A Multiple Detector Approach to Low-resolution FIR Pedestrian Recognition

Mirko Mählisch, Matthias Oberländer, Otto Löhlein, Dariu Gavrila and Werner Ritter
Dept. REI/AI, DaimlerChrysler AG, 89081 Ulm, Germany
[uni-ulm.maehlisch, matthias.oberlaender, otto.loehlein, dariu.gavrila, werner.r.ritter]@daimlerchrysler.com

***Abstract*— In this paper we present a recognition scheme which is both reliable and fast. The scheme comprises the simultaneous harmonized use of three powerful detection algorithms, the hyper permutation network (HPN), a hierarchical contour matching (HCM) algorithm and a cascaded classifier approach. Each algorithm is evaluated separately and afterwards, based on the evaluation results, the fusion of the detection results is performed by a particle filter approach.**

I. Introduction

A. Motivation

Driver assistance systems supporting the driver at night are of increasing interest in the car market. The first generation of non-warning night vision systems has been already introduced. Current research focuses on the second generation of night vision systems integrating a warning function. An indispensable feature of the second generation systems is the capability of reliably detecting vulnerable road users (VRUs), like pedestrians up to distances of 100m (330ft). Human and animal thermal radiation has its peak within the far-infrared (FIR, $\lambda = 6-15\mu m$) wave band, thus obstacles can act as radiation emitters and no illumination is necessary. In this contribution we used a microbolometer technology sensor, sensitive between 7 and 14 micrometers with spatial, temporal and spectral resolution of 164x129px @ 25Hz, 14bit. Automotive pedestrian detection from FIR

Fig. 1. preprocessed FIR-images

data is a challenge in many ways (Fig. 1). The segmentation process has to cope with moving objects in front of an itself moving background, eliminating any background subtraction strategies. Additionally, no analytical representation of pedestrian shape and/or texture can be found, as i.e. in case of traffic sign detection, where the objects to be segmented are of circular or triangular shape and are provided with a well known color distribution. Pedestrian movement causes shape inconsistency over time, too. Texture inconsistency comes from different clothing and is - especially in winter time - strongly dependent on how long the human already stayed outdoor.

This research was supported by NIRWARN, BMBF 01M3157B.

B. Related work

A.Broggi and T.Graf et. al. solved the problem of pedestrian recognition in FIR images using multiresolution texture symmetry, edge symmetry and edge density ROI extraction, together with a texture and shape correlation validation step, based on shaded 3-D pedestrian models [1]. Liu and Fujimura's strategy applies intensity thresholding, followed by a motion constraint computed from stereo data and aspect ratio/size discrimination [2]. In [3], F.Xu and K.Fujimura used intensity thresholds followed by a combination of support-vector-machine classification and Kalman-Filtering. H.Nanda and L.Davis introduced a probabilistic template matching on hot-spot ROIs [4]. The system of Tsuji et. al. consists of hot-spot analysis, stereo verification and ego-motion compensation [5]. We ourselves focused on FIR pixel classification so far [6]. An overview on detection features for infrared data is given in [7]. For a general overview of recent pedestrian systems see [8].

C. Paper structure

After the first introductory section, including related work and motivation, section two shortly introduces currently promising pedestrian segmentation methods, partially originating from visible wavelength image processing. Section three deals with our way to evaluate detector performance and presents the evaluation results for each method of its own, pointing out its strengths and weaknesses. Based on these results, in section four we introduce a novel detection system for FIR-data, with object list level output, performing the detection task in realtime on a 3 GHz Pentium IV. We close with an overview on problems still unsolved and suggest further development steps.

II. Pedestrian recognition methods

From examinations we know that schemes relying on intensity, uniformity or symmetry assumptions fail to reliably detect VRUs (compare fig.1 left and right). This is why our approach completely operates on pattern correlation and trained pattern classification. For the task of pedestrian

recognition we use a combination of the three powerful recognition methods, *Hierarchical Countour Matching*, *Cascade Classification* and *Hyperpermutation Networks*.

