Olmeda, D.; Armingol, J. L.; Escalera, A. (2013). Part based pedestrian detection based on logic inference. En: *Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013)*, The Hague, The Netherlands, October 6-9.

DOI: 10.1109/ITSC.2013.6728421

Part based Pedestrian Detection Based on Logic Inference

Daniel Olmeda, Jose Maria Armingol and Arturo de la Escalera

Abstract— **This paper presents an approach on detection of largely occluded pedestrians. From a pair of synchronized cameras in the Visible Light (VL) and Far Infrared (FIR) spectrum individual detections are combined and final confidence is inferred using a small set of logic rules via a Markov Logic Network. Pedestrians not entirely contained in the image or occluded are detected based on the binary classification on subparts of the detection window. The presented method is applied to a pedestrian classification problem in urban environments. The classifier has been tested in an Intelligent Transportation System (ITS) platform as part of an Advanced Driver Assistance Systems (ADAS).**

I. Introduction

Object recognition in images has become a very important topic in the fields of traffic infrastructure and driving assistance system. Applications such as traffic signs recognition, obstacle avoidance and traffic surveillance have gotten the attention of the industry for some time now. The case of people detection is an exceptionally relevant case, as it leads to a number of important applications, some of which strive for saving lives. Pedestrian recognition in images is geared toward a variety of applications, which include driver assistance systems [1] and autonomous robotic vehicles [2]. It is also useful in security, be it for automatic surveillance [3] or people counting [4]. Those applications can be extended to low visibility conditions by using FIR images.

In this paper a pedestrian detection scheme based on hyper spectral visual information is presented. The system has been tested in an experimental vehicle under normal conditions of urban traffic.

Most of the recent research in pedestrian detection for Intelligent Transportation Systems (ITS) is based on visible light (VL) images. FIR images share some key characteristics with their VL images counterparts. They both are 2D representations of a scene captured by redirecting electromagnetic waves by means of a lens, light in the first case and infrared radiation, which is proportional to the objects temperature, in the second. Some of the key ideas on pedestrian classification in VL images can be extended to work on FIR images, exploiting common characteristics of both, or adapt them to take benefit of the different kinds of information provided by FIR images. Furthermore FIR and VL images are complementary in the sense that the detection performance of both systems is optimum in opposite circumstances. The VL image based systems work best under conditions of high external luminosity (i.e. daylight). In the case of FIR images, the best performance is obtained under conditions of low temperature, which coincides with low illumination or night conditions.

This paper presents an approach to combining the results of pedestrian detection in images in the visible spectrum with far infrared images. Using a constrained sliding window approach, each pair of synchronized images is searched for pedestrians, using two rigid objects descriptors, HOG [5] in the case of images in the visible spectrum and HOPE [6] in the case of far infrared images. Detections are projected into the three-dimensional world, assuming that the detected pedestrians are standing on the pavement in front of the vehicle. Final detection is based on the output of a Markov Logic Network.

This approach is extended to part-based detection by training a classifier for each of the cells on which the descriptors are divided. Instead of applying predefined rules, the Logic Network seeks iteratively relations among all possible rules. The resulting weights allow inferring the presence of pedestrians from incomplete samples by looking for latent part structures, outperforming rigid models looking for complete objects, in the case of occluded pedestrian.

This paper is structured as follows. Section II includes a brief discussion on related work in pedestrian detection. It also covers an overview of descriptors and FIR image-based classification methods. Section III focus on the features used for detection, and the geometrical relation between detections in both images of the pair. The probabilistic inference of the final detection, based on individual detections on VL and FIR images is explained in section IV. In section V the part-based approach on occluded pedestrian detection is explained. In section VI experimental results are presented. The experimental results are further discussed in section VII.

II. Related Work

Detecting pedestrians in images is an area of great interest. In recent years, a large number of methods for the detection of pedestrians in images have been published. They are usually based on data from just one sensor, be it a stereoscopic rig or monocular cameras. Many of these methods are based on images in the visible spectrum, although some of them have been adapted for use also in the far-infrared spectrum.

In the area of pedestrian detection in VL images the HOG descriptor [5] has become a benchmark for pedestrian classification. Since its introduction, many features are using

*This work was supported by the Spanish Government through the Cicyt projects FEDORA (GRANT TRA2010-20225-C03- 01) and Driver Distraction Detector System (GRANT TRA2011-29454-C03-02), and by the Comunidad de Madrid through the project SEGVAUTO (S2009/DPI-1509).

