MVA '96 IAPR Workshop on Machine Vision Applications, November. 12-14, 1996, Tokyo, Japan

Applying a Dynamic Recognition Scheme for Vehicle Recognition in Many Object Traffic Scenes

Włodzimierz Kasprzak [*]
Laboratory for Artificial Brain Systems
Frontier Research RIKEN

Heinrich Niemann [†]
Knowledge Processing Research Group
Bavarian Research Center FORWISS

Abstract

An adaptive object recognition scheme for image sequences of many object scenes is described. The scheme is applied for traffic object recognition under ego–motion. The recursive estimation of object states is performed by an extended Kalman Filter with modified error estimation, which is a neural network learning process. This new feature allows to separate the judgment needed for selection of best measurement among competitive image segments and the measurement judgment required by the recursive estimator.

1 Introduction

Road traffic control [6] and driver support [7] is an attractive application field for image sequence analysis systems. A reliable obstacle detection and classification in images of many–object scenes is still a challenging problem [8]. The complex nature of the subject makes it necessary to apply a dynamic model–based image analysis scheme, constraining the classes of recognized objects [3]. Usually such scheme employs a Kalman filter (KF) for recursive estimation of hypotheses. But even the tracking of many hypotheses is not sufficient for a complete scene recognition, as tracking works in an object–centered manner. One selects these image features only which support the concrete hypothesis. This does not include the explanation of other features.

We describe an adaptive recognition scheme, which deals with many object scenes and provides mechanisms for selection of best subsets of tracked hypotheses. We also identify image measurement problems caused by the use of a Kalman Filter. The KF requires that individual measurements are statistically independent. This is often not the case in vision tracking systems, where one usually selects these segments which best fits the hypothesis expectations. Furthermore, KF does not require judgments of each single measurement, assuming every measurement in given channel to be equal probable. This does not allow to track highly nonstationary signals, i.e. to follow abrupt changes of the tracked signal. We propose a necessary modification of the recursive estimator, that follow the idea of self–adaptation of learning rates in unsupervised learning of neural networks [1], [2].

2 The adaptive approach

2.1 The recognition scheme

The proposed adaptive recognition scheme is summarized in Figure 1. The value, token (or object) to be recognized in the image is modeled as a *dynamic system*. Only the projections of this object can be observed in the image (so called *measurements*) but the inherited *state* of the object is unknown. A parallel tracking of many competitive (up to n) *hypotheses* is performed. Each tracking process is supported by the *recursive estimation* mechanism. The selection of inconsistency free solutions takes place as soon as a robust selection among the hypotheses is possible. A selected and tracked hypothesis is shifted to the *recognized* state, if its stabilization satisfies a given criterion.

The general adaptive recognition scheme has been implemented on three different data abstraction levels of an image sequence analysis system for traffic scenes under a moving observer [5].

In this paper we concentrate on the design of the measurement process and the recursive estimator, as required for 3–D object recognition [4].

2.2 Filtering a dynamic system

Let $\boldsymbol{s}(t)$ be the unknown *state (parameter)* vector at time t and let $\boldsymbol{m}(t)$ be an associated observable *measurement* vector. The dynamic behavior of both vectors may be modeled by the following non–linear dynamic system in discrete time:

$$\boldsymbol{s}(k+1) = \mathbf{f}(\boldsymbol{s}(k)) + \boldsymbol{v}(k), \quad (1)$$

$$\boldsymbol{m}(k) = \mathbf{h}(\boldsymbol{s}(k)) + \boldsymbol{w}(k). \quad (2)$$

where $\boldsymbol{f}[\boldsymbol{s}(t)], \boldsymbol{h}[\boldsymbol{s}(t)]$ are vectors of nonlinear functions, $\boldsymbol{v}(k)$ denotes the system noise and $\boldsymbol{w}(k)$ means the measurement noise.

