MVA'90 IAPR Workshop on Machine Vision Applications Nov. 28-30,1990, Tokyo

OBJECT MOTION IDENTIFICATION FOR OBJECT RECOGNITION

Vito Cappellini, Alberto del Bimbo, Paolo Nesi

Facolta' di Ingegneria, Dipartimento di Sistemi e Informatica
Universita' di Firenze, Via S. Marta 3, 50139 Firenze, Italy

ABSTRACT

Motion analysis has been used for a long time in vision to derive the 3-D shape of the moving object from the image sequence as well as to derive the motion law for prediction and tracking. However, motion can also be regarded as a property of the object and hence employed for improving recognition. This is particularly useful when the form provides little or no help in discriminating between different hypotheses. In this paper, an approach is presented in which motion descriptors are defined at different levels of abstraction. Coarse descriptors refer to the motion of rigid objects. Finer descriptors model the motion of composite objects with coordinated moving subparts. Autoregressive models are used to provide motion descriptor estimations.

1. INTRODUCTION

The analysis of dynamic scenes has a long history in computer vision. Many researchers addressed such a task in order to recover the 3D object structures or to derive the object motion for predicting and tracking the object temporal evolution. Several approaches are available in the literature, leading to solve possibly complex sets of linear [1], [2] or non-linear [3], [4], [5] equations. In most of them, typical assumptions are that the object is a rigid body and translational and rotational motions do not change in the observation window. The most noticeable techniques and results in this research area are reported in [6].
Object motion analysis has also been used for object classification purposes. In this case, the job is to reason about a set of properties describing the actual object motion. To this end important results have been obtained in the ALVEN system for ventricular motion analysis [7], [8].
A different approach has been followed in [9], [10], where motion features are regarded as properties of the object itself and used for improving classification. This can be helpful in situations where the form is not enough to perform recognition (e.g. the case of night processing of different light sources, or of undefined shapes like smoke and clouds, where shapes are of no use for recognition or the object forms are not fixed).
In this paper this approach is developed further. We want to define structures which model the actual object motion, and that can be profitably used in the classification task.
If the behavior is considered as a property of the object, several descriptions are possible depending on the level of abstraction at which the object is regarded.
Coarse descriptions refer to a view of the object as a single whole. Examples are views of objects as blobs or rigid bodies without components. In this case, motion parameters can be simply estimated by observing the centroid movement. In the present approach, the object motion is modeled through an autoregressive model. The multidimensional field of the possible coefficients represents the allowed object motions and is used as a reference in the recognition task.
Finer descriptions are related to modeling the movements of composite objects with subparts moving in coordinated motion. In particular, the following factors have to be taken into account:
- identification of the joints in the overall object body;
- modeling of the time-varying relationships among the subparts and of the constraints;
- modeling of the coordinated motions.

The reference model has to collect both the allowed motion patterns of the joints and the allowed pattern correspondences. Motion descriptors are stored as features in the object models. Each model has to give its confidence of matching the observed behavior.
Apart from these descriptions, we have to consider that objects a exhibit different behavior depending on the context in which they are observed. Therefore, a useful approach is to partition the set of the allowed motions according to the context information.
This has been proposed in [9], by adopting a specific object-based information model.
This paper is organized as follows: In Sect.2 the coarse model used to describe motion of simple objects is presented with a brief description of how behavior matching is supported. Sect.3 contains the corresponding description for the case of composite objects with subparts in coordinated motion. Some considerations related to confidence updating as time progresses are expounded in Sect.4. In Sect.5 an example of behavior-based recognition for simple objects is shown. Conclusions are given in Sect.6.

