–18.

American Society of Photogrammetry, 1966. Manual of Photogrammetry, 3 rd ed. Falls Church, VA.

Aristotle, 340 B.C. De motu animalium. Translation E.S. Foster, Harvard University Press, Cambridge, MA, 1945 (Progression of Animals).

Bernstein, N., 1967. The Coordination and Regulation of Movements. Pergamon Press, London.

Borelli, G.A., 1680. De motu animalium. Lugdunium Batavorum.

Borghese, N.A. and Ferrigno, G., in press. An algorithm for 3D automatic movement analysis by means of standard TV cameras. IEEE Trans. Biomed. Eng.

Borghese, N.A., Ferrigno, G. and Pedotti, A., 1988. 3D Movement Detection: A Hierarchical Approach. Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics. International Academic Publisher Pergamon Press, Beijing 100044 China, 1: 303–306.

Cheng, I.S., Koozekanani, S.H. and Fatchi, M.T., 1975. Computer-television interface system for gait analysis. IEEE Trans. Biomed. Eng. (BME), 22: 259.

Ferrigno, G., 1990. Metodiche ed algoritmi per l'analisi del movimento, Tesi di dottorato di ricerca, Politecnico di Milano, Italy.

Ferrigno, G. and Gussoni, M., 1988. Procedure to automatically classify markers in biomechanical analysis of whole-body movement in different sports activities. Med. Biol. Eng. Comput., 26: 321–324.

Ferrigno, G. and Pedotti, A., 1985. Elite: A digital dedicated hardware system for movement analysis via real-time TV signal processing. IEEE Trans. Biomed. Eng. (BME), 32: 943–949.

Fioretti, S., Germani, A. and Leo, T., 1985. Stereometry in very close-range stereophotogrammetry with non-metric cameras for human movement analysis. J. Biomech., 18: 831–842.

Hatze, H., 1988. High precision three-dimensional photogrammetric calibration and object space reconstruction using a modified DLT approach. J. Biomech., 21(7) 533–538.

Jarrett, M.O., Andrews, B.J. and Paul, J.P., 1976. A television computer system for the analysis of kinematics of human locomotion. IERE Conf. Proc., 34: 357–370.

Kenefick, J.F., Gyer, M.S. and Harp, B.F., 1972. Analytical self-calibration. Photogramm. Engng., 38: 1117–1126.

Lanshammar, H., 1981. Precision limits on derivatives obtained from measurement data. Biomechanics, VIIa: 586–592. Morecki et al. (editors), Polish Publisher's House.

Marey, E.J., 1902. The history of Chronophotography. Smithsonian Institution, Washington D.C.

Marr, D., 1982. Vision. Freeman, San Francisco, CA.

Marzan, G.T. and Karara, H.M., 1975. A computer program for direct linear transformation solution of the collinearity condition and some applications of it. Proc. Symp. Close-range Photogrammetric Systems. American Society of Photogrammetry, Falls Church, VA, pp. 420–476.

Mitchelson, D.L., 1975. Recording of movement without photography. In: Techniques for the Analysis of Human Movement. Lepus Books, London.

Muybridge, E., 1901. The Human Figure in Motion. Chapman & Hall, London.

Poggio, G.F. and Poggio, T., 1984. The analysis of stereopsis. Annu. Rev. Neurosci., 7: 379–412.

Taylor, K.D., Mottier, F.M., Simmons, D.W., Cohen, W., Pavlak, Jr. R., Cornett, D.P. and Haskins, B., 1982. An automated motion measurement system for clinical gait analysis. J. Biomech., 15: 505–516.

Ullman, S., 1979. The Interpretation of Visual Motion. M.I.T. Press, Cambridge, MA.

Whiting, H.T.A. (Editor), 1984. Human motor actions. Bernstein reassessed, North Holland, Amsterdam, 634 pp.

Winter, D.A., Greenlaw, R.K. and Hobson, D.A., 1972. Television computer analysis of kinematics of human gait. Biomed. Res., 5: 498–504.

Wolf, P.R., 1983. Elements of Photogrammetry, McGraw Hill, New York, NY.

Woltring, H.J., 1975. Single and dual axes lateral photodetectors of rectangular shape. IEEE Trans. El. Dev., 22: 580–581.

Woltring, H.J., 1977. Measurement and Control of Human Movement. H. Peters & J. Haarsma, Nijmegen, NL.

Woltring, H.J., 1980. Planar control in multi-camera calibration for 3D gait studies. J. Biomech., 13: 39–48.

Woltring, H.J. and Marsolais, E.B., 1980. Optoelectric (SELSPOT) gait measurement in two and three dimensional space. A preliminary report. Bull. Prosth. Res., 17: 46–52.

Wood, G.A. and Marshall, R.N., 1986. The accuracy of DLT extrapolation in three-dimensional film analysis. J. Biomech., 19: 781–785.

Robotica (1991) volume 9, pp 275–282

399

ELITE: A goal oriented vision system for moving objects detection

*N.A. Borghese, †M. Di Rienzo, †G. Ferrigno and †A. Pedotti

(Received in Final Form: 2 October 1990)

SUMMARY

A specially designed system for movement monitoring is here presented. The system has a two level architecture. At the first level, a hardware processor analyses in real-time the images provided by a set of standard TV cameras and, using a technique based on the convolution operator, recognizes in each frame objects that have a specific shape. The coordinates of these objects are fed to a computer, the second level of the system, that analyses the movement of these objects with the aid of a set of rules representing the knowledge of the context. The system was extensively tested on the field and the main results are reported.

The whole system can work as a controlling device in robotics or as a general real-time image processor as well as an automatic movement analyser in biomechanics, orthopedic and neurological medicine.

KEYWORDS: *ELITE*: Vision system; Moving objects; Robot control; Movement analyser

INTRODUCTION

The term "vision" refers to the complex set of functions which range from the mere image of the external world to the complete interpretation of the scene and its symbolic description. The design of artificial vision machines relies as closely as possible upon biological structures that are involved in the vision process in living beings.

Referring to the bidimensional vision, the entire visual process may be represented by a multi-level hierarchical structure.[1–3] At the first level, the image from the external world is sampled. At the second level, the entire scene is split into zones tentatively corresponding to separate objects. At the third level, the characteristics (position, colour, shape, texture, etc.) of each zone are extracted. At the fourth level, each zone is assigned to a class of objects (the object is recognized). Finally, at the fifth level, the sematic meaning of juxtaposition among various objects is extracted. The passage from one level to another implies a progressive abstraction of the visual information due to an increasing integration of the data coming from the image with the pre-existent knowledge collected by experience. This hierarchical model of vision is not only efficient from a computational point of view, but also seems to reflect the biological strategy of image processing.[4]

This "static" phase of image analysis is followed by a "dynamic" one during which the data obtained from each frame are related to the data collected in the previous frames in order to evaluate the temporal dynamic of the scene (interpretation of visual motion[5]).

The complexity of this model requires parallel elaborative structures to be implemented in real-time on a vision machine. Neural networks reveal a similar parallel computational architecture in living beings (Parallel Distributed Processing...PDP).[6]

Technological and theoretical limitations have, up to now, thwarted the development of devices able to perform the entire vision process in real-time. Nevertheless, several image processing systems performing part of this complex task are now available.

Context-free devices, based on different multi-processor architectures (SIMD, MIMD, pipeline structures[7–9]), reach a good degree of flexibility and generality but huge amounts of hardware and software are required to enable these systems to work in general environments.[10] This high entropy dramatically increases the cost of these devices and makes them suitable only for research on strategical applications. Actually, in many applications a general purpose image processor is not required, but there is a strong demand for a machine able to perform only specific tasks in a controlled environment ("Goal-Oriented" vision systems). In this case, only a sub-set of the entire visual function is used and the hardware of the system may be optimized for the application, dramatically reducing complexity and costs.[11]

In the frame of this "goal-oriented" devices, a new vision system was designed and developed at "Centro di Bioingegneria" of Milan. This device, *ELITE* (ELaboratore di Immagini TElevisive—TVimages processor), is a vision machine able to recognize objects of a predefined shape and to monitor their trajectories in real-time. The system was initially used to analyse human movements recognizing hemispherical passive markers which were positioned on moving subjects. The whole system has a two-level structure, as follows (Figure 1):

* Centro di Bioingegneria, Fondazione Pro Juventute, Politecnico Milano, Via Gozzadini 7, 20148 Milano; and Istituto di Fisiologia dei Centri Nervosi, CNR, Via Mario Bianco, 9 20131 Milano, (Italy).
† Centro di Bioingengeria, Fondazione Pro Juventute, Politecnico Milano, Via Gozzadini 7, 20148 Milano (Italy).

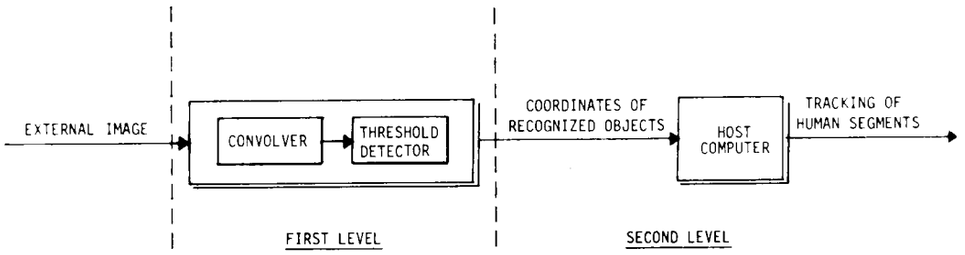

Fig. 1. Two level structure of the *ELITE* system.