[*] Address: 2–1 Hirosawa, Wako-shi, 351–01 Saitama, Japan. E-mail: kas@zoo.riken.go.jp

[†] Am Weichselgarten 7, D–91058 Erlangen, Germany. E-mail: niemann@forwiss.uni-erlangen.de

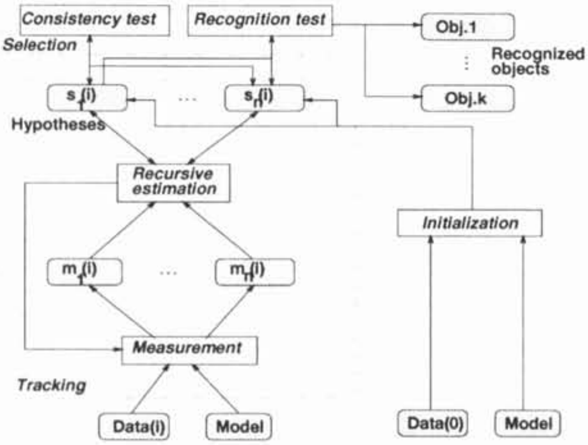

Figure 1: The adaptive recognition scheme.

```
1. Initialization: Application–dependent computation of initial estimation $s^*(k_0)$ and covariance matrix $P^*(k_0)$ of estimation error. Go to step 6.
FOR every next image $k > k_0$
    2. Detection of new measurement $m(k)$. The covariance matrix of the system noise $Q(k)$ is also available.
    3. Estimation of Kalman gain $K$ (see section 2.5).
    4. State modification (innovation):
       $s^*(k) = s^+(k) + K(k)\{m(k) - H(k)s^+(k)\}$
    5. Modification of matrix $P$ (section 2.5).
    6. Prediction of next state:
       $s^+(k+1) = F(k)\ s^*(k)$
    7. Prediction of next matrix $P$:
       $P^+(k+1) = F(k)P^*(k)F^T(k) + Q(k)$
    8. $k \leftarrow k+1$
```

Figure 2: The recursive object state estimator.

The goal of the *filtering* task is to make consecutive state estimations $s^*(t)$ on the basis of measurements $m(t)$ only. Unfortunately straightforward we can only calculate the measurements given the state vector, but not vice versa. For the filtering task of above system usually the so called *extended Kalman Filter* (EKF) is applied [9]. EKF works with instantaneous gradient values of above nonlinear functions. The state transition function $f(.)$ is linearized around the estimated state with its *Jacobi matrix* $F(k)$ and the projection function $h(.)$ is linearized around the predicted state with its *Jacobi matrix* $H(k)$:

$$F(k) = \frac{\delta f}{\delta s}|_{s^*(k)}, \ H(k) = \frac{\delta h}{\delta s}|_{s^+(k)}. \tag{3}$$

We apply a recursive estimator, which is similar to an EKF filter Figure 2. From the state modification it is evident that a crucial role in the minimization of the estimation error plays an appropriate design of the $K(t)$ matrix. But originally the gain matrix depended on the estimation error covariance matrix and on noise covariances only. There are two steps which we have modified with respect to our application. At first, the Kalman gain in EKF is estimated as (index k is omitted):

$$K(k) = P^* H^T \Big\{ H P^* H^T + R \Big\}^{-1}, \tag{4}$$

where $R(k)$ is the covariance matrix of the measurement. At second, the original modification equation of the error covariance matrix P is:

$$P^*(k) = P^+(k) - K(k)H(k)P^+(k). \tag{5}$$

Our main concern was, that both steps were independent from current *tracking error* $m(k) - H(k)s^+(k)$. Otherwise, it would be not trivial to set properly all the $n \times m$ parameters $K_{ij}(t) \in K(t)$ for all t by default. Hence, we need either a scheme for measurement judgment or a direct inclusion of the tracking error in the gain estimation equation.

2.3 The measurement step

During the measurement step at first it is important to detect and to select the current measurement vector $m(k)$ itself. It is clear that various image segments may have different certainty factors associated with them. The criteria for selection of the best one should be most independent from current tracked object hypothesis as possible. The search area in the image for new measurement may be related to the area explained by given object hypothesis. But other selection criteria should follow model specific assumptions and not the current hypothesis.

