8—12

Detection of Crowds of People by use of Wavelet Features and Parameter Free Statistical Models

D. Faulhaber, H.Niemann, P. Weierich
E-mail: faulhaber@forwiss.de
Bavarian Research Center for Knowledge Based Systems (FORWISS)
Research Group for Knowledge Processing
Am Weichselgarten 7, 91058 Erlangen, Germany

Abstract

Based on a sample of 8000 manually segmented video images acquired in a subway station, we examine if heads of pedestrians can be detected by use of wavelet features and parameter free statistical classifiers. In particular we address the question of which wavelet filter to use for feature generation and how to model class conditional densities of features. We show that a markov chain based approach in combination with HAAR-wavelets is a good compromise between high recognition rates and computational efficiency. The latter is required since developed algorithms are supposed to run on a dedicated low cost hardware.

1 Introduction

The research project "Intelligent Train Platform" is performed with participation of PLETTAC ELECTRONICS - a camera and security equipment provider, VAG Nuremberg - an operator of a public transport subway net and the Bavarian Research Center for Knowledge Based Systems (FORWISS). Our goals are to increase passenger safety and to minimize the time a train has to stop in a subway station. Among the tasks to be solved is the detection of arriving and departing trains, recognition of intrusions into the rail area and detection of people on the platform. This in conjunction with object tracking mechanisms can be used by a scene analysis module for interpretation of pedestrian and train traffic in a subway station. Minimization of train stop time is to be achieved by automatic recognition of the end of pedestrian traffic between train and platform - a procedure currently done by the train driver. The basis for our system are video images of which an example is given in figure 1. The field of view extends over approximately $50m$.

Ultimately the whole system has to be executable on a dedicated low cost hardware. Hence efficient, simple and reliable signal processing methods have to be developed. In the following we examine the suitability of wavelet features in conjunction with efficient parameter free statistical classifiers. In particular we address the question of which wavelet filter pair and class conditional density model to use. Our results are based on a sample of 8000 manually segmented video images in which 140 000 head centroids of pedestrians were labeled.

Our basic assumption is that pedestrians head's form a texture which is clearly discernible from scene background by use of wavelet features and parameter free statistical classifiers. It has been shown before (e.g. [5, 3]) that wavelet features can be used for texture discrimination. However it has not been tried to detect human heads with this approach.

2 Related works

Colmenarez et al. ([2]) used first order markov processes and grey value features to detect frontal human faces in photos. An optimal markov process with regard to a class discrimination measure was constrcuted with a minimum spanning tree approach. Very good results were obtained for a sample ten times smaller than ours. Figure 1 shows that in our case most pedestrians look away from the camera. This makes a frontal view approach infeasable.

Regazzoni et al. ([6]) assumed that any grey value change in the image was due to pedestrian movement. Hence substraction of subsequent video frames could be used to find pedestrians. Then the number of people in a crowd was estimated by use of edge detectors and Kalman filters. In our case there are at least two more reasons for grey value changes: i) Incoming and departing trains and ii) Lighting changes in outdoor subway stations. This means that a more sophisticated crowd detection approach must be developed.

3 Classification

Detection of crowds of people in images of a video stream can be reduced to the following question: For each input pixel at a position (x, y) of an image s from a video stream S it has to be decided whether it belongs to class Ω_0 = "scene background" or class Ω_1 = "head of pedestrian". It was chosen to do this by use of a Bayes Classifier to minimize the error rate. Hence the comparison

$$p_0(s)\ p(\mathbf{c}(s,x,y)|\Omega_0) > p_1(s)\ p(\mathbf{c}(s,x,y)|\Omega_1)$$

has to be performed for every pixel of the input image. $p_0(s)$ and $p_1(s)$ are the a priori probabilities of classes Ω_0 and Ω_1 in image s. $\mathbf{c}(s,x,y)$ is some feature vector at position (x, y) of the image s. $p(\mathbf{c}(s,x,y)|\Omega_\kappa), \kappa = 0, 1$ is the class conditional density of class Ω_κ.