1. First level

This level, entirely implemented by hardware, performs the "static" analysis of the scene, i.e. objects recognition. At this stage an image processor analyses in real-time, at a sampling rate of 100 Hz, each video frame coming from a set of CCD TV cameras, detects the objects having a predetermined shape in the scene, displays their 2D position on a TV monitor and sends their coordinates to a host computer or to a monitoring device. Recognition is performed by computing a bidimensional convolution between the image coming from the TV camera and a mask associated to the shape to be recognized. This level is essentially constituted by a parallel hardware convolver, a threshold detector and a generator of coordinates.

2. Second level

Here, a host computer performs the "dynamic" analysis of the scene. The information coming from the first level is compared with the results obtained in the preceding frames. This temporal matching, integrated by information extracted from a knowledge base, allows the tracking of recognized objects and the description of their trajectories.

This system has been found extremely flexible and it can be used in several fields where shape recognition is required, even far from the biomechanic area. Moreover, the usage of the convolution algorithm for processing the TV signal, makes *ELITE* not only a shape detector, but also a general purpose image processor.

This paper contains a discussion of the principles on which *ELITE* is based (analysis of the structure of the mask when the *ELITE* system is used for shape recognition and for other types of image processing), a description of the system and a brief evaluation of its performances.

THE CONVOLUTION PROCESS AND THE STRUCTURES OF THE MASK

Convolution is a powerful image processing tool. Enhancement of scene quality (by filtering) and extraction of image features (shape recognition, edge detection) are obtained by changing only the structure of the associated mask without any change in the algorithm. The particular structure of computations permits one to implement this operator directly on parallel VLSI hardware.[12] This parallel approach has a biological analogy in the spatial organization of the retinal photoreceptors which are interconnected in order to perform instantaneous convolutions of an external image.[3,13–15]

As described in further references, we recall the formula of the discrete bidimensional convolution between an image and a generical mask:

$$R(x, y) = \sum_{0}^{P-1} \sum_{0}^{Q-1} I(x-i, y-j) * M(i, j)$$

where:

$I(x, y)$ is the level of brilliance of the pixel (x, y).

$M(i, y)$ is the element (i, j) of the mask.

$R(x, y)$ is the value of the convolution computed on the pixel (x, y).

The mask is a $P \times Q$ matrix whose elements assume both positive and negative values.

In the next sections we will report some notes about the structure of the mask.

1. Shape recognition

The generic mask used for shape recognition is composed of three concentric zones (see Figure 2):

—Z1 (active zone), the inner one, with positive values, having a shape similar to the one to be recognized.

—Z2 (zone of indetermination), the middle one, filled with zeroes, encircling Z1 (this region is not always present).

—Z3 (inhibition zone), the most external, with negative values.

With a detector consisting of a convolver and a threshold detector (Figure 1), a correct recognition of a specific shape in a TV frame is obtained if the convolution integral between the image and the mask overcomes the threshold value T just in the part of the image containing the shape S, i.e. if:

$$R(x, y) \geq T \quad (x, y) \in S \tag{A1}$$

$$R(x, y) < T \quad (x, y) \notin S \tag{A2}$$

This event is verified if the active zone of the mask (Z1) has the same shape of the form to be recognized and if the weights of zones Z1 and Z3 are set according to the criteria hereafter described.

We define:

S1 area of zone Z1.

$Z1_i$ i-th element of unitary area of Z1.

$CZ1_i$ its weight (>0).

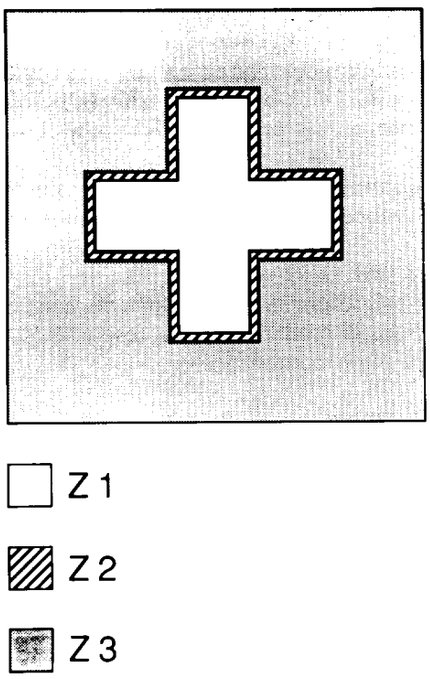

Fig. 2. Example of mask for shape recognition.

S3 area of zone Z3.
$Z3_i$ i-th element of unitary area of Z3.
$CZ3_i$ its weight (<0),
and assume intensities of brilliance having positive values ranging from 0 (black) to 1 (white). To simplify the analysis, consider each zone having weights of the same value. Z2 will not be considered for the moment.

The outline of an object in the frame coincides with the shape in the mask if and only if the following conditions are verified:

Condition I: the outline of the object does not exceed in any point the border of Z1;

Condition II: the area of the object is not smaller than the area of Z1.

Thus $CZ1_i$, $CZ3_i$ and T values must be chosen in order to satisfy relation (A) only when conditions I and II are simultaneously verified. We will state the values selecting rules when operating in different environmental situations.

2. White target on black background

Assuming a value of 0 for the background and a value of 1 for the object to be recognized, condition A1 will be satisfied if:

$$CZ3_i \geq \left(T - \sum_0^{S1}{}_j CZ1_j\right) \quad \forall i \qquad \text{(B)}$$

This condition is very restrictive; an object having a shape exceeding the mask size, will be rejected because the pixels of the object, which fall in the zone Z3 of the mask, will provide such a negative contribution to $R(x, y)$ that relation (A) will no longer be satisfied.

Conditions A2 will be verified if the threshold value T in (A) is:

$$T = \sum_0^{S1}{}_j CZ1_j \qquad \text{(C)}$$

Selecting $CZ1_i$, $CZ3_i$ and T in order to satisfy equations (B) and (C), relation (A) is verified only when the target object exactly matches the shape on the mask.

3. Grey target on grey background

In this case, the object and background values range from 0 to 1.

Relations (B) and (C) must be modified in order to perform the recognition in the scenes with lighter backgrounds and without white maximally target objects.

We define:
L1 the brilliance of a unit area element of the object.
L3 the brilliance of a unity area element of the background with $L1 > L3$.
Condition A1 will be satisfied if:

$$\left(\sum_0^{S1}{}_i CZ1_i\right) * L1 + \left(\sum_0^{S3}{}_j CZ3_j\right) * L3 \geq T \qquad \text{(D)}$$

Condition A2 will be satisfied if:

$$T = \sum_0^{S1}{}_i CZ1_i * L1 + \sum_0^{S3}{}_j CZ3_j * L3 \qquad \text{(E)}$$

As condition (E) satisfies also the relation (D), (E) is the only rule to be considered when selecting CZ1, CZ3 and T in this environment. This relation ensures a correct recognition whenever both the object and the part of background close to the object have homogeneous brilliances (this is usually the case of a small size mask). When this assumption cannot be accepted (the mask is large and the scene is not especially prepared), the relation (E) is used as well, but L1 and L3 must be regarded as unitary brilliance values, respectively, averaged over the zones Z1 and Z3. It must be remarked that in this case, relation (E) gives a good correct-recognition score but it can no longer guarantee an unfailing performance as in the previous cases. The reason will be discussed below.

Suppose, we intend to ideally project the mask on the part of the scene containing the object to be screened (see Figure 3).

We assume:
F1 the area of the object overlapping Z1.
F3 the area of the object overlapping Z3.
F (F1 + F3).
S1 the background area overlapping Z1.
S3 the background area overlapping Z3.
S (S1 + S3).
and LF1, LF3, LF, LS1, LS3, LS the related brilliances.
We have:

$$L1 = (LF1 + LS1)/S1.$$
$$L3 = (LF3 + LS3)/S3.$$

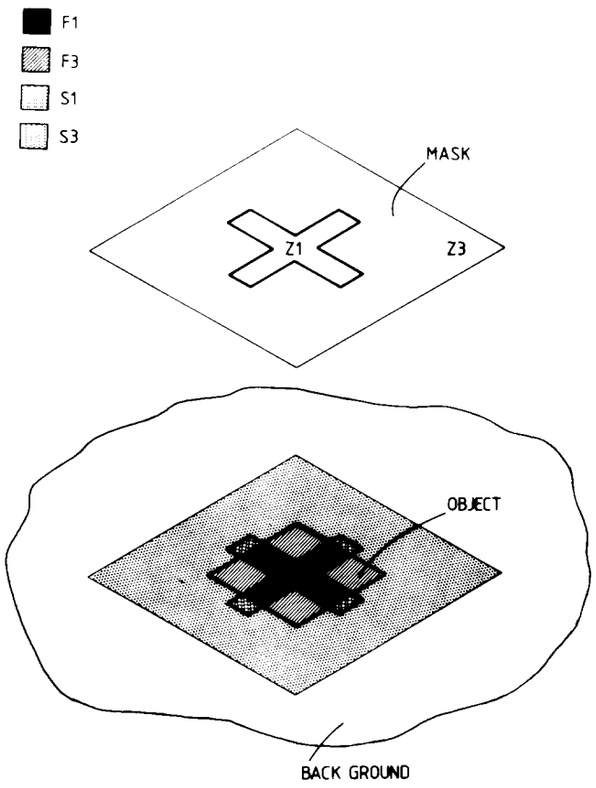

Fig. 3. Source of errors in shape recognition. The mask (upper part of the figure) is set to recognize a cross, the lower plane is a portion of a frame containing a square (target object). The shaded area is the projection of the mask on the image. For the meaning of each shaded region see text.