2. MOTION DESCRIPTORS FOR SIMPLE OBJECTS

Generally speaking, in a Cartesian coordinate system, as that in Fig.1, the 3-D motion from a generic point $P(t)\,(x,y,z)$ to $P' = P(t+\Delta t)\ \ (x',y',z')$, under translation and rotation is described as:

$$(2.1)\quad \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R_r \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix}$$

and the following relationships hold, between the projections of P and P' on the image plane:

$$(2.2a)\quad X' = L_f \frac{r_{11}X + r_{12}Y + r_{13} + \dfrac{\Delta x}{\Delta z}\dfrac{\Delta z}{z}}{r_{31}X + r_{32}Y + r_{33} + \dfrac{\Delta z}{z}}$$

$$(2.2b)\quad Y' = L_f \frac{r_{21}X + r_{22}Y + r_{23} + \dfrac{\Delta y}{\Delta z}\dfrac{\Delta z}{z}}{r_{31}X + r_{32}Y + r_{33} + \dfrac{\Delta z}{z}}$$

where:

a) r_{ij} are the elements of R_r and represent the entity rotational motion components, defined as a function of cos θ, sin θ, cos Φ, sin Φ, cos ψ, sin ψ;
b) $\Delta x, \Delta y, \Delta z$ represent the translational motion components;
c) (X,Y) and (X',Y'), are the coordinates of the projection of P and P' on the image plane normalized with respect to z and z'.
d) L_f is the camera focal length.

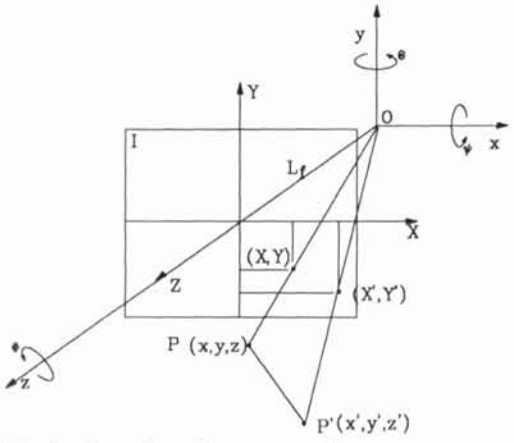


Fig.1 - Cartesian reference coordinate system for a simple moving object.

In general, the 3-D motion estimation problem can be subdivided in two steps. The first step is to estimate the object displacement in the image space by matching several points of the object in consecutive frames.

These points are used in order to solve for the unknown parameters $\Phi, \theta, \psi, \Delta x/\Delta z, \Delta y/\Delta z$, in the above equations, being Δz the scale factor.
As observations can be noisy, estimation techniques are used as a second step to derive motion descriptors with sufficient accuracy. A rigid object undergoing rotational and translational motions can be modeled as a discrete dynamic system. An extended Kalman filter was used to this end in [5]. Even a good precision is obtained, convergence is reached only with a great number of frames.
Here we was used an autoregressive model with minimum square error (MSE) filter to perform the estimation of the motion descriptors. Experimental results have proven that this technique performs well with faster convergence. In this case, the model is:

$$(2.3)\quad s(t) = a_1 s(t-\Delta t) + a_2 s(t-2\Delta t) + \ldots + a_n s(t-n\Delta t) + u(t)$$

where:

a) $s(t)$ is the vector of measurable parameters, and should include parameters such as $x_m, y_m, \dot{x}_m, \dot{y}_m, \Phi_m, \dot{\Phi}_m, \theta_m, \dot{\theta}_m, \psi_m, \dot{\psi}_m$,
b) a_i are unknown matrices (to be estimated) that should be regarded as descriptors of the actual object motion.

Estimated descriptors model the behavior of the object in the observation window. At each instant t the set $\hat{A}(t)$ of the coefficients a_i, is estimated through the MSE filter (see Fig.2), on the basis of the previous values and of the error $E(t-1)$, $E(t)$ being defined as:

$$E(t) = s_p(t) - s(t)$$

Some initial steps are usually needed to stabilize the coefficients and to initialize the MSE filter. However, the speed of convergence weakly depends on the initial guess for the unknown parameters.