If components c_k of feature vector

$$\mathbf{c}(s,x,y) = (c_k)_{k=0,\dots,n-1}$$

are statistically independent,$p(\mathbf{c}(s,x,y)|\Omega_\kappa)$ can be modeled as

$$p(\mathbf{c}(s,x,y)|\Omega_\kappa) = \Pi_{k=0}^{n-1} p(c_k|\Omega_\kappa). \qquad (1)$$

This is a product of n one dimensional densities which can be approximated using bin histograms with b bins.

Dependencies, both linear and nonlinear, of pairs of feature vector components can be modeled with stochastic automata M_κ using b finite states. An automaton M_κ for class Ω_κ is defined such that it accepts a feature vector $\mathbf{c}(s,x,y)$ with the following probability

$$p(\mathbf{c}(s,x,y)|\Omega_\kappa) = \Pi_{k=0}^{n-1}\ {}_\kappa p_{q_k(c_k),q_{k+1}(c_{k+1})}(k) \qquad (2)$$

This equation is made up of several steps: First each component c_k of $\mathbf{c}(s,x,y)$ is quantized into one of b discrete values by a corresponding quantization function $q_k(c_k) : \mathbf{R} \rightarrow \{0, \dots, b-1\}$. This gives discrete automaton states $u_k = q_k(c_k)$, $k = 0, \dots, n-1$. In the second step a transition probability ${}_\kappa p_{u_k,u_{k+1}}(k)$ is computed for every pair c_k, c_{k+1} of subsequent feature vector components. Finally all transition probabilities are multiplied to get the probability of acceptance for $\mathbf{c}(s,x,y)$.

Transition probabilities ${}_\kappa p_{u_k,u_{k+1}}(k)$ are estimated from a sample $\Omega_\kappa = \{\mathbf{c}_0, \dots, \mathbf{c}_r\}$ of labeled feature vectors in the following way:

$$\begin{aligned}
\mathbf{c} &= (c_0, \dots, c_k, \dots, c_{n-1}) \in \Omega_\kappa \\
{}_{max}c_k &= max\{c_k; \mathbf{c} \in \Omega_\kappa,\ k = 0, \dots, n-1\} \\
{}_{min}c_k &= min\{c_k; \mathbf{c} \in \Omega_\kappa,\ k = 0, \dots, n-1\} \\
q_k(c_k) &= \left\lfloor \frac{c_k - {}_{min}\,c_k}{{}_{max}c_k - {}_{min}\,c_k} \cdot b \right\rfloor
\end{aligned}$$

$$\begin{aligned}
{}_\kappa a_{u_k,u_{k+1}}(k) &= |\{\mathbf{c} = (c_1, \dots, c_n) \in \Omega_\kappa; \\
&\quad q_k(c_k) = u_k \wedge \\
&\quad q_{k+1}(c_{k+1}) = u_{k+1}\}| \\
{}_\kappa o_{u_k}(k) &= \sum_{j=0}^{b-1} {}_\kappa a_{u_k j}(k) \\
{}_\kappa p_{u_k,u_{k+1}}(k) &= \frac{{}_\kappa a_{u_k,u_{k+1}}(k)}{{}_\kappa o_{u_k}(k)}
\end{aligned}$$

Both equations (1) and (2) can be implemented in a very simple manner: For equation (1) a one dimensional bin histogram is needed. This can be implemented by looking up probabilities $p(c_k|\Omega_\kappa)$ in precomputed one dimensional tables. For equation (2) two dimensional tables are neccessary. Here probabilities ${}_\kappa p_{u_k,u_{k+1}}(k)$ are taken from position (u_k, u_{k+1}) of a precomputed look up table. Generalisations to more dimensions are possible ([2]). Stochastic automata are also known as "markov chains" or "markov processes".

