Pedestrian Identification in Infrared and Visible Images Based on Pose Keypoints Matching

Raluca Brehar
Computer Science Dep.
Technical University of Cluj-Napoca, Romania
Raluca.Brehar@cs.utcluj.ro

Tiberiu Marita
Computer Science Dep.
Technical University of Cluj-Napoca, Romania
Tiberiu.Marita@@cs.utcluj.ro

Mihai Negru
Computer Science Dep.
Technical University of Cluj-Napoca, Romania
Mihai.Negru@cs.utclu.ro

Sergiu Nedevschi
Computer Science Dep.
Technical University of Cluj-Napoca, Romania
Sergiu.Nedevschi@cs.utcluj.ro

ABSTRACT

The identification of persons in multi-spectral images that are not spatially aligned is a challenging process. A correct identification can improve the pedestrian detection task for machine vision applications. In this context we propose a person identification mechanism able to correctly find the same person in infrared and visible images. The main contribution of the paper is the keypoint based matcher that uses a deep learning based solution for finding relevant human pose keypoints and a local neigbour search for extracting the best candidates for person identification. For each of the two images, infrared and color we first perform pedestrian detection. Next, on each detected instance we extract the relevant keypoints for shoulders, hands and legs. A matching algorithm between the extracted keypoints is proposed in order to perform the identification of persons in the two images. We obtain an identification accuracy of 76% for pedestrians that have a medium height with respect to the image dimensions, and have a small occlusion degree.

CCS CONCEPTS

•Computing methodologies ~ Object detection

•Computing methodologies ~ Object identification

•Computing methodologies~Matching

•Computing methodologies~Activity recognition and understanding

•Computing methodologies~Image representations

KEYWORDS

Pedestrian identification, Sensor fusion, Infrared sensors, Convolutional Neural Networks

ACM Reference format:

Raluca Brehar, Tiberiu Marita, Mihai Negru, and Sergiu Nedevschi. 2019. Pedestrian Identification in Infrared and Visible Images Based on Pose Keypoints Matching. In CCVPR 2019 . ACM, New York, NY, USA, 5 pages.

1 Introduction

Detection of pedestrians in infrared and visible domains has been actively explored in the last years. With the emergence of benchmark datasets such as FLIR-ADAS [6], KAIST [9] that contain temporally aligned color and infrared images of the same scene new algorithms have been proposed for object detection with exact focus on vulnerable road users (VRU) detection.

The difficulties that arise in detecting the vulnerable road users are due to the complex scenarios of the urban traffic, or to the dynamic shape and appearance which these users have. Fusion based approaches provide good results [7] for cases in which the images are both temporally and spatially aligned, as is the case of images from KAIST [9]. A spatial alignment means that there is a one to one correspondence between a pixel in the infrared

image and a pixel in the color image. But when images are not spatially aligned the fusion of detections is more difficult due to the lack of transformation information among the two image spaces (see Figure 1). The spatial alignment information could be provided by cross calibration procedures, and by knowing the bidirectional transformation between the infrared coordinate system and the color based coordinate system. That is at any moment know for a point in infrared image what is the corresponding point in RGB image and vice versa. This process is known as image registration and can be used for multi-spectral object matching and identification. As depicted in Figure 1 in order to correctly identify a person in both IR and RGB images we need to know the exact mathematical dual transformation between the two image spaces.

Figure 1: Infrared and RGB are temporally aligned but no spatial information is provided

In this work we study the detection of pedestrians in temporally aligned infrared and color images and we propose an original mechanism for bidirectional pedestrian identification in IR and RGB images. Our proposed method performs a pose based key-point registration and uses a nearest neighbor matching approach.

The paper is structured as follows: section 2 presents other existing approaches for keypoint based image matching and image registration. Next we describe the proposed solution in section 3 and we show the evaluation and experimental results in 4.