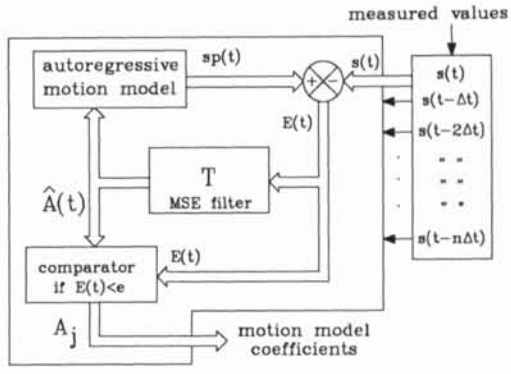


Fig.2 - Autoregressive motion model with MSE filter.

In order to perform recognition by behavior, the estimated descriptors are then compared with those stored in the object models. In particular, each model stores a set M of reference patterns. which describes all the allowed motions for the specific object, in the context selected. M is defined as:

$$M = \{A_j \mid j = 1, m\}$$

with: $A_j = \{a_i \mid i = 1, n\}$

and: $a_i = \{\alpha_{kl} \mid k = 1, p; \quad l = 1, p\}$

where m is the cardinality of the parameter field, n the order of the modeling motion equation and p the dimension of the space vector.
In practice, only the bounds of the multidimensional field M of the motion descriptors are stored into each model, in the form of a fuzzy membership function. Thus, the confidence $0 \leq \mu_m \leq 1$ of matching the observed motion, is computed by each model through the embedded fuzzy-matching procedures.

3. MOTION DESCRIPTORS FOR COMPOSITE OBJECTS WITH COORDINATED MOVING SUBPARTS

In this section we will discuss how composite objects with coordinated moving subparts can be described and how these descriptions should be used for recognition.
In this case, a distinction has to be made between the object body and the other moving subparts. We will assume that the object body corresponds to the subpart that exhibits a minor variance of its displacements. This classification can be easily obtained after few frames. In addition, joints connecting two or more subparts have to be considered. Every joint can be regarded as a point with several grades of freedom. It can be identified as the object point which remains fixed with respect to the other points of the moving subparts.
For the following discussion, several reference coordinate systems are defined (see Fig.3). One is centred in the centroid of the object body (center of motion). If the composite object has n subparts, $n-1$ coordinate systems are also defined, centred in the subpart joints. Finally an absolute coordinate system centred in the camera focus is defined.
As to the motion of the object body, the same approach followed in Sect.2 can be used for the evaluation of displacements and motion parameters, as well as for the estimation of the global motion descriptors.
However, some descriptors of the relative motion between subparts are also needed. These are a synthetic repre-

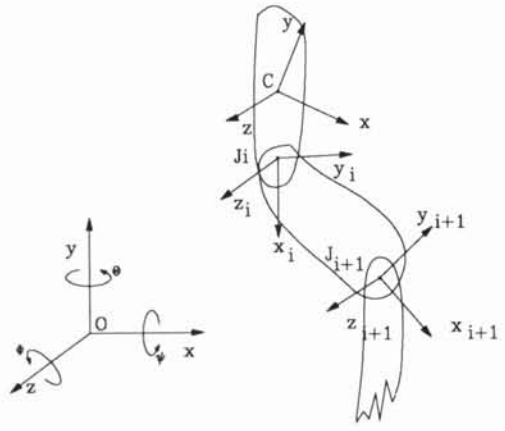


Fig.3 - Cartesian reference coordinate systems for a composite object with moving subparts.