The authors are with the Department of Systems Engineering, Universidad Carlos III de Madrid, C/ Butarque 15, Leganes, Madrid, Spain. dolmeda at ing.uc3m.es

local histograms [7] [8], including classifiers targeted to FIR images [6].

FIR images represent the heat emitted by people, so there is no need of external illumination in order to operate them. Some authors have developed classification methods based on the temperature distribution of the human body. Most systems take advantage of this feature and select regions of interest based on the distribution of the warm parts of the image [9]. A review of techniques pedestrian automotive infrared detection can be found in [10]. In these systems, the discriminating feature of pedestrians is the body shape, and the regions of interest are validated by correlation with predefined probabilistic models [11] and [12].

In spite of the number of methods available there are few studies in which the information extracted from VL and FIR images is used together. In [13] a review of fusion techniques is presented. In it, methods are differentiated into the following groups: pixel-based, feature-based and decision-level-based. The approach presented in this paper would be included in the latter.

Regarding the application of MLNs to pedestrian detection, Oliveira et al. presented in [14] a Lidar-based system, immune to partial segmentation of data. The system infers the relationship of the sub-segments and their context. In [15] their work is extended in a multisensory scheme that fuse visual information with Lidar data, based on spatial relationship of parts-based classifiers, via a MLN.

III. Pedestrian Detection

The detection is carried out independently in each of the two images. The algorithm has been tested using HOG features [5] in the VL dataset, as implemented in the Dollar toolbox [16]. Optimum parameters have been selected by grid-search cross-validation. In this work we have used 5×5 pixel non-overlapping cells. Within each cell, a 9-bin histogram of orientation between 0 and 2π radians is calculated. The model has been trained using the INRIA database [5].

Defining the shape of a pedestrian in FIR images is challenging using one dimensional point derivatives. This is due to the much wider infrared spectrum, compared with visible light. Another difficulty is that the sensitivity curve of an uncooled microbolometer sensor changes very quickly with minimum changes of its temperature [17]. To overcome these challenges, we proposed in [6] a contrast invariant descriptor for pedestrian classification in FIR images called HOPE. Basically, the HOPE descriptor encodes a grid of local oriented histograms extracted from the phase congruency of the images, which is computed from a joint of Gabor filters. The FIR model has been trained using the LSI Far Infrared Pedestrian Dataset.

For each detection in one of the two images, the position of the region of interest is projected into the ground ahead of the vehicle by $p_g = \zeta(r_i)$, where p_g are the two dimensional coordinates of the pedestrian position, whose reference point is between both cameras image centers. Assuming the images are properly rectified and pedestrians are standing on a flat ground, the projection function ζ can be simplified as a pinhole model, where the intrinsic parameter matrix M (eq. 1), and the extrinsic parameters W_V and W_F between the cameras and the reference point (eq. 2) are known:

$$M = \begin{pmatrix} \mathrm{f_x} & 0 & \mathrm{c_x} \\ 0 & \mathrm{f_y} & \mathrm{c_y} \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

Where f_x and f_y are the orthogonal focal lengths of the lens and (c_x, c_y) is the image center. The rotation and translation matrix W is:

$$W = \begin{pmatrix} R_{(1,1)} & R_{(1,2)} & t_x \\ R_{(2,1)} & R_{(2,2)} & t_y \\ R_{(3,1)} & R_{(3,2)} & t_z \end{pmatrix} \tag{2}$$

$$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \tag{3}$$

$$R_y = \begin{pmatrix} \cos(\beta) & 0 & -\sin(\beta) \\ 0 & 1 & 0 \\ \sin(\beta) & 0 & \cos(\beta) \end{pmatrix} \tag{4}$$

$$R_z = \begin{pmatrix} \cos(\gamma) & \sin(\gamma) & 0 \\ -\sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5}$$

$$R = R_x \cdot R_y \cdot R_z \tag{6}$$

$$H = M \cdot W \tag{7}$$

$$\begin{bmatrix} x \\ y \\ s \end{bmatrix} = H^{-1} \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{8}$$

$$p_g = \begin{bmatrix} x/s \\ y/s \\ 1 \end{bmatrix} \tag{9}$$

Where α is the pitch angle, β is the yaw angle, and γ is the roll angle. The point of contact of the region of interest with the ground if given in image coordinates (u, v)

Detection Matching: Positive detections in VL and FIR images are assigned to pedestrians in the three-dimensional world using the Munkres assignment algorithm [18], where for each positive detection in the VL range D_{VL} or the FIR range D_{FIR} a detection pair is created $d_i = (v, f)$, where $(v, f) \in \{\{0,1\}, \{0,1\}\}$.