4 Wavelet Features

For the ability of decomposing images into multiple resolutions and for efficiency reasons wavelet features were chosen. A multiscale approach should help to detect pedestrians at varying distances to the observation camera. It has for example been shown by Fatemi-Ghomi ([3]) that wavelet features can be used for texture discrimination. In addition Randen ([5]) reported that texture segmentation with wavelet features yields only a slight degredation in performance when compared to other feature computation approaches.

An introduction to wavelets can be found in Burrus ([1]) and Mallat ([4]). In the following Mallat's notation will be used to explain how local feature vectors $\mathbf{c}(s,x,y)$ were computed by picking up wavelet coefficients across scales.

Let Φ be a one dimensional scaling function and Ψ the corresponding one dimensional wavelet. Two dimensional scaling functions and wavelets are constructed by $\Phi(x,y) = \Phi(x)\Phi(y),{}_1\Psi(x,y) = \Phi(x)\Psi(y),{}_2\Psi(x,y) = \Psi(x)\Phi(y)$ and ${}_3\Psi(x,y) = \Psi(x)\Psi(y)$. Operators A_{2^j} and ${}_iD_{2^j}$ project the input signal $s(x,y)$ into wavelet decomposition subspaces of scale 2^j:

$$\begin{aligned}
A_{2^j}s &= (\langle s(u,v), \\
&\quad \Phi(u - 2^{-j}n, v - 2^{-j}m)\rangle)_{n,m\in\mathbf{Z}} \\
{}_iD_{2^j}s &= (\langle s(u,v), \\
&\quad {}_i\Psi(u - 2^{-j}n, v - 2^{-j}m)\rangle)_{n,m\in\mathbf{Z}}, \\
&\quad i = 1, 2, 3
\end{aligned}$$

where $\langle .,. \rangle$ denotes the inner product of two functions. The computation of A_{2^j} and ${}_iD_{2^j}$ with a

QMF Filter pair is explained in Burrus ([1]) and Mallat ([4]). As a result of the operators a resolution hierarchy of wavelet coefficients is obtained. Its tree structure is shown in figure 3. Feature vectors were formed by picking up coefficients across scales in this tree:

$$\begin{aligned}
\mathbf{c}(s,x,y) &= (s(x,y), \\
&\quad s_1(\lfloor\frac{x}{2}\rfloor,\lfloor\frac{y}{2}\rfloor), s_2(\lfloor\frac{x}{2}\rfloor,\lfloor\frac{y}{2}\rfloor), \\
&\quad s_3(\lfloor\frac{x}{2}\rfloor,\lfloor\frac{y}{2}\rfloor), s_4(\lfloor\frac{x}{2}\rfloor,\lfloor\frac{y}{2}\rfloor), \\
&\quad s_5(\lfloor\frac{x}{2^2}\rfloor,\lfloor\frac{y}{2^2}\rfloor), \ldots, s_8(\lfloor\frac{x}{2^2}\rfloor,\lfloor\frac{y}{2^2}\rfloor) \\
&\quad \ldots, \\
&\quad s_o(\lfloor\frac{x}{2^{t(o)}}\rfloor,\lfloor\frac{y}{2^{t(o)}}\rfloor)) \\
&= (c_k)_{k=0,\ldots,n-1}.
\end{aligned}$$

$t(o)$ says which operator $A_{2^{t(.)}}$ and ${}_iD_{2^{t(.)}}$ has been applied to the input signal $s(x,y)$. o is the maximal decomposition depth.

5 Results

To verify our approach a video stream of 8000 images was sampled at $5Hz$ and digitized to grayscale images of 352×288 pixels. In this stream 140 000 centroids of pedestrians head's were labeled manually. To do this efficiently a GUI was designed and implemented using JAVA. The first 4000 images were used for training purposes while the other half of the images served as a test set. Around each head centroid a bounding box of the head area was approximated with a rectangle. The size of the rectangle was chosen proportional to the image's scanline. This was convenient approximation of pedestrian's distance to the observation camera.