sentation of the relationships between the object center of motion and the joints.
These relationships are modeled according to general transformation matrices in homogeneous coordinates. In particular, if the generic joints J_i, J_{i+1} are chosen, the following equations hold:

$$J_i = F_{i,i+1}\ J_{i+1}$$

where:

a) $F_{i,i+1} = R_{i,i+1}\ T_{i,i+1}$

b) R is the composite rotation matrix for the generic rotations Φ, θ, ψ with respect to the z, y and x axis, respectively:

$$R = \begin{bmatrix} C\Phi C\theta & C\Phi S\theta S\psi - S\Phi C\psi & C\Phi S\theta C\psi + S\Phi S\psi & 0 \\ S\Phi C\theta & S\Phi S\theta S\psi + C\Phi C\psi & S\Phi S\theta C\psi - C\Phi S\psi & 0 \\ -S\theta & C\Phi S\psi & C\Phi C\psi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

being: $C_\psi = \cos(\psi)$ $S_\psi = \sin(\psi)$ $C_\theta = \cos(\theta)$
$S_\theta = \sin(\theta)$ $C_\Phi = \cos(\Phi)$ $S_\Phi = \sin(\Phi)$

c) T is the translation matrix for the generic translations dx, dy, dz with respect to the x,y and z axis, respectively:

$$T = \begin{bmatrix} 1 & 0 & 0 & dx \\ 0 & 1 & 0 & dy \\ 0 & 0 & 1 & dz \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Depending on the type of the joint, R and T have a different appearance.
If joints are rotational with respect to the z axis, and the x axis of the generic J_{i+1}-based coordinate system is constrained to be on the rotation radius of the joint J_i, (this hypothesis will be assumed throughout the rest of the paper), then $\theta = 0$, $\psi = 0$, $dy = 0$, $dz = 0$, hold, and hence the matrix F reduces to:

$$F = \begin{bmatrix} C\Phi & -S\Phi & 0 & dxC\Phi \\ S\Phi & C\Phi & 0 & dxS\Phi \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The relationship between the system of coordinates located in the center of motion and the absolute system is also modeled by the transformation F. Therefore, if $P_a(x, y, z)$ is a generic point in the absolute reference system, the corresponding P_c in the other coordinate system is given by:

$$P_c = F^{-1}\ P_a$$

The motion model for the relative movements between the joints has to take into account, for each joint, the set of the allowed patterns for the rotation angle Φ and, possibly, the angular speed $\dot{\Phi}$.
In our model, the DFT (Discrete Fourier Transformation) samples of the angles and of the angular speed, normalized with respect to their mean values (thus independent of the amplitude) are used. Both modules and phases are stored. In Fig.4 examples of patterns of the angles and of the corresponding DFT modules for the joints of a walking-man leg are reported.

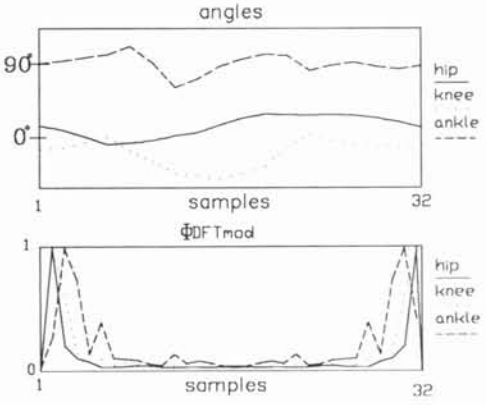


Fig.4 - Patterns of angles and of the DFT modules for the joints of a walking man leg.

Moreover, the ranges of the possible angle amplitudes and angular speeds are included in the model, as well.
In order to use the behavior for recognition observed patterns are normalized and operated with DFT for each joint. These are then compared with the descriptors stored in the models selected. Several matching steps are performed.
First, the angle and angular speed amplitudes are checked with the corresponding ranges in the models. If the confidence is large enough, the inspection of modules of the DFT samples is carried out. Finally, phases are compared and displacements are evaluated with respect to the reference values. If the displacements are the same for every joint, this is assumed as a measure of the coordination of motions.
Each model C at every sampled instant t, computes a global confidence which is given by:

