Pedestrian Detection in the Context of Multiple-Sensor Data Alignment for Far-Infrared and Stereo Vision Sensors

Raluca Brehar, Cristian Vancea, Tiberiu Mariţa, Ion Giosan, Sergiu Nedevschi
Technical University of Cluj-Napoca, Computer Science Department
Raluca.Brehar, Cristian.Vancea, Tiberiu.Marita, Ion.Giosan, Sergiu.Nedevschi@cs.utcluj.ro

***Abstract*—Multiple sensor systems are extremely used in autonomous driving for providing increased object detection accuracy. We present a multiple sensor based pedestrian detection system that combines aggregated channel features classifiers trained on images captured with two types of sensors: far infrared and stereovision sensors.**

We developed a spatio-temporal data alignment between the two sensorial systems. For the temporal alignment we used an original camera response timing model for free running cameras in order to align the infrared image with grayscale intensity images captured by trigger-based cameras. The spatial data alignment is done by performing the extrinsic parameters calibration of the infrared camera in the ego car world coordinate system which is also the reference for the stereovision sensors.

Based on the aligned sensorial data we developed a classification fusion mechanism for combining infrared and grayscale detections on a unified pedestrian detection stream. We obtain an increased accuracy showing that the two detectors complete each other.

I. Introduction

Pedestrian detection by day and by night is a very explored problem in the computer vision world. Most of approaches use visual data provided by either stereo [1] or monocular systems [2], [3], but there are also approaches that combine different types of sensors like visible cameras, near and thermal infrared cameras, RADAR, LASER scanner [4], [5], [6].

As most approaches use stereovision sensors for pedestrian detection, with this paper we enhance the power of stereo based pedestrian detection with the addition of a far-infrared pedestrian detection system. Our solution is motivated by the limitations of the visual spectrum during low lighting conditions (dawn, dusk, night), or difficult weather conditions (like heavy sun, strong shadows, fog, snow, strong rain). On the other hand, the multiple sensor fusion pedestrian detection is also motivated by the challenges of illumination variation in the environment, cluttered background characterizing the traffic scenes, pedestrian occlusions, and complex pedestrian appearance. The two types of sensors used in this paper are somehow complementary as the visual cameras capture the light reflected by the objects in the scene, while far infrared cameras capture the heat emitted by the objects.

A first original contribution of this paper is given by the proposal of a camera response timing model for free running cameras (in our case far IR camera) which was successfully applied for synchronizing with trigger-based cameras (stereo-system).

The second contribution of the paper is related to the development of an original calibration method of the extrinsic parameters of the far infrared camera without using any calibration object, which allows and ad hoc calibration every time the far IR camera is mounted on the test vehicle.

A third contribution of the paper is given by the combination of three information cues for pedestrian detection: these are depth information, monocular intensity features and far-infrared information. These information cues are fused using aggregated channel feature pedestrian classifiers. The Aggregated Channel Feature Classifier has been introduced by [7] and it has a good performance on monocular color images.

A fourth important contribution is the development of a dataset of train and test images that proves the effectiveness of our approach. The dataset contains far-infrared and intensity images aligned by means of a stereovision sensor. The dataset was annotated using a Caltech annotation toolbox [8].

II. Related Work

We revise the multiple sensor approaches for pedestrian detection in traffic scenes. A comprehensive analysis of color-, infrared-, and multimodal-stereo approaches is presented by [6]. They also describe their own solution that uses a custom rig formed of two color cameras and two infrared cameras arranged in stereo pairs. The solution describes a trifocal framework containing the color, disparity, and infrared images. They are combined into a single five-channel multispectral image. On their proposed dataset they obtain a 91.89% pedestrian detection rate for a 5% false positive rate. They use histogram of gradient, color, disparity and infrared channels. The classification of pedestrians is performed using Support Vector Machines.

Another approach based on far-infrared stereo vision is presented by [9]. The proposed system combines warm-area detection, with a step that detects cold areas that potentially contain a pedestrian. The cold area analysis is done based on edges and disparity. Pedestrians are validated using head morphological characteristics and also thermal properties.

A tetra-vision system containing four cameras is explored by [10] for detecting pedestrians by means of the simultaneous

use of two pairs of stereo systems: a far infrared and a visible spectrum one. They exploit the advantages of both far-infrared and visible cameras and by their combination they try to overcome the limitations of each system in part. Using visual stereovision they obtain a list of bounding boxes that potentially contain a pedestrian. Based on warm area detection and symmetry based analysis the results are further refined. A last validation step is also performed by searching for human shape characteristics based on head detection, shape detection, and active contours.

An approach for detecting and predicting the pedestrian motion with the purpose of avoiding imminent collisions has been developed by [11]. The approach combines a stereo vision system with a laser scanner. This combination provides an accurate positioning of the obstacles in the environment. Using the obstacle hypothesis their system identifies pedestrians based on polylines and leg pattern identification based on laser data combined with dense disparity maps and u-v disparity. The pedestrians are tracked within the detected areas by means of validation gates.

