Walking Pedestrian Recognition

Cristóbal Curio, Johann Edelbrunner, Thomas Kalinke, Christos Tzomakas, and Werner von Seelen

Institut für Neuroinformatik,
Lehrstuhl für Theoretische Biologie
Ruhr-Universität Bochum, 44780 Bochum, Germany

Contact Author: Cristóbal Curio
Address: Ruhr-Universität Bochum
Institut für Neuroinformatik
44780 Bochum
Germany
Telephone: +49-234-322-4231
Fax: +49-234-321-4209
Email: Cristobal.Curio@neuroinformatik.ruhr-uni-bochum.de
I. Introduction

In recent research many approaches for recognizing human shapes in indoor and outdoor scenes have been developed. Different sensors have been used, such as thermal sensors and CCD cameras. Although thermal sensors can yield accurate results, they defunct as soon as the environment’s temperature reaches a certain level so that objects cannot be reliably separated from the background. In vision-based approaches human movements can be detected by the subtraction of subsequent image frames as long as the camera does not move and the lighting conditions change slowly. In the case of a moving observer, the ego-motion implies additional motion in the background making the detection of independent motion a nontrivial task. In general, there are no common features for the recognition of pedestrians. In [1] a large number of features has been used to build up a neural classifier. Still, the number of features is too high to ensure fast computation and detection has mainly been restricted to rear and front views of pedestrians. In another approach [2] a time-delay neural network was trained to recognize pedestrians from spatio-temporal receptive fields. In that approach the initial detection is based on a stereo camera system. Stereo vision is useful for short and middle distance ranges. The classification is done by analyzing the raw intensity values where background texture influences the signal to noise ratio. Another system has been introduced which relies on an initial stereo segmentation step and further neural network processing [3] aiming to detect pedestrians in static images in the near field. Another stereo and intensity based system to detect pedestrians and basic actions in static background scenes is described in [4].

Here, a new method for walking pedestrian recognition is presented. It is characterized by a successive processing at various levels of description and their integration yields the final recognition. The method is restricted to the detection of pedestrians that cross the road (see Figure 1).

The system will be described in the following sections. A system overview can be found in Figure 2. The system consists of a general hypothesis generation part in which different kind of cues are dynamically integrated. This initial detection process will be described in more detail in section II. For additional temporal analysis, tracking algorithms being described more detailed in section IV will be applied to the generated object hypothesis. In sections V and VI a final decision about the objects being investigated using a walking model (refer section III) is described. In this recognition process, all processed cues are arranged for a consistent interpretation of the object hypothesis leading to a final classification.

II. Initial detection

The initial detection of pedestrians is performed by combining three information cues: at first the local image entropy (texture information) [5] is calculated and secondly models are matched based on contour information [6]. Additionally, to increase the detection performance for the short distance field inverse perspective mapping (binocular vision) is applied. Therefore, besides the monocular camera a binocular camera system observes the short distance field in front of the car. The feature of local image texture (local image structure) focuses the attention of succeeding processing on some distinct regions in the image. This ensures the restriction in model translation space. The scale space is limited by a model-based method, a template matching, using the camera geometry to adjust the scale.
of the models (herein the walk step) as shown in [7]. The template matching is performed in a multi-resolution approach (e.g. quad-tree partitioning, see also [8]). Different phases of the walking sequence are matched to given contour features using the Hausdorff-Distance [9] as a measure of similarity. A new feature integrated in the search algorithm compared to the one described in [7] is the adaptive threshold and the fractions of matched features, respectively) according to the distance to the observer. Since the perspective geometry yields a linear function for the object scale in relation to the vertical axis in the image plane, the matching parameters are adapted to the object scale, and thus to the distance of the object in the real world. Therefore, a higher tolerance can be achieved for near distances, and a more accurate match at far distances in parallel. For the contour model matching process the Hausdorff-Distance copes optimally with the quantization effect of the walking model, thus intermediate phases are also detectable by the model. The Hausdorff-Distance \( H(P, Q) \) for two point sets \( P \) and \( Q \) is

