REAL-TIME OBJECT RECOGNITION BASED ON ACTIVE VISION AND SEQUENTIAL ANALYSIS

V. Ortmann, R. Eckmiller

Department of Computer Science VI, Neuroinformatik,
University of Bonn
Bonn D-53117, F.R. Germany

ABSTRACT

An image processing system for the real-time object detection and recognition was designed on the principles of Active Vision and Sequential Analysis. The real-world visual tasks can be solved due to predictive control of the vision sensor. The Sequential Analysis allows real-time implementation of the system on the low cost DSP hardware. The system was implemented on the DSP TMS320C50 and requires 18-30 ms for the detection and recognition of the object.

1. INTRODUCTION

Signal processing tasks and especially image processing may be simplified and effectively implemented in real systems using active control of the processing algorithm. In image processing such fields as attention control or purposive vision deal with optimal organization of processing methods in order to achieve the best performance under constraints of processing time or system costs.

The processing time is closely connected with robustness of processing methods. The robust performance of signal processing systems depends on variation of the observation conditions in the real world. Such variations may be partly neglected through special compensators and invariant transformations, or by means of sensor adaptation.

Active Vision can be defined as control of all parameters of the vision sensor [1]. The control allows to achieve much better performance of the whole image processing system due to extension of dynamical ranges, increase of resolution and compensation of internal noises. The simple pan-tilt movement of the camera allows to observe objects placed outside of the camera angle, or using local brightness control it's possible to see the objects with highly different luminance simultaneously.

Camera control means the change of internal electrical parameters or movement of lenses. The time span required for such changes depends on the electromechanical performance of the camera. Movement of the camera from one position to another or changing the shutter may take some seconds. This time increases the whole processing time and may not be directly compensated by using of a faster processor.

An optimal parallel implementation is possible only on a real parallel hardware. The parallel hardware is more expensive than modern DSP or embedded PC systems and price/performance ratio does not allow it's application in most industrial systems. So we choose the standard sequential processors as core of the image processing system and focus on optimal organization of the algorithms.

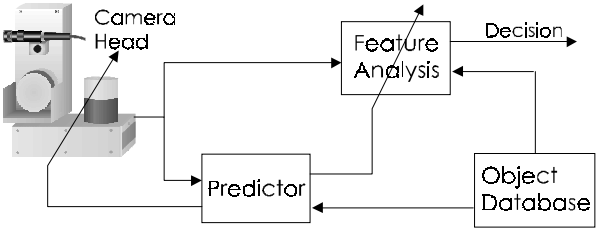

Figure 1 Structure of the recognition system

Performance of the designed algorithms is demonstrated on a visual recognition system. The structure of the system is presented in Figure 1.

2. RECOGNITION PROCESS

2.1 Feature Analysis

Assume $X(\theta,\alpha)$ is the feature vector describing the image under observation conditions θ and camera parameters α. Object template $X_0(\theta_0,\alpha_0)$ generated under optimal observation conditions θ_0 is contained in the object database. The components of X describe the stationary features of the object and may be detected with high probability under different observation conditions.

The matching procedure is based on the calculation of metric $L(X,X_0)$ and estimation of the current observation conditions θ. The variations of θ must be compensated as far as $L(X,X_0)$ is not invariant to the whole dynamical range of variations.

The multidimensional vector θ represents different observation conditions such as brightness, relative object position, rotation, scale, velocity of the movement. Some of the components can be estimated directly from X, others need a search procedure to find the moment estimation. The local combinatorial search was used in the system with the initial estimation of θ generated by the predictor.

The decision about the object appearance in the field of view of the camera may be defined as:

1. Object is present, if $L(X,X_0)<B$
2. Object is not present, if $L(X,X_0)\geq B$

2.2 Sequential Analysis

Each component of vector X is measured by a physical sensor independently from the other components and costs processing

time t^x_i. In order to minimize the whole processing time the sequential analysis is used.