A LIDAR sensor and an infrared camera are combined by [12] for detecting and classifying pedestrians based on their moving direction and relative speed. The two sensors are used for generating regions of interest in which pedestrians may appear. Within those regions 2D translation invariant features are extracted and then classified by means of support vector machine classifiers.

A laser scanner and a stereovision system are employed by [13] for detecting pedestrians in urban environments. The laser based pedestrian detector is composed of a distance based clustering process that separates the different clouds of points that represent each obstacle followed by a polyline based shape estimation. The pedestrian hypothesis is generated based on the comparison of the shape cluster with a pedestrian model. The laser based detector is combined with the u-v disparity based object detection provided by the stereovision system. The stereo-based pedestrian hypotheses are classified as being pedestrians based on the similarity between the vertical projection of the silhouette and the histogram of a normal distribution. The fused pedestrian detections are combined with context information (velocity and GPS information) in order to provide danger estimation.

A dual camera system combining visible light and thermal cameras is used by [14] for detecting pedestrians. They define a geometric transformation matrix that represents the relationship between the two cameras' axes. Two background images for visible and thermal images are constructed based on the pixel difference between an input thermal and pre-stored thermal background images. By means of background subtraction combined with shadow removal and morphological operations the regions of interest for the visible images are obtained and then are projected onto the thermal image. Based on the horizontal and vertical histograms pedestrians are identified in the two images.

Support Vector Machine classifiers trained on different local and global SURF features extracted on both visible and far-infrared images are described by [15]. They propose a two-stage recognition method in order to cope with the complexity of the system.

Different probabilistic based fusion schemes that combine information from visible and infrared images for classifying road obstacles based on SVM are approached by [16]. The approaches refer an early fusion method applied at the feature level, an intermediate fusion at kernel level and a late fusion scheme that combines detection scores from visible and infrared detectors.

III. Proposed method

The method we propose combines aggregated channel feature classifiers trained on two types of data one generated by an infrared sensor and another given by a stereovision system. The overall architecture of the system is described in Fig. 1 and 2. The first module performs the alignment of the infrared image with the grayscale image of the left camera from the stereovision system. This allows us to precisely find the correspondence of a point from the left image to the infrared image. Due to field of view variations, the borders of the left image will be ignored. Using stereovision we also obtain the 3D object hypothesis based on the depth map [17].

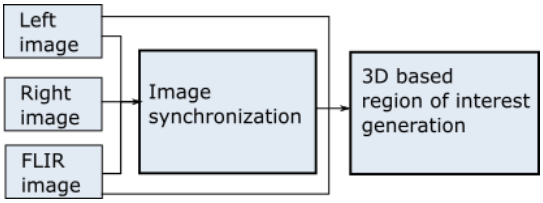


Fig. 1: Region of interest generation

The 3D object hypotheses are projected onto the left image but also on the infrared image. The region of interest of both images are scanned Aggregated Channel Feature based classifiers, and then the detections from the two images are fused as shown in Fig. 2.

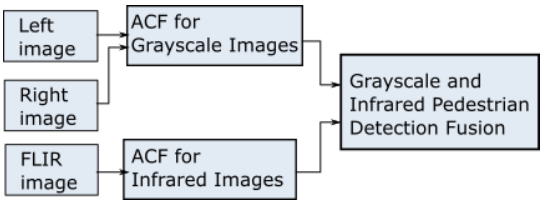


Fig. 2: Pedestrian classification fusion

A. Multiple sensors data alignment

The architecture of the stereo vision system complies with the one presented in [18]. The stereo sensor is based on two cameras mounted in a general configuration behind the windshield of a test car. The system provides a dense depth map generated on rectified images using a GPU implementation of the SGM approach[19]. High level perception methods

[20] implemented upon the low level sensorial data are used for detecting and tracking obstacles and associate classes to them (cars, pedestrians etc.). All the measurements of the stereovision sensor are reported in the ego-vehicle's 3D coordinate system. Its origin is the projection of the ego vehicle's front bumper mid-point on the ground, with the OZ axis pointing forwards (Fig. 3). The position and orientation of the stereo cameras relative to the ego vehicle coordinate system are defined by the extrinsic parameters: the translation vectors T_{CL} and T_{CR} and the rotation matrices R_{CL} and R_{CR}. The parameters are estimated with high precision using a calibration method specially developed for high accuracy automotive stereo vision [21]. Consequently a reliable dense depth map is available.