Once L1 and L3 have been determined, the false positives are due to LF3 and LS1, these being components not separable from LS3 and LF1, respectively.

When a recognition based on a tight matching with a predefined form is not applicable (not sharply defined shapes as well as very small shapes in which the quantization error introduces changes in the discretized shape, etc.) zone Z2 (the zone of indetermination) may be used. Such a procedure allows one to disregard, to some extent, the outline of the target object; the weights of the mask elements in this zone are zeroes and the values of the image pixels falling in this zone make no contribution to $R(x, y)$. Obviously, when a precise shape matching is required, the area of Z2 can be removed, thus letting Z1 and Z3 cover the entire mask area.

4. Resolution enhancement

Setting the mask of the recognizer of Figure 2 according to the relations (B) and (C) or (E), we obtain a single over-threshold point for each recognized object; this is the convolution function value $R(x, y)$ when the mask perfectly matches the object, x and y being the position of the mask. This couple of coordinates states the position of the target object in the frame. The coordinates are given with a resolution equal to the inter-pixel distance. By lowering the threshold from T to T1, a cluster of contiguous over-threshold points will be supplied for each recognized object. The centroid of the R values associated with these points permits one to estimate the actual position of the object with higher accuracy.

After experimental tests, the accuracy obtained by this method has been proved to be approximately 0.1 pixel.[16]

SPATIAL FILTERING

For this application, the mask elements must assume the values of the discretized impulse response of the spatial filter to be implemented.[17]

1. Edge detection

Marr and Hildreth[18,19] demonstrated that the edge detection problem may be solved by a procedure based on the convolution, called $\nabla^2 G$ "Zero Crossing Edge Detection", which is able to detect all the edges[20] of an image. By means of this algorithm, the image is purified from noise by a low-pass spatial filter having a Gaussian impulse response. Afterwards the second derivative of the filtered image is computed. This operation transforms the variations of brightness of the image into zero-crossings. The filtering and the derivative operation may be combined in a single convolution using the second derivative (Laplacian operator) of the Gaussian curve. The standard deviation of the Gaussian curve determines the cut-off frequency of the low-pass filter that is the width of the edges to be detected. The function $\nabla^2 G$ has a central positive region surrounded by a negative ring (just like the mask for shape recognition), but the weights are set in order to give a zero value whenever a sharp change of brilliance is encountered in the image independently from the shape of the object.

2. Oriented edges recognition

Oriented edges recognition is a particular kind of shape recognition; in this application the three zones of the mask are adjacent. The outlines of the zones have the same inclination of the edge to be detected and the weights of the mask are set in order to satisfy the same conditions required for shape recognition. Figure 4

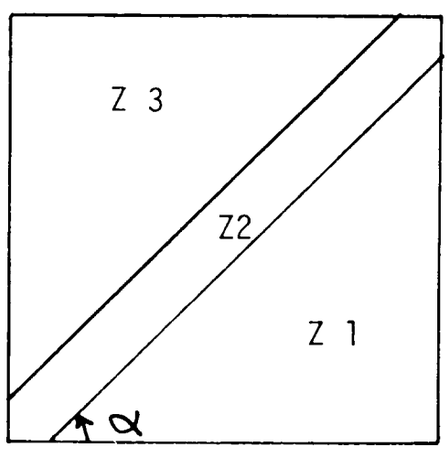

Fig. 4. Mask structure for recognition of bent edges. Angle alpha must be the same of the edge to be recognized.

shows a mask for the recognition of bent edges. In this case, the lighted portion of the object must lie on the right side of the edge.

STRUCTURE OF *ELITE* SYSTEM

The *ELITE* system consists of three parts:
—The interface to the Environment (ITE).
—The Fast Processor for Shape Recognition (FPSR).
—The Computer Level (CL).

A detailed hardware description of the *ELITE* system is reported in ref. 16.

1. The interface to the environment

The interface consists of devices required to collect the external images: the TV cameras and the markers.

Sampling rates of CCD TV cameras working at 50 and 100 Hz have been adopted. Shape distortions due to movements have been avoided by using an electronic shutter. The target of the TV camera is briefly exposed to freeze the movement. A specially designed strobe was realized by a ring of infrared LEDs encircling the TV camera's lens powered by an electronic board synchronized with the TV beam. The use of both these shuttering features also allows for the recognition of markers outdoors.

The markers are hemispherical plastic surfaces, covered by reflecting paper, with diameter about 1/200 of the field of view.

2. Fast processor for shape recognition (FPSR)

The TV signal is analysed by a specially designed hardware processor (the actual core of the system) called *FPSR* (*F*ast *P*rocessor for *S*hape *R*ecognition). It extracts in real-time from each TV camera frame, the position of the markers having the shape selected by the mask and sends to the computer the coordinates (x, y) and the convolution value $R(x, y)$ for each over-threshold point $P(x, y)$ of the cluster associated with the recognized marker. These tasks are performed by an A/D converter, a parallel convolver and a threshold detector. Each TV frame is sampled at 50 or 100 Hz and digitized in a 256×256 4 bits matrix of pixels.

The Hw convolver, working at a 5 or 10 MHz speed, performs a real-time convolution between the pixels matrix and the mask. The mask is constituted by a 6×6 matrix of elements of 4 bits each. The bidimensional convolution formula implemented is:

$$R(x, y) = \sum_{0}^{5}{}_{i} \sum_{0}^{5}{}_{j} I(x - i, y - j) * M(i, j)$$

The convolver is structured in five TV line delays and six line convolvers working in parallel (see Figure 5); the inner structure of each line convolver is fully pipelined (Figure 6). The output of the convolver is then processed by a threshold detector and for each over-theshold pixel, the correlation value R and its coordinates are sent to the computer level.

3. Computer level (CL)

This level includes a general purpose computer, its software and the hardware interface linking the *FPSR* to the computer. At this stage, the information from the image analysis is integrated with other knowledge linked with the specific application in order to evaluate the kinematics of the movement. As the calculations are performed by a standard serial computer, the output of this level can be obtained on-line with a computation time delay.

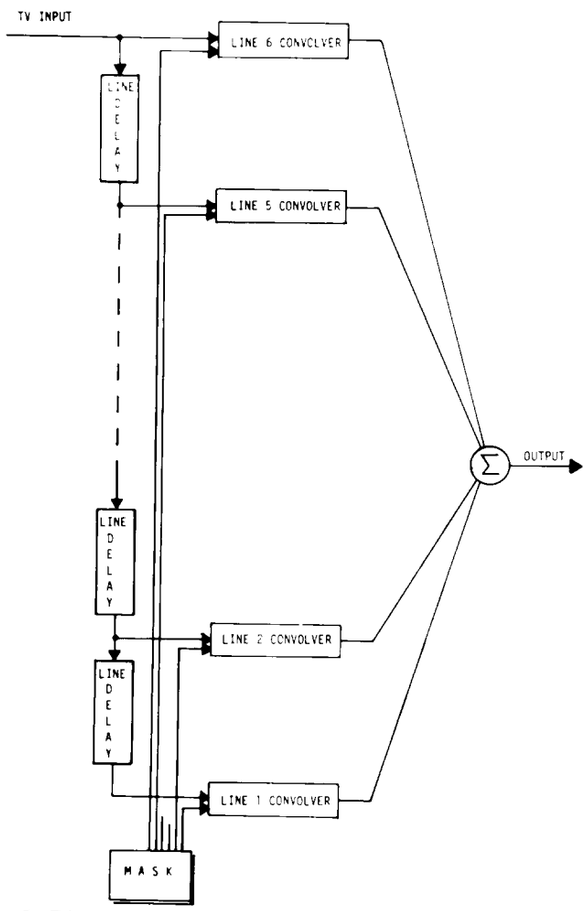

Fig. 5. Diagram of the image convolver.

Up to now several types of computer can be connected to the *ELITE* system; the Digital PDP 11 family: PDP 11/03, /23, /53, /73, /83, the HP1000 family: L, A700, A900, Olivetti M280, M380 and compatible computers.

The data transfer rate of the interface (supplied by the computer manufacturer) constitutes the only limitation to the number of markers that can be processed by the *ELITE* system in a single frame. This is because the data of the cluster of points associated with each recognized marker present on the frame, must be entirely transferred before the next frame is available (10 milliseconds for a 100 Hz TV camera).

As the data have been transferred from *FPSR* to the computer, the centroid position of the cluster of pixels belonging to the same marker is computed, and the optical and electronical distortions introduced by TV cameras are corrected. Subsequently, a "dynamic" analysis is performed; the markers are tracked identifying the respective trajectories and their 3D position is computed.

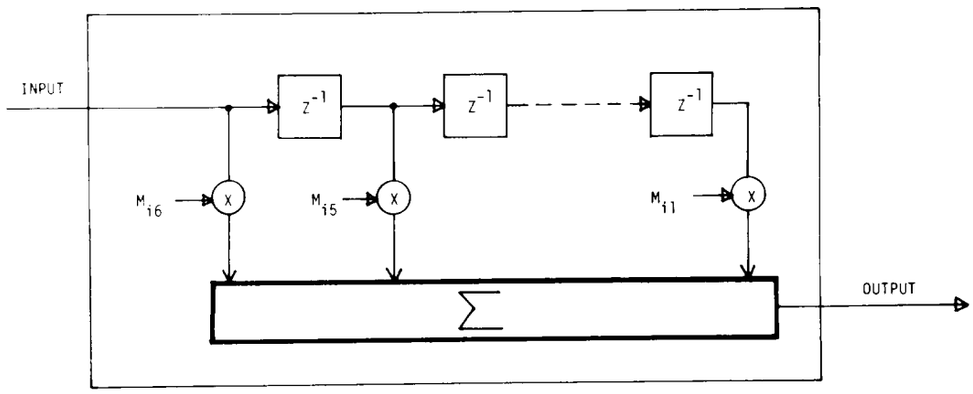

Fig. 6. Inner structure of a single line convolver.