The sequential algorithm of the object detection in the camera's field of view consists of the measurement of the i-th component of X and the calculation of metric $L_i(X,X_0)$ between object template X_0 and current observation. Due to theory of sequential analysis [2][3] after each estimation one of three decisions may be taken:

1. object is present, if $L_i(X,X_0)<B_i$
2. other components must be estimated, if $B_i \le L_i(X,X_0)<A_i$
3. object is not present, if $L_i(X,X_0)\ge A_i$.

Every decision is based on comparison of decision metric $L_i(X,X0)$ with two thresholds A_i and B_i, which depend on the noise of the sensor and the object itself. The A_i and B_i values can be estimated from the training set and define the bound in the space $P(L_i,i)$. The dynamical range of this space changes, as far as maximal number of X components or sample size depends on the object and variation of θ.

A recurrent RBF neural net is used in this design for dynamical estimation of A_i and B_i. The input of the net is vector X^i_0 and the outputs are the estimations of A_i and B_i. The net was trained to minimize the probability of the error of the first kind under fixed probability of the error of the second kind. The probabilities were estimated during the training process on an independent test set.

A significant acceleration in processing time is achieved due to adaptive choice of the X components depending on observation conditions θ.

Metric $L_i(X,X_o)$ and thresholds generate the trajectories in the 2D space $P(L_i,i)$ (Figure 2). The derivation $\partial L_i/\partial X_i$ is defined as informative importance of the i-th component of X. Depending on prior assumption about the probability of the object's presence in the field of view, the order of estimations is changed according to the ascend or descent of:

$$E\left(\left\|\frac{\partial L_i}{\partial x_i}\right\|\right)_{H_k},$$

where E(.) - expectation value,

$\|.\|$ - norm of derivation in $P(L_i,i)$

H_k - assumption (k=0,1) about the object's presence in the field of view.

Additional acceleration is achieved with the help of the sequential search method [4] for the calculation of $L_i(X,X_0)$. That means that the metric $L_i(X,X_0)$ must be monotone, so that $\forall X$, $\partial L/\partial X>0$ and must maximize the average SNR.

2.3 Camera Control

The system is based on a standard CCD-video camera , mounted on a pan-tilt unit and can be rotated in the range of ±60°. This camera has an interface for the control of internal parameters such as zoom, focus, gain, shutter, iris and camera position.

The probability of the error of the first kind can be significantly decreased through the compensation of variation of θ. The invariant properties of metric L have the optimal value in the predefined range around θ_0. Extension of this range needs either increase of computational complexity of the metric or increase of the search time for estimation of current θ. At the same time the dynamical range of a real CCD-sensor is much smaller as required for the real-world application. In order to minimize the processing time and to extend the applicability of the system the Active Vision approach is used for control of camera parameters and compensation of variation of observation conditions.

The estimation of optimal settings of parameters α uses the prediction of the current observation conditions, knowledge of transfer function of the camera and expectation of the SNR.

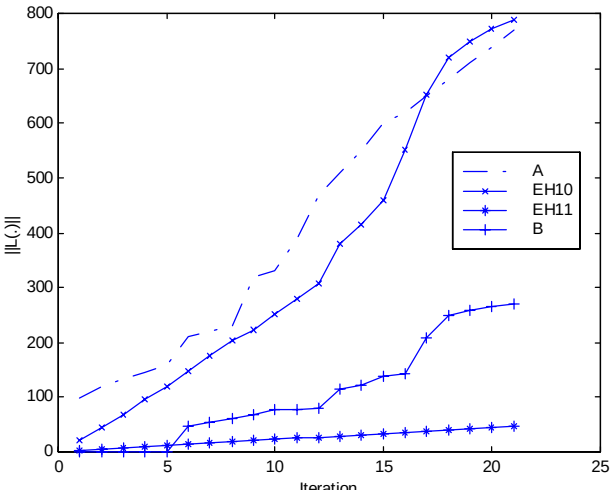

Figure 2 Dynamic of thresholds A,B and metric $L(X,X_o)$ under different prior assumptions about object EH11, EH10. (EH11 – detection of the object under assumption that object is present, EH10 – detection that object is not present under false assumption that the object is present)

2.4 Prediction

The predictor controls the feature analysis and the camera head. Predictive control is needed for optimization of feature analysis algorithm and compensation of the electromechanical delays of the camera.

