Model-based validation approaches and matching techniques for automotive vision based pedestrian detection

A. Broggi, A. Fascioli, P. Grisleri

Dip. Ingegneria dell'Informazione
Università di Parma
43100 Parma, Italy

T. Graf, M. Meinecke

Electronic Research
Volkswagen AG
Wolfsburg, D-38436, Germany

Abstract

Pedestrian detection is a challenging vision task, especially applied to the automotive field where the background changes as the vehicle moves. This paper presents an extensive study upon human body models and the techniques suitable for being used in a pedestrian detection system. Several different approaches for building model sets, such as synthetic, real, and dynamic sets are presented and discussed. Comparative results are reported with reference to a case study of a real system. Preliminary results of current research status are shown together with further developments.

1 Introduction

Pedestrian detection for automotive applications is currently one of the most hot and skilling research task in artificial vision. This kind of application is attractive for both manufacturers and users. The former are interested in selling high value-added products and the latter wish to buy safer vehicles. Several detection techniques and hardware systems have been investigated such as stereo [1], motion detection [2], and the use of far and near infrared sensors. Some of these approaches rely on the matching between features extracted from the image and features extracted from a single or a set of human models stored in a predefined or dynamically updated database.

1.1 Related works

In the following section some pedestrian detection systems are considered, each using different techniques. Being a very hot and new research topic, all these methods present situations in which their behavior is not satisfactory. Even the system we propose has some open issues that will be discussed in the concluding section.

[3] describes a system that detects and tracks pedestrians using a probabilistic template. The template is built in three different scales using a set of 1000 real images. The matching function is executed on three different probability maps and, after a threshold with a Bayesian classifier, local maxima are identified as pedestrians.

In [4] a pedestrian border template hierarchy is used in order to perform a coarse to fine detection. This hierarchy can automatically be constructed off-line from available example templates.

In [5] a two steps detection and tracking method is proposed: the detection is performed by selecting candidates searching for hot spots in the image and classifying them with a support vector machine, properly trained with a set of images.

A group of methods for detecting pedestrians use a shifting window of variable size: once the window has been extracted from the image, feature extractors operate on it using standard pattern classifiers for determining wether a feature is present or not. These types of detectors are distinguished by the classifier used. In [6] [7] Haar wavelet and PCA are used. In [2] each test window content must survive a cascade of classifiers, that first look for simple features and then for complex ones, thus reducing computational time. A pattern surviving to the whole classifier cascade is finally marked as being a pedestrian.

Other systems [8] combine the use of stereo vision with neural network classifiers or use stereopsys to obtain a V-disparity image [9] suitable for finding obstacle position. Stereo infrared systems have also been developed as explained in [10]. Hot spots are extracted from the images and a blob-level stereo triangulation is operated; optical flow is used to establish if an obstacle is stationary or moving and a final validation step is performed using an SVM pattern classifier.

1.2 Our contribute

The system presented in this paper is a three-step algorithm that operates on a single stream of images taken from a far infrared camera [11]. The algorithm uses considerations on

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

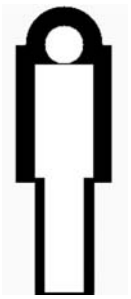

Figure 1: The simple model encoding the morphological characteristics of a pedestrian.

symmetries and vertical edges to extract all possible candidates to be classified as pedestrians in the analyzed image; then it filters out those who cannot be pedestrians due to their aspect ratio and particular shape features. The set of remaining candidates are matched to a set of models. Those that pass this test are classified as pedestrians and their feature such as distance and size are computed. The selection of models and the match itself are the focus of this study.

The approach to candidates validation is based on matching with shape models. This technique is promising since the extraction of the correct set of pedestrian candidates in an image using symmetries appear robust in most of the scenes.

This approach, as explained later in the paper, leads to appreciable results, especially from the false positives point of view: their value is significantly lower than that of other similar systems.

This paper deals with the different approaches that have been tested for validation, describes the current state of the research, and traces some lines for future developments showing preliminary results for new approaches currently under test.

Section 2 will describe the different investigated approaches: from the first static model, to the walking pedestrian model set, to dynamic models. Correlation techniques are also discussed in section 3. Section 4 shows the different methods used to evaluate the algorithms and correlation performance and section 5 concludes the work discussing current results and anticipating some guidelines for future work.