$$\mu(C,t) = \frac{w\Phi_{am}\ \mu\Phi_{am}(C,t) + w\Phi_{am}\ \mu\Phi_{am}(C,t) + w\Phi_{md}\ \mu\Phi_{md}(C,t) + w\Phi_{ph}\ \mu\Phi_{ph}(C,t)}{w\Phi_{am} + w\Phi_{am} + w\Phi_{md} + w\Phi_{ph}}$$

where:

a) $w\Phi_{am}$, $w\Phi_{cm}$, $w\Phi_{md}$, $w\Phi_{ph}$ are appropriately defined context-dependent weights;

b) $\mu\Phi_{am}(C,t)$ takes into account confidences evaluated for every joint with respect to the angle amplitudes;

c) $\mu\dot{\Phi}_{am}(C,t)$ takes into account confidences evaluated for every joint with respect to the angular speed amplitudes;

d) $\mu\Phi_{md}(C,t)$ takes into account confidences evaluated for every joint with respect to the DFT modules of angle amplitudes;

e) $\mu\Phi_{ph}(C,t)$ takes into account the differences between the phase displacements with respect to the stored DFT phases for the joints.

4. CONFIDENCE UPDATING

At each frame sample, for each observed object, there is a set of confidences $\mu(C,t)$ that are computed by all the selected models, being:

$$\mu(C,t) = \begin{cases} \mu(C,t) & \text{if } \mu(C,t) \geq \alpha \\ 0 & else \end{cases}$$

where α is a task-dependent threshold.
Progressing in time, as new confidences are evaluated, old ones have to be updated. This is made according to:

$$\mu'(C,t+\Delta t) = max[\ \mu_{int}(t+\Delta t,C)\ ,\ \mu_{avg}(t+\Delta t,C)\ ]$$

where:

a) $\mu_{avg}(t+\Delta t,C) = h\ [\mu(C,t),\mu(C,t+\Delta t)]$

b) $\mu_{int}(t+\Delta t,C) = \begin{cases} \mu(C,t) \cup \mu(C,t+\Delta t) \\ \qquad \text{if } \mu(C,t) \text{ and } \mu(C,t+\Delta t) \neq 0 \\ 0 \qquad\qquad else \end{cases}$

with h the fuzzy averaging function and $\cup$ defined according to:

$$\mu_{A \cup B}(C) = \max[\mu_A(C)\ ,\ \mu_B(C)]$$

The models that are always matched have maximum global confidences. Models build their final hypotheses incrementally. The understanding process terminates when some hypotheses with the required confidence level are reached.

5. AN EXAMPLE OF RECOGNITION BY BEHAVIOR

In the following, a simplified applicative example is presented, in which a dynamic analysis is carried out in order to classify vehicles coming out of paytoll stations on a highway, at night- time.
In this example, light blobs are tracked in a sequence of frames and, as shape cannot give useful information, only the entity behavior is used for recognition.
In this case only the coarse motion representation is employed, according to the approach described in Sect.2.
As the operating context is restricted, only models of objects that can be present at paytoll stations on highways at night-time are used for comparison. This limits the investigation to some vehicle types (e.g. cars, trucks), excluding others (e.g. bikes).
The context selected implements a simplified system of equations with respect to equations (2.1) and (2.2): the angle θ between the motion direction and camera axis, and the distance d between the vehicle trajectory and the TV- camera focus along the camera axis, are assumed to be known and, therefore, the grabbed space displacements at the time t, $s(t)$, can be directly adjusted according to:

$$s(t) = P(t) - P(t-1)$$

where $P(t)$ is:

$$P(t) = \frac{d}{\frac{N_{pixmax}}{N_{pix}(t)}\ \frac{L_f\ sin\theta}{W_{ccd}} - cos\theta}$$