IV. Logic Combination of VL and FIR Detections

Markov logic networks (MLN) are a first order knowledge base, based on a set of logic rules that define the occurrences of events or the relations or conditionality between them. Each logical formula f_i has an associated weight w_i, which is trained using discriminative learning [19] from a labeled database, and assumed to have equal prior probability. Thus each logical statement is no longer binding, but the events they that imply will be more likely to be true based on its weight.

The network models the joint distribution of the events. In this case, events are binary variables that answer whether each member of a detection pair is positive or not. If any of these formulas are true, the network implies that the detection is positive. The probabilistic inference that responds to the query *isPedestrian* is calculated on the minimum subset of events using the Lifted Belief algorithm [20]. In equation 11, the joint probability of the events given their responses to the logic formulation is defined as the normalized exponential product of the formula-weight.

$$P = \frac{e^{(\sum_i w_i f_i)}}{Z} \tag{10}$$

Where w_i are the weights, f_i are the set of logic formulae, and Z is the normalization factor.

For each detection pair $d_i = (v, f)$, v is true if there is a detection in the VL image, false otherwise. Likewise f is true if there is a detection in the FIR image, false otherwise. Detections are reprojected into the other image. If the region of interest falls outside of the borders of the image the queries *isVisibleVL* or *isVisibleFIR* is set to false, to true otherwise. Detection is inferred from the following logic rules f_i

$$\begin{aligned}
\forall i, \mathtt{isVL}(d_i) &\Rightarrow \mathtt{isPedestrian}(d_i)\\
\forall i, \mathtt{isFIR}(d_i) &\Rightarrow \mathtt{isPedestrian}(d_i)\\
\forall i, !\mathtt{isVisibleVL}(d_i) \wedge \mathtt{isFIR}(d_i) &\Rightarrow \mathtt{isPedestrian}(d_i)\\
\forall i, !\mathtt{isVisibleFIR}(d_i) \wedge \mathtt{isVL}(d_i) &\Rightarrow \mathtt{isPedestrian}(d_i)
\end{aligned}$$

V. Part-based detection

This approach can be extended to part-based pedestrian detection by adding to the MLN a new set of formulae. The detection window is divided into subparts and an SVM is trained for each one (Fig. 1). In the MLN, a positive detection of a part would imply that the query *isPedestrian* is true.

A. Parts SVMs

For each cell in the pedestrian descriptor an SVM classifier calculates the boundary between the pedestrian and background classes by searching the hyperplane that maximally separates the training set in a high-dimensional space of the descriptor. The SVM is trained with a representative subsample of the train dataset, on which pedestrians have a lateral occlusion between 0 and 50%. Pedestrian and non-pedestrian images are resized to have the same dimensions. The decision function in equation 11 is optimized so that $y_k(x)$ maximizes the distance between the nearest point (x_i) and the hyperplane. The linear discriminant function is:

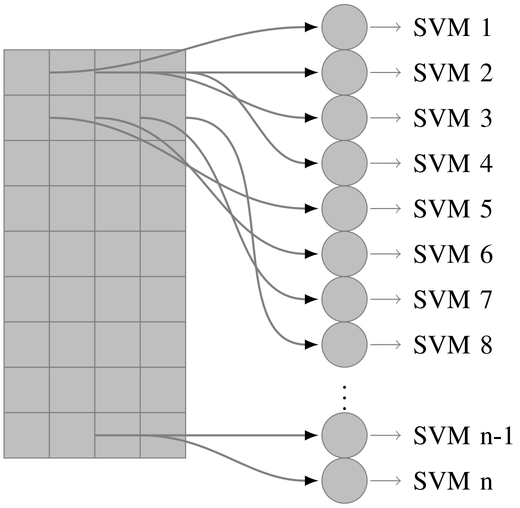


Fig. 1: Descriptor blocks of pedestrian parts. For each block an SVM classifier is trained.