The variation of parameters of observation conditions (position of the object, relative rotation and scale) is estimated with the help of local combinatorial search. The performance of different search algorithms depends on the initial assumption about the solution.

The nonlinear predictor, implemented as a recurrent multilayer neural network, was used for generation of the assumptions about possible values of θ components. The nonlinear predictor generates much better predictions [5] compared to such linear methods as Kalman-filter. The predictor uses the object template

X_o and the latest measured component of X as the input data for prediction of the θ components.

The step from the estimation of the (i-1)-th component to i-th is connected with the changes of the internal parameters of the camera. The changes of the electromechanical parameters require the time denoted as $t_{i-1,i}$, so that the whole processing time for estimation of N components is:

$$T = \sum_{i=1}^{N} t_i^x + \sum_{i=2}^{N} t_{i-1,i}$$

Minimization of t^x_i was achieved through prediction of the first assumption about observation conditions θ. The transition delays $t_{i-1,i}$ are compensated through the prediction of the next optimal setting of α before the current calculation of L_i is ready. The calculation of the prediction needs less time than the calculation of L_i , so that the camera is able to change it's parameters simultaneously with the calculation of L_i. It speedups the recognition, especially during the initialization phase.

Figure 4 Detection and recognition of the real object on a complex background.

3. RESULTS

3.1 Performance of the recognition system

The recognition system was tested on a set of the real world objects. The recognition statistics were estimated on 1042 images of 58 different objects. The objects were presented with approximately 16 images per object under different lighting conditions, background and variations of scale and rotation. The system is not able to recognize the object in the whole range of rotational and scale variations in the real time, thus the simulation results were used for testing. The probability of the error of the first kind is 0.06 if the probability of the error of the second kind was fixed at 0.02.

In order to compare the processing time of the system with other methods, the elastic-graph matching method [6] was implemented. This method shows very good performance in various applications and is used in a number of commercial systems. Both methods were simulated on PC.

The comparison shows that both methods are able to achieve the same recognition performance, but have different processing time.

EGM	Sequential method Initialization phase	Sequential method Predictive phase
220-860ms	80-720 ms	12-32ms

Table 1. Comparison of the processing time.

The analysis of the recognition process shows that sequential recognition algorithm is approximately 20 times faster than the application of parallel method on a sequential processor.

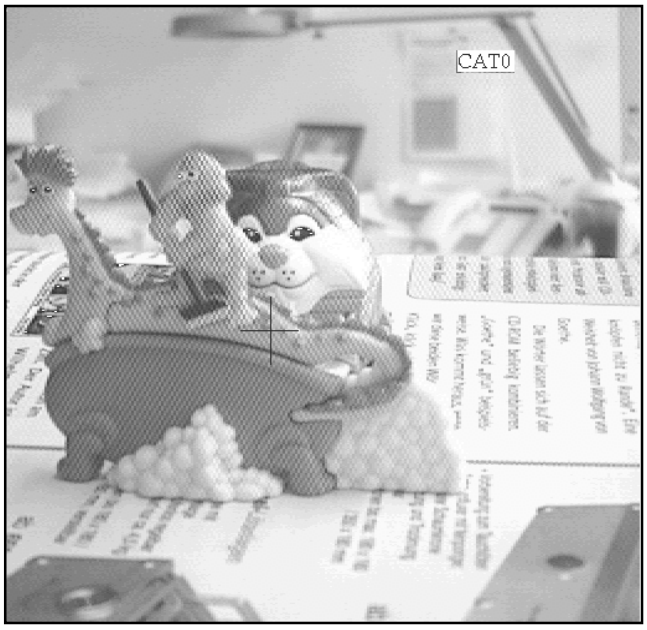

Figure 3 Detection and recognition of the partially covered object

The components of X describe stationary features of the object, allowing stable recognition of the object independent of the background (Figure 3).

The sequential analysis requires only partial matching between template and object. Due to this feature, the object can be recognized even if big part is covered by another object (Figure 4). The size of the observable part needed for stable recognition is dependent on the object and can be predicted for the defined observation conditions through direct modeling. After the

estimation of the observation conditions only 8-15% of the object was used for recognition.