2 Model-based validation approaches

The low-level phase based on vertical edges and symmetry provides interesting regions (bounding boxes) which are likely to contain pedestrians. These regions are validated through a higher level step based on shape and/or thermal patterns to remove candidates that do not feature a human shape. This section will present different methods that have been tested to remove false positives.

2.1 Simple models

Initially, a very simple morphological filtering was developed to discriminate bounding boxes actually containing pedestrians from false positives. That filter was based on a pattern matching with a simple binary mask representing a rough human shape, shown in figure 1. The model was resized according to the size of the bounding box, and then matched to the grey level original image. The mean value and variance were computed in the hot (white) region representing the human body, and in the cold (black) region representing the background. The obtained values were compared to a threshold in order to rate the match. The aim of the filter was to eliminate the candidates that did not match the model or were not sufficiently hot.

This very simple morphological model adapts to most standing pedestrians, but shows its limits with walking pedestrians or lateral views of crossing pedestrians. Moreover, the uniform color of the model does not fit the various texture of real pedestrians.

2.2 Different postures and clothing characteristics

The morphological match phase was then improved using a grey-scale 3D model of the human shape (see figure 2) and using a better pattern-matching algorithm. This new model can represent different postures and attitudes of the human shape and can be computer generated from different points of view to achieve a better adaptability to real situations. The model was scaled to the bounding box size and over-

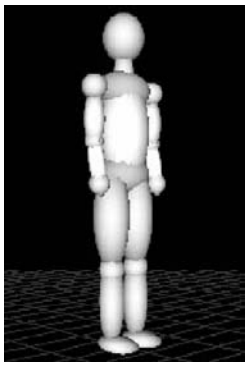

Figure 2: The 3D model encoding the morphological characteristics of a pedestrian.

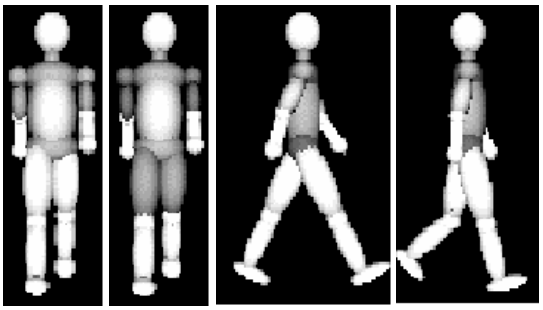

Figure 3: A few models representing different clothings, postures, and points of view.

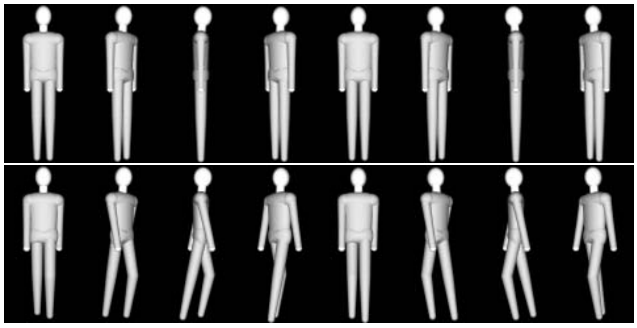

Figure 4: Examples of 8 points of view for a standing and walking pedestrian.

lapped to it using different displacements to cope with small errors in the localization of the box. The matching was implemented through a simple and fast cross-correlation function. The result was a percentage rating the quality of the match. A threshold was applied for the final evaluation.

The idea of generating the models at run-time and performing an exhaustive search for the best configuration was discarded, since it is time consuming and does not fit real-time criteria. A selection of pre-computed configurations was chosen.

The possibility to adapt the models to real images attributing different grey values to the body parts in order to encode different body temperatures was also considered and tested. In fact, generally head and hands are not covered by clothes and thus are warmer than the trunk or limbs both in winter and summer. Figure 3 shows some examples of models representing different clothings. Anyway, tests proved that these characteristics were difficult to match in real cases. Better results were obtained using a high number of different shapes without encoding thermal differences.

