ESTAT ACTUAL I PERSPECTIVES DE LA VISIÓ PER ORDINADOR

JOSEP AMAT
Membre de la Secció de Ciències i Tecnologia de l'Institut d'Estudis Catalans
Professor de la Universitat Politècnica de Catalunya

SUMMARY

Computer Vision, considered as the capability for image acquisition and processing started its development in the sixties. However, research in this area is still open and its possibilities nowadays are still far away from the operativity level required in many application fields.

This article exposes the main difficulties encountered when trying to build systems for correct scene interpretation, as well as the most used computer architectures and the kind of processing more frequently done.

Finally, the main current computer vision application fields are exposed as well as its future trends.

1. INTRODUCCIÓ

La visió per ordinador, considerada com la capacitat d'adquisició i tractament informàtic de les imatges obtingudes d'una escena, és una de les necessitats que apareix en el desenvolupament de la ciència i la tecnologia en els darrers anys, després que les necessitats del càlcul i la resolució de problemes molt diversos foren els impulsors de la informàtica.

La visió per ordinador comença a desenvolupar-se ja des del principi de la recent història de la informàtica. Les primeres experiències realitzades daten de la dècada dels seixanta, aplicades inicialment al reconeixement dels caràcters alfanumèrics i posteriorment al tractament d'imatges d'interès en el camp de la biologia (imatges obtingudes en microscòpia), la medicina (imatges radiològiques) o la física (anàlisi de la trajectòria de partícules).

En els darrers anys, han estat les aplicacions industrials en el camp de la robòtica i la inspecció les que han donat un fort impuls als sistemes de visió per ordinador.

Les possibilitats de la informàtica de poder efectuar el tractament d'imatges toparen ja inicialment amb la dificultat que comportava haver d'utilitzar ordina-

dors concebuts amb una arquitectura interna d'estructura seqüencial, tipus Von Neumann, que es basa en l'execució, pas a pas, de les instruccions contingudes en el programa i amb unes dades també obtingudes seqüencialment de la memòria. Aquesta arquitectura seqüencial que ha estat vigent fins fa molts pocs anys, actualment ha anat evolucionant cap a noves arquitectures que permeten la paral·lelització en la forma d'operar, cosa que ha fet augmentar considerablement la velocitat d'operació. Tot i això, encara es conserva l'estructura clàssica seqüencial pel que fa a l'obtenció de les dades de la memòria. Això representa un seriós inconvenient per als sistemes de visió, ja que l'anàlisi de la intensitat lluminosa d'un punt de la imatge, que en aquest context s'anomena *píxel*, no proporciona cap informació rellevant, sinó que es fa necessària l'anàlisi d'un píxel i els píxels d'un cert entorn.

Per altra part, les imatges obtingudes per una càmara de TV convencional, que proporcionen una resolució suficient per a la major part de les imatges que són d'interès en aquest mitjà de comunicació, però que fins i tot són de resolució insuficient per a certes aplicacions de tipus tècnic, estan formades per un mínim de 400 punts en cada una de les 625 línies, cosa que comporta haver d'obtenir més de 2.500.000 dades per imatge. Aquest elevat volum d'informació, que és obtingut de la càmera de TV cada 40 mil·lisegons, fa que hagi resultat molt difícil la seva memorització fins ben entrats els anys vuitanta, i que fins i tot amb les actuals generacions d'ordinadors, els temps de processament d'aquesta informació resulta tan elevat que fa molt restrictiva la seva aplicació, encara, en molts camps de les seves potencials aplicacions.

Per a reduïr aquests temps d'operació ha estat necessari desenvolupar ordinadors especialitzats, amb unes arquitectures més orientades a facilitar la resolució del problema de la visió pel que fa a la necessitat d'operar amb les dades corresponents a un cert nombre de píxels, que no pas en la disposició en paral·lel en l'execució de les instruccions.

2. ARQUITECTURES ESPECIALITZADES

Ja des de les primeres aplicacions en el camp de la visió per ordinador, a l'inici de la dècada dels anys seixanta, tot i operar amb imatges digitalitzades en formats molt més reduïts que els actuals, començaren a desenvolupar-se ordinadors especialitzats. Un dels primers, el Cellscan, desenvolupat per la Perkin-Elmer el 1961, disposava d'un processador matricial de 3 x 3 píxels, el qual permetia operar en paral·lel sobre un píxel i els vuit veïns, estructura que ha esdevingut clàssica en el camp de la visió per ordinador (figura 1).

Aquesta estructura, la de la utilització d'un processador matricial, resulta adecuada especialment per al tractament d'imatges, atès que les dades estan estructurades bidimensionalment. En aquesta línia es feren altres desenvolupaments utilitzant diferents processadors que operaven en paral·lel. Entre els computadors dotats de diferents processadors que operen en paral·lel poden

citar-se el Salomon, màquina desenvolupada per Westinghause el 1962 i que estava dotada ja d'una matriu de 16 x 16 processadors, encara que molt elementals. La família de computadors tipus CLIP desenvolupada a l'University College, en la versió IV, el 1976, disposava d'una matriu de 96 x 96 processadors.

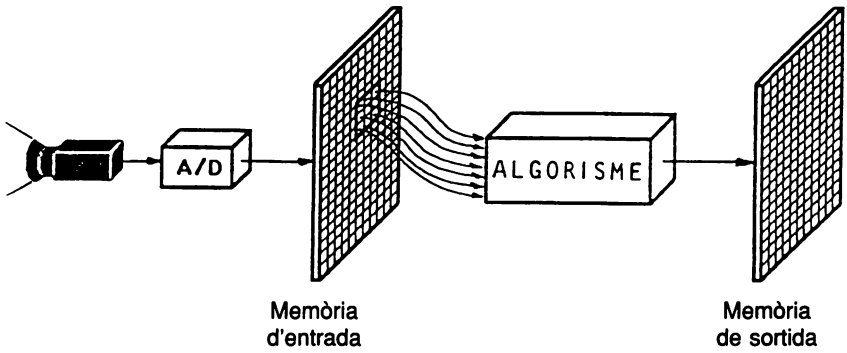

FIGURA 1. Estructura d'un operador matricial

Complementàriament a aquesta línia, la d'aconseguir una velocitat d'operació més elevada utilitzant processadors matricials capaços d'efectuar una mateixa instrucció sobre múltiples dades (arquitectures SIMD), començaren a utilitzar-se també arquitectures en sèrie, dissenyades per a realitzar seqüencialment diferents funcions, però a partir d'unes mateixes dades (estructures MISD). Aquestes estructures basades en una concatenació en sèrie de diferents processadors, poden ser de tipus *pipeline* o de tipus sistòlic. Les estructures sèrie són de tipus *pipeline* quan el temps d'operació de cada etapa processadora és el mateix, i normalment síncron amb el temps d'adquisició de la imatge, i s'obté així un flux continu de dades. En canvi, en les arquitectures sistòliques, el temps d'operació de cada processador no coincideix, i el flux de dades és discontinu. En aquest cas, el temps total d'operació és el corresponent al processador més lent. Un exemple d'aquesta altra orientació és el Cytocomputer desenvolupat per l'Environmental Research Institute of Michigan el 1979, que estava dotat de 80 etapes de processament en sèrie.