A. Chamfer Contour Matching

The Chamfer Contour Matching is a template correlation method, based on the object feature shape [9]. It detects pedestrians by comparing a huge database of possible pedestrian-silhouettes with subregions of the camera data's edge image. To perform this correlation, the distance-transform (DT) is applied to the edge-image first, using the popular Chamfer-distance. Afterwards, the database templates can be easily correlated with the edge image by laying them over the DT and averaging all distance-values of the DT below each template pixel. This average - the mean distance between silhouette and edge-image - is a good measure for similarity and can be thresholded for pedestrian detection. Here we use the hierarchical template-matching method introduced in [9] to overcome the extreme computational cost for brute-force correlating each of the templates on every image position.

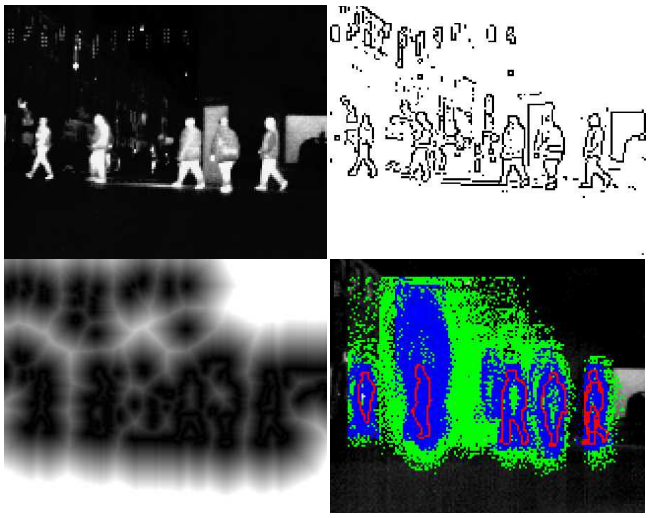

Fig. 2. brief overview of chamfer matching: upper left - preprocessed image, upper right - edge image, lower left - distance transform, lower right - three layer hierarchical template matching

B. Cascade Classification

The cascade detector, introduced by Viola and Jones [10] and modified by Wender and Loehlein [11], is a subwindow-classifier. To classify a subwindow, the detector uses a cascade of classifiers, successively increasing in computational cost. The cascade aims to reject as many subwindows as possible in the early computationally inexpensive stages and to classify only the remaining very low percentage with the computationally expensive ones. Subwindows passing all stages become detections (fig.3/left). Each classifier within a cascade is a scalar linear combination of a number of so called weak classifiers, build upon simple rectangular features, quickly computed from integral images. The decision about class membership in each stage is made by thresholding the linear combination value. The number of features increases from stage to stage, thus implying better discrimination at simultaneously raising computational cost. The count, type and positions of the features, the linear combination weights and the final decision thresholds within each stage are learned from pre-classified random sample images with the adaboost algorithm.

C. The Hyperpermutation Network

The Hyperpermutation Network (HPN) [12] generates a discrete confidence level for each pixel, indicating weather it probably belongs to class "background" or to class "object". Thresholding these levels leads to pixel classification. Low probability outputs discriminate background areas, high ones object areas and values in the middle can be interpreted as network indifference. To transform the original input image into the per pixel likelyhood image (Fig. 3/right), several HPN stages are successively computed on the image, meaning each stage processes the output image of the previous one. To compute the output value of each pixel within a stage, the HPN uses information of its local environment by scanning a random pattern of pixel values around it. This pattern was chosen randomly for each stage during training with growing spatial extent from stage to stage. This architecture implies feature extraction over small local image areas in the lower stages and inference on class membership, considering these features in larger image regions in the upper stages. The binary transformation function from the pattern pixel values to the target pixel value is done by lookup-tables, whose optimization is the network training task, described in detail in [13].