2.3 Use of a large model set

Most of the investigation was focused on using a high number of different shapes. Two degrees of freedom are sufficient to obtain a good match in most situations and are used to generate the complete matching set: posture and point of view. A third degree of freedom (size) is implicit in the matching process. A first set of 8 configurations obtained combining 4 points of view with 2 positions were initially tested but demonstrated to be not sufficiently reliable. A set of 72 configurations were finally chosen. They were obtained combining 8 different points of view with 9 positions (one standing and 8 walking). Figure 4 shows part of the 72 configurations. They were generated taking also into account the actual viewing angle, orientation, and height of the camera on the test vehicle. A new synthetic model, smoother than the previous one was generated to build the set, eliminating joints to be closer to real cases.

This filter proved to be effective in most cases both in the identification of pedestrians and in the exclusion of bounding boxes that do not contain humans. Anyway, the localization of pedestrians was difficult in some situations such as bikers, running people, or when the bounding box was not precise. In addition, false positives can be generated by tall and slim objects such as trees and poles (an example of false positive is displayed in figure 5). For this reason, further investigations on 3D models and the match function were performed.

2.4 Exploiting sensor characteristics

The 3D models used for the match are of paramount importance for a correct detection of pedestrians. The larger the model set, the higher the probability that a model representing each pedestrian to be validated is available. Unfortunately, only a reduced number of models can be used for matching since the match must be performed in real-time. Thus, a careful choice of the best set of models is mandatory. Some tests were performed using a different model set (see figure 6). A black border and white noise were added to the background in the image model. The underlying idea was to simulate

- the black halo around the pedestrian (this is a typical characteristic of the sensor when framing hot objects),
- and the noise present in the background.

Even if this idea seemed promising the results were not appreciably better than the previous ones. To improve results a different shaping of the background noise distribution (better than the uniform one) could be obtained measuring statistics in the border area of each bounding box under consideration. This idea is currently under test.

2.5 Fat models

The models used by the algorithm demonstrated to be appropriate in most cases, anyway they looked too slim and

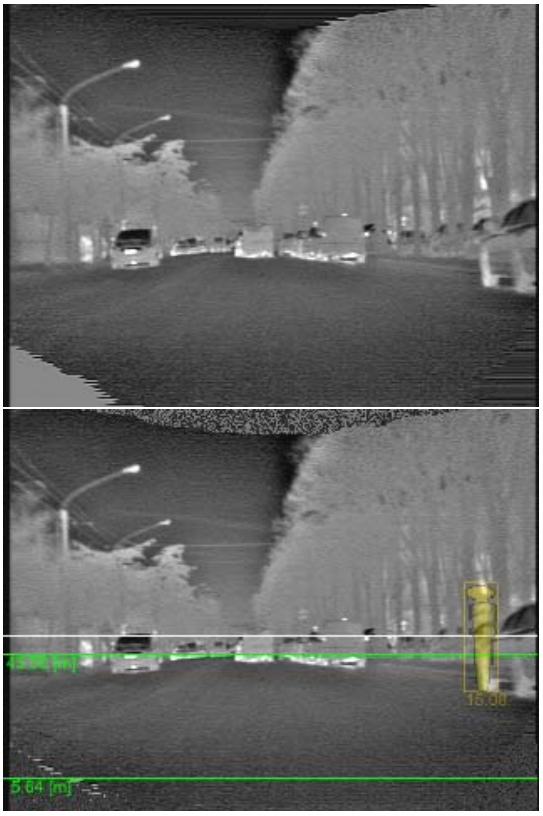

Figure 5: A false positive generated by a tree and a car. Top: original image. Bottom: image with the erroneously detected pedestrian.

stiff compared to real human shapes. Other models, more similar to a real human shape, were generated: the body and the limbs are fatter, the head is tilted forward for the lateral views, and the arms are bended. Moreover, a lighter color was used for the head and hands compared to the body (see figure 7). Some tests were made with these models as well: results improved but not substantially (the false positive number decreased of 0.03 per frame and the correct detection rate increased of 1%).

2.6 Real models

The collection and use of models taken from real images has also been considered. A set of real models taken from several pre-recorded video sequences has been manually generated. The set is representative of pedestrians with different clothes, posture, and positions. In this way the different heat emission of the body parts is also considered. The elements of the set have been selected considering also their frequency of match on image sequences different from the ones used to extract them (see section 4). Some examples are shown in figure 8.