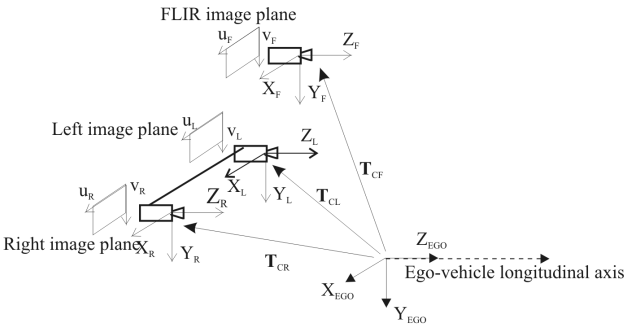


Fig. 3: Model of the multiple camera system

The far IR camera camera is mounted on the ego-vehicle's roof, positioned approximately above the stereo camera rig and aligned with the longitudinal axis of the car. Its extrinsic parameters relative to the ego-vehicles 3D coordinate system are defined by the T_{CF} translation vector and R_{CF} rotation matrix (their estimation method is described in the next section). Since all the sensors (cameras) have the extrinsic parameters estimated relative to an unique coordinate system (the ego vehicle 3D coordinate system) the alignment between the two sensorial systems (stereo and far IR) is performed by projecting the stereo reconstructed 3D points on the far IR image. This can be achieved using the projection matrix computed from the far IR camera intrinsic and extrinsic parameters:

$$\mathbf{P_F} = \mathbf{A_F} \cdot [\mathbf{R_{WF}} | \mathbf{T_{WF}}] \tag{1}$$

where:

- $\mathbf{A_F}$ is the internal matrix of the far IR camera encoding its intrinsic parameters (the focal length $[f_x, f_x]$ and the and principal point $[u_0, v_0]$):

$$\mathbf{A_F} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

- $\mathbf{R_{WF}}$ is the far infrared camera world-to-camera rotation matrix. $\mathbf{R_{WF}} = \mathbf{R_{CF}^T}$
- T_{WF} is the far infrared camera world-to-camera translation vector. $\mathbf{T_{WF}} = -\mathbf{R_{WF}}\mathbf{T_{CF}}$

The projection of a 3D point (expressed in homogeneous coordinates $\mathbf{XX_W} = [\mathbf{X_W}, \mathbf{Y_W}, \mathbf{Z_W}, \mathbf{1}]^\mathbf{T}$ relative to ego car world coordinate system) on the far infrared image will be a point $p[u, v]$:

$$s \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} = \mathbf{P_F} \cdot \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{3}$$

where: $[x_s, y_s]$ are the scaled image coordinates of $\mathbf{p}$ (with a ratio $s = z_s$). The image coordinates of $\mathbf{p}[\mathbf{u}, \mathbf{v}]$ can be obtained from the scaled ones using:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} x_s/z_s \\ y_s/z_s \end{bmatrix} \tag{4}$$

B. Multiple sensor data synchronization

The purpose of sensor synchronization is to obtain far infrared images aligned with the grayscale images provided by the video cameras of the stereo system. Our infrared camera is free running at a constant frame rate. The stereo cameras provide trigger-based synchronized video signal with insignificant relative delay times. We introduce a camera response timing model for free running cameras which relates the capture, the frame request and response moments for a particular frame. The timing model is shown in Fig. 4.

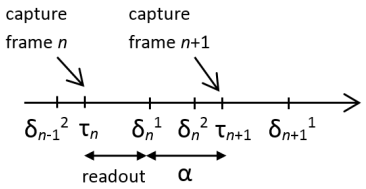


Fig. 4: Timing model for free running cameras

Considering the capture moment τ_n of a particular frame n any request for the image between δ^2_{n-1} and δ^1_n will be delayed until δ^1_n, and any request between δ^1_n and δ^2_n will encounter no delays. These parameters meet the following constraints:

$$\left|\tau_{n+1} - \delta^2_n\right| = \left|\tau_n - \delta^2_{n-1}\right| \tag{5}$$

The readout interval is defined as:

$$\delta^1_n - \tau_n = \delta^1_{n+1} - \tau_{n+1} \tag{6}$$

$$\tau_n < \delta^1_n < \tau_{n+1} \tag{7}$$

$$\delta^1_n \geq \delta^2_n < \delta^1_{n+1} \tag{8}$$

The frame period is defined as:

$$\tau = \tau_{n+1} - \tau_n \tag{9}$$

Given a constant frame period τ we define the *relaxing period* α as the time interval between releasing one frame and capturing moment of the next frame:

$$\alpha = \tau_{n+1} - \delta^1_n \tag{10}$$

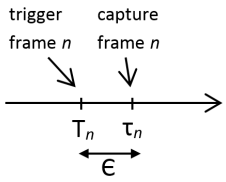


Fig. 5: Timing model for trigger-based cameras

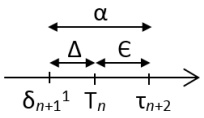


Fig. 6: Overlapped timing models

Considering the timing model for externally triggered video cameras depicted in Fig. 5 with known exposure time ϵ we were able to determine experimentally the relaxing period α for the infrared camera.

For this purpose we had to synchronize the execution thread with the moment δ_n^1 for each infrared frame. According to our proposed model for free running cameras shown in Fig. 4 any request for a frame is delayed until δ_n^1 or not delayed if the request is between δ_n^1 and δ_n^2. Due to this uncertainty period between δ_n^1 and δ_n^2 we perform two consecutive requests instead of one. According to 8 the first request will be released somewhere between δ_n^1 and δ_n^2 and the second request, which takes place immediately after the first response, will be released at δ_{n+1}^1. At the cost of losing two frames we managed to perform the synchronization on the 3rd frame by delaying the trigger moment of video cameras with a controlled Δ period as shown in Fig. 6.