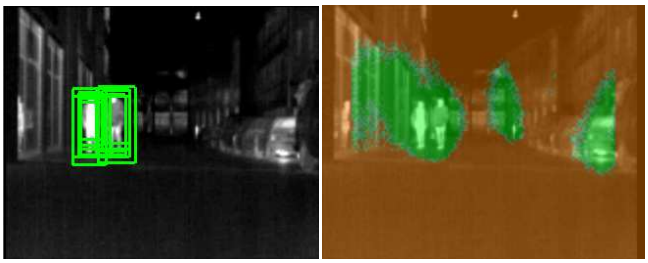

Fig. 3. left: subwindows passing the cascade classifier, right: HPN likelyhood image

III. Evaluation of Recognition Methods

A. Evaluation parameters

Corresponding to the shape templates' bounding boxes in chamfer template matching, correlating a template with an image position can be interpreted as classifying a subwindow, too. Therefore all presented methods can be evaluated with the basics of classifier evaluation. The membership of pattern P, representing rectangular subwindows for chamfer matching and cascade classification and pixels for HPN likelyhood estimation, to class "object" (K) or class "background" ($\bar{K}$) is determined by human pre-classification. The

classification algorithm "Cl" to evaluate, presents a class estimation $Cl(P)$, too. We distinguish four cases (see table I). Value "a" counts the number of true positives, hence pi-

TABLE I
CONFUSION MATRIX

	$P \in K$	$P \in \bar{K}$
$Cl(P) = K$	a	b
$Cl(P) = \bar{K}$	c	(d)

xels or subwindows belonging to a pedestrian and classified as pedestrian. Value "b" counts false positives, and value "c" false negatives, respectively. At least for subwindow classification, we can not use value "d", the true negative count, because the number of subwindows containing no pedestrian is almost infinite. Separating case "a" and "b" in subwindow classification additionally requires a "match operator", since we cannot expect pixel-accurate detections from the detectors. Tolerating some spatial imprecision up to a certain level was accomplished by defining a distance measure on two subwindows (B_1, B_2) upon the ratio of the area within the conjunction of the two boxes and the area of their disjunction:

$$cov(B_1, B_2) = \frac{A(B_1 \cap B_2)}{A(B_1 \cup B_2)} \in [0,1] \qquad (1)$$

This coverage measure can be transformed to a distance measure via $d(B_1, B_2) = cov(B_1, B_2)^{-1} - 1 \in [0, \infty]$. Thresholding this distance (or the coverage) leads to a binary scale invariant decision between case "a" and "b". Upon the confusion matrix entries, created by comparing human pre-classification and detector output, one can define the evaluation performance parameters. The sensitivity s (or detection rate) is an estimation of the probability, a real world object gets detected by the algorithm and is optimal at value one: $s = a(a+b)^{-1} \in [0,1]$. The precision p estimates the probability that a detection resulted from a real world object and not from clutter and is optimal at value one, too: $p = a(a+c)^{-1} \in [0,1]$. Another often used parameter in subwindow classification is the false alarm rate f per image, normalizing the false alarm count "b" on the image count "N", that is equivalent to the number of false alarms per time (transformation via measurement frequency): $f = b/N \in [0, \infty]$. In pixel classification the true negative number "d" is known from the image resolution, thus the false positive rate can be defined as $f = b(b+d)^{-1} \in [0,1]$. Assuming the number of real world objects $a+c$ and the number of detections $a+b$ grow proportional with time, the parameters p and f carry the same information - the quantity of false alarms.

Honest detector evaluation depends on the simultaneous presentation of both, detection probability and false alarm quantity, expressed by a vector function $(s, p)^T = f(\vec{x})$ on the varying algorithm parameter vector $\vec{x}$.

B. Evaluation results

We took our data set from urban as well as country scenes in the autumn and winter time at night and day. The data set was divided into learn and test set as usual. For evaluation set pedestrian and image quantities see table II. Additionally