\[
H(P, Q) = \max(h(P, Q), h(Q, P))
\]

with

\[
h(P, Q) = \max_{p \in P} \min_{q \in Q} \|p - q\|,
\]

where \( h(P, Q) \) is called Forward-Distance and \( h(Q, P) \) describes the Backward-Distance. Looking now at a subset of ordered minimal distances

\[
h_K(P, Q) = K^{th}_{p \in P} \min_{q \in Q} \|p - q\|
\]

with \( 1 \leq K \leq N \), \( N \): cardinality of \( Q \), the matching process, using the partial Hausdorff-Distance as a measure of similarity, is more robust against noise.

The sensor module for obstacle detection in the short distance field is based on the ‘Inverse Perspective Mapping’ (IPM) [10]. Here the image of the right camera is mapped onto the image of the left camera under the assumption that the vehicle moves on a horizontal plane and the parameters of the cameras are known. A difference image is calculated and the pixels which exceed a certain threshold represent an obstacle in the field of view. Texture, contour matching, and IPM information are fed into a temporal dynamic activation field in which a final decision about a reliable region of interest is made by using an activation threshold.

To integrate initial cues of different pre-processing algorithms, in this case texture, contour matching and IPM, that we figured out to be most appropriate to solve the detection problem (refer [11]), we developed a temporal dynamic activation field (DAF). This field is based on the image coordinate system and the image resolution. It mainly represents an occurrence probability distribution \( S(x, y) \in \mathbb{R} \) of object hypothesis. A dynamic integration over time ensures reliable results.

It should be noted, that the IPM algorithm is very sensitive to bumpings of the vehicle, since it is based on a fixed set of geometric acquisition parameters being temporarily violated. We assume that these disturbances, occurring on a relatively short time scale, can be filtered out by the dynamic activation field which kind of smoothes the integrated signal over time. Also, the IPM has been extended recently to model the effect of non-flat roads based on measurements of road lanes [12]. The effect of bumping has to be investigated more closely in our integration framework not only with respect to the IPM sensor module but also to all other sensor modules. It would be desirable to have an integration mechanism which can track and integrate all the modules information consistently in a way, so that bumping is not any more noise to the system but can also be detected.

Every detected image point induces an activation value in the DAF. The strength of image structure (texture), the contour matching values (model) and the reliability for the image area of a detected obstacle (IPM) are fed into the DAF. Additionally, the information of the road can be integrated as a pre-shape to sensitize this area of the DAF, i.e. to bias it. Especially, the flexibility to add new information cues is a great advantage of this method. Furthermore, the information of the object tracking module, described in section IV, can be incorporated to stabilize the results.

The elements of the DAF are activated by \( V_u(x, y) = \{ V_u(x, y) \in \mathbb{R} \mid 0 \leq V_u(x, y) \leq V_{\text{max}} \} \) provided by each input \( u \). This dynamics ensures an adaption over time. An exponential decrease of the activation is demanded and the updating of the field is done synchronously to the frame rate. The convergence and the choice of parameters bound the activation. Under the assumption that a point \((x, y)\) in the DAF will be activated for \( t \) frames by the same value \( V_{\text{max}} \) the maximum of activation \( S^\infty(x, y) \) is calculated by

\[
S^t(x, y) = (\ldots ((V_{\text{max}} + V_{\text{max}}) + V_{\text{max}}) \ldots) \hspace{1cm} (1)
\]
\[
\gamma(i) = V_{max} \sum_{i=0}^{t-1} \gamma^i \\
= V_{max} \frac{1 - \gamma^t}{1 - \gamma}.
\]

So for \( t \to \infty \) the limit of
\[
S^\infty(x, y) = \frac{V_{max}}{1 - \gamma}
\]
is reached. In the other extreme, if a point in the DAF is activated only for one frame, the activation at this point will be \( V_{max} \gamma^t = V_{max} e^{t \ln \gamma} \) after \( t \) frames later. The value decreases exponentially with time. One question of interest is after how many time steps \( t_{1/2} \) the activated will be half of the starting value:
\[
\Rightarrow t_{1/2} = \frac{\ln \left( \frac{1}{2} \right)}{\ln \gamma}
\]

