FUSION OF TEXTURE AND CONTOUR BASED METHODS FOR OBJECT RECOGNITION

Uwe Handmann and Thomas Kalinke

Institut für Neuroinformatik, Ruhr-Universität Bochum

44780 Bochum, Germany

Keywords: Machine vision, Data fusion, Object recognition

ABSTRACT

We propose a new approach to object detection based on data fusion of texture and edge information. A self organizing Kohonen map is used as the coupling element of the different representations. Therefore, an extension of the proposed architecture incorporating other features, even features not derived from vision modules, is straight forward. It simplifies to a redefinition of the local feature vectors and a retraining of the network structure.

The resulting hypotheses of object locations generated by the detection process finally are inspected by a neural network classifier based on cooccurence matrices.

INTRODUCTION

In order to build driver assistance systems or autonomous vehicles like [D$^+$94] does one of the main tasks to be solved is the detection, tracking and classification of objects. This processing stage should end up in a representation of the vehicle's environment depending on the actual task to be performed [vS$^+$97].

We propose a new approach to object detection based on data fusion of texture and edge information. A self organizing Kohonen map [Koh82] is used as the coupling element of the different representations. Therefore, an extension of the proposed architecture incorporating other features, even features not derived from vision modules, is straight forward. It simplifies to a redefinition of the local feature vectors and a retraining of the network structure.

The resulting hypotheses of object locations generated by the detection process finally are inspected by a neural network classifier based on cooccurrence matrices [HSD73].

We start with a short description of the feature extraction. Then, the generation of the input vector of the Kohonen map is described. The second part deals with the generation of the input data for the neural network classifier. We conclude with a real world experiment.

SEGMENTATION BY FUSION

In knowledge-based image processing systems it is necessary to combine texture- and contour-based methods on different processing levels in order to increase the performance, robustness and efficiency of the algorithms. In contrast to the application named *VISIONS* introduced by [Das94], here the central element of the fusion process is the learn-able coupling structure. In [Das94] a predefined structure controls the influence of the individual methods to the complete system. Our work introduces a coupling structure based on a neural network which learns to combine the individual algorithms according to the performance of the complete system (refer figure1). The approach chooses the pixel coordinate system as a common base for fusion. Its aim is to evaluate an object-background-separation of traffic scenes. In a preprocessing step information about contours and textures is used. A feature vector is created for each pixel derived from the results of the preprocessing algorithms. Finally a neural network learns the necessary coupling structure and how to combine the con-

tributions of the individual methods. We apply the structure to the segmentation of images. Detected image parts can be interpreted as a map containing danger spots for the driver assistance system and are further processed by a classifier.

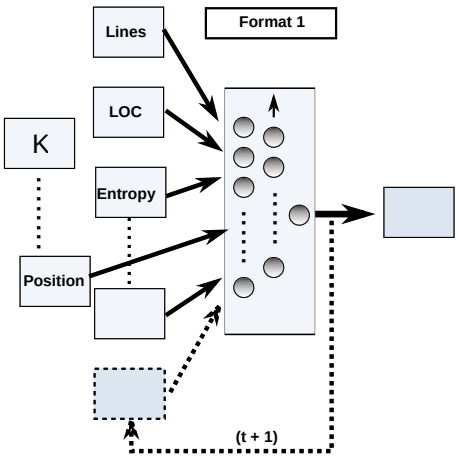

Figure 1: Model of fusion process.

REPRESENTATIONS

Texture and contour representations are determined by three methods:

- local orientation coding (LOC) [GNW95],
- polygon approximations of edges, and
- local image entropy [KvS96].

The LOC has been widely used as a feature for segmentation, tracking and classification. It codes the gradient information depending on the local orientation. Polygon approximations can be used to characterize different objects. The entropy of an image part gives an estimation about the contents of information.

Local Orientation Coding (LOC)

The 'raw' grey scale images are preprocessed by a differential method we call local orientation coding (LOC). The image features obtained by this preprocessing are bit strings each representing a binary code for the directional grey-level variation in a pixel neighborhood. In a more formal fashion the operator is defined as

$$\begin{aligned} b'(n,m) &= \sum_{i,j} k(i,j) \cdot u(b(n,m) \\ &- b(n+i,m+j) - t(i,j)), \\ (i,j) &\in \text{neighborhood} \end{aligned}$$

where $b(n,m)$ denotes the (grey scale) input image, $b'(n,m)$ the output representation, $k(i,j)$ a coefficient matrix, $t(i,j)$ a threshold matrix and $u(z)$ the unit step function. The matrices may have negative index values. The output representation consists of labels, where each label corresponds to a specific orientation of the neighborhood. For a N_4 and a N_8 neighborhood on regular square grids, suitable choices for the coefficient matrices are

$$\begin{bmatrix} 0 & 1 & 0 \\ 2 & R & 4 \\ 0 & 8 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 4 \\ 8 & R & 16 \\ 32 & 64 & 128 \end{bmatrix} \qquad \begin{matrix} n \\ \hookrightarrow \ m \ , \end{matrix}$$

where R is the reference position. This choice for N_4 leads to a set of labels $b'(n,m) \in [0,\ldots,15]$ corresponding to certain local structures. The choice of the coefficients and the formulation of the operator gives rise to some properties:

- Due to the unique separability of the sum into its components, the information of the local orientation is preserved.
- The approach is invariant to absolute intensity values.
- The search for certain structures in the image reduces to working with different sets of labels. For horizontal structures mainly the labels 1, 8 and 9 have to be considered.

An adaption mechanism for the parameters $t(i,j)$ of the coding algorithm yields a high level of flexibility with respect to lighting conditions [GNW95].

Polygon Approximation

The polygon approximation is calculated by a special hardware applying a *Sobel*-filter, some thinning and concatenation of the contour

points. Polygons of a length $l < 10$ are suppressed in order overcome background noise.

Entropy

The calculation is based on the information theory introduced by Shannon [Sha48]. A part of an image can be interpreted as a signal x_k of k different states with the entropy $E(x_k)$ determining the observer's uncertainty about this signal. It measures the contents of information. For every pixel the normalized histogram of a centered neighborhood is calculated as an estimation of the probability distribution function $p(x_k)$

$$E(x_k) = -\sum_k p(x_k) \log p(x_k).$$

Figure 2 shows the different representations used for constructing the feature input vector of the Kohonen map.

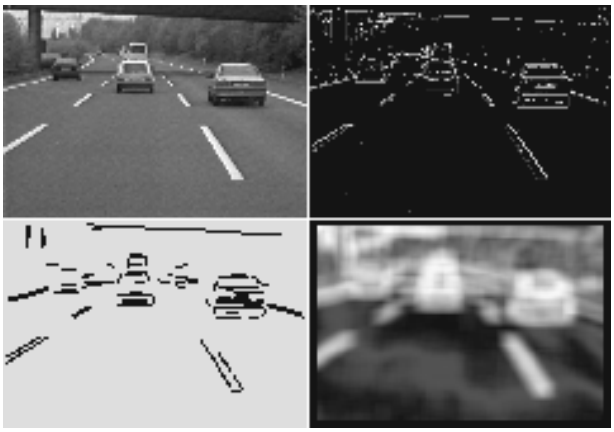

Figure 2: Representations: Original image, local orientation coding, approximated polygons and entropy (from left to right, top to bottom).

COUPLING STRUCTURE - A NEURAL NETWORK

Based on the three representations for each pixel a 12-dimensional input vector

$$\mathbf{u}(x,y)^T = (\mathbf{u_1}(x,y)^T, x, y)^T$$

of the Kohonen map is generated. Here, the 10-dimensional vector $\mathbf{u_1}(x,y)$ refers to the contour and texture information and the values of x and y represent the pixel's coordinate position. Vector $\mathbf{u_1}(x,y)$ is defined by

$$\mathbf{u_1}(x,y) = \sum_{(i,j)\in R} \mathbf{v}(i,j)$$

where R is a local neighborhood (e.g. 9×9) of (x,y) and $\mathbf{v}(i,j)$ is a binary vector. $(v_1(x,y), \ldots, v_4(x,y))^T$ code subsets of the LOC results, $(v_5(x,y), \ldots, v_9(x,y))^T$ code the value of the entropy and $v_{10}(x,y) = 1$, if (x,y) belongs to a polygon.

SEGMENTATION RESULTS

In a set of images the regions of the objects are labeled by hand and the input vectors of the Kohonen map are generated. The resulting training set consists of 55% vectors referring to objects. Figure 3 depicts the resulting Kohonen map after a coarse-to-fine training.

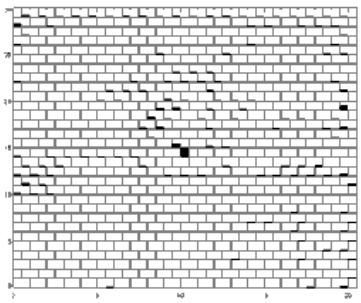

Figure 3: Kohonen map which learnt the coupling structure.

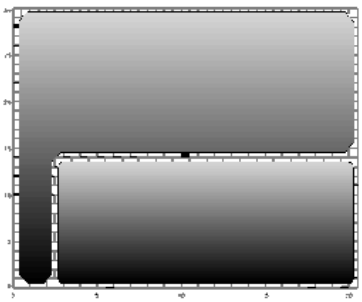

Figure 4: Labels: Objects and background.