2 Related Work

The problem of identifying pedestrians in images that are not spatially aligned can be regarded as a problem of matching multi-spectral points from images. This process in known as registration. The registration between multi-spectral images (i.e. visible images (VIS) and long-wavelength infrared images (LWIR) is a very difficult task due to the non-linear relationship between the intensities of correspondent pixels: the intensity of the LWIR image pixels is proportional with the objects temperature while the intensity if the VIS pixels is given by the objects color end reluctance. Moreover infrared images exhibit lower contrast, lower intensity gradients and less texture details especially due to the thermal diffusion effect. There are two main approaches for registering such images: stereo camera based and monocular camera based.

If depth information from a visible stereo system is available, and the relative extrinsic parameters between one of the visible cameras (i.e. left) and the infrared camera are known by calibration, the pixels correspondence between the left visible camera and the infrared camera can be computed by using the projection matrix [3] or the trifocal-tensor [4]. In the monocular camera case no depth information is available and a homographic relation between the visible and infrared image planes can be established only in some particular cases that are rarely met in practice: pure rotation, all points in the scene are co-planar or are at infinity [8]. The solution in this case is to perform registration only on some objects/ patches that can be approximated as planar surfaces and then register the pixels between these objects by computing a homographic like transformation using some matched keypoints. Such a solution is presented in [2] in 4 steps: image rectification, sparse disparity map computation using a multi-modal cost function based on Mutual-Information (MI) [12] and gradients similarity, generation of planar hypotheses describing surfaces by split and merge segmentation, and combination of the results using the random Markov fields theory and the graph-cut algorithm. Another approach that uses the MI maximization as matching cost function to register objects (pedestrians) by disparity voting over rectified multi-modal images is presented [11]. However MI based keypoints matching robustness is dependent on the window size. In [13] the authors present a method for multispectral registration of humans in surveillance scenarios using the LSS (Local Self Similarity) feature vector for computing pixels correspondences. The advantages of the LSS feature vector based pixels matching approach over the MI metric are: is usable in a global matching scheme and the measurement unit for LSS is a small image patch that contains more meaningful patterns compared to a pixel as used for MI computation [13]. In [15] an image patches similarity measure

based on Convolutional Neural Networks is presented. The authors propose three architectures: 2-channel, Siamese and pseudo-Siamese) that are all able to outperform manually designed descriptors (e.g., SIFT, DAISY) or other learnt descriptors in the visible spectrum. In [1] the authors use a similar approach but train the networks on pairs of 64x64 visible-near infrared image patches, and test the results on both visible-near infrared and visible-far infrared datasets. They show that the CNN based cross spectral similarity measure outperforms the state-of the art in cross-spectral image descriptors. Although the CNN patch matching approaches are state-of-art in terms of performance, in terms of run time are slower than conventional hand-made approaches.

3 Proposed Solution

The proposed linear processing pipeline is shown in Figure 2. First both infrared and color images that capture the traffic scene are fed to a pedestrian detector, which provides human proposals or detections.

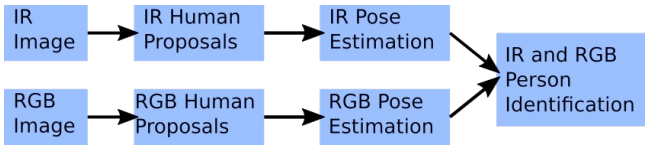

Figure 2: **Processing pipeline**

These proposals are further processed by a deep learning based pose estimation framework [5], [14]. From the estimated pose, the relevant keypoints are extracted and used to perform person identification based on a nearest neighbor approach in a weighted euclidean feature space.

3.1 Pedestrian Detection

The pedestrian detection process is based on [10]. It employs a YOLO[16] based pedestrian detector. The human proposals are fed to the spatial transformer network is used to generate improved regions of interest for the persons in both infrared and color images. As described by [10] the spatial transformer contains a localization network that by means of several hidden layers determines the parameters of the spatial transformation that is suitable for the input feature map. The obtained parameters are used to generate a sample grid which together with the feature map are fed to a sampling mechanism. This detection process is applied to both infrared and color images.