To determine the parameters \( \gamma \) and \( V_{max} \) we develop the values vice versa: The value \( t_{1/2} \) and the limit \( S^\infty(x, y) \) are chosen and the parameters result in
\[
\gamma = e^{\frac{\ln \left( \frac{1}{2} \right)}{t_{1/2}}} \quad \text{and} \quad V_{max} = (1 - \gamma)S^\infty(x, y).
\]

One should keep in mind that the occurrence of the objects at the same point in the image is very unlikely so that the maximum of the DAF will be notably smaller than \( S^\infty(x, y) \). Measurements \( V_u \) resulting from different information cues can be integrated flexibly by the DAF in the following manner:
\[
V = \sum_{u=1}^{U} \lambda_u V_u,
\]

where \( u = \{ u \in \mathbb{N}^+ \mid 1 \leq u \leq U \} \), \( U \in \mathbb{N}^+ \) and \( \lambda_u = \{ \lambda_u \in \mathbb{R}^+ \mid 0 \leq \lambda_u \leq 1 \} \) with \( \sum_{u=1}^{U} \lambda_u = 1 \). In this application \( V_1 \) is provided by the strength of local image structure, \( V_2 \) by the matching values of models and \( V_3 \) by the reliability of obstacle detection of the IPM. So far the weights \( \lambda_u \) have been determined heuristically. We think of an optimization of the system with respect to these parameters in our future research.

Figure 3 and 4 shows the results of the template matching, texture calculation and IPM processing using the temporal dynamic activation field for the initial detection of pedestrians.

III. A MODEL FOR HUMAN WALKING

For the detection of pedestrians in urban sceneries, characteristic cues are needed that distinguish human shapes from other objects. A typical property of pedestrians is walking. The periodical relative limb movement is a feature that does not apply to rigid obstacles. Since the detection of independent movement is an ill-posed problem [13], we transfer the problem of motion detection solely to the detection of certain phases (or gaits) of the human walking process according to a synthetic human walking model. We restrict the detection of pedestrians to the lower part of their body (i.e., hip and legs), since the torso of the body typically shows high variance in its appearance (e.g., due to the arm’s movement). Furthermore, only the outline of the lower part of the body is used for matching, because the inner contour cannot always be detected (e.g., for women wearing skirts). By this way, we achieve an increase in the time efficiency, since fewer model-features are used and a greater flexibility of the model is obtained. Thus, the model resembles approximately a down-faced V-form in different deformations corresponding to the various gaits. This walking model is described as follows.

A lot of studies on human walking can be found in biomechanical papers. Studying the human kinematics (see also [14], [15]) linear models can be used for the human limbs motion. Herein, the motion of the thighs is modeled as a pendulum with a frequency \( f_w \), while the calves swing with the double frequency \( 2f_w \). In contrast to common
Fig. 4. Examples of initial pedestrian detection with the temporal dynamic field. Images show regions extracted using entropy, model contour and stereo cues. In the upper two images the DAF result is mapped into the left image of the binocular camera system. At the bottom, two images show the results of the DAF mapped in the monocular camera image.

approaches, our model generating human walking is not based on a mechanical process governed by Newton’s law; it is purely mathematical. The parameters of this model are the lengths of the thigh $l_1$ and the calf $l_2$, the maximum angle $\alpha_1$ between the thighs, and the minimum angle $\alpha_2$ between the thigh and the calf. Additionally, the hip and knee angles must satisfy certain constraints. For example, the maximum of the angle $\alpha_2$ between the thigh and the calf must not be greater than 180 degrees. The velocity $v_w$ of a walker is described by the step lengths $l_w$ and the frequency $f_w$: $v_w = l_w \cdot f_w$. By increasing the velocity, people prefer to increase the step length instead of the frequency [16]. In our model the frequency is fixed at $f_w \approx 2Hz$. This frequency is in good agreement with the walking of crossing pedestrians. Also, all other parameters of this model are tuned according to the anatomy and kinematics of an average sized human body. Typical phases of a gait are generated by sampling the synthetic kinematic model equidistantly in time, as shown in figure 5. From these discrete phases of the human walking model, only a subset of the gaits shown in Figure 5 where the legs do not occlude each other (d-l) are used for the detection of pedestrians in order to reduce the false alarm rates.