TABLE II
EVALUATION SET QUANTITIES

	images	pedestrians
learn set	3853	4688
test set	1589	1703
total	5442	6391

to varying the algorithm parameters while generating the performance measures, we varied the imprecision tolerance for subwindow classification, giving us an idea of the localization performance. Besides generating an overall measure, we discretized the pedestrian heights and calculated separate measures for each height interval, providing us information of how well the algorithm will detect objects in different distances w.r.t. the mounted optics. From figures 4/5 we learn that cascade detection is more robust against object downscaling than chamfer matching. The fact that feature shape looses discrimination the smaller the objects become, due to raster data representation, explains this behavior. Additionally the architecture of the weak learners in cascade detection allows simultaneous processing of both, shape and texture features leading to more precise discrimination of small objects. As a consequence of these results, we apply chamfer template matching only within short distances, by removing the smaller shape templates. On the other hand, in the height intervals where chamfer matching is not penalized by image raster effects, it does much better localization than cascade classification. This is observable from nearly constant good performances at low imprecision tolerance, where cascade detection performance breaks down, e.g. because subwindows are normalized to a fixed size in cascade classification, disobeying the true pedestrian aspect ratio. To transform the performance statements of figures 4/5 into detection ranges we use the camera projection equations. Figure 7 shows the obstacle distances in world coordinates for each discrete object height step in image coordinates we used to evaluate the detectors. If tolerating low localization precision, we are able to detect pedestrians up to 75 meters. HPN pixel classification output was compared to naive intensity thresholding, by varying the HPN probability threshold and the intensity threshold. Sensitivity over precision (Fig. 6/top) shows that HPN pixel classification is far away from optimal discrimination, but in fact much better than intensity thresholding. However, in contrast to the other approaches the image content independent computational cost and the general execution speed, resulting from its arithmetic free architecture, are advantages of HPN classification. Although the high probability pixel clouds diffusely disperse around real pedestrians (fig. 3/right), corrupting any direct detection method, like

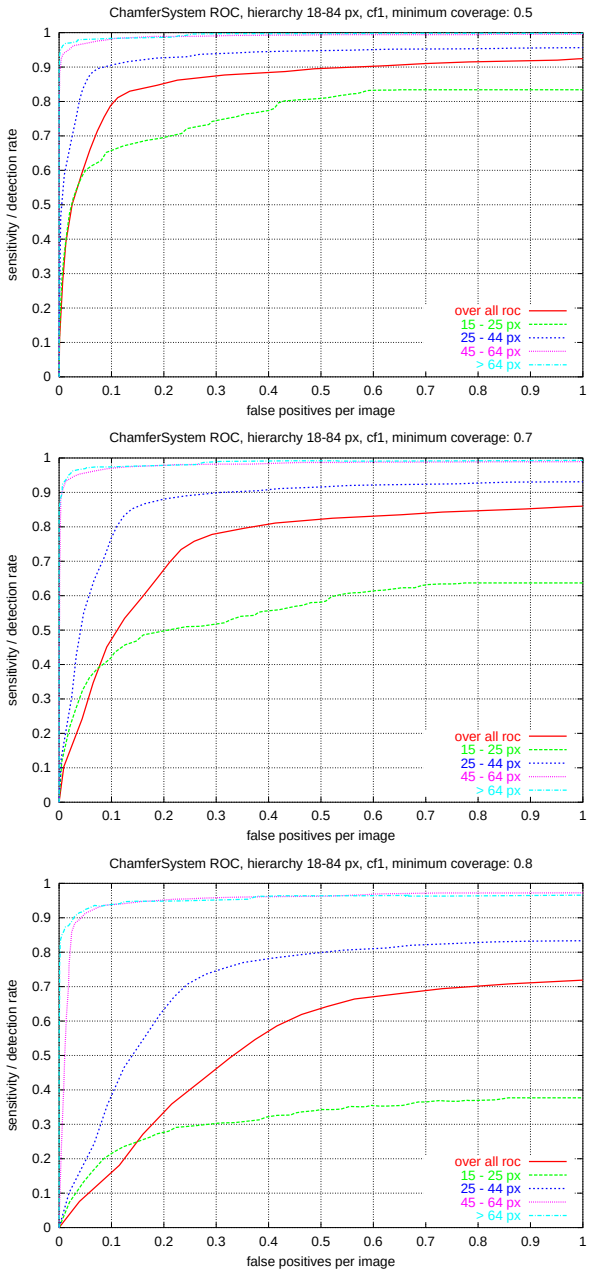

Fig. 4. left column: ROC detection performance of hierarchical contour matching with top-down increasing hit accuracy demand, broken down into different object heights