According to the two labels (objects, background) the map can clearly be divided into two regions (figure 4).

Figure 5 shows an example of the segmentation result of our data fusion approach. Here, the class of the vectors $\mathbf{u}(x,y)^T$ of the original image in figure 2 is determined by the label of the Kohonen map's representation vector with minimal Euclidean distance.

Figure 5: Results: Segmentation by fusion.

TEXTURE-BASED CLASSIFICATION

As shown in figure 5 the image information has been restricted to some aspects (image blobs). The fusion process enhances the performance of the individual algorithms and concentrates the attention of further image processing to structural features. Due to the fact that the knowledge about the scene estimates initial ROIs by calculating the center of the blobs under constraints of object sizes a classification task has to be solved in order to build up a representation of the world for prediction and behavior in future. A tracking based on cross correlation (Ccor)

$$\mathrm{Ccorr}(I_{ROI}(t), I_{ROI}(t-1)) = \frac{\mathrm{Cov}(I_{ROI}(t), I_{ROI}(t-1))}{\sqrt{\mathrm{Var}(I_{ROI}(t), \mathrm{Var}I_{ROI}(t-1))}}$$

is used to track the initial regions $I_{ROI}(t-1)$ over time t. The tracking results are finally classified.

A lot of different classifiers have been introduced for traffic scene analysis. In [GNW95] or [Bra94] classifiers depending on contours and contour-models are described. They suffer under inefficient calculation of contour features which are unusable, if the objects are too small thus, they are situated in the long distance field. Furthermore rotations of objects are difficult to cope with because e.g. LOC-features will change and in case of a model-based approach [NWvS95] a further degree of freedom for the elastic model has to be permitted.
Therefore in this application a texture-based measurement was chosen which is rotation- and scaling invariant. Usually texture calculation is partitioned into structural and statistical methods. In the case of car and truck classification the statistical models have to be preferred because a structural description is not flexible enough to cover all different types of objects. The well known cooccurrence matrices introduced by [HSD73] are chosen as a feature vector for solving the classification task.

Cooccurrence Matrix

Cooccurrence matrices are second order statistics. Haralick defined 14 different measures which have been commonly used for texture classification tasks [HS92]. The matrices concentrate the statistical image information under the constraints of different angles and distances. In a region of interest $I_{ROI}(x,y)$ of size $M \times N$ and a maximal number of different grey-values Q the cooccurrence matrix $P(i,j)$ is calculated (as illustrated in figure 6) for a given direction α and for a given distance d by

$$P_{d,\alpha}(i,j) = \frac{num}{denom},$$

where the numerator (num) is defined by num = [Number of pairs $(x,y),(x',y')$, satisfying (d,α) and $I_{ROI}(x,y) = i$ and $I_{ROI}(x',y') = j$] and the denominator ($denom$) is defined by $denom$ = [Number of all pairs $(x,y),(x',y')$].

In order to establish rotation invariance the sum $S(i,j) = \sum_\alpha p_\alpha(i,j)$ is calculated for all possible angles $\alpha \in \{0, 45, 90, 135\}$ and d is kept constant. Scaling invariance is implicated in the calculation role. As long as the background of an object does not differ to much a restricted translation invariance is given, too.

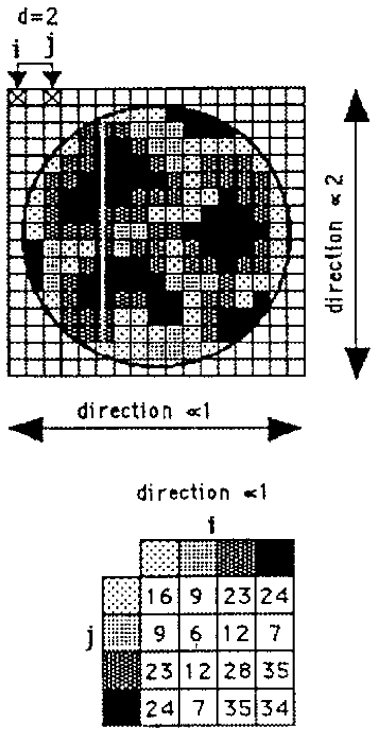

Figure 6: Calculation of the cooccurrence matrix.

Classification

In a lot of applications features of the cooccurrence matrices like energy, entropy, contrast, correlation and so on [HSD73] were used for classification processes. But every reduction of the dimensionality implicates reduction of information as well. Therefore, the matrix itself is used for classification. It is the task of the neural network to extract the necessary feature by itself. Due to the fact that the matrix is symmetric only the non redundant part of the matrix is extracted to build the characteristic vector $S_1(i,j)$.