Ja a la dècada dels vuitanta, els esforços per aconseguir més velocitat de processament s'encaminaren al desenvolupament d'arquitectures massivament paral·leles, capaces d'efectuar múltiples instruccions sobre múltiples dades (estructures MIMD). La miniaturització progressiva dels microcomputadors facilita el fet d'arribar al límit: la de dedicar un processador a cada píxel d'una imatge. Dins aquesta línia cal destacar el desenvolupament del computador Zmob a la Universitat de Maryland el 1982, dotat de 256 processadors amb una potència ja considerable. Al principi del 1983, la Good-Year Aerospace lliura a la NASA el Computador MPP *(Massive Paral·lel Processor)* amb 128 x 128 processadors,

orientat al tractament d'imatges. I ja més recentment, es desenvolupà la versió V de la sèrie de Connection-Machine en el Massachussets Institute of Technology (MIT). Aquest computador estava constituït per mòduls de 8 K processadors. Cada un d'aquests processadors és ja un computador relativament potent de 8 bits, dotat de processador de punt flotant i amb 256 Kb de memòria local, amb què conjuntament amb la memòria del host, supera els 2 Gb de memòria disponible.

Els esforços efectuats per poder disposar de màquines orientades al tractament d'imatges cada vegada més potents (figura 2), han aconseguit en molts casos resultats satisfactoris, especialment pel que fa a les aplicacions científiques, però aquests computadors de preus molt elevats, no han proporcionat solucions rendibles als problemes que la indústria planteja en el camp de la classificació, la inspecció i el control de qualitat.

Els sistemes industrials de visió per ordinador actualment disponibles, són fruit d'un compromís entre la seva eficiència i velocitat d'operació i per l'altra part del seu cost. Aquest compromís s'ha assolit utilitzant majoritàriament els ordinadors cada vegada més potents tipus PC (en versió industrial) que incorporen un o més processadors tipus matricial, formant una estructura *pipeline* que opera sobre el propi bus del computador. Aquest mòdul addicional és utilitzat per a adquirir les imatges de vídeo i efectuar un primer preprocessament.

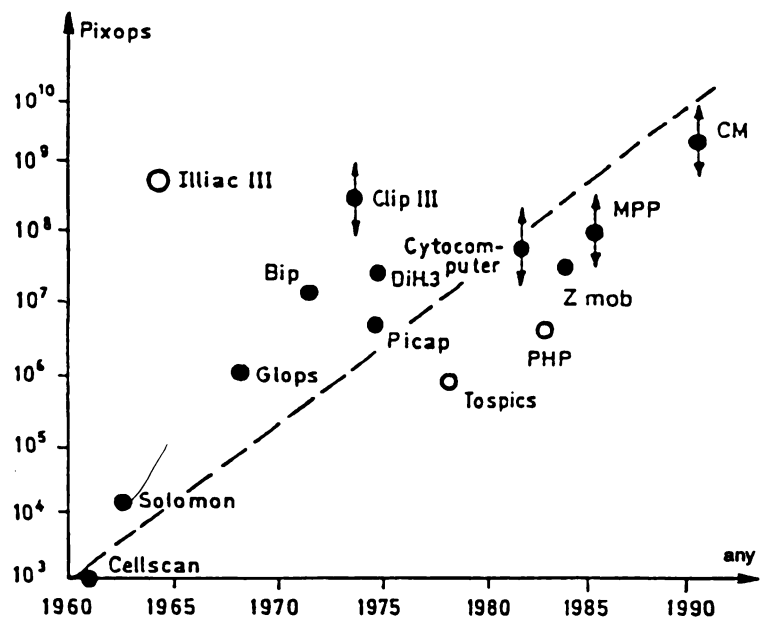

FIGURA 2. Capacitat de diferents màquines especialitzades desenvolupades per al tractament d'imatges.

Per a efectuar aquest preprocessament cal aplicar un algorisme sobre les dades corresponents a un cert entorn d'un píxel, de cada píxel de la imatge i atès que el senyal és obtingut de la càmera de TV en sèrie, línia per línia, cal disposar d'una memòria intermèdia de *n-1* línies i de *n* píxels per tal d'obtenir en paral·lel les dades d'aquests n x n píxels. D'aquesta manera, l'ordinador disposa, a més de la imatge digitalitzada, d'una nova imatge tractada i que és obtinguda simultàniament amb la que és adquirida amb un processador *pipeline,* encara que amb un cert retard. Aquest retard és molt inferior al temps que seria necessari per a obtenir aquesta imatge tractada utilitzant el processador en sèrie de què estan dotat els computadors convencionals.

També són utilitzades arquitectures paral·leles estructurades en forma piramidal (figura 3), o en forma matricial, utilitzant els *transputers,* que són un tipus de processadors que estan dotats d'uns canals de comunicació en sèrie d'alta velocitat d'operació, que faciliten la formació d'aquestes configuracions.