Recognition rates $\theta_0, \theta_1, \theta$ for a videostream S consisting of images $s \in S$ were computed by

$$\begin{aligned}
\theta_0 &= \frac{\sum_{s\in S} cbg(s)}{\sum_{s\in S} cbg(s) + ibg(s)} \\
\theta_1 &= \frac{\sum_{s\in S} cfg(s)}{\sum_{s\in S} cfg(s) + ifg(s)} \\
\theta &= \frac{\sum_{s\in S} cfg(s) + cbg(s)}{\sum_{s\in S} cfg(s) + cbg(s) + ifg(s) + ibg(s)}
\end{aligned}$$

where $cbg(s)$ counted the number of correctly found background pixels of class Ω_0 in image s and $ibg(s)$ the number of incorrectly classified background pixels of class Ω_0. $cfg(s)$ and $ifg(s)$ are the corresponding measures for class Ω_1, i.e. foreground pixels. In addition θ_{hit} shows the percentage of head regions for which at least one pixel from the head's bounding box area was classified correctly.

From tables 1 and 2 we see that the markov chain based approach of equation (2) outperforms the histogram based approach of equation (1) because higher foreground recognition rates θ_1 can be achieved when stochastic automata are used. We concluded to use the HAAR-Wavelets for future experiments since it yielded a good compromise between recognition rates and computational efficiency.

Our implementation takes $1s$ on a *Sun Ultra Sparc III* with $296Mhz$ to classify one image. Future work will have to concentrate on increasing background recognition rates θ_0. Current rates are not practical yet. Figure 2 shows that there are still too many false classification results in background class Ω_0.

References

[1] C. Burrus, R. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms.* Prentice Hall Publishers, New Jersey, 1998.

[2] A. Colmenarez and T. S. Huang. Face detection with information-based maximum discrimination. In *CVPR 1997*, Puerto Rico, June 17-19.

[3] N. Fatemi-Ghomi. *Performance measures for Wavelet-based Segmentation Algorithms.* PhD thesis, Centre for Vision, Speech and Signal processing, School of Electronic Engineering, Information technology and Mathematics, University of Surrey, 1997.

[4] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[5] T. Randen. *Filter and Filter Bank Design For Image Texture Recognition.* PhD thesis, Norwegian University of Science and Technology, Stavanger College, 1997.

[6] C. Regazzoni and A. Tesei. Distributed data fusion for real-time crowding estimation. *Signal Processing*, (53):47–63, 1996.

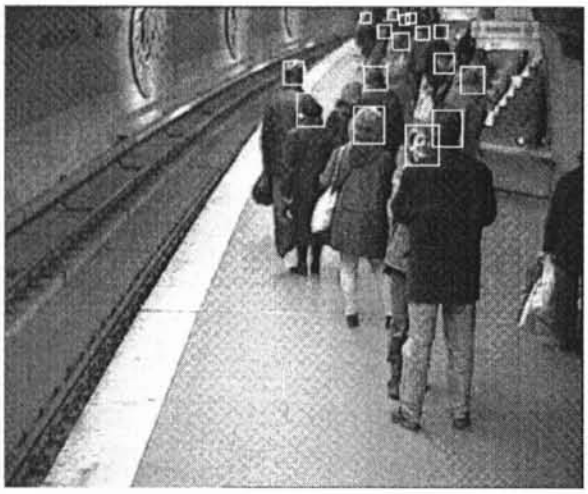

Figure 1: manually segmented input image

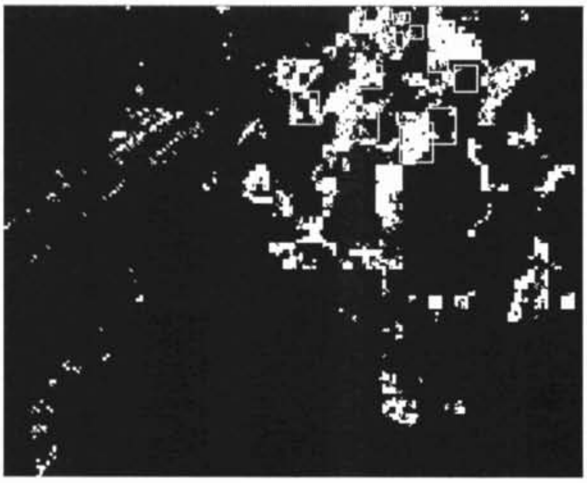

Figure 2: A classification result of the markov chain based approach,$b = 50$,$t(o) = 5$