$$y_k(x) = w^T \cdot \Phi(x_k) + b_k = \sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\sum_{l=1}^{m} w_{ijl}\Phi_l(c_{ij}) + b_k \tag{11}$$

Where w is normal to the hyperplane, b_k is the classification bias of part k, and $\frac{b_k}{||w||}$ is the perpendicular distance from the hyperplane to the origin. $\Phi(c_{ij}) \in \mathbb{R}^m$ is the corresponding histogram of gradients with m cell bins, at pixel c_{ij}. Φ is the kernel function that is used to project the samples. In this evaluation a linear kernel has been used. The sample is assigned to one of the two classes by thresholding the decision function, where a sample with a score of $y(x) > b_k$ is classified as a pedestrian and as background otherwise.

Each pedestrian part classifier $(y(p_{g_i}))$ is evaluated by the area under the ROC curve (a_{uc}). Only those with an a_{uc} over a threshold (thr) are used to calculated the MLN weights. For each pedestrian classifier with $a_{uc} >$ thr, the bias is selected by calculating the optimal operating point of the ROC curve, that is, the one on which the curve intersect with the line with slope

$$S = \frac{\mathrm{c}(P|N) - \mathrm{c}(N|N)}{\mathrm{c}(N|P) - \mathrm{c}(P|P)} \cdot \frac{N}{P} \tag{12}$$

Where c is the cost of misclassifiying one sample as a member of the opposite class., and P=TP+FN and N=TN+FP are the total number the positive and negative samples, respectively.

B. Fixed set of First Order Logic rules

The presented approach is extended by adding one rule for each part classifier of a detection pair d_i. If a part is classified as pedestrian-part the network imply that the sample is a pedestrian.

$$\forall w, \texttt{isPart}_i(w,t) \quad \Rightarrow \quad \texttt{isPedestrian}(w), \forall i \in n_p$$

Where n_p are the set of parts that have classifiers with an area under the roc curve $a_{uc} > \text{thr}_i$, $\texttt{isPart}_i(w,t)$ is true if part i of window w if classified as positive, and t is the kind of image this part belong to (i.e. VL or FIR). Figure 2 shows the relative classification weight of pedestrian parts on the FIR database.

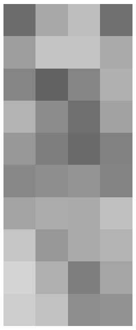


Fig. 2: Relative classification weight of pedestrian parts scaled in $w_p = \{0, 1\}$.

C. Structure Learning

The set of logic rules can be automatically learned from the ground truth database, along with their respective weights. The relation between each possible combination of clauses is tested and added to the set of rules if there exist statistical significance. This approach is used to detect incomplete pedestrian, be it because they are occluded of because only part of it falls inside the image. Occlusion percentage is calculated as depicted in Fig. 3. The result is a MLN trained to find latent structures of heavily occluded pedestrians. In this implementation, clauses are found using a beam search algorithm [21].

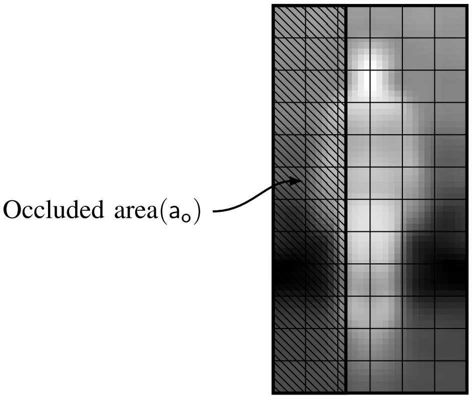


Fig. 3: Occluded area of the region of interest.

VI. Experimental Results

A. Results of the logic combination of VL and FIR detections

Detection in VL and FIR images has been evaluated using the Pascal Criteria [22]. Detections are considered true or false positives based on the area of overlap with ground truth bounding boxes. To be considered a correct detection the score of the SVM has to be $y_k(x) > 0$ and the overlap area a_o of the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 50% by the equation 13.

$$a_o = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \tag{13}$$

For each pedestrian, it is usual that more than one detection appear in the neighbourhood around the ground truth bounding box. If two or more detections match the same ground truth bounding box, only the one with the higher score would be considered a true positive. Other overlapping detections are considered false positives. To minimize the number of repeated detections, a greedy non-maximum suppression (NMS) algorithm, pairwise max (PM) suppression [23], is applied to all bounding boxes. It selects iteratively detections with higher scores than their neighbourhood, discarding detections with lower scores over an overlapping percentage. This overlap is again calculated with equation 13. In the case of Logic Inference, results have been evaluated by manually classifying each detection pair as a positive or a negative.