The experimental setup uses both infrared and video cameras capturing images of a fast moving object emanating heat. The delay Δ is increased incrementally until the moving object has similar position in both grayscale and infrared images. Once Δ is stabilized for a fixed exposure ϵ we are able to compute the constant α:

$$\alpha = \Delta + \epsilon \tag{11}$$

Once the relaxing period α is known we can afford a variable exposure ϵ_n by changing the delay time Δ_n of the video trigger for each pair of captured images:

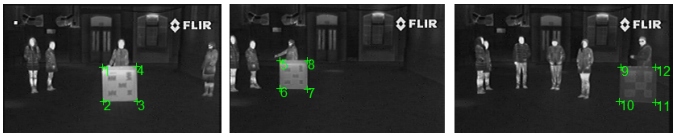
$$\Delta_n = \alpha - \epsilon_n \tag{12}$$

C. Calibration of the far IR camera parameters

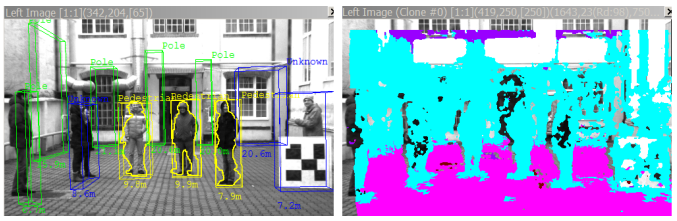
The intrinsic parameters of the far IR camera are inferred from the data sheet of the manufacturer. Since the camera can be used in 2 resolutions two sets of intrinsic parameters are generated:

- QVGA (native): f_x=f_y=410, u_0=160, v_0=120 [pixels]
- VGA (upscaled): f_x=f_y=620, u_0=320, v_0=240 [pixels]

For the extrinsic parameters estimation an approach similar to [21] could be used by replacing the control points (detectable in visible light) with ones that have a thermal



(a) Contol points selection - far IR image



(b) Left image with obstacles detection and 3D dense map - stereo sensor

Fig. 7: Typical experimental scenario used for the extrinsic parameters calibration of the far IR camera.

footprint. But this would require a complicated calibration setup which should be reproduced every time the far IR camera is mounted on and off the car roof. In our case the far IR camera does not have a permanent mounting socket on the car and is removed after every driving experiment (due to security reasons). For such experimental condition a much simpler calibration methodology was proposed which can be easily performed every time the far IR camera is mounted on the car:

1) Acquire synchronized pairs of stereo and far IR images in a static scenario with uniquely selectable image features in both sensorial image spaces. Preferred features are those that are easy to localize (i.e. corner features) in both sensors: they are visible in the stereo sensor images and also have a visible thermal footprint in the far IR image.
2) Manually select at least 6 points with 3D information from the stereo system (using the left stereo image + depth map): $\mathbf{P_i}(\mathbf{X_i}, \mathbf{Y_i}, \mathbf{Z_i})$. Manually find their corresponding points on the far IR image (a procedure to compute their sub-pixel coordinates using a zoom in technique is employed): $\mathbf{p_i}(\mathbf{x_i}, \mathbf{y_i})$. An example of such a calibration scenario is presented in Fig. 7 along with a numerical example of the selected features' values in Table I.
3) The extrinsic parameters of the far IR camera ($\mathbf{T_{CF}}$ and $\mathbf{R_{CF}}$) relative to the ego-vehicle coordinate system are estimated by minimizing the projection errors of the 3D points $\mathbf{P_i}(\mathbf{X_i}, \mathbf{Y_i}, \mathbf{Z_i})$ against the selected image points: $\mathbf{p_i}(\mathbf{x_i}, \mathbf{y_i})$ using the Gauss-Newton iterative method [21].

D. Aggregated Channel Feature Classification

The Aggregated Channel Feature is a model for multiple resolution image feature approximation that works fast and has good performance. The model can be applied to a generic object detector and it was tuned for pedestrian detection in the Aggregated Channel Features (ACF) framework [7], [8]. The

TABLE I: Numerical example of the selected features values for the calibration scenario from Fig. 3.

	far IR control points $\mathbf{p_i}$		3D coordinates $\mathbf{P_i}$ (stereo sensor)		
ID	x_i [pixel]	y_i [pixel]	X_i [mm]	Y_i [mm]	Z_i [mm]
1	153.25	86.92	-323	-926	7751
2	153.99	128.94	-304	85	7775
3	195.64	128.76	700	57	7733
4	195.27	85.82	689	-924	7899
5	93.16	81.03	-2165	-876	10258
6	93.02	115.14	-2160	117	10070
7	127.13	115.71	-1148	101	9941
8	126.98	80.89	-1106	-911	9539
9	237.83	87.84	1689	-905	7898
10	236.49	129.85	1661	88	7732

idea of ACF is to replace the actual feature computation at every image scale in an image pyramid with feature approximation by extrapolation from nearby scales. [7] prove that for a broad family of features this process does not reduce the performance of the detection. Their proposed approximation yields considerable speedups with negligible loss in detection accuracy.