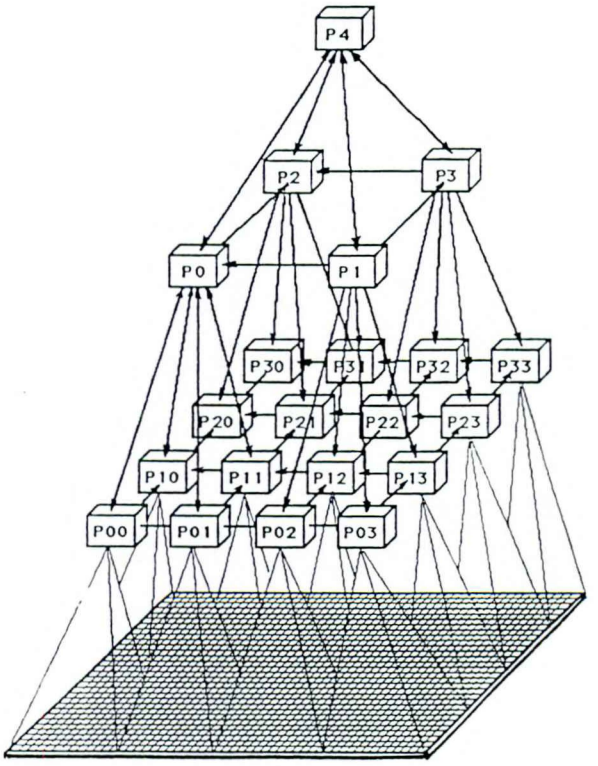

FIGURA 3. Processament en paral·lel d'una imatge utilitzant una estructura piramidal de processadors.

Finalment, també són utilitzades les xarxes neuronals com a estructures paral·leles d'unitats de processament d'estructura variable, que són reconfigurables a partir d'un procés d'aprenentatge. Aquests processadors neuronals proporcionen bons resultats en determinades aplicacions en què, pel tipus d'escena canviant amb el temps difícilment programable, permet autoadaptarse fins a assolir uns resultats desitjats, com pot ser el cas de la lectura i la interpretació de caràcters manuscrits.

3. TIPUS DE PREPROCESSAMENT D'IMATGES

Segons els objectius del sistema de visió, caldrà fer un pretractament específic de la informació digitalitzada de cada imatge amb un processador de tipus matricial, a fi d'obtenir, amb el mínim retard possible, les característiques més significatives i útils de la imatge. A partir d'aquestes característiques, es podrà fer un tractament posterior que és necessari per a visualitzar, localitzar, reconèixer objectes o interpretar l'escena de treball. Aquest processament posterior, podrà ja ésser realitzat eficientment amb un computador convencional atesa la gran reducció del volum d'informació aconseguida amb l'etapa prèvia de preprocessament.

Els algorismes de pretractament d'imatges utilitzats més correntment són: els de millora del contrast, el binaritzat, el filtrat i el d'obtenció de contorns.

3.1. Intensificació d'imatges

La digitalització de la imatge, consistent a passar d'un senyal analògic corresponent a la lluminositat de cada punt a un codi binari expressat en r bits, permet considerar 2^r nivells de lluminositat. En imatges de color són binaritzats tres senyals corresponents als tres colors bàsics, però cal remarcar que en imatges en blanc i negre, que són les utilitzades en la major part d'aplicacions industrials, aquest nombre r és de 8 bits, cosa que representa operar amb una resolució molt superior a la sensibilitat de l'ull humà.

Tot i això, en algunes aplicacions en què la interpretació de la imatge no és realitzada per l'ordinador, sinó visualment, cal efectuar un realçament de la imatge per a poder fer visible a l'ull febles diferenciacions de nivells de gris que poden resultar rellevants. Aquest realçament pot ésser obtingut efectuant una conversió analògico-digital no lineal, de forma que es produeixi una concentració més gran del nombre dels nivells de grisos sobre una determinada gamma d'aquest color, mentre es produeix una concentració menor en la resta de gammes considerades menys rellevants, cosa que produeix una rehistogramació del nombre de píxels per cada nivell de gris d'una imatge.

Aquesta rehistogramació és efectuada aplicant al valor de la variable intensitat I de cada píxel (x, y) de la imatge $I(x, y)$ una transformació no lineal de la

forma $R = f(I)$, i s'obté una nova imatge $R(x, y)$ en què s'han realçat certs detalls de les gammes clares o fosques d'una imatge, per exemple d'una radiografia. La rehistogramació posa de manifest detalls que són continguts en la informació obtinguda pel sensor, però que, sense aquesta transformació que permet obtenir una nova imatge, no seria possible observar a simple vista.

3.2. Millora del contrast

Un altre tipus de pretractament que també es pot efectuar sobre una imatge, és l'aplicació d'un operador multivariable de la forma $L(x, y) = f(I_1, I_2 ... I_n)$, en què cada valor $L(x, y)$ és obtingut no segons el seu propi valor $I(x, y)$ sinó segons un cert nombre de valors $I_1, I_2 ... I_n$ corresponents a les dades de la lluminositat de n píxels situats a l'entorn de cada píxel de la imatge.

Això permet utilitzar un algorisme de transformació basat, no en els valors de les intensitats, sinó en els dels increments d'intensitat, per exemple, de la forma $G(x, y) = [I(x, y) - I(x + 1, y)] + [I(x, y) - I(x, y + 1)]$ i s'obté així un gradient d'intensitats tant en el sentit vertical com en l'horitzontal de forma discreta (algorisme de Roberts). Aquesta aproximació del gradient

$$G = \left| \frac{\partial I}{\partial X} \right| + \left| \frac{\partial I}{\partial Y} \right|$$

podria ésser millorada calculant el mòdul del vector gradient de la forma:

$$G = \sqrt{\left(\frac{\partial I}{\partial X} \right)^2 + \left(\frac{\partial I}{\partial Y} \right)^2}$$

A partir d'aquesta nova imatge gradient $G(x, y)$, es pot aconseguir millorar notablement el contrast d'una imatge, formant una imatge transformada obtinguda de la forma: $L(x, y) = \alpha\, G(x, y) + (1-\alpha) I(x, y)$, en què α és el nivell de gradient suposat a la imatge $I(x, y)$ que permet intensificar els contrastos febles de determinades imatges, i millora així la capacitat de reconeixement i interpretació dels elements d'una escena (figura 4).

Aquest gradient obtingut de forma discreta a partir d'un entorn de 2 x 2 píxels, també podria éser obtingut en entorns més amplis, de 3 x 3 píxels i fins i tot de 5 x 5 píxels, utilitzant unes matrius de ponderació, G_h i G_v. D'aquestes matrius, les més utilitzades són les corresponents a l'algorisme de Sobel i que són definides com:

$G_h =$

1	2	1
0	0	0
−1	−2	−1

$G_v =$

−1	0	1
−2	0	2
−1	0	1

i que permeten el càlcul tant del mòdul del gradient com del seu argument.