Figure 4 represents in a DET curve the detection performance of VL and FIR descriptors, as well as the results of logic inference.

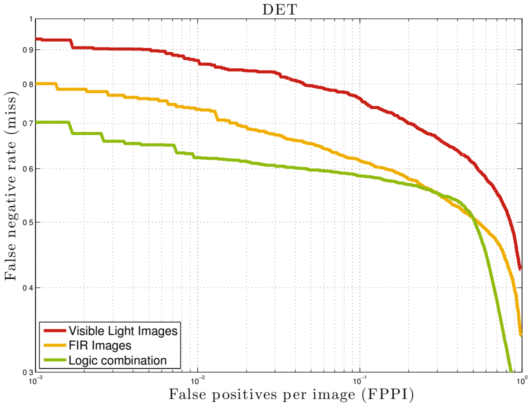
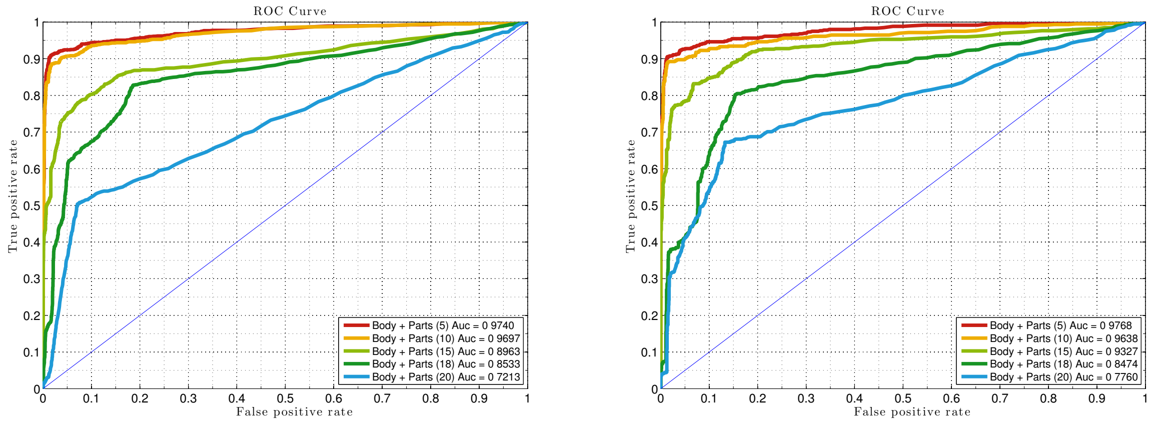


Fig. 4: Detection DET Curves obtained on the VL and FIR images, and the logic combination of both.

B. Results of the part-based models

The following results are based on set of bounding boxes containing occluded pedestrian, cropped from the full-sized images, and resized to a common size. The test dataset has been divided into five subsets based on the percentage of lateral occlusion. Figure 5 represents the true positive rate plotted against the false positives per image (FPPI) on the



(a) Roc curve of logic inference based on the fixed set of rules for different percentages of occlusion.

(b) Roc curve of logic inference of latent structure for different percentages of occlusion.

Fig. 5: Roc curves for occluded pedestrians using a full model and part based detection based on logic inference. Legend states occlusion and area under the curve (auc).

FIR database. Those are the results of applying the part-based model, where Fig. 5a are results of using the fixed set of logic rules, and Fig. 5b represent the results of the latent structure. For reference, Fig. 6 represents the roc curve of full body classification on the FIR database for different percentages of occlusion. It can be appreciated that, for large values of occlusion, classification degrades. In both part-based detection approaches, the classification results get significantly better for large values of occlusion, while it is slightly worse than the full-body approach for samples on which most or all of the pedestrian is visible. That being said, detection rates of largely occluded pedestrian are still not comparable with full body samples.