AdaBoost learner build on top of level-two decision trees is employed. The classification score of the AdaBoost learner is a linear combination of weighted weak learner responses $h_1, h_2, \dots h_T$: $f(x) = \sum_{t=1}^{T} w_t h_t(x)$. Each weak learner response is weighted by w that is proportional to the error of the weak learner.

A cascade of such composite ensemble is used. The cascade has four stages and each stage has the same positive training set, while the negatives for each stage are the false positives of the previous stage. Each weak learner is a decision tree. The number of weak classifiers in the stages is $256, 512, 1024, 2048$.

In the proposed method we train two ACF based pedestrian detectors: one for the grayscale intensity images and another for the infrared images. For the grayscale intensity image we train the ACF framework using ten channels, namely LUV, gradient magnitude, gradient orientation histogram containing six bins. Let D_g be the detector that results.

For the infrared intensity image we train the ACF framework on infrared images and we use the gradient magnitude, the gradient orientation histogram with six bins and the LUV channels. Denote with D_i the resulting classifier.

Next the two classifiers are applied in parallel and the results for an image I of each detector are fused using the following:

$$F(I) = \alpha D_g(I) + \beta D_i(I) \tag{13}$$

where α and β are parameters which we use to weight the influence of each classifier. For example in low lightning driving scenarios, or when the grayscale images are over-saturated the infrared classifier will have a greater weight than the grayscale classifier. On the other hand when the temperature difference between the environment and the pedestrians or other objects of interest is reduced causing a very low contrast on infrared features the grayscale classifier will prevail.

The 3D object points of object hypothesis generated by the stereo sensor are projected on both the grayscale and the far IR images. The fusion between the grayscale and far IR detections is performed based on an overlapping predicate (OR) on their 2D image projections (as shown in Fig. 8). The grayscale/stereo detections and the far IR detections (given by orange bounding boxes) are fused on the far IR image – as shown in the left parts of Fig. 8 .

IV. Experimental Results

A. Assessment of the far IR camera parameters calibration procedure

The accuracy of the data alignment procedure depends on the extrinsic parameters which are used in the projection matrix (1). As evaluation metric the 2D projection error of a set of 3D points generated with the stereo sensor vs. their manually selected correspondences on the 2D far IR image was considered.

$$\epsilon = \begin{bmatrix} x_{GT} - x_{PR} \\ y_{GT} - y_{PR} \end{bmatrix} \tag{14}$$

where:

- $[x_{GT}, y_{GT}]$ are the coordinates of the manually selected 2D control points on the far IR image (ground truth - GT)
- $[x_{PR}, y_{PR}]$ are the 2D projections of the 3D control points selected from the stereo sensor.

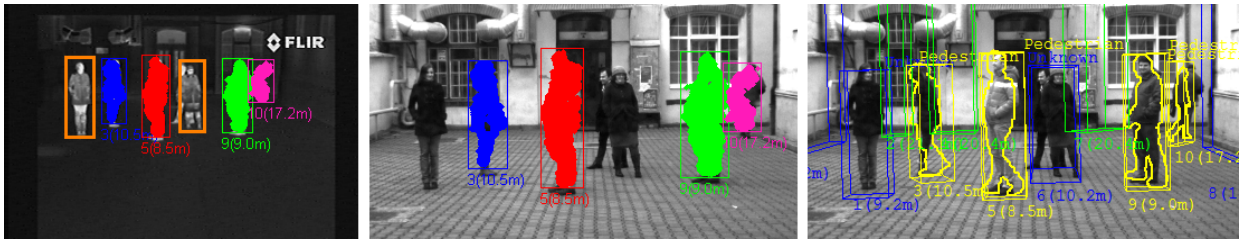
In Fig. 9 the qualitative assessment of the projection errors is shown in a visual form (green points are the projected points while red ones are the GT points manually selected on the far IR images). This is also observable in Fig. 8.

In Table II the corresponding numerical errors in terms of the (14) metric are shown. Two scenarios are analyzed: in Scenario 1 the calibration experiment from Fig. 7 was evaluated using far IR images acquired in QVGA resolution. The results labeled Scenario 2 were obtained performing a new calibration experiment with the far IR camera mounted in a new position using the established methodology and evaluation in real traffic conditions on VGA resolution images.

TABLE II: Numerical evaluation of the projection errors.