3.2 Pose estimation

As presented by [14] the result of the sampling process from the spatial transformer network STN is provided to a single person pose estimation (SPPE) convolutional neural network. Both STN and SPPE are fine tuned together [14] by comparing the output of SPPE with the labels of center-located ground truth poses. Figure 3 shows the poses obtained for the IR and color images that capture the same scene.

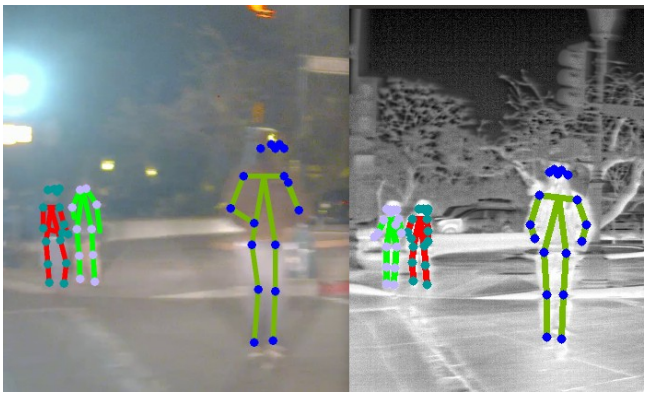

Figure 3: **RGB and Infrared pose estimation obtained with [14]**

3.3 Pose based matching algorithm

For each human proposal provided by Spatial Transformer Network we obtain a set of keypoints that describe the pose. As we can notice from Figure 3 and from our experiments the keypoints in the head area are not relevant in the matching process because the pedestrians are too far from the camera and the details on the head such as eyes and ears are detected less precisely. In our person identification model we use 12 keypoints on the shoulders, elbows, wrist, hip, knee and ankle plus a keypoint that corresponds to the middle of the shoulders. For each set of keypoints that corresponds to a human proposal we create a normalized euclidean feature space. First we extract the bounding box of the keypoints by computing the minimum and maximum coordinates. Hence we obtain a rectangle with parameters : (x_t , y_t w, h).

For each of the keypoints we perform a normalization in order to map them to an aligned

feature space. Because we need to match keypoints from infrared images and color images, that are different in size and were recorded with cameras having different parameters a keypoint K(x, y) part of an image having width and height, is normalized using:

$$x_n = x/width; y_n = y/height \quad (1)$$

In order not to match pedestrians that have relatively different locations in the infrared and color images, as shown in Figure 4 we have to ensure a local image search. We choose to match only the points that are close to each other in the normalized feature space of the infrared image with respect to the normalized feature space of the color image, that is the points having a distance less than a given threshold value. As one can notice from Figure 4 a person in the infrared image (bounded with the green rectangle) can be matched with a person in the color image (bounded with the blue rectangle) because their posture is quite similar. This is in fact a wrong match, because the person should be matched with the pedestrian in the color image marked with a dotted green line. Hence we reject all matches that are made between person proposals which are too far away from each other in the normalized feature spaces.