4. Software processing

The coordinates of the over-threshold points are supplied by *FPSR* to the host computer in the same order in which the TV camera beam scans the frame (left-right, up-down). When rotatory movements are monitored, the mutual position of the markers may change and the order in which the markers are scanned may reverse; a marker classification is therefore necessary.

This procedure is based on a trajectory prediction for the single marker and on body modelling as an articulated structure of hinges connecting rigid links.[21,22]

The markers must be manually classified and assigned to the model points for the first two frames; then the software automatically provides for markers classification for all frames.

Three-dimensional reconstruction is carried on by a generalized stereophotogrammetric algorithm; for each marker the distance between the straight lines through image point and perspective centre of each TV camera is computed,[21] and the 3D position is set in the middle of this segment. The algorithm is very fast and allows a free positioning of TV cameras.

EXPERIMENTAL VERIFICATION OF THEORETICAL PERFORMANCES

General image processing (filtering, contour finding, oriented edge recognition) have been verified on the *ELITE* system. The performances of the system have shown a good agreement with the theoretical expectations even if, obviously, some limitations arose because of the reduced dimensions of the physical mask (6×6 elements) and the number of bits forming the single mask element (4 bits).

In Figure 7 a cinematographic projector is shown, in Figure 8 its edges have been recognized in real-time by the *ELITE* system. Modifying the weights of the correlation mask, we can recognize only horizontal edges (Figure 9), vertical edges or the edges forming a certain

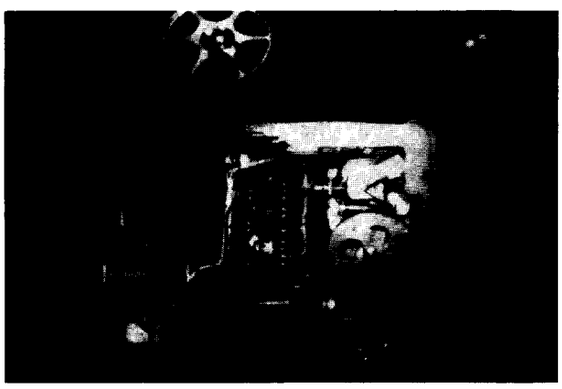

Fig. 7. Original image from which the subsequent figures are derived.

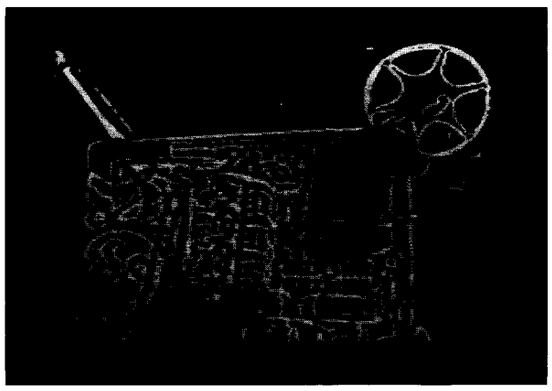

Fig. 8. Contour detection performed using a mask recalling the $\nabla^2 G$ function.

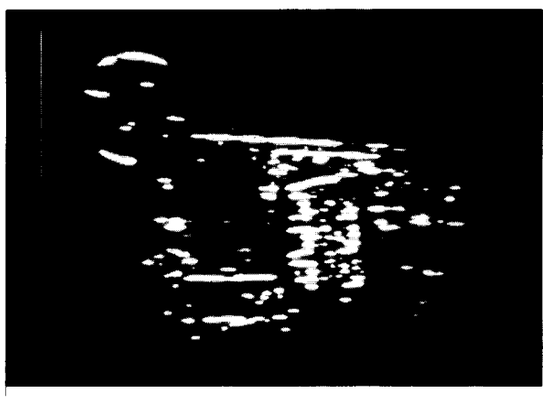

Fig. 9. Horizontal edges detection.

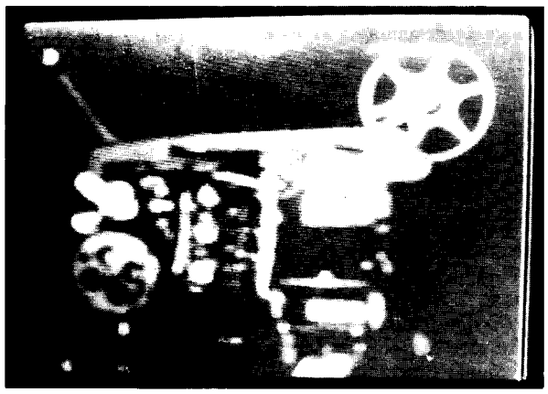

Fig. 10. Low pass filtered image.

Fig. 11. Subject to which markers have been attached.

Fig. 12. Real-time markers detection; on monitor the position of each marker is reported in real-time.

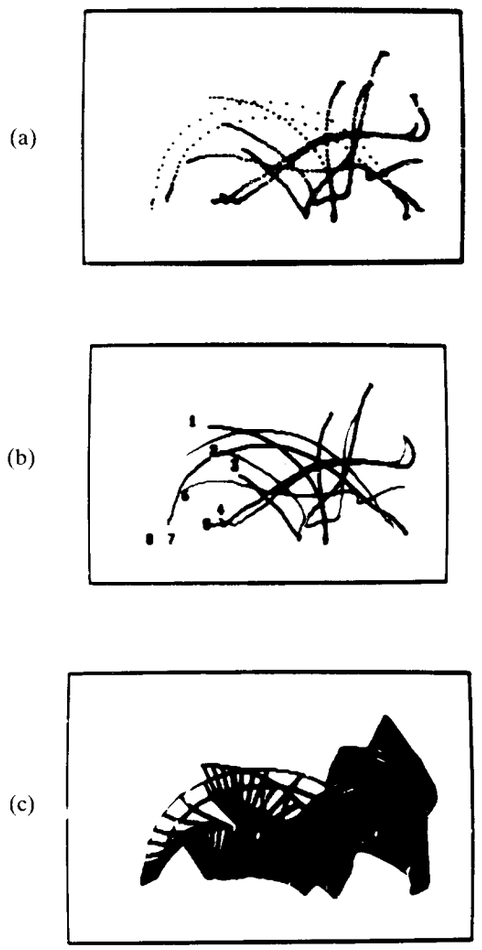

Fig. 13. Plot of the recognized points before classification (a), trajectories (b) and flick diagram (c) obtained after the classification phase.

angle with the horizontal plane, as well as filter the image with a low-pass filter (Figure 10).

Figures 11 and 12 show real-time shape detection in movement analysis. On a monitor (Figure 12) the position of the hemi-spherical markers attached on a running girl (Figure 11) is reported in real-time.

The results of software automatic on-line processing are reported in Figures 13a, 13b and 13c as an example of classification and restoration of a flick exercise.

CONCLUSIONS

The high reliability of the *ELITE* system for the automatic analysis of complex articulated movements has been assessed during four years of analysis of the performances of human subjects. In this period, the system has been heavily used either by technicians or non-specialized people in several lighting and environmental conditions, showing great reliability thanks to the shape detection real-time algorithm; its accuracy, tested in the field, proved to be of more than 1 part in 2500 of the field of view.

After a successful period of tests, the system is now commercially available; moreover, the *ELITE* system, although designed to recognize hemispherical markers, can also be used when detection of particular features is required, and also as a general purpose real-time image processor.