The quality of state estimation is the result of combining the actual estimation error, expressed by $P^*(k)$, with current $R^*(k)$ and the system error, expressed by the covariance matrix $Q(k)$ of $v(k)$. Usually the measurement error is assumed to be modeled by a constant matrix $R(k)$. This assumption is not acceptable neither for vision purposes nor for nonstationary signals. Due to the tracking error independent design of the Kalman gain, the estimator may not be able to follow highly non–stationary signals, i.e. to track abrupt changes. After some tracking time the covariance matrix P is relatively stable, resulting in low covariance values and from that moment the state estimation is only little affected by next measurements. But some non–stationarity must be expected in our application as both rapid movement changes of objects may appear and wrong individual measurements may occur.

2.4 Designing the recursive estimator

For recursive estimation of the gain and the state covariance matrix a scheme is proposed, that is similar to the self–adaptation of learning rates in neural network learning, first proposed in [1].

Let us consider a neural network described in matrix form as: $y(t) = W(t)x(t)$, where $y(t)$ is the output vector, $W = [w_{ij}] \in R^{n \times n}$ is the synaptic weight matrix, $x(t)$ is the measurement vector. A learning rule means the minimization of expected error, i.e.

$$\frac{dw_{ij}(t)}{dt} = -\mu_{ij}(t) f_{ij}(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{x}), \tag{6}$$

where $\mu_{ij}(t) > 0$ is a local adaptive learning rate, and f_{ij} is some loss function.

In a recent paper [2], it was proposed that each synaptic weight w_{ij} has its own (local) learning rate $\mu_{ij}(t)$ and that this rate is adjusted during the learning process according to a set of differential equations. These rules can be written in discrete time form as:

$$v_{ij}(t+1) = (1-\delta) v_{ij}(t) + \delta f_{ij}(t), \tag{7}$$

$$\eta_{ij}(t+1) = (1-\eta_{ij}(t)\delta_1)\eta_{ij}(t) + \alpha\eta_{ij}(t)|v_{ij}(t)| \tag{8}$$

where δ, δ_1 and α are some positive scalars. δ and δ_1 control the stabilization speed of the expected error $\boldsymbol{v}(t)$ and learning rate $\boldsymbol{\eta}(t)$ matrices, whereas α controls the influence of expected error in one iteration and is normalized by the maximum expected error.

In our estimator, for measurement judgment we use the current tracking error $\boldsymbol{e}(k) = \boldsymbol{m}(k) - \boldsymbol{H}(k)\boldsymbol{s}^+(k)$ with its covariance matrix $\boldsymbol{R}_e(k) = \mathrm{E}\{\boldsymbol{e}(k)\boldsymbol{e}^T(k)\}$. To assure a minimum error threshold the default matrix $\boldsymbol{R}(t)$, corresponding to measurement noise is also added. Thus the estimation of matrix $\boldsymbol{P}^*(k)$ is given as:

$$\boldsymbol{P}^* = (1-\delta)\boldsymbol{P}^+ + \delta \boldsymbol{P}\boldsymbol{H}^T(\boldsymbol{R}_e + \boldsymbol{R})(\boldsymbol{H}\boldsymbol{P}^+\boldsymbol{H}^T)^{-1} \tag{9}$$

and the estimation of a single gain element is:

$$\begin{aligned} K_{ij}(k+1) &= (1 - K_{ij}(k)\delta_1)K_{ij}(k) + \\ &+ \alpha K_{ij}(k) \textstyle\sum_l P_{il}(k) H_{lj}(k). \end{aligned} \tag{10}$$

3 The object recognition application

The processing structure of model based vehicle recognition in images of traffic scenes consists of object initialization, recursive object estimation with measurement and final object consistency and selection tests.

By the initialization of an object hypothesis we mean the detection of a segment *group*, the detection of an object class, and the initialization of an appropriate *state vector* elements of the object hypothesis. For example, as seen in Figure 3 there are two bounding boxes provided for one contour group: the overall box and the included first contour box. The combination of these two detected image bounding boxes with the model-based restrictions about the length-to-width and height-to-width ratios makes the direction hypothesis possible.

