#### 1

# Pedestrian Detection for Autonomous Vehicle Using Multi-spectral Cameras

Zhilu Chen and Xinming Huang, Senior Member, IEEE

Abstract-Pedestrian detection is a critical feature of autonomous vehicle or advanced driver assistance system. This paper presents a novel instrument for pedestrian detection by combining stereo vision cameras with a thermal camera. A new dataset for vehicle applications is built from the test vehicle recorded data when driving on city roads. Data received from multiple cameras are aligned using trifocal tensor with precalibrated parameters. Candidates are generated from each image frame using sliding windows across multiple scales. A reconfigurable detector framework is proposed, in which feature extraction and classification are two separate stages. The input to the detector can be the color image, disparity map, thermal data, or any of their combinations. When applying to convolutional channel features, feature extraction utilizes the first three convolutional layers of a pre-trained convolutional neural network cascaded with an AdaBoost classifier. The evaluation results show that it significantly outperforms the traditional histogram of oriented gradients features. The proposed pedestrian detector with multi-spectral cameras can achieve 9% logaverage miss rate. The experimental dataset is made available at http://computing.wpi.edu/dataset.html.

*Index Terms*—Multi-spectral camera, autonomous vehicle, pedestrian detection, machine learning

# I. INTRODUCTION

Automatic and reliable detection of pedestrians is an important function of an autonomous vehicle or advanced driver assistance system (ADAS). Research works on pedestrian detection are heavily depended on data, as different data and methods may yield different evaluation results. The most commonly used sensor in data collection is a regular color camera, and many datasets have been built such as the INRIA person dataset [1] and the Caltech Pedestrian Detection Benchmark [2]. Thermal cameras have also been considered lately, and different methods of pedestrian detection were developed based on the thermal data [3]. It is worth investigating whether the methods developed from one type of sensor data are applicable to other types of sensors. A method may not work anymore since the nature of data has changes, e.g., finding certain hot objects by intensity value threshold on thermal image is not applicable to a regular color image. Some methods such as gradient and shape based feature extraction may still be applicable since an object has similar silhouettes in both color and thermal images. In addition, data from different sensors may contain complementary information and combining them may result better performance. Multiple cameras can form stereo vision, which provides additional disparity and depth

information. An example of combining stereo vision color cameras and a thermal camera for pedestrian detection can be found in [4].

The data collection environment is also very important. Unlike static cameras for surveillance applications, cameras mounted on a moving vehicle may observe much more complex background and distance-varied pedestrians. Therefore, it calls for different pedestrian detection algorithms from the surveillance camera applications. To use multiple sensors on a vehicle, a cooperative multi-sensor system need to be designed and new algorithms that can coherently process multi-sensor data need to be investigated. The contributions of this paper are listed as follows:

- A multi-spectral camera instrument is designed and assembled on a moving vehicle to collect data for pedestrian detection. These data contain many complex scenarios that are challenging for detection and classification. The experimental dataset is made available at http://computing.wpi.edu/dataset.html.
- The multi-spectral data are aligned using trifocal tensor. It is then possible to combine features from different sources and compare their performance.
- A machine learning based algorithm is employed for pedestrian detection by combining stereo vision and thermal images. Evaluation results show satisfactory performance.

The rest of the paper is organized as follows. Section II provides a summary of related work. Section III describes our instrumental setup for data collection. In Section IV, we propose a framework that combines stereo vision color cameras and a thermal camera for pedestrian detection using different feature extraction methods and classifiers. Performance evaluations are presented in Section V, followed by further discussion in Section VI and conclusions in Section VII.

#### II. RELATED WORK

There are many existing works on pedestrian detection. The Caltech Pedestrian Detection Benchmark [2] has been widely used by the researchers. It contains frames from a single vision camera with pedestrians annotated. Based on the CVPR2015 snapshot of the results on the Caltech-USA pedestrian benchmark, it was stated in [5] that at ~95% recall, the state-of-the-art detectors made ten times more errors than the human-eye baseline, which is still a huge gap that calls for research attentions. Overall, the detector performance has been improved as new methods were introduced in recent

This work was supported by the U.S. NSF under Grant CNS-1626236. The authors are with Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA. The corresponding author is Xinming Huang (e-mail: xhuang@wpi.edu).

years. Traditional methods such as Viola–Jones (VJ) [6] and Histogram of Oriented Gradients (HOG) [1] were often included as the baseline. A total of 44 methods were listed in [7] for Caltech-USA dataset, and 30 of them made the use of HOG or HOG-like features. Channel features [8] and Convolutional Neural Networks [9]–[11] also achieved impressive performance on pedestrian detection. The Convolutional Channel Features (CCF) [12], which combines a boosting forest model and low level features from CNN, achieved as low as 19% log-average miss rate (MR) on Caltech Pedestrian Detection Benchmark. Despite the progressive improvement of detection results on the datasets, color cameras still have many limitations. For instance, color cameras are sensitive to the lighting condition. Most of these detection methods may fail if the image quality is impaired under poor lighting condition.