being $N_{pix}(t)$ the measured distance (in pixels) between the observed point and the vertical axis measured at the time t, N_{pixmax} the resolution of the sensor, L_f the camera focal length, W_{ccd} the TV camera CCD sensor width. W_{ccd}, L_f and N_{pixmax} are hardware-dependent factors.
In this case a second order autoregressive equation with null input (from (2.3)) is used, where a_i and $s(t)$ are scalar values:

$$s\ (t) = a_1\ s\ (t - \Delta\ t) + a_2\ s\ (t - 2\ \Delta\ t)$$

The second order motion equation has been adopted for the sake of simplicity; however, it has been proven that higher order equations do not give significantly better results in this application. For each motion coefficient a_i, the estimated value $\hat{a}_i$, is evaluated using the MSE filter as:

$$\hat{a}_i\ (t) = \frac{det_i\ G}{det\ G} \qquad\qquad i = 1,\ 2$$

where :

a) G is the 2×2 matrix of the elements g_{ij} defined as:

$$g_{ij} = \sum_{k=0}^{n_o - 1} s\ (t - (k+i)\Delta\ t)\ s\ (t - (k+j)\Delta\ t) \qquad i = 1,\ 2\ ;\quad j = 1,\ 2$$

being n_o the number of observations.

b) $det_i\ G$ is the determinant of the matrix G where the column i-th has been replaced with the column vector V with elements v_i :

$$v_i = \sum_{k=0}^{n_o - 1} s\ (t - (k+i)\Delta\ t)\ s\ (t - (k+3)\Delta\ t) \qquad i = 1,\ 2$$

Coefficients are stabilized in few steps.
The following approximating membership functions ($\mu_{CAR}(a_1,a_2)$, $\mu_{TRUCK}(a_1,a_2)$) with elliptic sections have been derived by observing the car and truck behavior in the *'night-time at paytoll station'* context (5.1a, 5.1b respectively).
These fuzzy-membership functions have been put into the *'night-time_at_paytoll_station'* models of cars and trucks, respectively, and represent their typical behavior in that context.
The membership function for a truck has a different profile from that of a car. As can be argued, an overlapping is presented between the spaces of car and truck coefficients. Obviously, in the case of uniform motion, there is no way to discriminate between cars and trucks.
The recognition was made under the assumption that light blobs are rigid objects, and the displacements measured between frames are small, so that we can keep track of spatial tokens from one frame to the next.
Fig.5/a shows one of the grabbed raw-images in the frame sequence. Fig.5/b depicts the corresponding binary scene after histogram filtering. Fig.5/c presents the results of the segmentation. Estimated motion descriptors and confidences are shown in Fig.5/d for the observed moving objects. In this figure, the matching with the reference descriptor fields of car and truck models is also displayed.

6. CONCLUSIONS

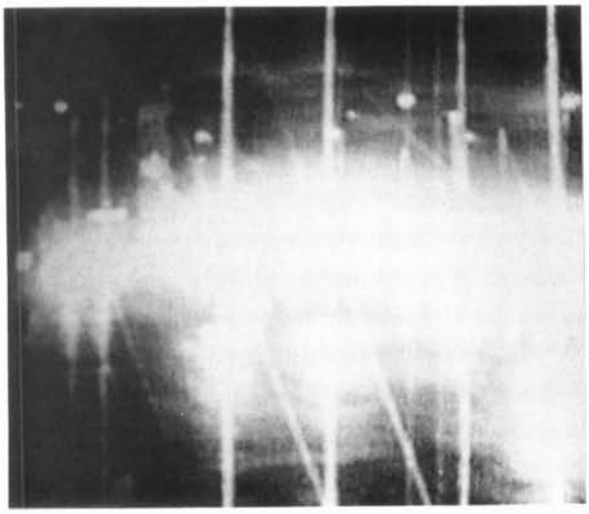
In this paper, the problem of considering motion as a property of the object and of using this in order to improve the recognition process was discussed. Several issues have been addressed regarding definition of motion descriptors, the description of motion at different abstraction levels and support for the object models.
An approach was presented in which motion descriptors for the observed object are estimated through an autoregressive