During object tracking a *recursive* estimation of each valid hypothesis is performed after a new measurement $\boldsymbol{m}(k)$ for it was detected. For tracking error detection, a predicted 3-D wire-frame model (3-D bounding box or model edges) are projected onto the image and they are matched with the new image features (groups or edges).

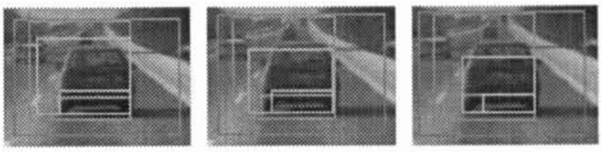

Figure 3: Adaptation of the measured segments and the object hypothesis onto the real shape in consecutive images (the outer double box denotes the image search area, the smaller one the object prediction and the bright one comes from the measured data).

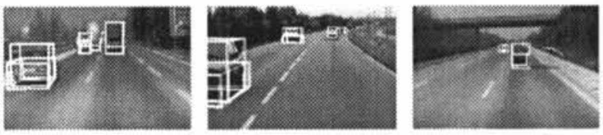

Figure 4: Examples of selected vehicle hypotheses in three image sequences.

The measurement in our application means the use of problem dependent or model based methods for the detection of appropriate image features, structures, road parameters, or object parameters. The matching of model feature with the next image features can be performed on two ways: the measured points are derived from contour groups or from line segment groups. In both cases the modification is based on the differences between projected model points and significant points of measured data group.

In Figure 3 an example of one projected hypothesis and measurement of a real vehicle data in the image sequence is given. The measurement variance is nearly independent from the distance of the selected segment group from the expected object hypothesis. There are many alternative projections of the object hypothesis generated in the image, due to a limited variability of the state parameters.

A hypothesis is either in its *tracking* or in one of its *recognition* phases. A tracking phase is given if the tracking time of this hypothesis is lower than T_{min} or its variance is greater than the $maximum\_var$. Otherwise the hypothesis is in one of its three recognition phases. These phases are closely related to the use of object specialization levels (from *Shape* over *Fine* to *Type*).

The consistency test takes place between pairs of hypotheses. If the tracking times of two competitive hypotheses are both larger than T_{min} or one of them is in the recognition phase (i.e. tracking time $> T_{min}$ and its variance $< maximum\_var$), then the consistency test among them is performed.

4 Computer simulation results

We tested the system on several image sequences of length 125–250 images, usually containing 3–5 vehicles (Figure 4).

The projection conditions have been determined by the use of a camera with the relation of the focal length to pixel width of 708 (after sub-sampling) and with the height position over road of ca. 1.70$[m]$.

Object	Vis	Gen	Gen/Vis.	Track	Tr./Vis.	Recog.	Re./Vis.
l. car	125	125	100.0	125	100.0	122	97.6
bus	125	122	97.6	125	100.0	120	96.0
tanker	125	105	84.0	88	70.4	35	28.0
truck	125	123	98.4	121	96.8	65	52.0
r. car	62	51	82.2	50	80.6	40	64.5
far car	125	37	29.6	28	22.4	17	13.6

Table 1: Qualitative evaluation of vehicle recognition in one image sequence. *Vis* means the number of images in which an object was visible, *Gen* - image number when at least one hypothesis for given object was generated, *Track* - the time a hypothesis was in the tracking phase, *Rec* - the time a hypothesis was in the recognized phase.

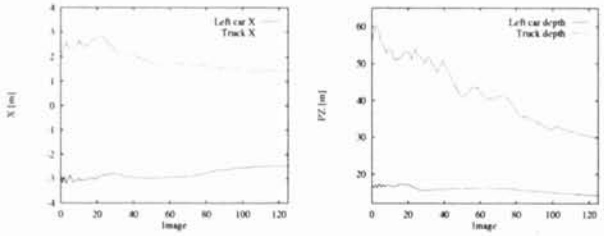

Figure 5: The estimated side positions (left) and depths (right) of the car and truck.