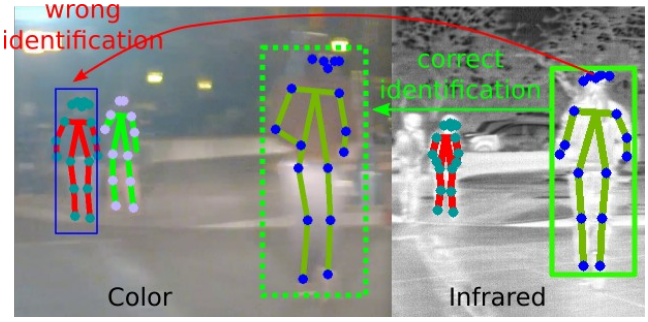

Figure 4: **Color and Infrared wrong person identification example**

The matching score used for person identification comprises the keypoints and also the width and height of the pedestrian bounding boxes. All these coordinates are considered in the normalized feature space. For a person in infrared image, the set of keypoints is: $K^{ir} = (x^{ir}_1, y^{ir}_1), (x^{ir}_2, y^{ir}_2) \ldots (x^{ir}_{13}, y^{ir}_{13})$. We also consider (w^{ir}, h^{ir}) the with and the height of the bounding box. The corresponding feature vector is also computed for the rgb images. The equation that returns the matching score is:

$$\sqrt{\left(x_1^{ir} - x_1^{rgb}\right)^2 + \ldots + \left(x_{13}^{ir} - x_{13}^{rgb}\right)^2} + \left|w^{rgb} - w^{ir}\right| + \left|h^{rgb} - h^{ir}\right| \quad (2)$$

It finds the closest points in the normalized Euclidean feature space of the two images. This corresponds to a one near neighbor search.

4 Experimental Results

All our experiments have been done on the FLIR-ADAS dataset [6] because it contains infrared and color images with annotations for pedestrians in the infrared space. We also enhance the annotations with bounding box pedestrian coordinates for color images and we provide a unique identifier to each person in the dataset.

For evaluating our solution we have divided the pedestrian annotations based on their height, into three classes:

• Large pedestrians: the detections in the color image have a height greater than half of the image height, respectively in the infrared image the detections have a height greater than half of the infrared image height.

• Medium pedestrians: are considered to be the detections which have a height in the range height image /2...h/4 (half image height and a quarter of the image height) for both infrared and color images.

• Small pedestrians: are the detections having a height of bounding box smaller than 0.25% of the image height in both infrared and color images.

Since the proposed person identification is based on a CNN detector, we briefly analyze the detection results. A detection is considered correct if the intersection over union (IoU) ratio between the detection and the ground truth bounding box is greater than 0.5. The accuracy of the YOLO-v3 based pedestrian detector is shown in Table 1, for both color and infrared images. We consider that at least 80% of the pedestrian body is visible.

As it can be noted from Table 1 best detection results are for medium size pedestrians. An important aspect in the detection process is the occlusion factor. If persons are occluded partially the body part keypoints cannot be found hence the average matching accuracy decreases.

Pedestrian Type	Average Accuracy IR	Average Accuracy RGB
Lage	72%	73.2%
Medium	82%	85%
Small	69%	68.54%

Table 1: **Detection accuracy**

In order to evaluate the accuracy of the identification we count how many correct matches are done by our proposed algorithm, when the human proposals are present in both infrared and color images. The results are shown in Table 2.

Pedestrian Type	Identification accuracy	Remarks
Large	72%	Occlusion less than 20%
Medium	76.2	Occlusion less than 20%
Small	67%	No occlusion

Table 2: **Person Identification Accuracy**

Best results are obtained in the case of medium pedestrians for which the occlusion factor is less than 20% (see Figure 5). For small pedestrians the identification accuracy is smaller because some of them are not visible by night (dark clothes) and too far away. A similar case is the one of thermal occlusions, when infrared pedestrians are not detected.

Figure 5: **Results - medium pedestrians**

In figure 6 we present the case where two small pedestrians are correctly matched, while the third which is even smaller and is barley visible in the color space, due to its dark clothing, is not matched with its correspondent in the infrared image.

Figure 6: **Results - small pedestrians**

In the case of large pedestrians, which have an occlusion less than 20% the identification is satisfactory (72% are identified correctly). The problem is with occlusions, as it can be noticed in figure 7 in which the person on the left is not matched because it has an occlusion of over 40% in the infrared image.