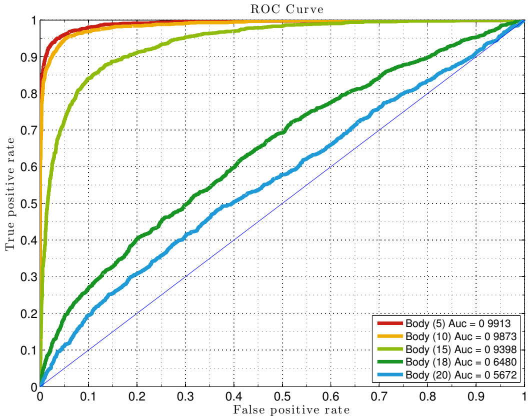


Fig. 6: Roc curve of full body classification for different percentages of occlusion. Legend states occlusion and area under the curve (auc).

Figure 7 represents the roc curves of each individual SVM trained on a part of the pedestrian. Only roc curves with an area under the curve $a_{uc} > thr$ are plotted. It is to be noticed that the classification performance of any of the parts is evidently inferior, when compared with the logic inference approach.

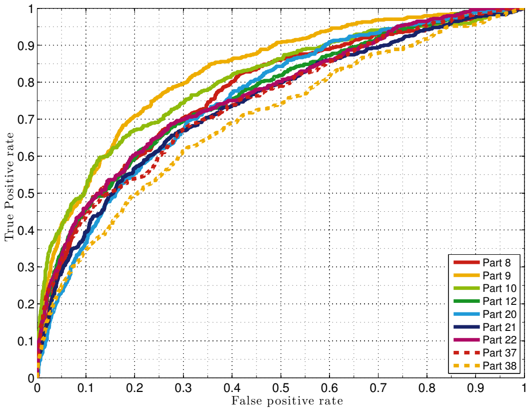


Fig. 7: Roc curves of the best performing part classifiers. Only parts classifiers with an area under the curve $a_{uc} > thr$ are plotted.

VII. Conclusions and Discussion

This paper presents a method of combining pedestrian detections on VL and FIR images based on a Markov Logic Network. Detection is inferred by training the network with a set of logic rules. This approach is extended to object detection in images under heavy occlusion based on the binary classification on subparts of the detection window.

From the experimental results reported in the previous sections it can be concluded that the combination of detections from cameras sensitive to different wavelengths helps to achieve better results. Specifically, the number of false positives is reduced in every case. The sequence in Fig. 8 shows an example of the output of the combination of both sensors.

The part-based method proposed increases the detection rate of largely occluded pedestrians. However, for full pedes-

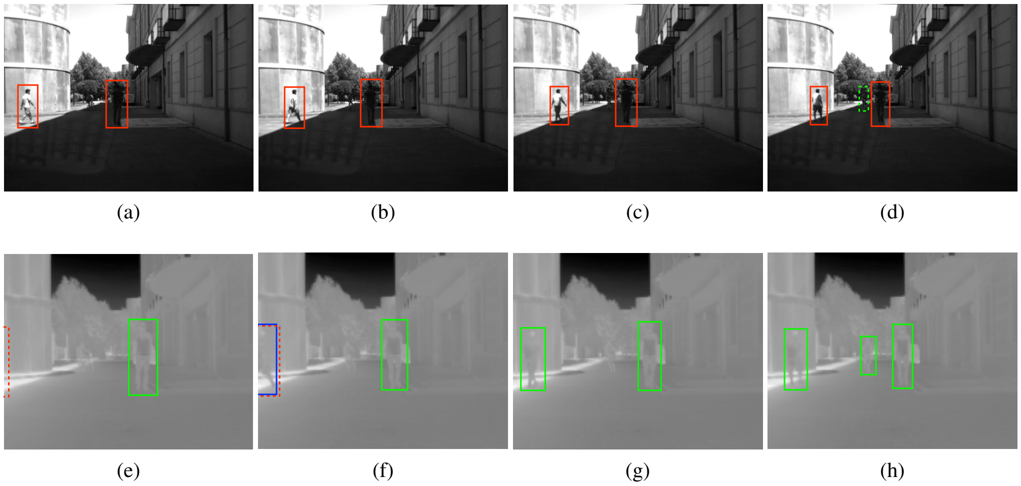


Fig. 8: Example sequence with color-coded detections. Only detection pairs with a positive output by the MLN are plotted. Positive detections in the VL images are plotted as a red rectangle, and as a green rectangle in the FIR images. If one of the detections in a pair is negative, the positive one is projected into that image and plotted using a dashed line. Occluded pedestrians are plotted as a blue rectangle.

trians, results degrade slightly. From these results, it may be concluded that this approach would be especially useful for pedestrians not entirely contained in the image, using a holistic approach for regions of interest fully contained in the image.