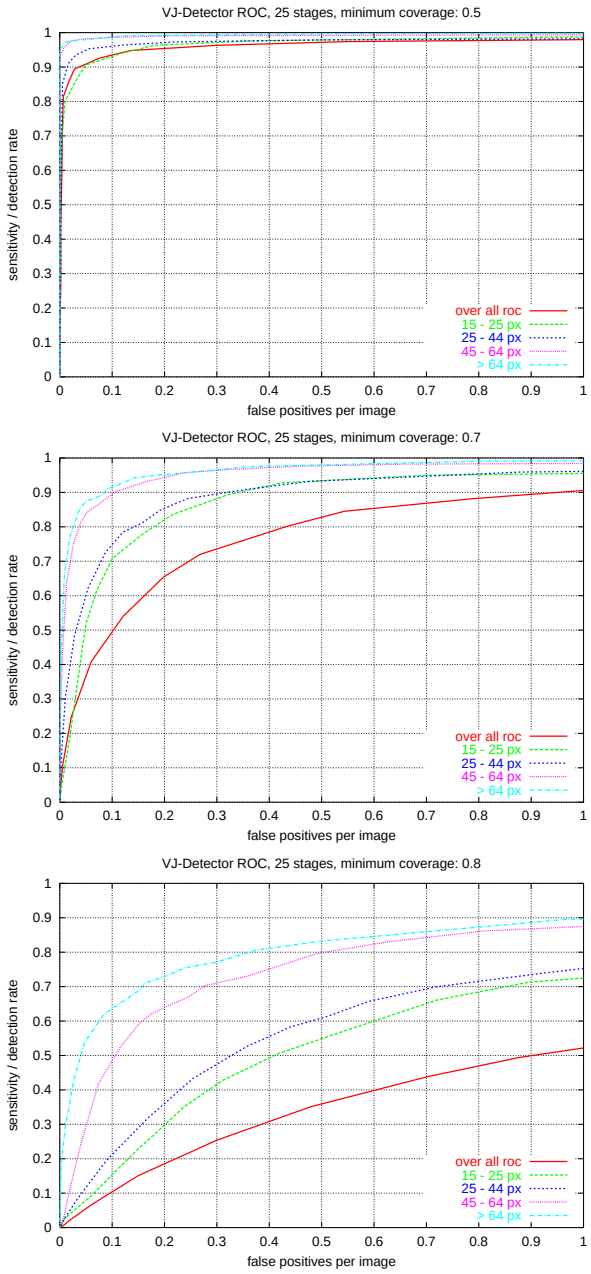

Fig. 5. right column: ROC detection performance of cascade subwindow classification with top-down increasing hit accuracy demand, broken down into different object heights

i.e. binary connected components clustering [14], we learn from the 90%-sensitivity kneepoints in fig. 6, that we can operate the HPN with an average true negative rate of approx. 96%. This observation motivates a new possibility of HPN usage within a detection system, described in the next section.

IV. Realtime pedestrian detection

A. Detector Combination

The chamfer template matching method, as well as the cascade detector, cannot be applied in real-time, if every possible image position has to be correlated or classified. That is why ROI-Extraction, as fast method for background skipping, comes into play. The common solution for this problem is the ”flat-world-assumption” (FWA, fig. 8), meaning the camera is looking down on a planar world and, additionally, every object in the world has to be standing on that plane. Together with the standard camera transformation from world to image coordinates, this constellation implies a restricted search area in the image plane for each object, depending on its height. In practice this assumption holds only for short distances because of the unknown