References

1. D. Marr, "Early Processing of Visual Information" *A.I. Memo* 340 (M.I.T. Artificial Intelligence Laboratory, Cambridge, 1975).
2. D. Marr, *Vision*, (Freeman, San Francisco CA, 1982).
3. M.S. Livingstone and D.H. Hubel, "Connections between leyer 4B of Area 17 and the thick cytochrome Oxidase Stripes of Area 18 in the Squirrel Monkey" *J. Neuroscience* **7** (11), 3371–3377 (1987).
4. D.C. Van Essen and J.H.R. Maunsell, "Hierarchical Organization and Functional Streams in the Visual Cortex" *Trends in Neurosciences TINS* **6,** No. 9, 370–375 (1983).
5. S. Ullman, *The Interpretation of Visual Motion*, (M.I.T. Press, Cambridge, MA, 1979).
6. D. Rumelhart, J. McClelland and the PDP Research Group, *PDP Parallel Distributed Processing* vols. 1, 2 (MIT Press Cambridge, MA, 1986).
7. J. Kittler and M.J.B. Duff (eds.), *Image Processing System Architectures* (Research Studies Press, Letchworth, UK, 1985).
8. M.J.B. Duff and S. Levialdi (eds.), *Languages and Architectures for Image Processings* (Academic Press, New York, 1981).
9. K.S. Fu and T. Ichikawa (eds.), *Special Computer Architectures for Pattern Processing* (CRC Press, London, 1982).
10. B.L. Bullock, *Computer Vision Systems* (Academic Press, New York, 1978).
11. R.A. Jarvis, "Application-oriented Robotic Vision, a review" *Robotica* **2,** part 1, 3–15 (1983).
12. H.K. Nishihara and N.G. Larson, "Towards Real-Time Implementation of the Marr-Poggio stereo-matcher" *Proc. Image Understanding Workshop* (edited by Lee Baumann, 1981).
13. H.B. Barlow and R.W. Levick, "The mechanism of directionally selective units in rabbit's retina" *J. Physiol.* **178,** 477–504 (1965).
14. D.R. Williams, "Seeing through the photoreceptor mosaic" *Trends in Neuroscience TINS* **9,** No. 5, 204–211 (1986).
15. C. Koch, T. Poggio and V. Torre, "Computations in the vertebrate retina: gain enhancement, differentiation and motion discrimination" *Trends in Neuroscience TINS* **9,** No. 5, 204–211 (1986).
16. G. Ferrigno and A. Pedotti, "*ELITE*: A Digital Dedicated Hardware System for Movement Analysis Via Real-Time TV Signal Processing" *IEEE Trans. Biom. Eng.* **BME 32,** 943–949 (1985).
17. A.V. Oppenheim and R.M. Schafer, *Digital Signal Processing* (Prentice Hall, Englewood Cliffs, N.J., USA, 1975).
18. D. Marr and E.C. Hildreth, "Theory of Edge detection" *Proc. Roy. Soc. London* **B 207,** 187–217 (1980).
19. E.C. Hildreth and C. Koch, "The analysis of Visual Motion: from computational theory to neuronal Mechanisms" *Ann. Rev. Neurosci.* **10,** 477–533 (1987).
20. A.L. Yuille and T.A. Poggio, "Scaling Theorems for Zero Crossing" *IEEE Trans. on Patt. Anal. and Mach. Intell.* **PAMI-8,** No. 1, 15–25 (1986).
21. N.A. Borghese, G. Ferrigno and A. Pedotti, "3D Movement Detection: A Hierarchical Approach" *Proc. 1988 IEEE International Conference on Systems, Man, and Cybernetics*" 1 (International Academic Publishers, Pergamon Press, Beijing 100044, China, 1988) pp. 303–306.
22. N.A. Borghese and G. Ferrigno, "A Knowledge based system for automatic movement tracking" **In**; *Expert Systems; theory and applications, Proceedings IASTED International Symposium Geneva, CH (June,* 1987) (Acta Press, Calgary, Alberta, Canada T2M 4L8, 1987) pp. 225–227.

ELITE: A Digital Dedicated Hardware System for Movement Analysis Via Real-Time TV Signal Processing

GIANCARLO FERRIGNO AND ANTONIO PEDOTTI

Abstract—The system illustrated in this paper has been designed and developed particularly for automatic and reliable analysis of body movement in various conditions and environments. It is based on real-time processing of the TV images to recognize multiple passive markers and compute their coordinates. This performance is achieved by using a special algorithm allowing the recognition of markers only if their shape matches a predetermined "mask." The main feature of the system is a two-level processing architecture, the first of which includes a dedicated peripheral fast processor for shape recognition (FPSR), designed and implemented by using fast VLSI chips. The second level consists of a general purpose computer and provides the overall system with high flexibility. The main characteristics are: no restriction on the number of markers, resolution of one part in 2500, and a 50 Hz sampling rate independent of the number of markers detected. The prototype has been fully developed, and preliminary results obtained from the analysis of several movements are illustrated.

INTRODUCTION

IN different areas of medicine and human science, the analysis of movement plays an important role: in neurophysiology—in order to achieve a better understanding of the basic mechanisms of movement control and the related strategies [1], [11]; in orthopaedics and motor rehabilitation— for more detailed and quantified functional diagnosis and therapy assessment [2], [10]; in neurology—moreover—the early diagnosis of cerebral lesions appears to be possible by detecting slight deviations from the norm which are not evident by simple visual inspection [14]. Back pain can be prevented by examining and quantifying load and movement during constrained work and, consequently, removing risk factors by a better design of the work place [5].

Athletic performance can also be improved by breaking down the movement into the elementary components and identifying the suitable reference models arising from the observation of outstanding athletes [12], [13].

For these reasons, much effort has been devoted to setting up easily workable equipment for movement analysis. Well known are Muybridge's studies [8] based on cameras. More recently, standard cameras and high-speed cameras, associated with suitable devices for facilitating frame reading and digitizing the meaningful coordinates, have been largely used [15]. The length of time required for manual reading, and the inherent imprecision, precluded wider use of these procedures which were effectively bound to the research laboratories. In the last few years, transducer advanced technology and microelectronics made possible the development of new systems for automatic movement analysis. The Selspot [16], [17], [19] uses light emitting diodes (LED) which must be fixed on the subject's chosen points, associated with a lateral effect photodiode as transducer. IR LED's are also used in the COSTEL system [7], the light of which after optical processing can be detected by linear CCD arrays. In both cases, the synchronization between the LED's flashing and transducing system is performed by telemetry. Wires and the necessary hardware must therefore be carried by the subject.

The problem of using active markers also arises with the TV systems which recognize markers by their brightness, the markers having to be brighter than the background. For this reason, some methods require light sources placed on the subject [3], [4], while the VICON system [6] uses IR lamps close to the lens of the camera to light reflective targets. In this case, particular care must be used to keep the subject in the shadow or in an environment without other IR sources.

In this frame, the ELITE (elaboratore di immagini televisive) system, which will be illustrated in this paper, possesses several features which make it particularly suitable for the applications cited at the beginning.

Operatively, it uses small hemispherical passive markers which are placed on the relevant points of the body. The standard TV camera used in the present version limits the sampling rate to 50 Hz (European standard) which has proved satisfactory for most applications. However, the system is designed to work with other inputs (i.e., CCD cameras or high-speed cameras), thereby allowing for a higher sampling rate.

The main innovative feature is represented by the two-level architecture of the processing system which includes the specially designed peripheral fast processor for shape recognition (FPSR). This device, which constitutes the core of the ELITE system, has been specially designed and its implementation has been made possible by using

Manuscript received July 17, 1984; revised June 18, 1985.

The authors are with the Centro di Bioingegneria, Dipartimento di Elettronica, Politecnico di Milano & Fondazione Pro Juventute Don Gnocchi, Via Gozzadini, 7 Milano, Italy.

0018-9294/85/1100-0943$01.00 © 1985 IEEE

very fast VLSI chips which have only become available on the market in the last few years.

It processes the TV image in real time and it uses a dedicated algorithm to recognize markers only if their shape matches a determined "mask."

Thanks to the algorithm used, the resolution is up to one part in 2500, with a standard analog TV camera (312 lines). Moreover, it recognizes markers more reliably, even in a noisy environment, including daylight, and there is no restriction on the number of markers which can be simultaneously detected. These features will be detailed in the following paragraphs. The first prototype of the system has been completely developed at the Bioengineering Center of Milan and, since September of 1983, it has been used to investigate selected human motor activities. Several results obtained during this preliminary activity will be illustrated in order to show the high reliability of the system.

Architecture of the System

The whole system has been designed to perform the following operations:

1) to recognize (in the environment) the presence of one or more objects of a predetermined shape (markers);

2) to compute the x and y coordinates of the marker centroids;

3) to perform the previous operations in real time (within 20 ms of the frame duration);

4) to classify the marker, that is, to attribute each marker to the proper point on the basis of a suitable model of the body which must be analyzed (system depending);

5) to perform routine data processing for

- distortion correction by calibrating procedures
- reconstruction of point trajectories by best fitting techniques
- three-dimensional (3-D) analysis by stereometric techniques when two or more cameras are used simultaneously;

6) to develop further data processing specifically devoted to the problem approached (i.e., a calculation of angular speeds, curvature ratio, correlation among several variables), generally based on a mathematical model of the system depending on the problem.

The operations 1) and 2) are necessary to allow recognition based on the marker shape and dimension. These characteristics are the most relevant of the system and make good performances possible, even in the presence of reflexes or noisy background.

The real-time processing is essential in order to avoid the storage of the large amount of data necessary to describe a sequence of images. Moreover, it allows the immediate assessment of the recognition quality.

The operations 1), 2), and 3) require very high-speed processing (approximately 5–9 ns/operation) which has been obtained by a specially designed full parallel hardware device. The operations 4), 5), and 6) are carried out by a general purpose computer.

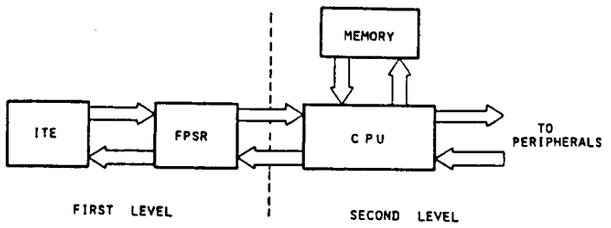

Fig. 1. ELITE two-level block diagram. The first level comprises the interface to the environment (ITE) and the special fast processor for shape recognition (FPSR). The second level (CPU) is implemented by a general purpose computer.

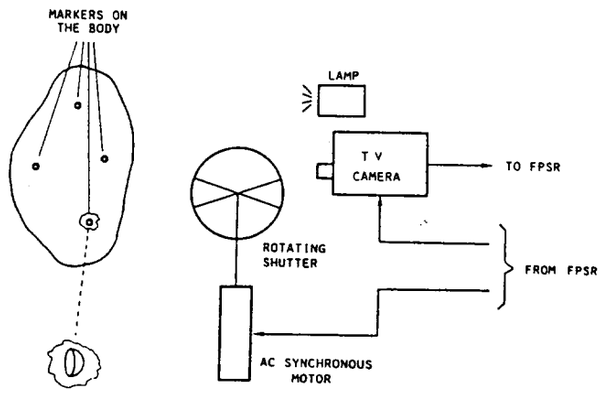

Fig. 2. Interface to environment (ITE) block scheme. The ac synchronous motor of the rotating shutter is driven by a sinusoid in phase with the vertical synchronism of the TV camera. Both are provided by the FPSR.