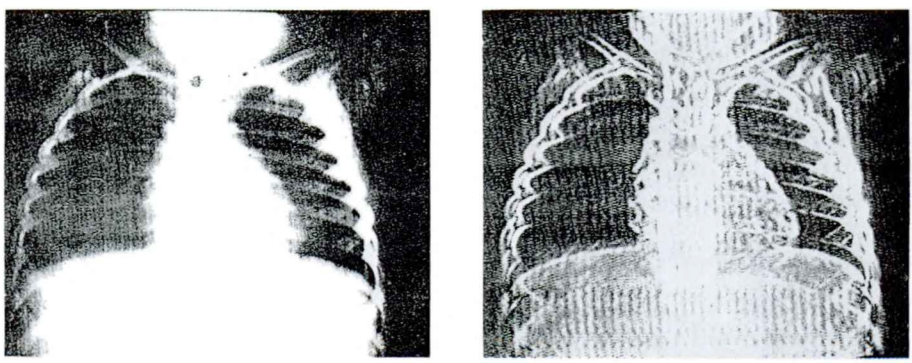

FIGURA 4. Millora del contrast d'una radiografia a partir de l'obtenció dels gradients.

Un augment similar del contrast d'una imatge es pot aconseguir igualment aplicant un filtre passa-alt H a la transformada de Fourier d'una imatge. Si $H(u, v)$ és la transmitància d'aquest filtre passa-alt i $F(u, v)$ és la transformada de Fourier de la imatge $I(x, y)$, la imatge contrastada obtinguda seria:

$$L(x, y) = F^{-1}[H(u, v) \cdot F(u, v)]$$

però el processament d'imatges basat en les transformades de Fourier té l'inconvenient de necessitar un temps de càlcul molt superior respecte a l'aplicació de les matrius de convolució tal com les descrites.

3.3. BINARITZACIÓ

Per a fer possible al utilització de la visió per computador a la indústria, cal aconseguir no sols la seva rendibilitat econòmica sinó també la seva viabilitat pel que fa al temps de processament. Per a aconseguir aquests objectius, en els entorns industrials es procura simplificar al màxim la complexitat de l'escena. Això s'aconsegueix, en gran part, amb les tècniques d'il·luminació i amb l'elecció

del fons més adequat, a fi d'aconseguir que tots els píxels dels objectes a visualitzar arribin a ser de major (o menor) nivell de lluminositat, que tots els del fons de l'escena, incloent les possibles ombres. Quan això s'aconsegueix, es posa de manifest en la forma que presenta l'histograma del nombre de píxels de la imatge per cada nivell de gris. En aquest cas l'histograma presenta dues regions separables (figura 5(a)), entre les quals és possible definir un llindar h en què la imatge binària obtinguda (d'objectes foscos sobre fons clar) és de la forma:

$$B(x, y) = \begin{cases} 1 \text{ si } I(x, y) \leq h \\ 0 \text{ en cas contrari} \end{cases}$$

en què B(x, y), correspon precisament a la imatge binària de l'objecte considerat. (figura 6 (b) i (c)).

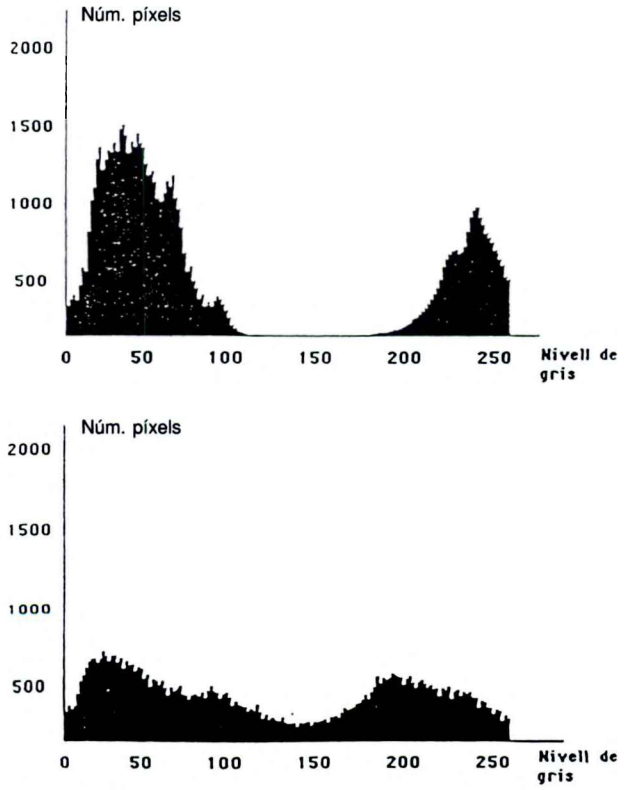

FIGURA 5. Diferents histogrames obtinguts d'una escena segons el contrast entre l'objecte i el seu fons.

Quan l'escena, amb l'ajut del sistema d'il·luminació, permeti diferenciar d'una forma binària entre els elements buscats de l'escena i el seu fons, s'haurà aconseguit una segmentació molt eficient, que permetrà ja aplicar sobre aquesta imatge binària els algorismes de localització, reconeixement o classificació que cada aplicació requereixi.