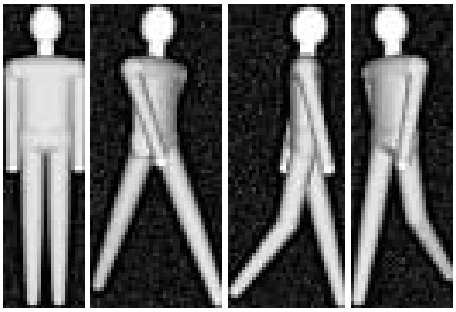

Figure 6: Models with noisy background.

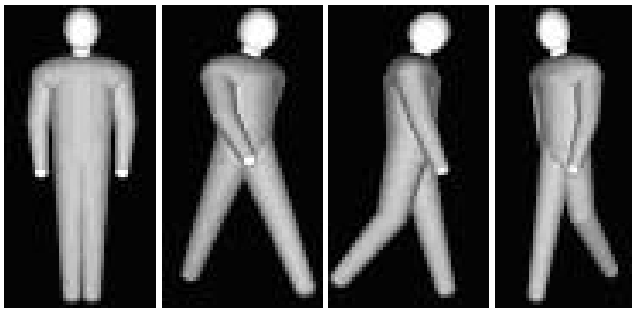

Figure 7: Fatter models increasing similarity to real cases.

Promising results have been obtained with a set of 50 real models, however it is reasonable to think that performance might be substantially improved by enlarging the set. Further investigations are under development.

In order to build a large set an automatic procedure would be desirable. Since the method used to measure performance is based on the match with ground truth data obtained by a manual annotation of pedestrians in each frame of a video sequence (see section 4.2), the set of real models could also be automatically generated using this information. The model set that is generated by this automatic tool contains an extremely large number of candidates, therefore the set needs to be reduced.

Figure 8: Some examples of real models.

Usually, in order to achieve good correlation results, background should be removed, this operation can be performed manually by a human operator or automatically. Unfortunately, automatic background removal is not trivial since background texture may be very complex. To simplify this process it is possible to extract pedestrian from ad-hoc image sequences acquired with a uniform background which can be easily removed by an automatic system.

2.7 Dynamic models

Once a bounding box has been validated as a pedestrian through the match with a model (synthetic or real), the box itself can be used as a model for the next frame implementing a sort of visual tracking. In fact, the appearance of a pedestrian is very similar in consecutive frames, therefore the cross-correlation is very high if the model to be compared to the pedestrian is represented by a shot of the same pedestrian taken a very short time before. In this case, a better match for the background is also found.

To implement this idea, a list of dynamic models is created and appended to the static one using the pedestrians validated in the current frame. The complete list of static plus dynamic models will be used for the next frame. Dynamic models have a one-frame lifetime and are matched first. In case a match is not found, static models are then used. Since dynamic models present a good correlation, a higher threshold is used for them.

The number of dynamic models is small, and they are very likely to be matched in the next frame as well. When this happens the match with the whole static model set is not performed, thus saving computational time. A performance increase of at least 25% has been obtained. However this value strongly depends on the used threshold.

Moreover, since a dynamic model derives from a pedestrian appearing in the previous frame, it makes sense to match it only to that candidate found in the current frame which can reasonably correspond to the same pedestrian. A criterion based on the position in the image gives poor results due to vehicle motion (an object can appear in very different positions of the image in two consecutive frames in case the vehicle is yawing). Conversely, dynamic models are considerate only for candidate bounding boxes not too different in size from them. This allows to save additional time. Obviously, the knowledge of data about the vehicle ego-motion could help in restricting the range of possible choices for the bounding boxes in the next frame. However, this would entail labelling and tracking objects not yet classified as pedestrians. Tracking objects at this stage presents some drawbacks: the number of objects may be high and, most important, since it is not confirmed that the object is a pedestrian, the tracking model could fail. For these reasons, the choice made was to perform tracking after the following validation step [12].

Anyway, the main disadvantage of the use of dynamic models is an appreciable increment of false positives. In fact, when an object is erroneously validated as a pedestrian, a dynamic model is created for it and it will probably be recognized in the next frames as well. This negative effect may be attenuated improving the robustness of the match with static models, but it is still an open problem.

Results computed on a test video sequence shown a +3% detection rate improvement against an increment of 0.03 false positives per frame.