In order to estimate the statistics of even small objects a reduction of the grey-level dynamics is performed. The range of $Q \in \{0, \ldots, 255\}$ is reduced to a 4-bit range $Q \in \{0, \ldots, 15\}$ by a bit shift operation (extracting the highest 4 bits). Another method introduced in [BY95] implies a dynamic adaption to the actual grey-level distribution. It is the goal to keep the influence of the background to the grey-level reduction as small as possible. Therefore, the bit shift operation is selected to process all elements in the same manner without dependency on the background distribution.

CLASSIFICATION RESULTS

The cooccurrence matrices were calculated from a classification database. A neural network was chosen as a classification structure. A multi-layer perceptron with one hidden layer using the quick-prop-learning algorithm was trained. The net contains 136 input neurons, 5 hidden neurons and 3 output neurons for the classes car, truck and background. The input vectors are calculated from a classification database containing a wide spectrum of different objects. They were taken from different image sequences acquired by various camera under different points of view to cover large spectrum of objects. The training set contains about 680 examples: 300 cars, 200 trucks and 180 images of the background. All objects had different scalings starting from 40×20 pixel up to 400×200 pixels.

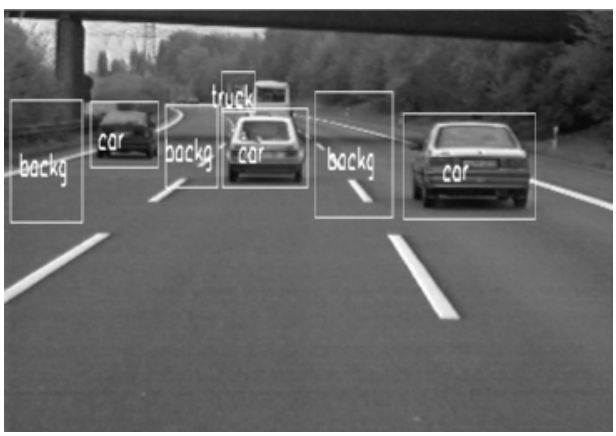

Figure 7: Classification results.

Figure 7 shows the classification results of a preprocessed image of figure 5. A classification of the background is added. The performance of the classifier scales with the size of the Rois, thus if structural details get lost the differentiation of cars and trucks becomes more difficult. A stabilization over time improves the results.

CONCLUSION

The proposed method combines two main aspects: initial segmentation and classification. The main part of the initial segmentation is done by the learnable coupling structure - a Kohonen map. The fusion of different types of data, textures and contours, provides a good measurement for objects. The final classification completes the application. A stable description of the image scene can be given. Lacks of the final classifications which can occur in noisy and small parts of the background have to be rejected by a stabilization over time. Nevertheless, the combined approaches are able to cope with most of the possible arrangements of vehicles, so that an environmental representation can be provided.

ACKNOWLEDGMENT

The authors want to thank Martin Werner for proof reading the paper.

REFERENCES

[Bra94] Michael E. Brauckmann. *Visuelle Automobilerkennung zur Fahrzeugführung im Straßenverkehr.* VDI Verlag, 1994.

[BY95] P. Bhattacharya and Y.-K. Yan. Iterativ Histogram Modification of Gray Images. *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-25(3):521–523, 1995.

[D+94] E.D. Dickmanns et al. The Seeing Passenger Car 'VaMoRs-P'. In *Proceedings of the Intelligent Vehicles '94 Symposium, Paris, France*, pages 68–73, 1994.

[Das94] B. V. Dasarathy. *Decision Fusion.* IEEE Computer Society Press, Los Alamitos, 1994.

[GNW95] C. Goerick, D. Noll, and M. Werner. Artificial Neural Networks in Real Time Car Detection and Tracking Applications. *Pattern Recognition Letters*, 1995.

[HS92] Robert M. Haralick and Linda G. Shapiro. *Computer and Robot Vision*, volume I. Addison-Wesley, Reading, Massachusetts, 1992.

[HSD73] R.M. Haralick, K. Shanmugan, and I. Dinstein. Textual features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6), 1973.

[Koh82] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[KvS96] T. Kalinke and W. von Seelen. Entropie als Maß des lokalen Informationsgehalts in Bildern zur Realisierung einer Aufmerksamkeitssteuerung. In *Mustererkennung 1996*, pages 627–634, 1996.

[NWvS95] D. Noll, M. Werner, and W. von Seelen. Real-Time Vehicle Tracking and Classification. In *Proceedings of the Intelligent Vehicles '95 Symposium, Detroit, USA*, pages 101–106, 1995.

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.

[vS+97] W. von Seelen et al. Image Processing of Dynamic Scenes. Internal Report IRINI 97-14, Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum, Germany, July 1997.