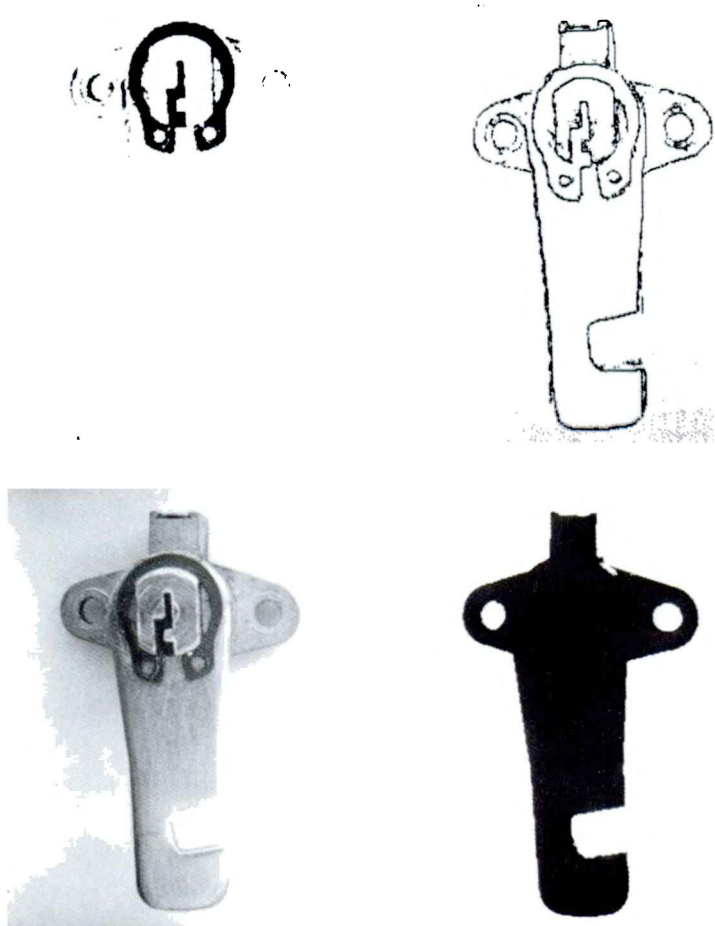

FIGURA 6. Binarització d'una imatge: *a)* Imatge original
b) Imatge binaritzada amb un llindar h_1
c) Imatge binaritzada amb un llindar h_2
d) Imatge dels contorns

Quan la binarització de l'escena, utilitzant un llindar h de nivell de gris determinat, no pot aconseguir separar tots el punts d'un element de l'escena, pel fet que la imatge presenta un histograma separat en dues regions definides, perquè hi ha una superposició entre els nivells de grisos corresponents als píxels de l'objecte i als del fons, es pot procedir a obtenir un llindar variable per zones o al llarg de tota la imatge $h(x, y)$. Aquest llindar s'obté de l'anàlisi d'histogrames parcials obtinguts descomposant la imatge en un cert nombre de regions.

Si la simplificació de la imatge a partir de la seva binarització no és possible, moltes vegades es recorre a obtenir els contorns dels objectes, definits com els píxels frontera entre nivells de grisos diferenciables en una imatge.

3.4. Obtenció de contorns

Atès que una de les informacions més rellevants sobre la forma d'un objecte és el seu contorn, és un preprocessament orientat a l'obtenció dels contorns qui permet reduir apreciablement la complexitat de la imatge i per tant, el temps necessari per a efectuar el reconeixement de formes.

Si el contrast entre els objectes que apareixen sobre una escena i el seu fons és suficient, els contorns es poden considerar com aquell conjunt de punts que defineixen una frontera entre zones de diferent nivell d'il·luminació. Aquests píxels frontera podran ésser obtinguts a partir de la imatge dels gradients $G(x, y)$ i de considerar que $G(x, y)$ és un contorn si el valor del gradient en aquest punt és superior a un valor prefixat:

$$C(x, y) = \begin{cases} 1 \text{ si } G[I(x, y)] \leq K \\ 0 \text{ en cas contrari} \end{cases}$$

i s'obté així una imatge binaritzada a partir d'una imatge gradient, en què generalment el nombre de píxels de contorn $C(x, y) = 1$ és molt reduït (figura 6 (d)).

Aquesta informació compactada es útil en moltes aplicacions industrials que han d'assolir una elevada velocitat d'operació tal com el reconeixement i la classificació d'objectes, o la localització i el càlcul d'orientació, per a la manipulació automàtica d'objectes mitjançant robots, i és un pas previ per assolir un nivell més alt de percepció: la interpretació de les escenes.

4. OBTENCIÓ D'INFORMACIÓ TRIDIMENSIONAL

Moltes aplicacions de la visió per computador, permeten operar amb les imatges bidimensionals obtingudes com a projecció d'una escena sobre la superfície retiniana del sensor utilitzat. En algunes aplicacions però, és necessari

obtenir una certa informació tridimensional de l'escena, ja sigui per a l'obtenció d'informació quantitativa sobre les dimensions d'un objecte a l'espai, o per a resoldre problemes d'ambigüitat que es poden produir en disposar únicament d'una imatge plana corresponent a escenes tridimensionals.

L'obtenció d'aquesta informació en el món animal i en particular, en l'home, es realitzada per estereovisió, és a dir, a partir de la correlació entre dues imatges preses des de punts de vista amb un cert desplaçament.

La correlació entre les dues imatges estereoscòpiques en molts casos és un problema difícil d'abordar, atesa l'existència de dos tipus de problemes. Per una part, és freqüent trobar àmplies zones de l'escena sense elements característics, que proporcionin punts de referència per efectuar l'aparellament del punts homòlegs entre ambdues imatges, com es dóna en el cas de superfícies en l'espai uniformes. Per altra part, es troben també situacions contràries, és a dir, la d'amplies zones de la imatge riques en característiques que també són abundants en l'entorn del punt de correspondència, com és el cas de les fulles d'un arbre en una escena natural, o la retícula repetitiva d'un mallat en un entorn industrial.

Tant enfront d'un com en l'altre tipus de dificultat, el problema encara no està resolt satisfactòriament amb els recursos informàtics actuals. Sorprèn en canvi la facilitat i velocitat d'operació del sistema humà de percepció, que es pren normalment com a pauta a seguir en el desenvolupament dels algorismes de tractament d'imatges.

Atesa la impossibilitat actual de poder disposar de computadors que permetin efectuar en paral·lel la correlació entre les dues imatges, tal com es produeix en la visió estereoscòpica animal, un dels mètodes utilitzats consisteix a efectuar la correlació únicament sobre l'entorn d'alguns píxels més característics de la imatge, com poden ser els vèrtex o els contorns, i obtenir per triangulació la informació sobre la profunditat. D'aquesta manera, s'obté en un temps raonable, informació tan sols de les coordinades d'un nombre molt reduït de punts, com a informació complementària a la imatge 2D obtinguda per una de les dues càmeres.

Quan aquesta informació no és suficient i cal disposar d'un nombre més gran de dades 3D, es recorre a identificar punt per punt de l'escena, o d'una àrea d'interès de l'escena, mitjançant un raig làser. D'aquesta manera, de cada punt il·luminat per làser sobre l'escena, se n'obté la posició del punt projectat sobre la retina de la càmera, que, juntament amb les dades corresponents a la distància entre càmera i làser i l'angle α de deflexió del làser, permet calcular per triangulació la profunditat Z de cada punt il·luminat.