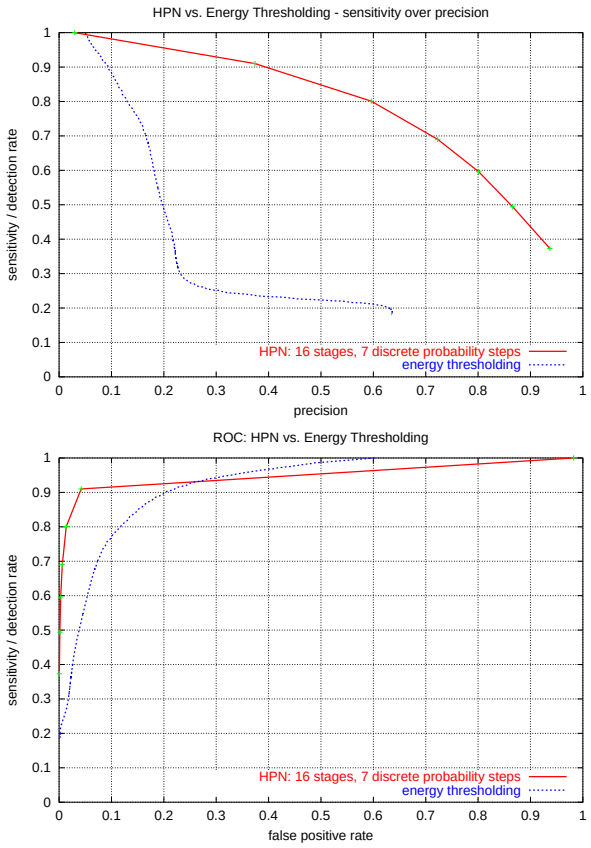

Fig. 6. HPN (red) vs. intensity threshold (blue); top: sensitivity over precision; bottom: sensitivity over false alarm rate

territory. Additionally, a real car is shaking on the road, due to unevenness and road holes. We solved this problem by combining a relaxed FWA and the HPN likelihood-image. At first, the object search tunnel, provided by the FWA is enlarged vertically, to compensate the territory uncertainty and the vehicle shaking. The increase in computational cost, linked to this relaxation, is compensated by skipping each image region that passes the relaxed FWA from being applied to chamfer matching and cascade classification, if the HPN indicates a high probability for background within its rectangular area. To compute the mean HPN activity for a rectangular image region we used the integral image method (eq. 3). The HPN integral image II_{HPN} is computed once per frame (eq. 2).

$$II_{HPN}(x,y) = \sum_{i=0}^{x}\sum_{j=0}^{y} HPN(i,j) \tag{2}$$

$$\begin{aligned} a_{HPN}(x_0,y_0,x_1,y_1) =& (II_{HPN}(x_1,y_1) - II_{HPN}(x_0,y_1) \\ & - II_{HPN}(x_1,y_0) + II_{HPN}(x_0,y_0))/ \\ & ((x_1-x_0)\cdot(y_1-y_0)) \end{aligned} \tag{3}$$

A threshold operation on the mean HPN activity performs the necessary binary decision. The chamfer template matching and the cascade detector are both applied to the remaining search areas. The bounding boxes of the matching contour-templates and the positively classified subwindows are combined into the detection box cloud.