2.8 3-blocks models

One of the approaches under evaluation is the use of models divided into 3 parts. In the image of a pedestrian three sections can be individuated: head, trunk and arms, and legs as shown in figure 9. The detection of head and legs is easier with respect to the detection of the trunk, due to their small area, low texture content, and very specific shape. On the other hand, trunk and arms may have a high variance of size and texture and a less defined shape. Thus computing the matching value with three different weights and weighting head and legs more than the body part leads to better results.

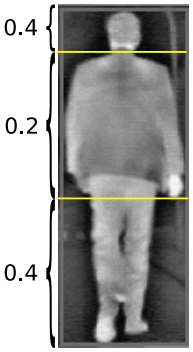

Figure 9: The three parts of the pedestrian image. Numbers represent the weight used in the correlation computation.

Images in the IR domain are characterized by the absence of texture. Thus, once the head has been found, it is not necessary to repeat this search for every model, and time can be saved in order to test more legs and body configurations. This approach is currently under investigation. Preliminary results on a test sequence seems to be promising showing a considerable increment of 5% in the detection rate versus a neglectable increase of 0.02 false positives per frame. However this technique still need further development.

3 Matching methods

In this section several considered approaches regarding the match function, are discussed considered. The first tested ideas were:

- matching contours instead of grey-level pixels (see figure 10),
- matching binary images (see figure 11).

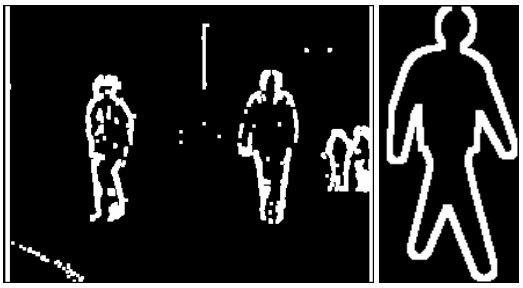

Figure 10: Matching contours: an original image and a model.

Figure 11: Matching binary images: an original image and a model.

In this case the algorithm correlates models and bounding boxes binarized according to the following threshold:

$$threshold = \frac{\sigma_1\sigma_2}{\sigma_1+\sigma_2} ln\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}$$

were μ is the average, σ the standard deviation, underscript 1 refers to the area inside the pedestrian (estimated using the model) and underscript 2 to the background. This correlation approach proved to produce good results where the difference between the pedestrian and the background brightness is high. Unfortunately, results are not satisfactory when pedestrians and background feature a similar grey level.

Other methods have been tried to match models and pedestrians in the boxes. An idea was to compute the variance in the box and use it to discriminate between background and pedestrians, which should have a higher variance. Figure 12 shows a false positive that this method is able to eliminate. However, this method does not work properly when pedestrians stand close to large hot objects, and also small hot objects can cause false positives.

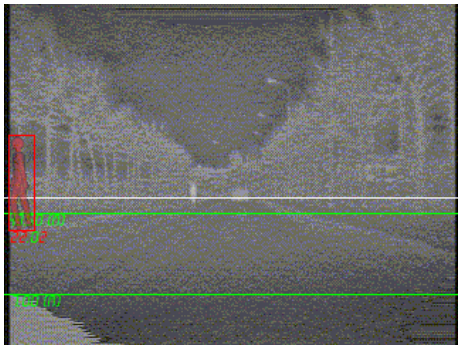

Figure 12: A false positive that can be eliminated using variance.

Another idea was to build the grey level histogram of the box and compare it to a reference histogram which characterizes a pedestrian. Unfortunately, the real signature of pedestrians varies according to the clothing. For example, figure 13 shows how a pedestrian wearing a jerkin may have a very dark trunk.

A further method investigated was to compute the distance between the average grey tone of the pedestrian area and the background one. The method, even if very simple, works quite well when the pedestrian is highly contrasted with respect to the background, however may fail in different cases.

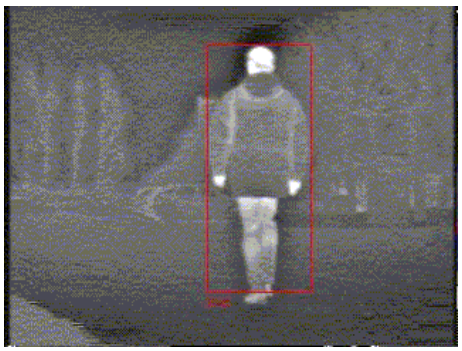

Figure 13: An example of clothing masking the heat emitted by the body.