Aquest mètode, però, és limitat per la quantitat el temps que es requereix per a escombrar amb el làser tota la imatge.

En cas d'objectes en moviment o de captació d'imatges des de càmeres en moviment, també es pot obtenir la informació de la profunditat per l'anàlisi del flux òptic de cada punt de l'escena. Per a càlcul de la profunditat cal efectuar l'anàlisi de les trajectòries obtingudes dels diferents punts de la imatge a partir d'imatges succesives.

5. PROCESSAMENT D'IMATGES

La interpretació de les escenes, objectiu final de la visió per computador, és un problema complex al qual sols s'ha pogut donar solució en cas d'escenes molt simplificades.

El problema de la interpretació correcta d'una escena radica en el fet d'obtenir, a partir del sensor, la càmera de TV, una imatge plana d'un entorn tridimensional, per tant, amb una quantitat d'informació molt parcial.

Aquesta informació parcial de l'escena visualitzada, condueix en general a una ambigüitat molt elevada, que cal resoldre d'acord amb el context i l'experiència. Una imatge d'una peça isolada sobre una taula de treball, per exemple, pot resultar inidentificable, però pot ésser identificada fàcilment en presència de la resta de peces que integren un determinat equip. Aquest efecte de context lògicament és donat també per uns coneixements previs adquirits, l'experiència, que en els sistemes de visió caldrà adquirir amb una etapa prèvia d'aprenentatge.

Per altra part, la interpretació de les escenes és dificultada no sols pel fet de disposar d'una informació molt reduïda a causa de la projecció o de les ocultacions, sinó que ve acompanyada també per una certa quantitat de falsa informació de l'entorn, produïda pel sistema d'i·luminació, que dóna lloc a les ombres i els reflexs.

Així doncs, el procés d'interpretació de les escenes, que constitueix un segon i més alt nivell de processament dins d'un sistema de visió per ordinador, correspon al nivell intel·ligent, que ha de ser capaç de distingir entre la informació extreta de la imatge en les anteriors etapes de nivell baix de preprocessament, com poden ser els contorns, i que pot consistir en:

- dades sobre elements rellevants dels objectes a visualitzar
- dades d'elements irrellevants dels objectes a visualitzar
- dades corresponents a elements de l'escena visualitzada aliens a l'objecte a visualitzar
- dades no corresponents a cap objecte contingut en l'escena (com poden ser les ombres).

Aquesta complexitat fa que el procés previ a la interpretació de l'escena, i que consisteix en la segmentació en regions i el posterior reconeixement dels objectes que hi apareixen, només hagi pogut ésser resolt de forma parcial amb l'ajut de sistemes adequats d'il·luminació i utilitzant un fons que permeti aconseguir un contrast suficient amb els elements rellevats de l'escena que cal visualitzar i identificar.

5.1. SEGMENTACIÓ D'IMATGES

La segmentació de les imatges és la tècnica de formació de regions diferenciables en l'escena i corresponents a diferents objectes o parts dels objectes que

hi apareixen. Aquesta segmentació es basarà, per una part, en el fet que els diferents píxels d'una mateixa agrupació tindran unes propietats o característiques semblants, i per altra, que hauran d'ocupar posicions contigües.

L'anàlisi de característiques es pot basar en criteris estrictament locals, com pot ésser el nivell de gris de cada píxel, o el color, o pot basar-se en una anàlisi més àmplia, considerant un cert entorn de cada píxel. En aquest cas, la característica a considerar pot ésser la de la textura, considerada con una propietat estadística associada a la regió de la imatge obtinguda o d'una imatge transformada.

La identificació i classificació de textures tals com llis, rugós, densitat de línies rectilínies o curvilínies, repetitivitat o aleatorietat, per exemple, poden ser quantificades dins d'un entorn, que ha d'ésser prou extens com per a poder discriminar entre textures diferents i prou reduït com per a poder encaixar dins els límits de cada part d'un objecte. Aquest entorn, que se sol denominar *texel*, cal definir-lo en molts casos segons una anàlisi prèvia de la imatge.

Cada regió és formada per un creixement entorn d'un punt inicial, mentre l'anàlisi de característiques efectuat no permeti detectar una variació apreciable en travessar un contorn prèviament definit.

En els sistemes de visió per ordinador aplicats a la indústria, es procura que aquesta segmentació pugui ésser feta per anàlisi local de característiques, cosa que s'aconsegueix utilitzant un fons adequat. En altres camps d'aplicació de la visió, això no és fàcilment possible, com succeeix en el camp de la biologia, la medicina, en què cal supervisar aquesta fase consistent en una segmentació assistida, prèvia a un reconeixement i una classificació.

5.2. Reconeixement

Un cop aconseguida la segmentació de l'escena entre un o diferents elements considerats rellevants i els fons o altres elements considerats irrellevants, es pot iniciar la fase de reconeixement dels objectes a partir de les seves formes aparents.

El reconeixement de formes es pot basar en tècniques de classificació d'acord amb l'anàlisi de característiques geomètriques, o amb l'obtenció de descriptors semàntics, i té com a objectiu poder operar fiablement amb independència de la posició i orientació dels objectes visualitzats.

Els sistemes basats en l'anàlisi de característiques geomètriques són molt diversos, i s'estenen des dels més simplificats, com pot ésser la simple identificació basada en un paràmetre únic però prou selectiu, com podria ésser la superfície, fins a sistemes basats en criteris de proximitat multiparamètrica.

En el cas, per exemple, d'ésser considerades dues característiques A i B, cada objecte és reconegut per la menor proximitat a cada model definit, ja sigui per les seves coordinades A i B i o per estar situat dins una àrea definida a l'entorn de cada model.

Aquestes àrees de pertinença a una classe K, poden ser definides per uns límits cartesians, de la forma

$$A_{min} < A_i < A^k{}_{max}$$
$$B_{min} < B_i < B^k{}_{max}$$

o per altres tecnologies com poligonals, el·lipsoides o d'altres, que permetin contenir les coordinades de qualsevol objecte de la mateixa classe.

La definició dels límits de pertinença ha de tenir en compte el pitjor dels casos, especialment en cas d'objectes de la mateixa classe però no necessàriament idèntics, o en cas d'objectes idèntics, la dispersió de mesures que produeix en els processos de digitalització i preprocessament o les possibles fluctuacions en les condicions d'il·luminació de l'escena.