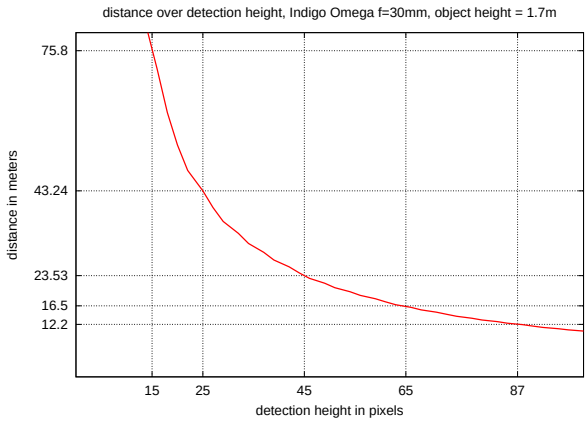

Fig. 7. detection ranges for f=30mm optics, assuming a pedestrian size of 1.70m

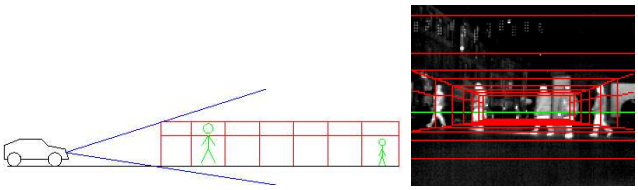

Fig. 8. Flat-World-Assumption

B. Detection Filtering

Often multiple Chamfer matching and cascade detection boxes spread around real pedestrians (i.e. fig. 3/left). Besides cleaning the measurements from this noise, the filtering component compensates short detector malfunctions and is capable of associating detections of different images to the same real world object, providing information for high level situation analysis. We used bayesian estimation on a state vector, consisting of position and size of the pedestrian bounding box in image coordinates. As mentioned, detector fusion is accomplished by sequential (or cooperative) fusion of a HPN-ROI stage with a box detector stage, built from a parallel fusion of chamfer template matching and cascade classification. Multiple object filtering was solved by a standard multiple filter approach. The oncoming data association problem was solved by usage of a threshold operation on the already introduced box distance measure (eq. 1). For state prediction we used the second order motion model

$$\begin{pmatrix} \vec{x}_{t+1} \\ \vec{x}_t \end{pmatrix} = \begin{pmatrix} A & B \\ I & 0 \end{pmatrix} \begin{pmatrix} \vec{x}_t \\ \vec{x}_{t-1} \end{pmatrix} + \begin{pmatrix} \vec{w}_t \\ 0 \end{pmatrix} \tag{4}$$

$$A = \left[1 + \frac{t_{+1} - t}{t - t_{-1}}\right] \cdot I \tag{5}$$

$$B = -\frac{t_{+1} - t}{t - t_{-1}} \cdot I \tag{6}$$

where t_{+1}, t, t_{-1} are three successive measurement points in time. Since we did not identify the detectors measurement noise and the pedestrian movement process noise to be gaussian so far, we used the particle filter approach [15] to solve the state estimation problem. The measurement likelyhood function is defined on the set of associated detection boxes Z with help of eq.1:

$$p(Z|x) \propto \frac{|Z|}{\sum_{i=0}^{|Z|-1} d(z_i, x)^2} \tag{7}$$

V. Conclusion

The joint operation of parameter optimized methods as fast background skipping by pixel classification, followed by the more discriminating subwindow classification and a filtering algorithm, enabled us to design a reliable realtime pedestrian detection system (fig.9).

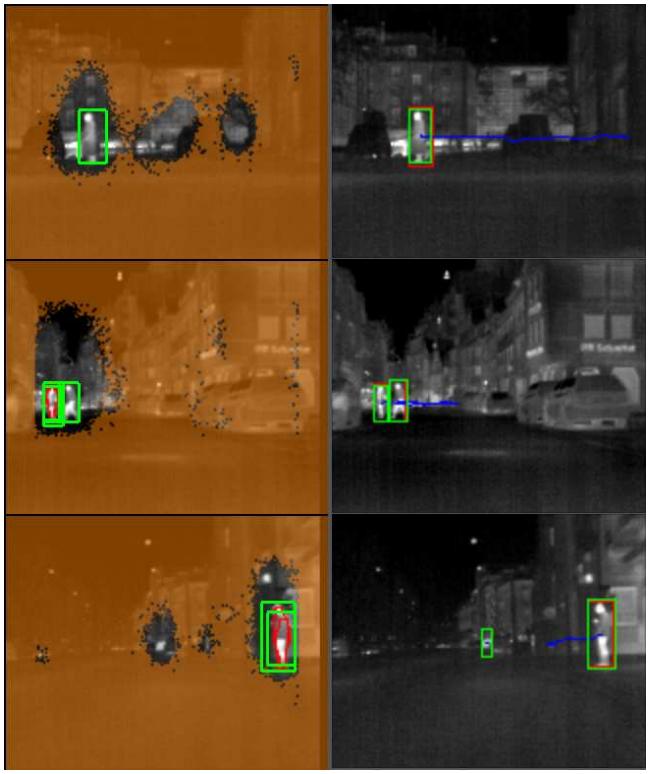

Fig. 9. pedestrian detection and filtering; left: orange - HPN ROI extraction, green - cascade detections, red: Chamfer matching detections; right: red - predicted state, green: corrected state, blue: trajectory