The results here presented are based on a general-purpose VL-FIR database, which contains images shot at different temperatures and light conditions. As future work, the authors propose training two models, one for daylight conditions and one for night driving. It is to be expected that, while driving at night, FIR detection would have higher weights, while VL detections would have higher weights on daylight conditions.

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*, pp.1239–1258, vol.32, no 7, 2010

[2] C. Premebida and U.J.C. Nunes, "Fusing LIDAR, camera and semantic information: A context-based approach for pedestrian detection," *The International Journal of Robotics Research*, 2013

[3] H. Kakiuchi, T. Kawamura, T. Shimizu, K. Sugahara, "Bypass methods for constructing robust automatic human tracking system," *Integrated Computer-Aided Engineering*, pp. 41–58, vol. 17, no. 1, 2010.

[4] C. Zeng and H. Ma, "Robust Head-Shoulder Detection by PCA-Based Multilevel HOG-LBP Detector for People Counting," *International Conference on Pattern Recognition*, pp. 2069–2072, 2010.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893 vol. 1, 2005.

[6] D. Olmeda, A. de La Escalera, and J. Armingol, "Contrast invariant features for human detection in far infrared images," *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pp. 117–122, 2012.

[7] G.Rogez, J. Rihan, C. Orrite-Urunuela and P.H.S. Torr, "Fast Human Pose Detection Using Randomized Hierarchical Cascades of Rejectors," *International Journal of Computer Vision*, 2012.

[8] S. Hussain, and B. Triggs, "Feature Sets and Dimensionality Reduction for Visual Object Detection," *British Machine Vision Conference*, 2010.

[9] M. Bertozzi, A. Broggi, M. del Rose, M. Felisa, A. Rakotomamonjy, and F. Suard, "A pedestrian detector using histograms of oriented gradients and a support vector machine classifier," *IEEE Intelligent Transportation Systems Conference*, 2007.

[10] R. O'Malley and M. Glavin,"A Review of Automotive Infrared Pedestrian Detection Techniques,"*Signals and Systems Conference. IET.*, 2008.

[11] D. Olmeda Reino, A. de La Escalera, and J. Armingol, "Far infrared pedestrian detection and tracking for night driving," *Robotica*, 2011.

[12] H. Nanda, "Probabilistic template based pedestrian detection in infrared videos," *IEEE Intelligent Vehicle Symposium*, 2002.

[13] J. Thomanek, M. Ritter, H. Lietz and G. Wanielik, "Comparing Visual Data Fusion Techniques Using FIR and Visible Light Sensors to Improve Pedestrian Detection," *IEEE International Conference on Digital Image Computing Techniques and Applications*, pp. 119–125, 2011.

[14] L. Oliveira and U. Nunes, "Context-aware pedestrian detection using LIDAR," *IEEE Intelligent Vehicle Symposium*, pp. 773–778, 2010.

[15] L. Oliveira, U. Nunes, P. Peixoto, M. Silva and F. Moita, "Semantic fusion of laser and vision in pedestrian detection," *Pattern Recognition*, pp. 3648–3659, vol. 43, no. 10, 2010.

[16] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 304–311, 2009.

[17] W. L. Fehlman II and M. K. Hinders, "Passive infrared thermographic imaging for mobile robot object identification," *Journal of Field Robotics*, 2009.

[18] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial & Applied Mathematics*, pp. 32–38, vol. 5, no. 1, 1957.

[19] P. Singla and P. Domingos, "Discriminative Training of Markov Logic Networks," *Proceedings of the 23rd national conference on Artificial intelligence*, pp. 868-873, 2005

[20] P. Singla and P. Domingos, "Lifted first-order belief propagation," *Proceedings of the 23rd national conference on Artificial intelligence*, pp. 1094–1099, vol. 2, 2008

[21] S. Kok and P. Domingos, "Learning the structure of Markov logic networks," *International Conference on Machine Learning*, pp. 441–448, 2005

[22] M. Everingham, L. Van Gool, C.K.I. Williams,J. Winn and A. Zisserman., "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, pp. 303–338, vol. 88, no. 2, 2010

[23] P. Felzenszwalb, D. McAllester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," *IEEE Conf. Computer Vision and Pattern Recognition*, 2008