Scenario 1: QVGA res.			**Scenario 1: QVGA res.**		
ID Fig. 9(a)	ϵ_x	ϵ_y	ID Fig. 9(b)	ϵ_x	ϵ_y
1	-0.17	1.76	1	1.15	-0.61
2	0.97	1.54	2	1.64	6.50
3	0.81	1.22	3	0.14	1.77
4	1.10	0.99	4	1.54	1.56
5	1.09	1.56	5	-0.01	1.58
6	-0.13	1.46	**RMS error**	1.13	3.18
7	-0.02	0.94	ID (Fig. 9(c))	ϵ_x	ϵ_y
8	1.76	1.02	1	2.09	-0.27
9	1.81	0.69	2	-0.01	-2.30
10	1.21	0.84	3	1.65	-1.26
RMS error	**1.09**	**1.25**	**RMS error**	**1.54**	**1.52**

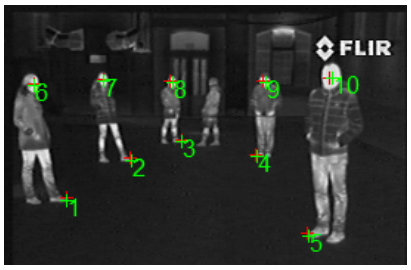
For the static scenarios and QVGA resolution images the RMS error of the projections was inside the 1.25 pixels range. For the real traffic scenarios and VGA resolution images the errors were inside the 2.5 pixels range (which is obvious due to the doubled pixel size / focal length) with some out-of-the-



(a) Static scene: left - far IR image, middle - grayscale image, right – 3D hypotheses from the stereo sensor

(b) Dynamic scene left - far IR image, middle - grayscale image, right – 3D hypotheses from the stereo sensor

Fig. 8: Projection of 3D points on the grayscale and on the infrared image



(a) Projection error with QVGA images (Scenario 1)

(b) Projection error with VGA images (Scenario 2)

(c) Projection error with VGA images (Scenario 2)

Fig. 9: Scenarios used to evaluate the data alignment accuracy.

range errors. The projection errors are mainly influenced by the manual correlation process of the feature points used in the tests but can also occur due to motion blur like effects in the thermal footprint in the case of dynamic scenarios.

B. Camera Synchronization

Our system uses one FLIR PathfindIR infrared camera and two synchronized JAI video cameras. PahfindIR has a 19mm focal length and provides 320×240 PAL (25Hz) images. The experimental setup used to detect the relaxing period α of PathfindIR consists of a table fan having a colored band and a heat emanating electrical circuitry attached to one of its wings. This way the wing position is clearly visible in both grayscale and infrared images. The exposure time of grayscale cameras was fixed and we ran the table fun at different speeds. We started recording frames while changing the delay Δ of the trigger incrementally. We filtered out un-synchronized frames and obtained a relaxing period of 23ms for synchronized frames. Similar results were obtained when we repeated the tests for different exposure values.

C. Pedestrian Detection

For assessing the performance of the trained pedestrian detectors we use the log-average miss rate computed on a dataset we have created. The dataset contains about 2000 annotated frames for both intensity and infrared cameras. The images contain about 3000 pedestrians.

We perform a comparative evaluation as follows:

1) For a given threshold value of the grayscale pedestrian detector we compute the true positive rate and the false positive rate when it is evaluated on grayscale images.
2) For a given threshold value of the infrared pedestrian detector we compute the true positive rate and the false positive rate when it is evaluated on infrared images.
3) For the given threshold value used for grayscale and infrared detectors we compute the fused image of detections and count how many pedestrians were detected correctly (either in grayscale or in infrared), along with the rate of false positives.

The results are shown in Table III. The threshold value for either of the detectors represents the classification score above which detections are considered as being positive.

TABLE III: Numerical evaluation of the pedestrian detection.

Missed Ped. IR	**TP IR**	**FP Per Image IR**
35%	65 %	0.9
Missed Ped. Gray	**TP Gray**	**FP Per Image Gray**
25%	75%	0.7
Missed Ped. Combined	**TP Combined**	**FP Per Image Combined**
10%	90%	1

From our experiments we show that the two detectors complete each other and the overall result is improved. Yet, the combined solution accuracy depends on the projection error from the stereo system onto the FLIR images. This is influenced by the accuracy of the 3D data provided by the stereo vision system and by the quality of the estimated extrinsic parameters of the far IR sensor (mainly dependent on the calibration scenario used).

The true positive rate of the combined classification scheme is about 90%. A disadvantage of the method is the propagation of the false positives in both images and we plan to enhance our detection mechanism with specific constraints for false positive.

In what follows we present some detection results by which we show how the two approaches complete each other. First we show cases of images in which the infrared detector detects the pedestrians, while the grayscale detector fails:

Secondly we show cases of images in which the grayscale detector detects the pedestrians, while the infrared detector fails:

In Fig. 10 and Fig. 11 the number above the detections represents the classification score of the AdaBoost classifier.

V. Conclusion

We have developed an integrated system that performs a spatio-temporal data alignment between two types of sensors: a monocular far infrared camera and a stereovision sensor. This allows to combine the 3D information of the obstacles with their thermal footprint. This approach augments the capabilities of the pedestrian detection algorithm, especially in situations when the detection with only one sensor is incomplete or less accurate.

The temporal synchronization is done using an original camera response timing model for free running cameras that aligns the infrared image with grayscale intensity images captured by trigger-based cameras. The spatial synchronization is done by performing the calibration of the extrinsic parameters of both sensors relative to the same coordinate system. The data alignment allows to correlate image features between the grayscale images of the stereovision sensor and the infrared image with the addition of the 3D information. Based on the aligned sensorial data we developed a classification fusion mechanism for combining infrared and grayscale detections on a unified pedestrian detection stream and we study the influence of the infrared pedestrian detector on the grayscale detections and vice-versa.