All the functions previously described are accomplished by the architecture shown in Fig. 1. The fast processor for shape recognition (FPSR) performs cross-correlation processing on the incoming digitized TV signal, recognizes the markers and computes their coordinates, allowing a data reduction of about 1000. The FPSR unit is doubly connected to the "interface to the environment" (ITE) because it not only receives the input data, but also provides the necessary signals for synchronization.

The FPSR consists of dedicated digital hardware implementing the algorithm used to detect the markers on the scene to compute the coordinates of their centroids and to send them to the central processing unit (CPU). The CPU is a general purpose computer which performs the remaining operations.

The Interface to the Environment (ITE)

As illustrated in Fig. 2, the interface to the environment consists of four different components: markers, shutter, TV camera, and a lamp. Because the recognition of the markers is accomplished by using information about their shape, the shutter is required to avoid shape distortions due to the nonsimultaneous reading of information by the electronic beam of the TV camera. The shutter acts as a stroboscope and leaves the lens open for approximately 1.5 ms, which is a duration short enough to detect movements at relatively high speeds. The shuttering is performed by a disk with a transparent narrow window on a dark opaque surface. The disk is driven by a synchronous

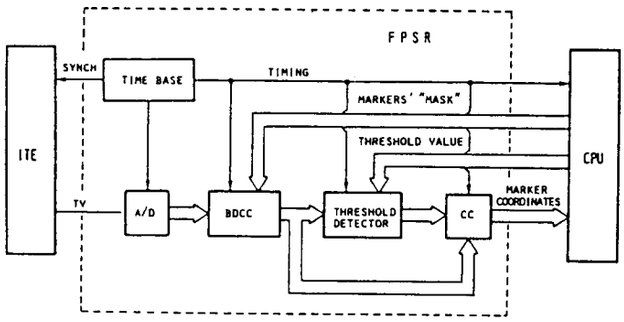

Fig. 3. Fast processor for shape recognition (FPSR) block scheme (see the text for further details).

motor and is powered by a sinusoidal current in phase with the vertical synchronism of the TV camera. Both the sinusoid and the vertical synchronisms are generated by the FPSR.

The markers are recognized on the basis of their shape; thus allowing any shape to be chosen by simply adjusting the "mask" of the FPSR (squares, crosses on black or white). In practice, hemispheric reflective markers are used for the following reasons:

- they can be easily fixed to the body;
- their image does not change if they rotate on their axis of symmetry;
- their image does not significantly change if they rotate on the other two axes up to approximately $\pm 70°$;
- the reflective material increases the contrast, thus improving recognition reliability.

For these reasons, they are well detected even for nonplanar movements. The marker dimensions can be varied depending on the size of the segments analyzed, on the amplitude of the movement, and on the distance from the TV camera.

For instance, in gait analysis (subject moving 7 m from the TV camera), diameters of 5 mm appeared to be the best choice. A highly sensitive and low-persistence TV camera, equipped with an Ultricon tube, has been chosen in order to avoid ghost images. An image every 20 ms is considered, and the analog signal is sent to the FPSR for real-time processing.

Fast Processor for Shape Recognition (FPSR)

The fast processor for shape recognition (FPSR) has been specially designed and implemented via hardware to perform peripheral processing of the tremendous quantity of data coming from the TV camera (Fig. 3). It constitutes the first level of the system (first degree of intelligence) and operates under the direct control of the CPU. The analog TV signal is digitized at a sampling frequency of 5 MHz, corresponding to 256×256 useful pixels matrix. Sixteen gray levels are considered and coded by 4 bits. The input data, therefore, is 256×256 matrix of 4 bits pixels every 20 ms. The CPU provides the "mask" related to the shape to be detected.

The shape detecting algorithm (SDA), essentially based on a bidimensional cross correlation (BDCC) between the actual digitized image and the predetermined "mask," is implemented by a parallel hardware structure allowing the real-time processing. The output from the FPSR are directly the "r" couples of horizontal and vertical coordinates of the "r" markers detected which are delivered to the CPU (second level) for further elaboration.

Functional Description of Shape Detecting Algorithm (SDA)

Two-dimensional representation of the TV image is easily obtainable by naming the pixels by their row and column indexes, that is,

$$P_{11}, P_{12}, \cdots, P_{1N}, \cdots, P_{21}, \cdots, P_{ij}, \cdots, P_{MN}$$

where we have considered an $M \times N$ matrix of pixels. The total TV signal contains both the marker shape and the background scene which, for the purpose of the present analysis, is considered as noise. The method used for the extraction of the first from the second is obtained by BDCC. If we define

$$A(x, y) = \text{marker shape} \tag{1}$$

$$A'(x, y) = \text{correlation function which is the predetermined "mask"} \tag{2}$$

$$T(x, y) = \text{total TV signal} \tag{3}$$

$$F(x, y) = \text{background signal} \tag{4}$$

we have for the total signal

$$T(x, y) = A(x, y) + F(x, y) \tag{5}$$

with the condition

$$F(x, y) = 0 \qquad \text{where} \quad A(x, y) \neq 0. \tag{6}$$

The cross correlation between the functions $T(x, y)$ and $A'(x, y)$ is given by [9]

$$R_{A',T}(h, k) = \int_{-\infty}^{\infty}\int A'(x, y)\, T(x + h, y + k)\, dx\, dy \tag{7}$$

by substituting (5) into (7), we obtain

$$R_{A',T} = R_{A',A} + R_{A',F} \tag{8}$$

where

$$R_{A',A} = \int_{-\infty}^{\infty}\int A'(x, y)\, A(x + h, y + k)\, dx\, dy \tag{9}$$

and

$$R_{A',F} = \int_{-\infty}^{\infty}\int A'(x, y)\, F(x + h, y + k)\, dx\, dy. \tag{10}$$

In the case of sampled signal and of a limited dimension of the shape to be recognized, (7) can be written as fol-

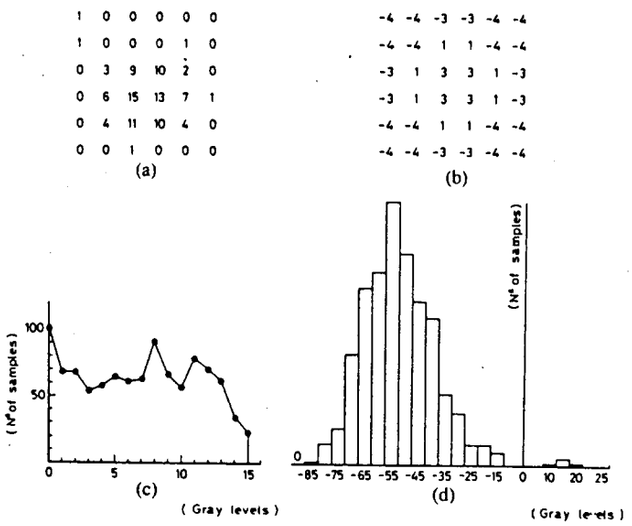

Fig. 4. (a) Marker shape (matrix representation). (b) "Mask" shape (matrix representation). (c) Distribution of the gray values in the simulated scene. (d) Distribution of the cross-correlation values after the elaboration of the scene.

lows:

$$R_{A',T}(h, k) = \sum_{0m}^{M'} \sum_{0n}^{N'} A'(m, n)\, T(m + h, n + k) \quad (11)$$

and

$$R_{A',A} = \sum_{0m}^{M'} \sum_{0n}^{N'} A'(m, n)\, A(m + h, n + k) \quad (12)$$

$$R_{A',F} = \sum_{0m}^{M'} \sum_{0n}^{N'} A'(m, n)\, F(m + h, n + k) \quad (13)$$

where M' and N' represent the width and the height of the "mask," able to contain the shape of the marker, which must be detected. The mask is therefore composed of $M' \times N'$ pixel with $M' < M$ and $N' < N$.

The problem is now reduced to the determination of the function $A'(x, y)$ that maximizes the term $R_{A',A}$ of (8) and minimizes the $R_{A',F}$. In this way, the maxima of the function $R_{A',T}$ will correspond to those of $R_{A',A}$. That is, A' must have low correlation with F and high correlation with A.

Fig. 4(a) shows the digitized marker shape as it appears after the A/D converter. It is characterized by a white spot surrounded by a black background.

As the $T(x, y)$ is a nonnegative function, a general criterion to determine the A' function is to choose a duplicate of the shape to be recognized, and to subtract a constant from each pixel, so that the mean value becomes negative. This condition gives a negative response at zero space frequency (180° phase shift), i.e., a positive spatial plane is turned into a negative one and, in general, a rejection of shapes different from the markers is obtained. Moreover, this choice criterion can be used in self-learning procedures where more complex shapes must be recognized. The A' shape, chosen for our particular case, is reported in Fig. 4(b).

In order to test the efficiency of the considered A' mask, simulation techniques have been adopted. The TV signal was simulated either as white noise or as a sum of particular shapes. The simulated signal was cross correlated with the A' function, and the results were printed and examined. To reproduce the operating conditions, where small markers are present in the TV scene, a 6 × 6 pixel mask was considered. The outcomes are reported in Fig. 4(c) and (d).

In this case, the peak of the total signal noise plus the useful signal was 15. To create the worst condition, the peak of the noise was chosen 100 percent, higher than the marker signal, the peak of which was chosen 70 percent of the maximum.

So the ratio between signal and background noise was lower than unity. Moreover, a Gaussian noise with zero mean and an SNR of 4 was superimposed on the useful signal. The gray value distribution before the processing is reported in Fig. 4(c). It is evident that, in this case, a threshold operation is not able to discriminate the marker from the background.