Quan és necessari considerar més de tres paràmetres, com poden ser: perímetres, dimensions màximes o mínimes, nombre d'arestes, nombre de vèrtexs, nombre de forats, etc., el mètode de la mínima distància a un model és igualment aplicable, però en aquest cas dins un hiperespai de n dimensions. En aquest cas, els límits d'un model poden ser definits per hiperplans donats pels límits de cada una de les n característiques.

En moltes aplicacions en què s'utilitza un elevat nombre de característiques, no totes elles són igualment discriminants. Per evitar els temps de càlcul de les característiques no definitòries d'un objecte, s'utilitzen uns arbres de decisió de camins variables per a cada model, cosa que permet minimitzar els temps de decisió del sistema.

Un altre camí seguit per efectuar els reconeixements de formes és el d'utilitzar uns *descriptors semàntics.* Aquests descriptors caracteritzen un objecte mitjançant una gramàtica o qualsevol altra forma que es basi en propietats locals d'un objecte. Així, per exemple, podria ser definit un alfabet a partir d'un conjunt de segments rectilinis, corbes, vèrtexs, etc., que permeti expressar la forma del contorn dels objectes a reconèixer. Aquest sistema, que opera per superposició de cada símbol definit sobre els contorns de la imatge, requereix formalment un temps de càlcul superior als sistemes basats en característiques geomètriques, però, per contra, resulta també eficient en casos de visibilitat parcial d'un objecte, en cas de produïr-se superposicions o ocultacions. Alguns d'aquests sistemes permeten, a més de ser invariants amb la posició i l'orientació de l'objecte en el pla, arribar a ser també tolerants amb els canvis d'escala.

En cas del reconeixement d'objectes tridimensionals a l'espai, normalment s'opera considerant per a cada objecte tants models com posicions estables possibles pugui presentar, cosa que repercuteix en un temps de processament apreciablement més llarg.

5.3. Interpretació d'escenes

El processament d'imatges d'alt nivell finalitza amb la interpretació de l'escena a partir dels objectes que hi han estat reconeguts, la seva posició relativa i un cert coneixement previ del context. Aquest coneixement del context és

imprescindible per a aconseguir una correcta interpretació de les escenes, ja que la informació obtinguda és en molts casos parcial i dóna lloc a una ambigüitat molt elevada. La dificultat que comporta aconseguir resoldre aquest problemes d'ambigüitat, i d'identificar correctament el context de cada escena, fa que els sistemes actuals de visió per ordinador no puguin ésser d'aplicació genèrica, i esdevenir així sistemes de visió artificial, sinó que han de ser desenvolupats per a aplicacions molt específiques.

Així doncs, en cada camp d'aplicació, el procés d'interpretació, el constituïrà el conjunt d'algorismes necessaris per a assolir els nivells d'operativitat i fiabilitat esperats en cada cas. Aquests algorismes poden ésser més o menys complexos, segons els tipus d'escena, ja que aquestes poden ésser molt estructurades i repetitives, com es dóna majoritàriament en els entorns industrials, o poden aparentar ésser molt poc estructurades i variables, tal com es dóna en les escenes naturals.

En els casos més simplificats i repetitius aquests algorismes poden ésser relativament senzills i fiables, i quedar reduïts a una decisió boleana, tipus afirmatiu o negatiu. És el cas, per exemple, dels sistemes dedicats a la detecció, la localització o la classificació d'objectes.

En altres casos, no és suficient efectuar una detecció i un reconeixement dels objectes, sinó que cal detectar possibles defectes de tipus i característiques molt diverses. En aquest cas, els algorismes utilitzats són d'una capacitat de discriminació més gran i de tipus més global.

En ordre creixent de dificultat, es presenten les escenes amb objectes no necessàriament aïllats entre si, en què poden produïr-se ocultacions. En aquest cas, solen utilitzar-se algorismes basats en la verificació de les diferents hipòtesis que poden ésser formades a partir d'indicis obtinguts pel reconeixement de característiques locals, o per estratègies probabilístiques.

En determinats camps també es poden obtenir escenes formades per objectes que han d'ésser considerats iguals, però que no són necessàriament idèntics, com es dóna, per exemple, en el control de qualitat o la manipulació automatitzada de fruites. Un problema semblant el constitueix el dels objectes que tot i ésser idèntics, a causa de la seva posició en l'espai, i per efecte de la perspectiva, són vistos en formes aparents molt diverses. En aquests casos, el reconeixement que permeti la interpretació de l'escena es basa en l'extracció d'unes característiques que resultin més invariants amb la mida i les formes aparents i en la utilització d'algorismes simbòlics més que en algorismes de tipus numèric.

Finalment, el nivell de més dificultat en la interpretació d'escenes es dóna en entorns naturals que ens són més habituals. Aquest problema no resolt és motiu de recerca per un gran nombre de centres, recerca que en molts casos es orientada a l'estudi i emulació del procés de la visió humana, tot i les limitacions imposades pels recursos informàtics a utilitzar.

Una manera d'enfocar aquest procés d'interpretació es basa en la conjunció d'un procés de tipus ascendent, des de l'extracció d'unes característiques de la imatge que a partir d'unes superfícies i uns volums permetin configurar uns

models geomètrics dels objectes reconeguts, i d'un procés paral·lel de tipus descendent. Aquest segon procés en sentit descendent (figura 7), parteix d'una hipòtesi inicial d'una escena, que es va redefinint d'acord amb la conjunció amb les dades obtingudes del procés ascendent, i que defineix un conjunt d'objectes presumiblement integrants de l'escena, i que són donats per la base de dades corresponent als coneixements previs adquirits, i que condueixen a un conjunt d'elements bàsics constituents d'aquests objectes que contínuament es comparen, en sentit horitzontal, amb els que es van obtenint en sentit ascendent de l'escena.