Figure 7: **Results - large pedestrians**

5 Conclusions

The paper presents a mechanism for identifying pedestrians in color and thermal images that are temporally aligned (they capture the scene at the same moment of time). The proposed model uses a set of reference keypoints that define the posture of the pedestrians and are relevant for shoulders, torso and legs. The pedestrians and their relevant keypoints are detected using deep learning methods. Based on these keypoints a normalized feature space is created for both infrared and color images and the matching is done via a nearest neighbor approach. The results show that an identification of 76% is made for pedestrians having an occlusion degree of less than 20% in both infrared and color images and their height being in the range greater than a quarter of the image height and less than half of the image height. The method can be extended to other relevant objects from traffic images.

ACKNOWLEDGMENTS

The results presented in this paper were partially supported in the framework of the following research projects: GnaC 2018 ARUT grant ”Detectia obiectelor in imagini monoculare termale FIR pentru viziune pe timp de noapte”, research Contract no. 3091/05.02.2019, with the financial support of the Technical University of Cluj-Napoca, partially supported in the framework of ”Multispectral environment perception by fusion of 2D and 3D sensorial data from the visible and infrared spectrum MULTISPECT”, CNCS-UEFISCDI, PN-III-P4-ID-PCE-2016-0727, grant no. 60/2017 and partially supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0917, contract 21PCCDI/2018, within PNCDI III.

REFERENCES

[1] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo. 2016. Learning Cross-Spectral Similarity Measures with Deep Convolutional Neural Networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 267–275. https://doi.org/10.1109/CVPRW.2016.40

[2] O Barrera, Felipe Lumbreras, and Cristhian Aguilera. [n.d.].Planar-Based Multispectral Stereo. In In Proc. Quantitative InfraRed Thermography. https://doi.org/10.21611/qirt.2012.172

[3] R. Brehar, C. Vancea, T. Maria, I. Giosan, and S. Nedevschi. 2015. Pedestrian detection in the context of multiple-sensor data alignment for far-infrared and stereo vision sensors. In 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). 385–392. https://doi.org/10.1109/ICCP.2015.7312690

[4] Z. Chen and X. Huang. 2019. Pedestrian Detection for Autonomous Vehicle Using Multi-Spectral Cameras. IEEE Transactions on Intelligent Vehicles 4, 2 (June 2019), 211–219. https://doi.org/10.1109/TIV.2019.2904389

[5] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In International Conference on Computer Vision.

[6] FLIR. [n.d.]. FLIR Thermal Datasets for Algorithm Training.https://www.flir.com/oem/adas/dataset/.

[7] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. 2019. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion 50 (2019), 148 – 157. https://doi.org/10.1016/j.inffus.2018.11.017

[8] Richard Hartley and Andrew Zisserman. 2003. Multiple View Geometry in Computer Vision (2 ed.). Cambridge University Press, New York, NY, USA.

[9] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. 2015. Multispectral Pedestrian Detection: Benchmark Dataset and Baselines. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. CoRRabs/1506.02025 (2015). arXiv:1506.02025 http://arxiv.org/abs/1506.02025

[11] Stephen J. Krotosky and Mohan M. Trivedi. 2007. Mutual information based registration of multimodal stereo videos for person tracking. Computer Vision and Image Understanding 106, 2 (2007), 270 – 287. https://doi.org/10.1016/j.cviu.2006.10.008 Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

[12] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. 1997. Multimodality image registration by maximization of mutual information. IEEE Transactions on Medical Imaging 16, 2 (April 1997), 187–198. https://doi.org/10.1109/42.563664

[13] Atousa Torabi and Guillaume-Alexandre Bilodeau. 2013. Local self-similarity-based registration of human ROIs in pairs of stereothermal-visible videos. Pattern Recognition 46, 2 (2013), 578 –589. https://doi.org/10.1016/j.patcog.2012.07.026

[14] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose Flow: Efficient Online Pose Tracking. In British Machine Vision Conference.

[15] Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to compare image patches via convolutional neural networks. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015), 4353–4361.

[16] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. ArXiv , 2018