As future work we envision an improvement of the acquisition frame rate of the fused sensorial system. Currently the proposed synchronization method discards two out of three consecutive frames and reduces the maximum running frame rate from 25Hz to 8.33Hz. This can be avoided through parallelism by creating a monitoring thread responsible for acquiring frames continuously without being interrupted by other processing tasks.

We also plan to improve the correlation mechanism between grayscale and infrared images not only by projecting the 3D points from the stereo system to the infrared system, but also by using a stereo-correspondence mechanism directly between one of the grayscale images and the infrared image based on epipolar constraints. This would also require a more accurate calibration methodology in a more controlled scenario.

Acknowledgment

This work has been supported by UEFISCDI (Romanian National Research Agency) in the national research project Cooperative Advanced Driving Assistance System Based on Smart Mobile Platforms and Road Side Units (SmartCoDrive), project no. PNII-PCCA 18/2012.

References

[1] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[2] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[3] R. Benenson, M. Omran, J. Hosang, , and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *ECCV, CVRSUAD workshop*, 2014.

[4] T. Gandhi and M. Trivedi, "Pedestrian collision avoidance systems: a survey of computer vision based recent studies," in *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, Sept 2006, pp. 976–981.

[5] ——, "Pedestrian protection systems: Issues, survey, and challenges," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 413–430, Sept 2007.

[6] S. Krotosky and M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 4, pp. 619–629, Dec 2007.

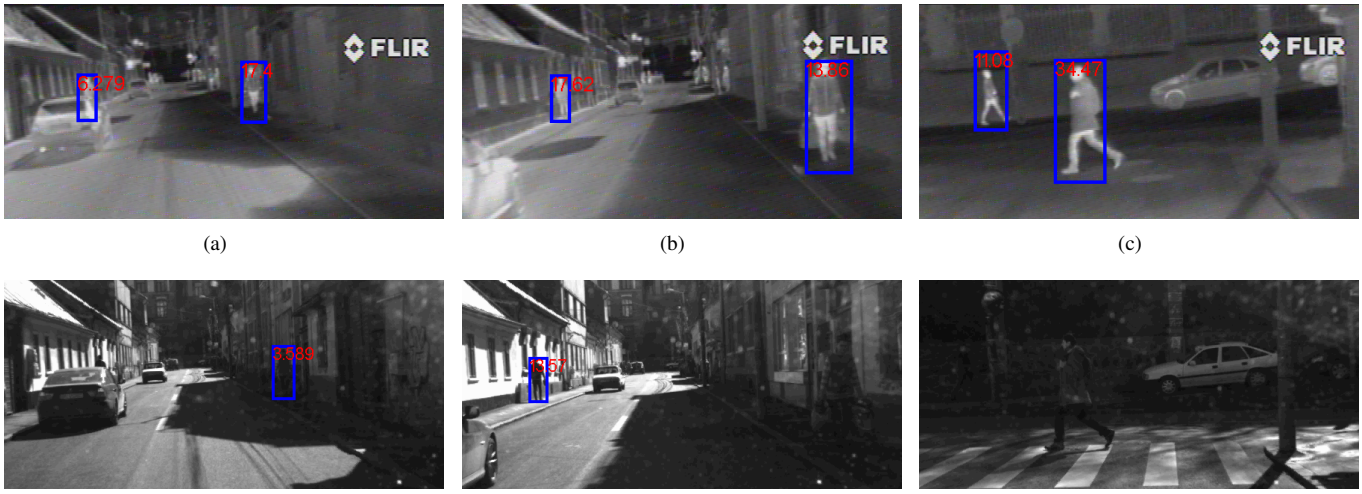
[7] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 8, pp. 1532–1545, Aug 2014.

[8] P. Dollár, "Piotr's Image and Video Matlab Toolbox (PMT)," http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[9] M. Bertozzi, A. Broggi, C. Caraffi, M. Del Rose, M. Felisa, and G. Vezzoni, "Pedestrian detection by means of far-infrared stereo vision," *Comput. Vis. Image Underst.*, vol. 106, no. 2-3, pp. 194–204, May 2007. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2006.07.016

[10] M. Bertozzi, A. Broggi, M. Felisa, S. Ghidoni, P. Grisleri, G. Vezzoni, C. Gmez, and M. Rose, "Multi stereo-based pedestrian detection by daylight and far-infrared cameras," in *Augmented Vision Perception in Infrared*, ser. Advances in Pattern Recognition, R. Hammoud, Ed. Springer London, 2009, pp. 371–401. [Online]. Available: http://dx.doi.org/10.1007/978-1-84800-277-7_16