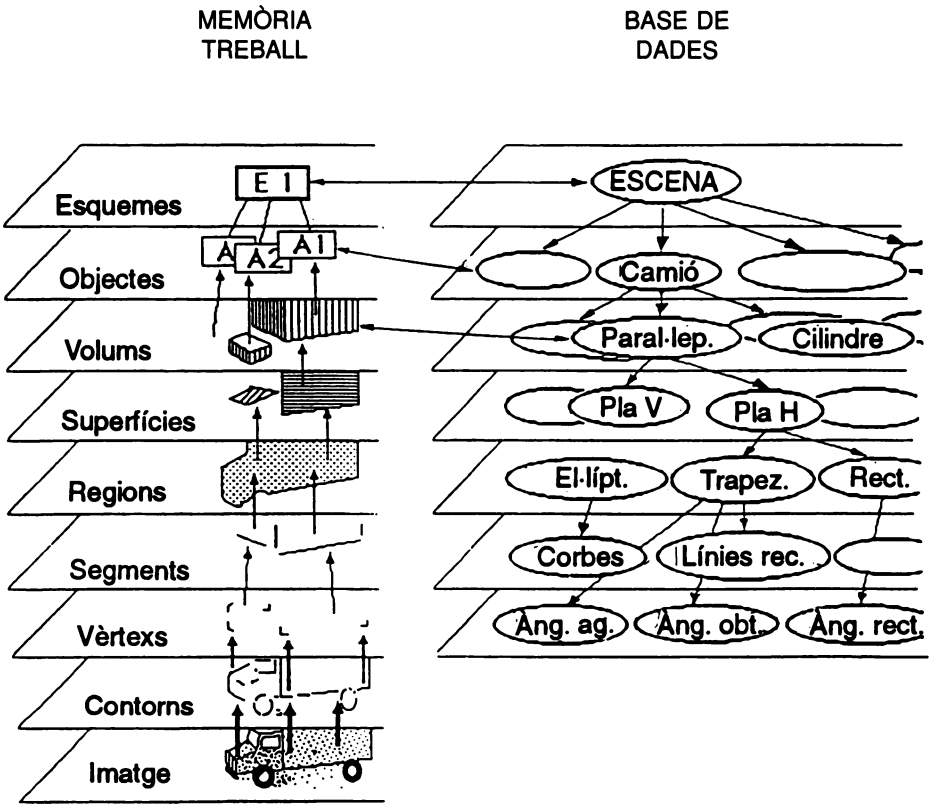

FIGURA 7. Procés seguit per a la interpretació d'una escena.

6. APLICACIONS

Els sistemes de visió per ordinador tenen un ampli abast d'aplicacions, que han anat creixent en els últims anys en poder assolir uns temps d'operació millors i uns costos cada cop més reduïts. Això ha permès la forta expansió actual dels sistemes de visió en aplicacions industrials en què el factor temps és essencial per a poder ésser introduït en les línies de producció.

Els sistemes de visió actualment poden classificar-se en dos grans grups. Un primer grup, el formarien els sistemes que no requereixen operar en temps real: el tractament d'imatges, i un segon grup estaria integrat pel conjunt d'aplicacions que operen en temps real.

6.1. Aplicacions del tractament d'imatges

El tractament d'imatges és el camp de la visió per ordinador que té per objectiu aplicar els algorismes necessaris per a obtenir noves imatges transformades, que posin de manifest característiques o aspectes més rellevants normalment interpretables visualment per l'usuari. Entre aquestes aplicacions, destaquen la biologia, la medicina o la mineralogia.

Les principals necessitats d'aquests sistemes són la resolució i capacitat de tractament d'imatges en color, essent, en canvi, l'aspecte temps en molts casos menys rellevant.

6.2. Aplicacions industrials

Entre les aplicacions industrials que més han introduït la visió en els darrers anys hi ha el de *la inspecció,* tant en aspectes de tipus quantitatius, com en aspectes més qualitatius, per assegurar la qualitat de la producció no solament en els aspectes més fàcilment parametritzables, com pot ser la presència, la quantitat o unes dimensions (figura 8), sinó també en els aspectes relatius a la correcta execució d'una tasca.

Un altre camp d'aplicacions de la visió a la indústria és el del *reconeixement i classificació d'objectes.* L'aplicació de les tècniques de reconeixement de formes, permet resoldre els problemes d'automatització en entorns on apareixen objectes heterogenis.

Alguns treballs experimentals han donat ja alguns resultats positius en determinades escenes i aplicacions, però encara són d'execució excessivament lenta i de fiabilitat reduïda com per a ser utilitzats en camps tals com el guiatge de robots autònoms, la conducció automàtica en autopistes o les aplicacions en vigilància i supervisió.

Les tècniques del reconeixement de caràcters i de les marques permeten també implementar sistemes informatitzats d'identificació, que s'utilitzen d'una manera complementària als massivament utilitzats codis de barres.

Per últim, els sistemes de visió també són utilitzats en la indústria per al *guiatge i control* d'altres màquines, especialment els robots en tasques de manipulació i muntatge.

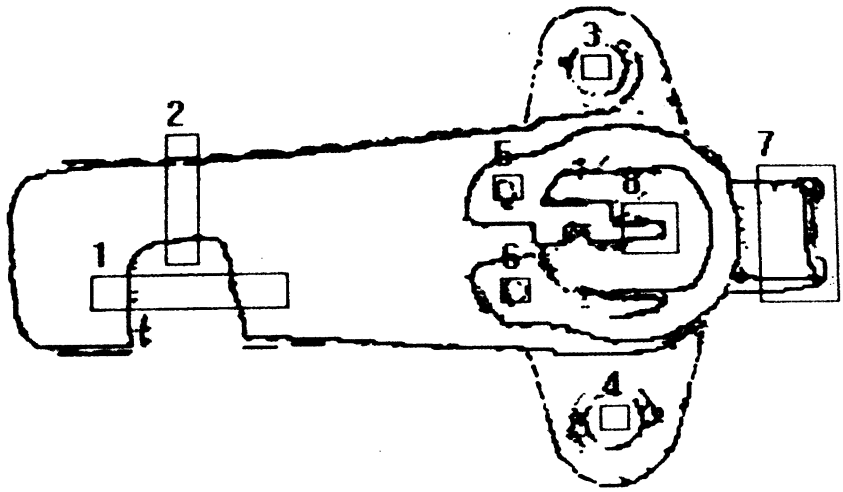

FIGURA 8. Inspecció basada en la mesura dimensional i paramètrica.

En els propers anys, la visió per ordinador veurà encara una més forta penetració en la indústria amb sistemes de cost més baix, que podran operar amb imatges en color i amb capacitats d'apreciació tridimensionals. Igualment, és de preveure la incorporació en un pròxim futur de la visió en el sector de serveis, especialment en tasques de control d'accessos i vigilància.

(Original rebut per a publicació el dia 15 de juny de 1994)