By using the described algorithm, the distribution of cross-correlation values obtained after processing is reported in Fig. 4(d). In this case, the positive values are relative to the useful signal, while the background noise has been shifted in the negative area. It is now very simple to set an adequate threshold in order to achieve reliable detection and minimize the false alarm occurrence.

Algorithm Implementation

The cross-correlation summation has been implemented by a full parallel processor. The new building block availability for correlation allows an easy implementation of BDCC algorithm. The cross-correlated signal is compared to predetermined threshold value and the overthreshold point coordinates are considered as a probable marker component.

The processing is performed at 5 MHz clock rate, i.e., a cross-correlation value is computed every 200 ns. The clock frequency affects both the basic resolution of the system (256 × 256) and the sample frequency (50 Hz).

The system resolution is increased by calculating the centroid of the overthreshold points relative to each marker on line, producing a more efficient data reduction. The block scheme of the processor is shown in Fig. 5. The *line correlators* are standard building blocks, the *line delays* delay a TV line in order to realize a bidimensional correlator. The centroid calculation block (CC) utilizes a look-up table to accept a candidate marker and calculates its centroid coordinate offset.

Centroid Calculation (CC)

Once the threshold detection has been performed, the centroid of the overthreshold point of the BDCC is calculated. This operation allows a remarkable increase of the resolution (approximately 10 times) and precision of the system.

The points over the threshold form a cluster like that

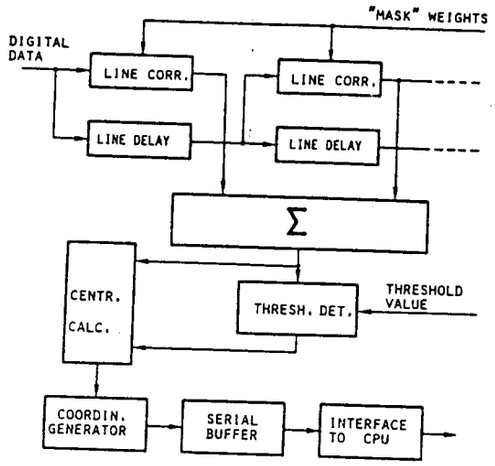

Fig. 5. Block diagram of the bidimensional cross correlator.

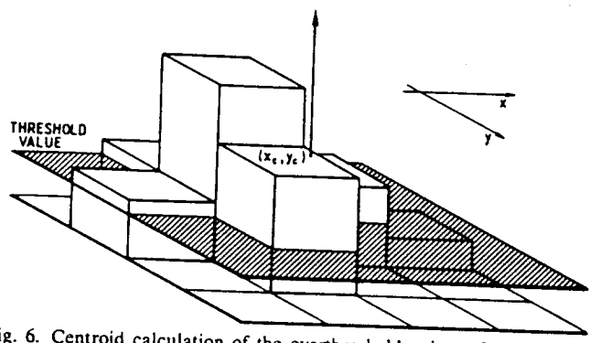

Fig. 6. Centroid calculation of the overthreshold points of the cross correlation function. The centroid is in the position of coordinates x_c, y_c.

shown in Fig. 6. The centroid calculation is performed by using the following formulas:

$$x_c = \frac{\sum_i x_i \sum_j R_{ij}}{\sum_{ij} R_{ij}} \qquad y_c = \frac{\sum_j y_j \sum_i R_{ij}}{\sum_{ij} R_{ij}} \tag{14}$$

where x_c, y_c are the centroid coordinates and R_{ij} is the value of the BDCC of the overthreshold point P_{ij} of coordinates x_i, y_j.

Spatial Accuracy with Speed

The general assessment of the spatial accuracy has been performed both in static and dynamical conditions. For the static case, a marker was placed on a tangent screw. In Fig. 7(a), the coordinates provided by the ELITE system are reported versus the screw displacement. The standard deviation of the distances between the experimental points and the regression line is $\sigma = 0.06$ pixels.

The maximum error detected was 0.12 pixels. The error value has been assumed

$$E = 1.64\sigma = 0.98 \text{ pixels.}$$

This assumption was made because, in normal distribution, the 95 percent of the values is bound within 1.64σ if the mean value equals zero as in this case.

The hypothesis of normality is supported by the fact that the error E is the sum of the errors E_i of the single pixels

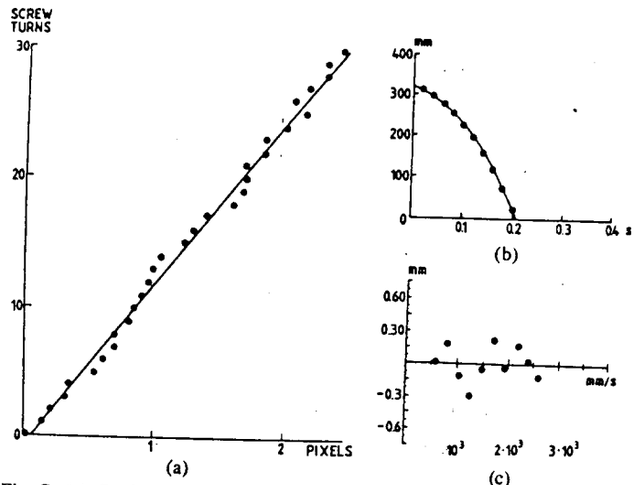

Fig. 7. (a) Static accuracy of ELITE system: coordinates provided by ELITE system versus screw displacement. (b) Experimental points of a falling body fitted with a parabola. (c) Deviations of experimental points from the fitting parabola versus body speed.

from which the centroid (x_c, y_c) coordinates are computed. These errors are uniformly distributed and practically uncorrelated, leading to the conclusion that the distribution of the stochastic variable E approximates a normal one because of the central limit theorem. The dynamic accuracy has been tested by analyzing a free falling body. A steel ball with marker shape was used for this purpose. The experiment was repeated 12 times.

In Fig. 7(b), the measured points and the best fitting parabola are illustrated for one of the experiments, while in Fig. 7(c), errors are reported as a function of speed. Ten different speeds were considered ranging from 10 to 1500 pixels/s (corresponding approximately to 0.10 : 15 m/s, if 2.5 m × 2.5 m field is considered).

For each speed, the standard deviation was computed from the 12 tests performed. No significant variation of σ versus speed was found [as also shown in Fig. 7(c)]. The mean value of σ for all the considered speeds was 0.09 pixels. Therefore, the corresponding error is $E = 1.64\sigma = 0.15$ pixels.

It must be pointed out that, in a dynamic condition, the marker crosses the all field, therefore, the resultant value of E also includes the calibration errors.

Central Processing Unit (CPU)

After the first-level processing, the information (marker coordinates) is transferred to the CPU for the second-level elaboration. This latter process comprehends the identification of the set of markers (marker classification), calibration, and calculation of various physical quantities. The algorithm for marker classification varies with application and it is based on a mathematical model of the system. The performed calibration algorithm allows fast position correction by using a polynomial expression. Physical quantities calculated are angles, velocities, accelerations, torques, and they vary with the particular application. The second-level elaboration is articulated in four main blocks.

Phase I must be executed necessarily in real time. The

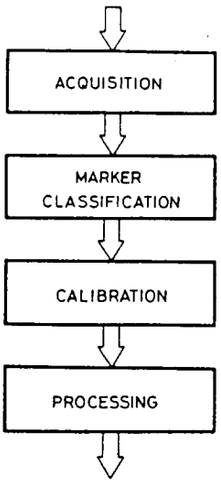

Fig. 8. Flowchart of the four phases of the data processing performed at the second level.

Fig. 9. The developed prototype of the ELITE system. The hardware illustrated includes the dedicated FPSR: the input is the analog TV signal, the output are the coordinates of the markers; the microswitches of the front panel allow selection of the choice of the shape; the grid for calibration is appearing on the monitor screen. A marker is placed on each line crossing during the camera calibration.

computer does not need to be particularly powerful. In fact, during this phase, it only has to collect data from the FPSR and store them. A PDP 11/03 laboratory computer equipped with floppy disks and an A/D converter for contemporaneous collection of other data (i.e., from the force plate) has been used.

Phases II, III, and IV are executed by the computer. Special software has been developed for graphic output (stick diagrams and other diagrams).

Acquisition

Data collection is performed on line by macrosubroutines via a standard interface bus (IEEE-488). The data collected are stored on a vector in the central memory and transferred into disks after the completion of the acquisition. It should be also possible, if necessary, to increase the number of markers by using a DMA, and the acquisition time by a simultaneous mass storage of data.

Calibration

The calibration is necessary to transfer the distorted camera coordinates into the undistorted world coordinates. A grid of (5 × 4) markers is used for this purpose. Several examples of global calibration in two and three dimensions are shown in [18]. In the ELITE system, the two correlation phases (scaling and distortion correction) have been merged in a unique calibration procedure specifically oriented to local error correction.

As the number of squares in the grid increases, the local correction approaches the global case. A transformation of the kind (14 a, b) is applied to the coordinates of the point to be calibrated.

$$X_{jk} = A_{j1}x_{jk} + A_{j2}y_{jk} + A_{j3}x_{jk}y_{jk} + A_{j4}x_{jk}^2 + A_{j5}y_{jk}^2 + A_{j6} \quad (14a)$$

$$Y_{jk} = B_{j1}ty_{jk} + B_{j}i2y_{jk} + B_{j3}y_{jk}x_{jk} + B_{j4}x_{jk}^2 + B_{j5}y_{jk}^2 + B_{j6} \quad (14b)$$

where A_{jl} and B_{jl} are coefficients depending on camera and real coordinates of the Jth grid square vertex; x_{jk}, y_{jk} are the camera coordinates of the point k belonging to the jth square; and X_{jk}, Y_{jk} are the global coordinate estimation.