IV. TRACKING OF OBJECT HYPOTHESIS

The goal of tracking the torso of the body is to eliminate the effect of ego-motion on the observed scenery and to guarantee a qualitatively good assessment of relative limb motion within the walking pedestrian. Therefore, both the induced torso motion and the observer motion have to be decoupled from the relative limb motion.

To track objects whose characteristic motion appearance is to be analyzed as described in the next section, various tracking algorithms have been successfully employed, e.g. [9], [17], and [18]. A similar system approach exploiting the periodic motion cue is suggested in [19]. In that work the generation of initial object hypothesis rely on an independent motion cue assuming planar background conditions and the objects of interest are small compared to the background. Afterwards, tracked regions are searched for periodic motion pattern with the help of a self similarity measurement over time. In urban environments we are dealing with 3D scenes with changing or mixed motion models of the environment which makes the independent motion detection a difficult task. An overview of methods and a suggestion to make the independent motion cue in complex 3D environments more feasible is suggested in [20].

For the application herein, the dynamical verification of the initial hypothesis from a moving vehicle requires robust tracking. The algorithms show satisfying robustness in keeping track of objects in front of moving backgrounds due to the moving observer. All of them are data driven. Basically, they find those regions in successive frames showing similar feature distributions as in previous frames. To obtain high robustness they are adaptive by dynamically updating the tracked feature sets to suppress background noise and taking slightly changing appearances of the objects into account. In [17] and [9] tracking is based on contour feature sets whose similarity is effectively measured with the Hausdorff-distance between successive frames. In [18] tracking is based on the evaluation of statistical texture cross-entropy measurements.

Also, an elastic graph matching algorithm applied here shows high robustness in the tracking process, and is briefly
described as follows. It is based on a stochastic search algorithm. To account for velocity and model scaling changes of the moving object and to reduce the search space between successive frames a Kalman filter with an underlying simple accelerated movement model is used for prediction of the new search region.

A discrete grid with a fixed resolution captures an object, based on the gray scale value information on the corresponding nodes. The search of the object in the next frame (object scaling, horizontal and vertical translation) is performed by independent Gaussian distributed stochastic variables. This search strategy is a special case of simulated annealing [21]. Here it is assumed that small deviations from the Kalman-predicted state variables are more likely to occur than large ones. The similarity measurement between model and the actual image data is based on an Euclidean metric. The displacement and scaling at which the Euclidean distance is smaller than a threshold determines the new position and new scaling of the object for the next tracking step. The gray scale values on the nodes in the actual frame are used to find the new position and scale of the object in the following frame.

V. Motion cues

As an additional cue motion of the relative limb movement is used. The characteristic sequence of movement patterns evolving from the relative limb movement between realigned images enables one to distinguish walking pedestrians from other tracked parts in the image not showing this characteristic motion pattern. The filter process is described by the following two equations,

\[
A_t = (1 - w) I_t + w A_{t-1}
\]

and

\[
\Delta_t = I_t - A_{t-1}.
\]

At time \(t\), \(A_t \in \mathbb{R}\) describes the temporarily with some exponential memory factor \(w \in \mathbb{R}\) \((0 \leq w \leq 1)\) weighted incoming gray scale images \(I_t\), which are locally smoothed. Whereas the motion energy distribution \(\Delta_t \in \mathbb{R}\) is used for further processing. Those areas with high motion disturbance \(\Delta_t\) correspond to large changes in gray scale information and therefore are highly related to phases of a walking cycle with large limb movement. For illustration see Figure 6.