Interesting properties of the predictive estimation of θ is the ability to use the internal memory of the predictor for the detection of the another object. The observation conditions change relative slow in compare to the processing time, that allows the predictor to use the old information about environment when a new object will detecting. This feature speedups the initialization phase of the detection of new objects. The Figure 5 shows significant decrease of the initialization phase of the detection of the second object.

Active Control of the camera parameters allows extending of the dynamical ranges of the system. Lighting conditions can change in the range of approximately 50dB without any loss of recognition performance. Controlling the zoom parameters of the camera the objects can be detected in the distance range from 0.1m till 6m, together with rotation of the pan-tilt unit the space angle of 180° can be scanned during a search of the object.

3.2 Real-Time Implementation

The described recognition system was implemented on a DSP system based on TMS320C50-80MHz and a gray level framegrabber. The system performs detection of the object, control of the camera and motor control of the pan-tilt unit. The DSP system was chosen as a low cost solution for industrial applications and can be easily integrated into other industrial image processing system.

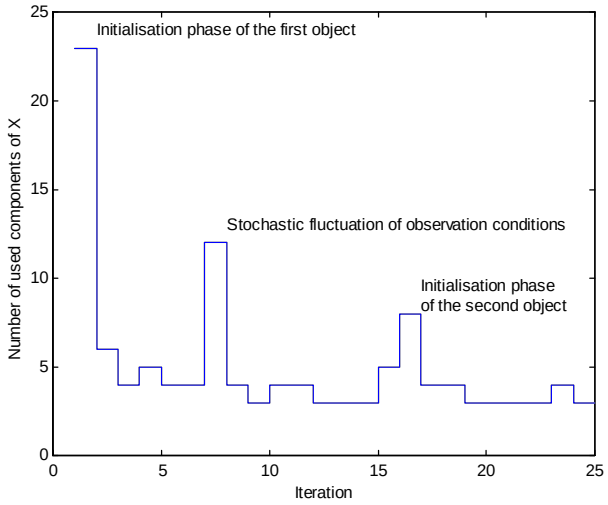

Figure 5 Optimization of the number of used components of X(θ,α)

The actual processing time depends on the processing stage and difference between θ and θ_0. The average processing time for the initialization phase is about 180-1200ms. When the first estimation of the observation conditions θ is done, the predictor is able to predict the optimal sequence of measurements and the processing time decreases drastically to 18-30 ms.

Various industrial applications require only estimation of some specific parameters of the object and in this case the consequent decrease of the processing time is possible. For example, if only the position of the object is interesting the processing time will reduced to 0.5-1.2 ms and free processor resources are available for detection of the next object.

4. SUMMARY

Application of the sequential analysis in combination with the active vision principles has allowed the real-time implementation of the visual recognition system on the low cost DSP hardware. The adaptive estimation of decision thresholds was implemented by the recurrent RBF neural net. Significant speedup was achieved through the application of the neural network predictor for the estimation of the initial conditions of the local combinatorial search algorithm and for the compensation of the internal delays of the vision sensor. Comparison with other recognition systems shows, that the application of described methods allows 20 times acceleration of the recognition process. So that detection and recognition of real objects on the complex background takes 18-30 ms on the low cost hardware.

5. REFERENCES

[1] Swain M., Stricker M. *Promising directions in active vision.* Technical Report CS 91-27, University of Chicago, 1991.

[2] Wald A. *Sequential Analysis,* Wiley publications in statistics, 1966

[3] Garvey T. D. *Perceptual strategies for purposive vision*, Technical Note 117, SRI International, September 1976.

[4] Barnea D.I., Silverman H.F. *A Class of Agorithms for Fast Image Registration*, IEEE Trans. Computers, C-21, pp. 179-186, February 1972.

[5] Ortmann V, Eckmiller R. *Neural Network Visual Tracking System*, Proc. of ICANN'97, Lausanne, Switzerland, pp. 817-822, Springer - Verlag, 1997.

[6] Patent PN4406020, *Verfahren zur automatische Erkennung von Objekten*, Zentrum für Neuroinformatik GmbH, 44801 Bochum, DE, 1995