(5.1a) $\mu_{CAR}(a_1,a_2) =$

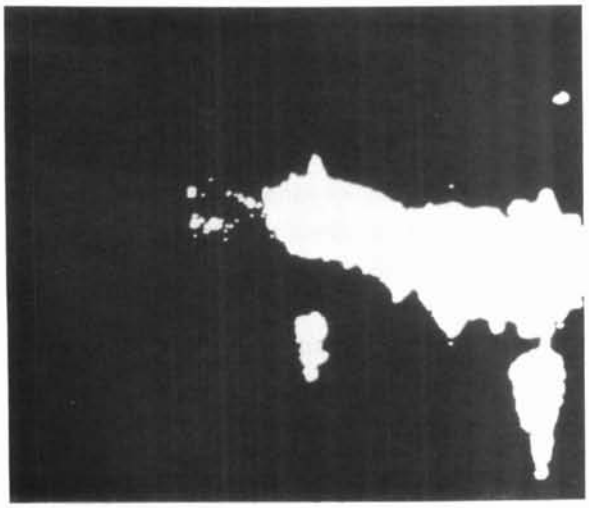
$$= \begin{cases} \dfrac{(a_1-0.65)^2}{0.619} + \dfrac{(a_2-0.35)^2}{1.052} + 1.606(a_1-0.65)(a_2-0.35) + 0.5 & \text{if } \left(\dfrac{(a_1-0.65)^2}{0.619} + \dfrac{(a_2-0.35)^2}{1.052} + 1.606(a_1-0.65)(a_2-0.35)\right) < 0.033 \\ \dfrac{0.001}{\dfrac{(a_1-0.65)^2}{0.619} + \dfrac{(a_2-0.35)^2}{1.052} + 1.606(a_1-0.65)(a_2-0.35)} & \text{if } \left(\dfrac{(a_1-0.65)^2}{0.619} + \dfrac{(a_2-0.35)^2}{1.052} + 1.606(a_1-0.65)(a_2-0.35)\right) > 0.15 \\ 1 & \text{if } 0.033 \le \left(\dfrac{(a_1-0.65)^2}{0.619} + \dfrac{(a_2-0.35)^2}{1.052} + 1.606(a_1-0.65)(a_2-0.35)\right) \le 0.15 \end{cases}$$

(5.1b) $\mu_{TRUCK}(a_1,a_2) =$

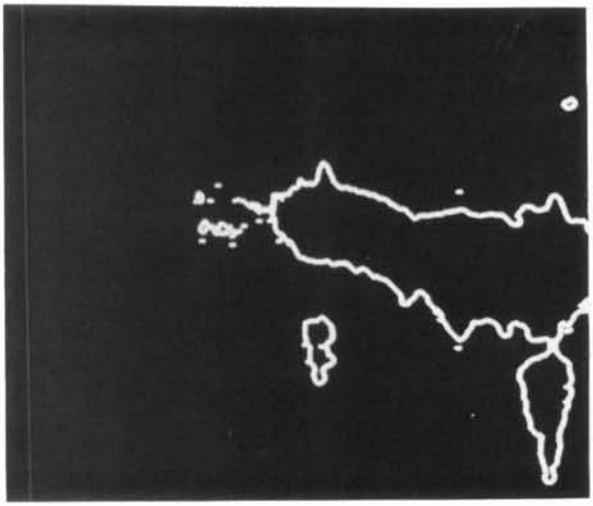
$$= \begin{cases} \dfrac{0.001}{\dfrac{(a_1-0.8)^2}{0.583} + \dfrac{(a_2-0.2)^2}{1.118} + 1.98(a_1-0.8)(a_2-0.2)} & \text{if } \left(\dfrac{(a_1-0.8)^2}{0.583} + \dfrac{(a_2-0.2)^2}{1.118} + 1.98(a_1-0.8)(a_2-0.2)\right) > 0.035 \\ 1 & \text{if } \left(\dfrac{(a_1-0.8)^2}{0.583} + \dfrac{(a_2-0.2)^2}{1.118} + 1.98(a_1-0.8)(a_2-0.2)\right) \le 0.035 \end{cases}$$