A qualitative evaluation is based on the hypothesis visualization in image sequences as shown in Figure 4 and on the tracking and recognition rate data as given in Table 1. In the first scene six moving vehicles are visible, one of them in the half number of images only. Four vehicles were detected very well and they were tracked in 95–100 % of the images. The image projections of these vehicles are from the interval of $20 \times 20[pixel^2] - 50 \times 70[pixel^2]$. The only partly visible vehicle, whose image size is about $10 \times 12[pixel^2]$, was detected in average in every third image only. Similar recognition quality was observed for other four sequences. After the hypotheses have been tracked successfully for some time, they change to the recognition phase.

The estimated position parameters for the left car and the truck in the center in first image sequence are shown in Figure 5. An acceptable depth estimation can be observed for such objects, which are at least projected to image regions of size $30 \times 30[pixel^2]$. This means for present projection conditions a depth of $60[m]$ only, but by increasing the image resolution only slightly, much larger depths will be estimated properly.

Figure 6 shows the dynamics of the estimation variances and gain parameters for the depth and side position of two tracked vehicles in first image sequence. It can be observed that for truck hypothesis several times the depth tracking error was relatively high, but was immediately compensated. This speedup was made possible by self–adaptive increase of corresponding gain parameter.

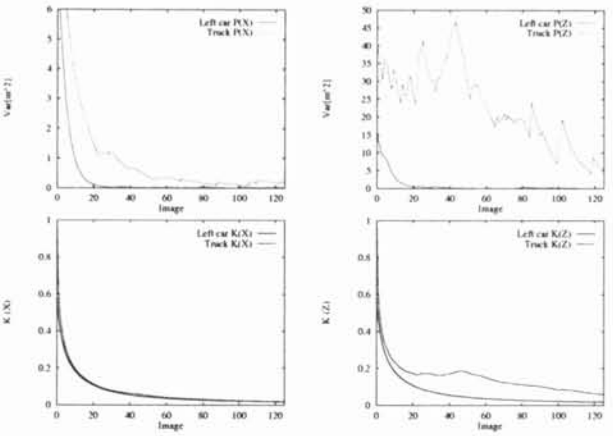

Figure 6: Estimation variances and gains corresponding to two tracked hypotheses.

5 Summary

Different from tracking single objects, which is mainly a stabilization task and a hypothesis–driven processing, in presented approach we provide a scene oriented explanation and assure real tracking of nonstationary measurements. Important aspect is the separation of measurement judgment for the purpose of selection among competitive measurements from measurement error, required for gain estimation. This allowed the design of an recursive estimation, which very well follows nonstationary measurement signals.

References

[1] S. Amari. Theory of adaptive pattern classifiers. *IEEE Trans. on Elect. Comput.*, 16(3):299–307, 1967.

[2] A. Cichocki, S. Amari, M. Adachi, W. Kasprzak. Self–adaptive neural networks for blind separation of sources. *ISCAS'96*, 2, pp. 157–160, IEEE, 1996.

[3] E.D. Dickmanns, V. Graefe. Applications of dynamic monocular machine vision. *Machine Vision and Applications*, 1:241–261, 1988.

[4] W. Kasprzak. Ground plane object tracking under egomotion. *International Archives of Photogrammetry and Remote Sensing*, 30(5W1):208–213, 1995.

[5] W. Kasprzak, H. Niemann, D. Wetzel. Adaptive estimation procedures for dynamic road scene analysis. *Proceedings ICIP–94*, , vol. I, pp. 563–567. IEEE Computer Society Press, Los Alamitos, CA, 1994.

[6] D. Koller, K. Daniilidis, H.-H. Nagel. Model–based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision*, 10(3):257–281, 1993.

[7] I. Masaki. *Vision-based Vehicle Guidance*. Springer, New York, Berlin, Heidelberg etc., 1992.

[8] U. Regensburger, V. Graefe. Visual recognition of obstacles on roads. *Proceedings IROS '94*. , pp. 980–987, IEEE/RSJ/GI Munich, Germany, 1994.

[9] J.J. Wu, R.E. Rink, T.M. Caelli, V.G. Gourishankar. Recovery of the 3-d location and motion of a rigid object through camera image. *International Journal of Computer Vision*, 3:373–394, 1988.