The features of this calibration technique are the efficiency, the speed in calculating coefficients A_{ij} and B_{ji}, and the speed in calibrating each point, so that it can be frequently repeated at the beginning of each experimental session.

Processing

The processing phase is performed in order to extract information of general interest from raw data. Generally, data filtering is needed before further processing. In fact, raw data are affected by a certain quantization noise whose power is (15) [9]

$$\sigma^2 = \frac{a^2}{12} \quad (15)$$

where "a" is the quantization interval. Generally, when a marker is moving, this noise has a high-frequency spectrum, so that it can be eliminated by using a low-pass or smoothing filter. For nonperiodic events, time domain filtering is used. After filtering, many physical quantities can be calculated, depending on the specific application.

Applications and Results

The prototype of the system (shown in Fig. 9) has been widely used for analyzing different types of movements and special programs have been developed for lower limb movements, hand movements, and jaw articulation movements.

The complete results of these studies will be illustrated in other papers. Here, a synthetic description of marker classification procedure adopted for limb movements will be reported.

Marker Classification in Lower Limb Case

Having collected the data, each marker must be classified, that is, attributed to the corresponding point of the body analyzed.

It is necessary to recognize the marker pattern by verifying the agreement between a given model and the pattern detected. For the analysis of the lower limb, markers

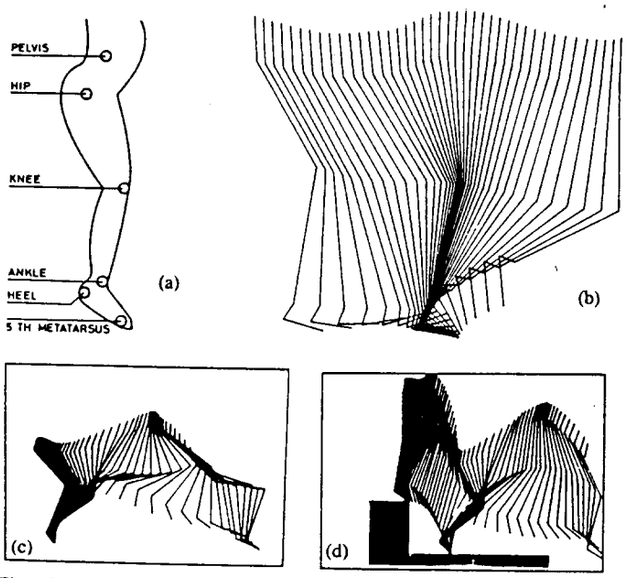

Fig. 10. (a) Example of disposition of markers on the subject (complete 6 marker set). (b) Stick diagram of a subject in normal level walking obtained by ELITE. (c) Stick diagram of a subject jumping forward. (Only 5 markers have been used in this case.) (d) Stick diagram of a subject jumping from a platform.

have been placed on the pelvis, hip, knee, ankle, heel, and metatarsus [see Fig. 10(a)].

In the first frame, all markers are named under the hypothesis that the pelvis marker has the highest y coordinate. The hip marker is recognized as the nearest to the pelvis one. The knee marker is the furthest from the centroid of the last four. Once the first three markers have been assigned, the centroid of the last three is computed and the marker furthest from this is the one of the metatarsus. The distances between the knee and the last two markers are computed, the heel marker is the furthest, while the ankle one is the nearest. If the hypothesis that the upper position is the pelvis marker proves true, correct marker assignment is performed for the first frame. Similar considerations on the leg structure, combined with tracking procedures, allow for the marker classification, even if the initial hypothesis of the pelvis y coordinate is not true during the movement.

If a marker disappears for a few frames, its coordinates are reconstructed by interpolation. Fig. 10(b)–(e) shows three examples of the high degree of reliability of the system: the first is a step during normal level walking, the second is a subject jumping forward, and the last is the stick diagram of a subject jumping from a platform. The optimal detection of the movement in spite of the high speed involved must be emphasized.

Conclusions

The intensive experimental use of the ELITE system has shown its reliability and efficiency in kinematic analysis of human movement. An improvement of the system should be possible by increasing the sampling rate by using CCD or a high-speed tube camera.

The ELITE system could also be used as an image processing system for industrial applications, standalone, or integrated into an automated factory. Compared to analogous systems already developed, two features appear particularly interesting. The high speed could be usefully adopted in special applications or for improving the reliability of the recognition. The two-level structure provides a high degree of flexibility and allows the development of self-learning procedures.

(For the device described here, the patent application number 11.06/615 442 was forwarded to the U.S. Department of Commerce—Patent and Trademark Office—on May 30, 1984.)

References

[1] N. A. Bernstein, *The Coordination and Regulation of Movements*. Oxford, England: Pergamon, 1967.

[2] S. Boccardi, A. Pedotti, R. Rodano, and G. C. Santambrogio, "Evaluation of muscular moments at the lower limb joints by an on line processing of kinematic data and ground reaction," *J. Biomech.*, vol. 14, no. 1, pp. 35–45, 1981.

[3] I. S. Cheng, "Computer television analysis of biped locomotion," Ph.D. dissertation, Dep. Elec. Eng., Ohio State Univ., Columbus, 1979.

[4] I. S. Cheng, S. H. Koozekanani, and M. T. Fatchi, "Computer-television interface system for gait analysis," *IEEE Trans. Biomed. Eng.*, vol. BME-22, p. 259, May 1975.

[5] S. Cantoni, D. Colombini, E. Occhipinti, A. Grieco, C. Frigo, and A. Pedotti, "Posture analysis and evaluation at the old and new workplace of a telephon company," in *Ergonomics and Health in Modern Offices*, E. Grandjean, E. Taylor, and Francis, London and Philadelphia, 1984.

[6] M. O. Jarrett, B. J. Andrews, and J. P. Pau, "A television/computer system for the analysis of human locomotion," in *IERE Conf. Proc.*, vol. 34, 1976, pp. 357–370.

[7] T. Leo and V. Marcellari, "On line microcomputer system for gait analysis data acquisition, based on commercially available optoelectronic devices," in *Biomechanics VII*. Baltimore, MD: University Park Press, 1979.

[8] E. Muybridge, *The Human Figure in Motion*. London, England: Chapman & Hall, 1901.

[9] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[10] A. Pedotti, "Simple equipment used in clinical practice for evaluation of locomotion," *IEEE Trans. Biomed. Eng.*, vol. BME-24, pp. 456–461, 1977.

[11] A. Pedotti, V. V. Krishnan, and L. Stark, "Optimization of muscle-force sequencing in human locomotion," *Math. Biosci.*, no. 38, pp. 57–76, 1977.

[12] A. Pedotti, C. Frigo, and R. Rodano, "Optimization of motor coordination in sport, An analytical and experimental approach," in *Biomechanics and Performance in Sport*, Baumann, Ed. Schorndorf, West Germany: V. K. Hofmann, Bundesinstitut fur Sportwissenschaft, vol. 40, 1983, pp. 145–160.

[13] A. Seireg, "Optimum control of human movement trajectories, forces and constraints," in *Biomechanics and Performance in Sport*, Baumann, Ed. Schorndorf, West Germany: V. K. Hofmann, Bundesinstitut fur Sportwissenschaft, vol. 40, 1983, pp. 161–171.

[14] C. Terzuolo, F. Lacquaniti, and P. Viviani, "The law relating the kinematic and figural aspects of drawing movements," *Acta Psychologica*, vol. 54, nos. 1–3, pp. 115–130, Oct. 1983.

[15] D. A. Winter, R. K. Greenlaw, and D. A. Hobson, "Television computer analysis of kinematics of human gait," *Comput. Biomed. Res.*, vol. 5, pp. 498–504, 1972.

[16] H. J. Woltring, "New possibilities for human motion studies by a real time light spot position measurement," *Biotelemetry*, vol. 1, pp. 132–146, 1974.

[17] ——, "Single and dual axis lateral photodetectors of rectangular shape," *IEEE Trans. Electron Devices*, vol. ED-22, pp. 580–581, 1975.

[18] ——, "Calibration and measurement in 3-dimensional monitoring of human motion by optoelectronic means," *Biotelemetry*, vol. 3, pp. 65–97, 1976.

[19] H. J. Woltring and E. B. Marsolais, "Optoelectronic (SELSPOT) gait measurement in two and three dimensional space. A preliminary report," *Bull. Prosth.*, vol. 17, pp. 46–52, 1980.

Giancarlo Ferrigno was born in Pozzuoli (Napoli), Italy, on March 3, 1958. He received the doctoral degree in electronic engineering in 1983 at the Politecnico of Milan, Italy.

Since then he has been working at Centro di Bioingegneria (Pro Juventute Fnd.—Politecnico of Milan) on systems and methodologies for posture and movement analysis related to sport, rehabilitation, and neurophysiology.

Antonio Pedotti was born in Voghera, Italy, on March 21, 1944. He received the doctoral degree in electronic engineering from the Politecnico of Milan, Italy, in 1968.

Since 1968 he has been a Researcher of the National Research Council. Presently he is Professor at the Engineering Faculty, Politecnico of Milan and Director of the Bioengineering Center of Milan, Milan, Italy. His current activities in research concern the analysis of biological systems, with special regard to posture and locomotion medical informatics and technological developments.

IEEE TRANSACTIONS ON

BIOMEDICAL ENGINEERING

NOVEMBER 1985 VOLUME BME-32 NUMBER 11 (ISSN 0018-9294)

A PUBLICATION OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY

PAPERS

COMMUNICATIONS

BOOK REVIEWS