(a)

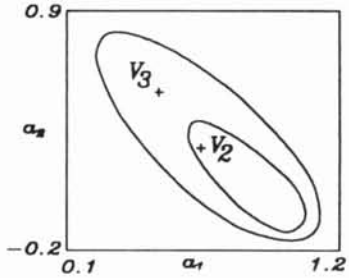


(b)



(c)

18th	a_1	a_2	μ_{car}	μ_{truck}
V_1	——	——	——	——
V_2	0.65	0.22	0.51	1
V_3	0.51	0.50	0.51	0.019



(d)

Fig.5 - Recognition by behavior of vehicles moving out of a highway paytoll station (18th frame); a) raw image, b) binarized image, c) segmented image, d) estimated motion descriptors and matching.

system and then compared with those stored in the object models. These are in the form of a fuzzy-membership function which represents all the allowed motions for the object, in the context selected. Coarse motion descriptors are provided for global object motion evaluation, while a finer description is given when the object is regarded as composed of coordinated moving subparts.
Experimented results have proven that this approach can be profitably used in those cases in which the form provides little or no help for recognition.

REFERENCES

[1] R. Y. Tsai, T. S. Huang, "Uniqueness and Estimation of Three-Dimensional Motion Parameters of a Rigid Objects with Curved Surfaces", *IEEE Transaction on PAMI*, Vol.6, n°3, pp.13-27, June 1984.

[2] J. Weng, T. S. Huang, N. Ahuja, "3-D Motion Estimation, Understanding, and Prediction from Noisy Image Sequences", *IEEE Transaction on PAMI*, Vol.9, n°3, pp.370- 389, May 1987.

[3] T. S. Huang, C. H. Lee, "Motion and Structure from Orthographic Projections", *IEEE Transaction on PAMI*, Vol.11, n°5, pp.536-540, May 1989.

[4] J. W. Roach, J. K. Aggarval, "Determining the Movement of Objects from a Sequence of Images", *IEEE Transaction on PAMI*, Vol.2, n°6, pp.554-562, 1980.

[5] T. J. Broida, R. Chellappa, "Estimation of Object Motion Parameters from Noisy Images", *IEEE Transaction on PAMI*, Vol.8, n°1, pp.90-99, Jan. 1986.

[6] H. Shariat, K. E. Price, "Motion Estimation with More Than Two Frames", *IEEE Transaction on PAMI*, Vol.12, n°5, pp.417- 434, May 1990.

[7] J. K. Tsotsos, J. Mylopoulos, H. D. Covvey, S. W. Zucker, "A Framework for Visual Motion Understanding", *IEEE Transaction on PAMI*, Vol.2, n°6, pp.563-573, Nov. 1980.

[8] J. K. Tsotsos, "Representation Axis and Temporal Cooperative Processes", in *Vision Brain and Cooperative Computation*, M. A. Arbib and A. K. Hanson, eds., MIT Press, Cambridge MS, pp.361-417, 1987.

[9] V. Cappellini, R. Cecchini, A. Del Bimbo, P. Nesi, "Object-Based Information Modeling for Pattern Recognition and Motion Analysis", *Proc. V European Signal Processing Conference*, EUSIPCO'90, Barcelona, September 12-21, 1990.

[10] V. Cappellini, A. Del Bimbo, P. Nesi, "Integrating Object-Oriented Programming Paradigm Concepts in Designing a Vision and Pattern Recognition System Architecture", *Proc. IEEE 10th Int. Conference on Pattern Recognition*, Atlantic City, NJ, June 1990.