As can be seen from Figure 6, the simulation intends to describe only the relative limb motion. The generated motion pattern is free of noise compared to the real world scene, where imperfect realignment of the tracked body’s main torso and artificially introduced background movements obscures slightly the motion data. In few recent works, e.g. [22] and [19], local periodicity measurements of space-time activity distributions have been performed for gesture recognition. Relative motion is used here as a hint to support or reject an object hypothesis created by the initial detection described in above sections. The detection in this work is mainly based on a model data base of a walking cycle (see Figure 5) with spread legs, whereas motion significantly carries information in phases of large movements for crossing legs (see figure 5). Figure 6 shows the evolution of the projection of the motion distribution to the ground over time by integrating the motion energy column wise. In the upper image pair the temporal evolution of the periodic pattern for a walking pedestrian in real world is shown, and in the lower image pair for the simulated case.

In both figures the dark areas correspond to the “head” of the direction of a limb and the light areas to the “tail”, while the background, corresponding to no motion, is ren-
It should be noted that pedestrians moving in various directions from other viewpoints also show periodic motion pattern, but they look differently. The amplitude of the maximum walking step reduces while the pedestrian turns away from the pure crossing direction. Interesting work has been done to classify walking gaits and in determining roughly walking directions by using prototypical motion models [23]. Obviously, it’s hard to detect typical motion patterns from walking pedestrians perpendicular to the image plane.

**VI. Recognition**

As described in sections III to V the detection of pedestrians is twofold. On the one hand a person is detected by its shape (see section III). On the other hand the periodic motion of the legs is typical to detect walking persons (see section V). For a final classification a combination of both is necessary to be most reliable in a large spectrum of different scenarios.

For the classification, matching values have been calculated by summing up the activation of the temporal dynamic field determined by model matching and texture analysis in the tracked area of the legs. Maxima of this curve indicate that the leg position reliably belongs to the models of Figure 5. Additionally, the motion features for the classification are determined by integrating the amount of motion shown in Figure 6 in three overlapping windows (see figure 7). In detail (see Figure 7), in the left window the motion energy of the front leg, in the right window the motion energy of the back leg, and in the middle window the motion energy of the area below the torso is measured by summing up. As shown in Figure 7 all three values show a periodic behavior over time for the ideal model, but due to background noise and uncertainties for real world data in determining the optimal integration space only the strong motion activation in the middle window is reliable. Comparing the both motion images of Figure 6 the typical motion behavior of the human walking gait is only guaranteed in the white frames, the middle window of Figure 7.

Figure 8 shows that the assumption of periodic change of a reliable model matching process (see Chapter II) and a motion detection process holds. We used the term performance to express the complementary contribution of the model matches and the motion information to the DAF at each instance of time. At this point of the project it is not wise to provide any data on detection and false alarm rates, since we just want to introduce the idea of a new flexible system architecture for detecting crossing pedestrians in urban environments out of a moving car. A computationally fast cross-correlation between one period $T_p \approx 14$ frames of model motion data and a window of the same size of real motion data of Figure 9 is done. In order to provide unambiguous results a whole period of model motion data is necessary.

In Figure 9 the cross-correlation is plotted for all possible time shifts. Generally, by determining the maxima of the correlation curve a recognition of motion periods is performed. In the worst case this takes about two periods to ensure stable results. In a similar manner a detection of the periodic contour-model matching values is done to stabilize the performance of the whole classifier and to shorten the number of time steps necessary for classification. In the long distance image starting from figure 3 the pedestrian is detected, tracked, and classified over time as shown in Figure 10.
Fig. 9. The solid line shows the cross-correlation values between an average motion model and the corresponding real motion (dashed line) over time. The vertical lines indicate the detected maxima of correlation. After 19 frames (approximately one second) the pedestrian has been classified reliably.

Fig. 10. Example of a scene during a typical recognition process. A person is initially detected, tracked and classified.