We do not go so far as to claim our system detects pedestrians at 100% accuracy (see roc-diagrams in section three), but we showed how strong evidence can be derived from even very low resolution image data, where simple assumption-based methods fail. To drive the object detection certainty against its maximum, we will examine two possible ways in the future: First, our detection algorithms are subject of permanent development, increasing their discrimination performance. Especially interesting is the question of various detector output correlation, rating the benefit of parallel detector fusion, that was not examined so far. Therefore we assumed some degree of independence of the detector outputs empirically observed from mutual exclusive true positives of the two subwindow approaches. Second, we start intensive investigations on sensor fusion systems, i.e. combining FIR-data with NIR or LIDAR data. Another problem during our tests emerged from the preprocessing stage. FIR sensor data seems to be strongly dependent on daylight, weather and seasonal conditions. This can be solved by either advanced preprocessing algorithms, presenting nearly constant images to the other stages, w.r.t. contrast and brightness, or by training specialized detectors for different environmental conditions. In the future we aim to quantitatively evaluate the performance of the whole detection system, including data fusion and tracking, but this requires object identity labeling of the random sample images and a track-based evaluation scheme.

References

[1] A.Broggi, A.Fascioli, M.Carletti, T.Graf, and M.Meinecke, "A multi-resolution approach for infared vision-based pedestrian detection," in *IEEE Intelligent Vehicles Symposium 2004*, Parma, Italy, 2004.

[2] X.Liu and K.Fujimura, "Pedestrian detection using stereo night vision," in *IEEE Transactions on Vehicular Technology*, November 2004, pp. 1657–1665.

[3] F.Xu and K.Fujimura, "Pedestrian detection and tracking with night vision," in *IEEE Intelligent Vehicle Symposium*, Versailles, France, Juni 2002.

[4] H.Nanda and L.Davis, "Probabilistic template based pedestrian detection in infrared videos," in *IEEE Intelligent Vehicle Symposium*, Versailles, France, Juni 2002.

[5] T.Tsuji, H.Hattori, M. Watanabe, and N. Nagaoka, "Development of night vision system," in *IEEE Intelligent Vehicle Symposium*, Tokyo, Japan, 2001.

[6] U. Meis, M. Oberlaender, and W. Ritter, "Reinforcing the reliability of pedestrian detection in far-infrared sensing," in *IEEE Intelligent Vehicle Symposium*, Parma, Italy, 2004.

[7] Y.Fang, K.Yamada, Y.Ninomiya, B.Horn, and I.Masaki, "Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection," in *Procs. IEEE Intelligent Vehicles Symposium 2003*, Columbus, USA, June 2003, pp. 505–510.

[8] D.M.Gavrila, "Sensor-based pedestrian protection," in *IEEE Intelligent Systems, vol.16, nr.6*, 2001, pp. 77–81.

[9] D.M.Gavrila and V.Philomin, "Real-time object detection for "smart" vehicles," in *Proceedings of IEEE International Conference on Computer Vision*, 1999, pp. 87–93.

[10] P. Viola and M. Jones, "Robust real-time object detection," in *Second international Workshop on statistical and computational Theories of Vision-Modeling, Learning, Computing, and Sampling*, Vancouver, Canada, July 13 2001.

[11] S.Wender and O.Loehlein, "A cascade detector approach applied to vehicle occupant monitoring with an omnidirectional camera," in *IEEE International Conference on Intelligent Vehicles*, June 2004, pp. 14–17.

[12] M. Oberlaender, "Hyperpermutation networks - a discrete approach to machined perception," in *Third Weightless Neuronal Networks Workshop*. University of York, 1999.

[13] T. Schwarz, "Optimierung Rückgekoppelter Hyperpermutationsnetzwerke," Ph.D. dissertation, Christian-Albrechts-Universität Kiel, 2000.

[14] U. Meis, H. Neumann, and W. Ritter, "Detection and classification of obstacles in night vision traffic scenes based on infrared imagery," in *IEEE International Conference on Intelligent Transportation Systems*, Shanghai, October 2003, pp. 1140–1145.

[15] B.Ristic, S.Arulampalam, and N.Gordon, *Beyond the Kalman Filter*. Artech House Publishers, 2004.