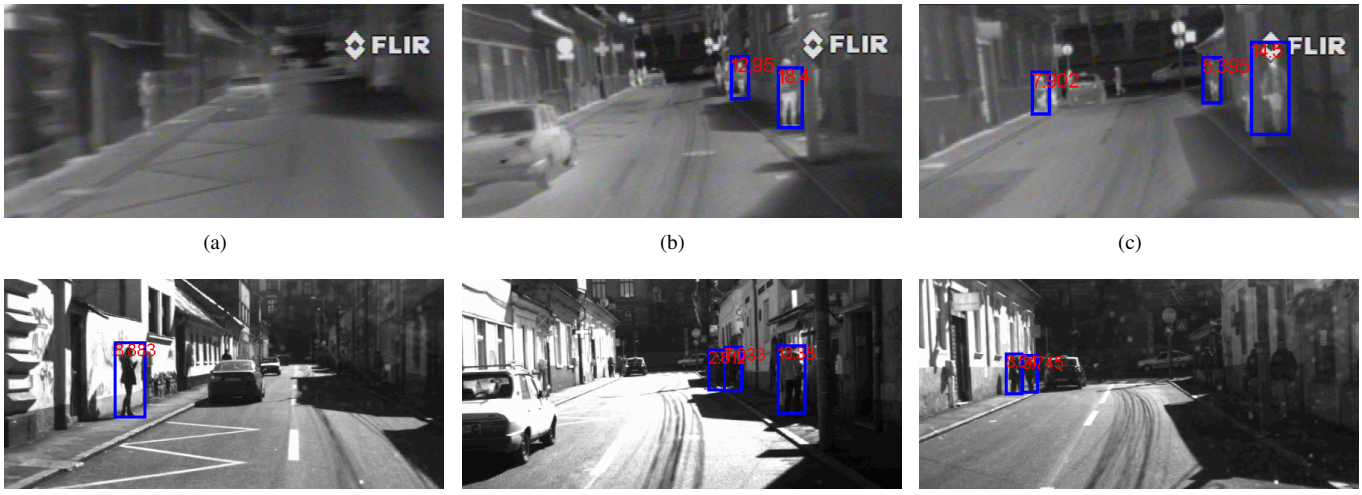
[11] B. Musleh, F. Garca, J. Otamendi, J. M. Armingol, and A. De la Escalera, "Identifying and tracking pedestrians based on sensor fusion and motion stability predictions," *Sensors*, vol. 10, no. 9, p. 8028, 2010. [Online]. Available: http://www.mdpi.com/1424-8220/10/9/8028



(a) (b) (c)

(d) (e) (f)

Fig. 10: Good detections in infrared that are missing in grayscale



(a) (b) (c)

(d) (e) (f)

Fig. 11: Good detections in grayscale that are missing in infrared

[12] A. Pérez Grassi, V. Frolov, and F. Puente León, "Information fusion to detect and classify pedestrians using invariant features," *Inf. Fusion*, vol. 12, no. 4, pp. 284–292, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.inffus.2010.06.002

[13] F. Garcia, A. de la Escalera, J. Armingol, J. Herrero, and J. Llinas, "Fusion based safety application for pedestrian detection with danger estimation," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, July 2011, pp. 1–8.

[14] J. H. Lee, J.-S. Choi, E. S. Jeon, Y. G. Kim, T. T. Le, K. Y. Shin, H. C. Lee, and K. R. Park, "Robust pedestrian detection by combining visible and thermal infrared cameras," *Sensors*, vol. 15, no. 5, p. 10580, 2015. [Online]. Available: http://www.mdpi.com/1424-8220/15/5/10580

[15] B. Besbes, S. Ammar, Y. Kessentini, A. Rogozan, and A. Bensrhair, "Evidential combination of svm road obstacle classifiers in visible and far infrared images," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*, June 2011, pp. 1074–1079.

[16] A. Apatean, A. Rogozan, and A. Bensrhair, "Visible-infrared fusion schemes for road obstacle classification," *Transportation Research Part C: Emerging Technologies*, vol. 35, no. 0, pp. 180 – 192, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X13001563

[17] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 10, no. 3, pp. 380–391, 2009.

[18] S. Nedevschi, T. Marita, R. Danescu, F. Oniga, S. Bota, I. Haller, C. Pantilie, M. Drulea, and C. Golban, "On-board 6d visual sensor for intersection driving assistance," in *Advanced Microsystems for Automotive Applications 2010*, ser. VDI-Buch, G. Meyer and J. Valldorf, Eds. Springer Berlin Heidelberg, 2010, pp. 253–264. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-16362-3_25

[19] C. Pantilie, I. Haller, M. Drulea, and S. Nedevschi, "Real-time image rectification and stereo reconstruction system on the gpu," in *Parallel and Distributed Computing (ISPDC), 2011 10th International Symposium on*, July 2011, pp. 79–85.

[20] S. Nedevschi, M. Tiberiu, R. Danescu, F. Oniga, and S. Bota, "On-board stereo sensor for intersection driving assistance architecture and specification," in *Intelligent Computer Communication and Processing, 2009. ICCP 2009. IEEE 5th International Conference on*, Aug 2009, pp. 409–416.

[21] T. Marita, F. Oniga, S. Nedevschi, T. Graf, and R. Schmidt, "Camera calibration method for far range stereovision sensors used in vehicles," in *Intelligent Vehicles Symposium, 2006 IEEE*, 2006, pp. 356–363.