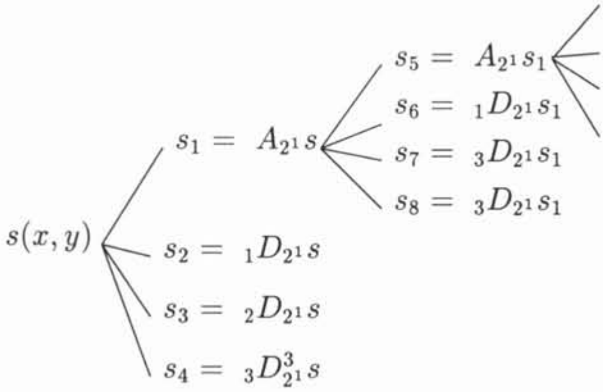

Figure 3: Waveletdecomposition of an input signal $s(x, y)$

	Test			
Wavelet	$\theta_1$in%	$\theta_0$in%	θin%	θ_{hit}in%
Haar	67.1	81.7	80.9	99.9
Daub,4-Tap	**69.3**	83.8	83	99.8
Daub,6-Tap	66.7	85.1	84.1	99.7
Daub,8-Tap	66.1	84.6	83.5	99
Coif,6-Tap	65.6	83.3	82.2	99.9
John,8-Tap	64.1	82.9	81.8	99.9
John,12-Tap	65	82.9	81.9	99.8
Andrew,10-Tap	65.6	82.2	81.3	99.9
Andrew,8-Tap	62.5	85.2	83.9	99.8
Smithb,8-Tap	60.6	83.8	82.5	99.7
Gop,8-Tap	56.1	**86.5**	**84.8**	98.9
Zhu,4-Tap	61.9	81	79.9	99.7
Simmon.,9-Tap	63.9	83.6	82.6	99.2
Daub,9-Tap	65.2	82.6	81.6	99.1
Villase.,9-Tap	65.2	82.6	81.6	99.1
Zhu,5-Tap	62.2	82.4	81.2	99.3

Table 1: Wavelet based recognition of heads with bin histograms, $b = 50$, $t(o) = 5$

	Test			
Wavelet	$\theta_1$in%	$\theta_0$in%	θin%	θ_{hit}in%
Haar	71.6	81.7	81.1	99.8
Daub,4-Tap	**74.4**	82.2	81.8	99.8
Daub,6-Tap	70.9	83.5	82.8	99.7
Daub,8-Tap	68.3	84.2	83.3	99.6
Coif,6-Tap	72.2	82.2	81.7	99.6
John,8-Tap	67.6	83	82.1	99.7
John,12-Tap	67	83.2	82.3	99.7
Andrew,10-Tap	68.2	82.7	81.9	99.8
Andrew,8-Tap	67.4	84.2	83.2	99.6
Smithb,8-Tap	62.7	83.8	82.6	99.7
Gop,8-Tap	60	**86**	**84.5**	99.5
Zhu,4-Tap	65.3	82.4	81.4	99.5
Simmon.,9-Tap	62.8	84.6	83.4	99.6
Daub,9-Tap	64.5	84	82.9	99.6
Villase.,9-Tap	64.5	84	82.9	99.6
Zhu,5-Tap	64.2	83.4	82.2	99.5

Table 2: Wavelet based recognition of heads with markov chains, $b = 50$, $t(o) = 5$