A specific tool was developed to compare different matching algorithms, which allows to implement and test different correlation functions. The tool loads all models that should be tested, and the frame on which the test should be carried out. Models can be changed, made larger or smaller, and they can also be stretched to adapt to the pedestrians in the image. The main advantage of the tool is that it provides a visual output for the result of the match: the model can be superimposed on the frame and moved, while the tool dynamically updates an image containing the correlation between the model and the portion of the frame where it is positioned. Looking at this image makes it possible to understand which portions of the model have a good matching, and which ones do not. Moreover, every time the model is compared with the frame, a value describing how good is the whole match is computed. It is then represented in another window by a pixel whose position is the same as the central pixel of the model, and whose brightness depends on the computed value. This image then shows which position gives the best matching.

The matching function that demonstrated the better behavior is the cross-correlation between grey-level images.

4 *Performance*

4.1 Models usage statistics

A tool to evaluate matching statistics for each model was developed. A histogram is computed representing for each model the relative number of times it was selected as the best matching model. This allowed to analyze the matching distribution for synthetic models on a long test sequence (about 4100 images): as shown on the left part of figure 14 most models are seldom or never selected, while a small number of models are matched most of the times. This consideration suggests the set of model be reconsidered.

Statistics have also been measured for fatter models on the same test sequence. A better frequency distribution can be observed in this case (see the right part of figure 14).

4.2 ROC Curves

In order to evaluate algorithm performance, a tool, described in [13], has been developed. An operator manually annotates on a file the pedestrians position in each frame using a graphical interface studied for minimizing the number of operation to be performed to accomplish the task. The algorithm under test uses a library to write another file containing the processing results. The two files produced are then analyzed by a statistics computation tool that extracts the correct detection rate (CDR) and the number of false positives per frame (FPN).

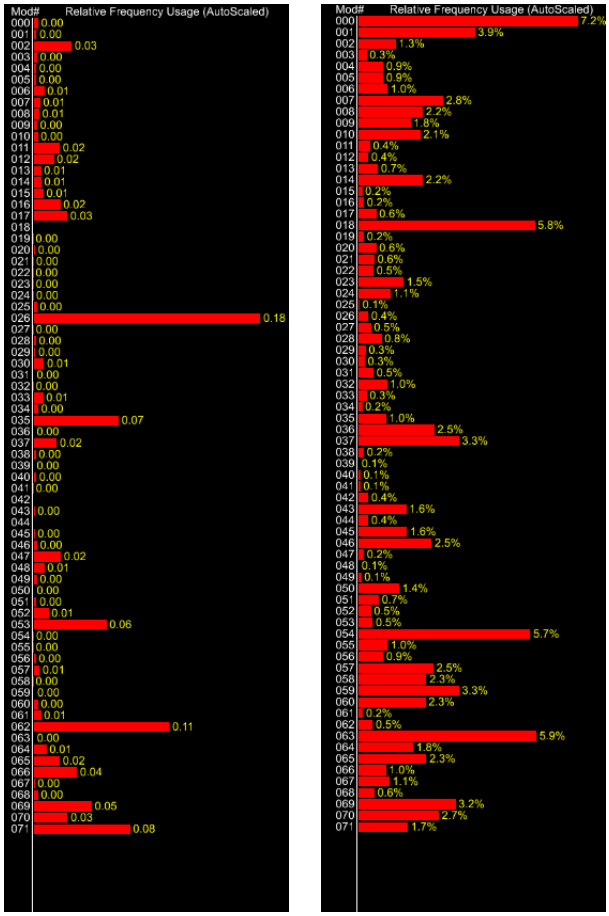

Figure 14: The usage statistics for the set of synthetic models: on the left the original model set, on the right the fatter models set.

Using this tool it is possible to compute the pair (CDR, FPN) for different values of the algorithm parameters in order to obtain ROC curves. This allows to fine tune the parameters and assess the performance of the system.