VII. SUMMARY AND OUTLOOK

We have presented a new architecture for pedestrian recognition in urban environment with moving observer. The initial detection is based on the integration of texture information, template matching, and the IPM. Texture information reduces the search space only to structured regions. The camera geometry supports the estimation of object scales in the image. A synthetic human walking model was generated for the localization of pedestrians. The template matching is based on the Hausdorff-Distance. The IPM ensures the detection of obstacles in the short distance field where the model cannot sufficiently cope with variations of human shapes. Furthermore, motion cues are extracted by utilizing robust trackers that aim the verification of the hypotheses. Additionally, the motion cues are used in combination with the contour model-features for a complete recognition in the DAF. After an initial detection (see dynamic feature integration in Figure 11) is accomplished, the complementary nature of the techniques described herein along with the robustness of the proposed trackers are able to finally recognize a pedestrian.
motion pattern recurrent neural networks will be applied for further analysis in the future.

**Acknowledgments**

This work was partially supported by the German BMBF (VESUV 18S811B9).

**References**


In Figure 12 the motion energy of the tracked regions rendered in Figure 10 is shown. On the right side of each image the cross-correlation values of motion integration and model matching (see Figure 8 and Figure 9) is represented by the left and right bar, respectively. The bar on the left side indicates the classification result. If the bar exceeds a certain threshold a pedestrian is recognized.

To improve the classification process of the dynamic limb motion pattern recurrent neural networks will be applied for further analysis in the future.

**Acknowledgments**

This work was partially supported by the German BMBF (VESUV 18S811B9).

**References**


**Cristóbal Curio** received his M.S. degree in Electrical Engineering and Information Sciences, Ruhr-Universität Bochum, Germany, in 1998. He is currently a Ph.D student at the Institute of Neuroinformatik, Bochum, Germany, in the Pattern Recognition and Scene Analysis group. From 1995 to 1996 he has been an exchange student at Purdue University, Electrical Engineering and Computer Sciences, Indiana. His research interests include biomedical engineering, ultrasound, object recognition and scene analysis.

**Johann Edelbrunner** received the M.S. degree in physics from the University of Graz, Austria, in 1997. He is currently a Ph.D student at the Institute of Neuroinformatik, Ruhr-Universität Bochum. His research interests include image processing, object recognition and statistical signal processing.

**Thomas Kalinke** graduated (Diplom-Ingenieur) in 1993 and received his doctor degree in 1999 at the Department of Electrical Engineering and Information Sciences, Ruhr-Universität Bochum, Germany. During 1993, Thomas Kalinke was an Erasmus student at the TIMC-IMAG laboratory the University Joseph-Fourier of Grenoble, France. Since 1994 he works at the Institute of Neuroinformatik in the Pattern Recognition and Scene Analysis group as a Research Scientist.

**Christos Tzomakas** graduated in 1993 at the Computer and Engineering and Informatics Department of the University of Patras, Greece. He received his doctorate degree in 1999 at the Department of Electrical Engineering and Information Sciences, Ruhr-Universität Bochum, Germany. During 1993, Christos Tzomakas was an Erasmus student, from 1995 to 1998, he was a DAAD student and in 1998 and 1999, he was working as a Research Scientist at the Institut of Neuroinformatik, Ruhr-Universität Bochum, Germany.

**Werner von Seelen** is the Head of Department of Theoretical Biology within the Institute of Neuroinformatik at the Ruhr-Universität Bochum and Director of the Centre for Neuroinformatik. He graduated (Diplom-Ingenieur) and received his doctorate degree in the Faculty of Electrical Engineering at the Technische Universität Hannover. Professor von Seelen joined the Fraunhofer Institute and headed a group for Biological Information Processing. After a two year period with the Batelle Institute (nonlinear optics) he took up a position as Professor at the Department of Biophysics at the Universität Mainz. Since 1989 he has been Chairman for Theoretical Biology at the Ruhr-Universität Bochum. His current research focuses on System Theory, Brain Research and a neural architecture for autonomous systems. Professor von Seelen has written numerous research articles and was the editor/co-editor of several books. He is a member of the editorial board of Neural Networks, was President of the German Society of Cybernetics (three years) and is a member of several scientific societies. In 1995, he was awarded the “Kupfmüller Ring” and in 1998 the “Karl-Heinz Beckurts” prize for his outstanding scientific contributions in the field of interdisciplinary research.