All the performance figures presented in the paper are expressed in relative terms with respect to the original algorithm performance measured on a 4111 frames test sequence framing a variety of different urban scenes. In fact, during statistics computation it has been observed that absolute results are heavily dependent on the selected sequences. A very long (30 minutes or more) sequence for which ground truth is available, is supposed to be necessary in order to supply consistent numbers to the reader.

5 Conclusions and future works

In this paper several techniques and approaches to human models suitable to be used in the validation phase of a real

pedestrian detection system have been presented. The usage of rough 2D and 3D models seems to be not satisfactory due to the inaccurate matching with bounding boxes containing real candidates. Lot of information is contained in the borders of the framed pedestrian, thus a model that takes in great account this feature, coupled with an appropriate matching function is better than a plain correlation on pixel values.

The usage of models for candidates validation has also shown some intrinsic problems:

- it is difficult to obtain an exhaustive model set that gives good results on very different scenes. This is especially true for real models extracted from specific sequences; results are promising, especially looking at models' frequency usage, but the set needs to be very large and thus time consuming and not suitable for real-time applications.
- the correlation function used to perform the match is the most critical choice, since it controls both the accuracy and the time spent for each matching test.

Another approach that is currently under test concerns the usage of models with additional information enclosed within. These information must also be provided by the algorithm in order to enforce the recognition process. Metadata such as the head, hands, or feet position, can be added to the model bitmap in order to obtain a more robust matching and a shorter matching computation time.

A further development that could be investigated is the join of rendered images with dynamic models. The images can be produced starting from a realistic 3D model of a pedestrian animated with a 3D-modeler.

References

[1] Xia Liu and Kikuo Fujimura. Pedestrian Detection using Stereo Night Vision. *IEEE Trans. on Vehicular Technology*, 53(6):1657–1665, November 2004. ISSN 0018-9545.

[2] Paul Viola, Michael J. Jones, and Daniel Snow. Detecting Pedestrians using Patterns of Motion and Appearance. In *Procs. IEEE Intl. Conf. on Computer Vision*, pages 734–741, Nice, France, September 2003.

[3] Harsh Nanda and Larry Davis. Probabilistic Template Based Pedestrian Detection in Infrared Videos. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[4] Dariu M. Gavrila and J. Geibel. Shape-Based Pedestrian Detection and Tracking. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[5] Fengliang Xu, Xia Liu, and Kikuo Fujimura. Pedestrian Detection and Tracking With Night Vision. *IEEE Trans. on Intelligent Transportation Systems*, 6(1):63–71, March 2005.

[6] Constantine Papageorgiou, Theodoros Evgeniou, and Tomaso Poggio. A Trainable Pedestrian Detection System. volume 38, pages 15–33, June 2000.

[7] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based Object Detection in Images by Components. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(4):349–361, April 2001.

[8] Liang Zhao and Charles Thorpe. Stereo and neural network-based pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 1(3):148–154, September 2000.

[9] Massimo Bertozzi, Alberto Broggi, Roland Chapuis, Frédéric Chausse, Alessandra Fascioli, and Amos Tibaldi. Shape-based pedestrian detection and localization. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2003*, pages 328–333, Shangai, China, October 2003.

[10] Xia Liu and Kikuo Fujimura. Pedestrian Detection using Stereo Night Vision. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2003*, pages 334–339, Shangai, China, October 2003.

[11] Massimo Bertozzi, Alberto Broggi, Alessandra Fascioli, Thorsten Graf, and Marc-Michael Meinecke. Pedestrian Detection for Driver Assistance Using Multiresolution Infrared Vision. *IEEE Trans. on Vehicular Technology*, 53(6):1666–1678, November 2004. ISSN 0018-9545.

[12] Emanuele Binelli, Alberto Broggi, Alessandra Fascioli, Stefano Ghidoni, Paolo Grisleri, Thorsten Graf, and Marc-Michael Meinecke. A Modular Tracking System for Far Infrared Pedestrian Recognition. In *Procs. IEEE Intelligent Vehicles Symposium 2005*, Las Vegas, USA, June 2005. In press.

[13] Massimo Bertozzi, Alberto Broggi, Paolo Grisleri, Amos Tibaldi, and Michael Del Rose. A Tool for Vision based Pedestrian Detection Performance Evaluation. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, pages 784–789, Parma, Italy, June 2004.