Thermal cameras can be employed to overcome some limitations of color cameras, because they are not affected by lighting condition. Several research works using thermal data for pedestrian detection and tracking were summarized in [3]. Background subtraction was applied in [13] for people detection, since the camera was static. HOG features and Support Vector Machine (SVM) were employed for classification [14]. A two-layered representation was described in [15], where the still background layer and the moving foreground layer were separated. The shape cue and appearance cue were used to detect and locate pedestrians. In [16], a window based screening procedure was proposed for potential candidate selections. The Contour Saliency Map (CSM) was used to represent the edges of a pedestrian, followed by AdaBoost classification with adaptive filters. Assuming the region occupied by a pedestrian has a hot spot, candidates were selected based on thermal intensity value [17] and then classified by a SVM. In addition, both Kalman filter prediction and mean shift tracking were incorporated for further improvement. A new contrast invariant descriptor [18] was introduced for far infrared images, which outperformed HOG features by 7% at  $10^{-4}$  FPPW for people detection. The Shape Context Descriptor (SCD) was also used for pedestrian detection in [19], followed by AdaBoost classifier. The HOG features were considered not suitable for this task because of the small size of the target, variations of pixel intensities and lack of texture information. Probabilistic models for pedestrian detection in far infrared images was presented in [20]. The method in [21] found the head regions at the initial stage, then confirmed the detection of a pedestrian by the histograms of Sobel edges in the region.

Stereo vision can provide additional information such as disparity map to better detect people in the frame. RGB-D cameras were used for indoor people detection or tracking in [22], [23], and stereo thermal cameras were used in [24] for pedestrian detection. and the image pixel registration was done using 3D point cloud. The combination of stereo vision cameras and a thermal camera was used in [4]. Trifocal tensor was used to align the thermal image with color and disparity images. Candidates were selected based on disparity, and HOG features were extracted from color, thermal and disparity images. Concatenated HOG features were then fed to radio basis function (RBF) SVM classifier to obtain the final decision. An indoor people detection system using stereo vision cameras and a thermal camera was presented in [25]. Instead of trifocal tensor, 3D point cloud projection was used for image point registration between thermal and color images.

For ADAS applications, pedestrian detection is often challenging because the camera is moving with the vehicle, and the pedestrians are often very small on images due to the distance and image resolution. Several pedestrian detection research works were summarized in [26], including the use of color cameras and thermal cameras, as well as sensor fusion such as radar and stereo vision cameras. A benchmark for multispectral pedestrian detection was presented in [27] and several methods were analyzed. However, the color-thermal pairs were manually annotated and it is unclear if any automatic point registration algorithms were used. Furthermore, more sophisticated applications or systems can be built upon pedestrian detection, such as pedestrian tracking across multiple driving recorders [28] and crowd movement analysis [29].

### III. DATA COLLECTION AND EXPERIMENTAL SETUP

#### A. Data Collection Equipment

To collect on-road data for pedestrian detection, we design and assemble a custom test equipment rig. This design enables the data collection system to be mobile on the test vehicle as well as maintaining calibration between data collection runs. The completed system can be seen in Figure 1.



Figure 1: Instrumentation setup with both thermal and stereo cameras mounted on the roof of a vehicle.

The stereo vision cameras called ZED StereoLabs are chosen for providing color images as well as disparity information. The ZED cameras can capture high resolution side-by-side video that contains synchronized left and right video streams, and can create a disparity map of the environment in real-time This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIV.2019.2904389, IEEE Transactions on Intelligent Vehicles

3

using the graphics processing unit (GPU) in the host computer. Furthermore, an easy to use SDK is provided, which allows for camera controls and output configuration. In addition, the on-board cameras are pre-calibrated and come with known intrinsic parameters. This makes image rectification and disparity map generation easier. The rectified images and the disparity map can be obtained by using the SDK, and point correspondence between the 2 stereo images can be calculated as  $x_{left} = x_{right} - disparity(x_{right}, y)$ , where  $(x_{left}, y)$  is the point location in the left image,  $(x_{right}, y)$  is the point location in the right image, and disparity() is the disparity value at the given location.

The thermal camera is called FLIR Vue Pro, which is a long wavelengths infrared (LWIR) camera. The IR camera is an uncooled vanadium-oxide microbolometer touting a  $640 \times 512$  resolution at a full 30 Hz and paired with a 13 mm germanium lens providing a  $45^{\circ} \times 35^{\circ}$  field of view (FOV). This IR camera has a wide  $-20^{\circ}$  to  $50^{\circ}$  operation range which allows for rugged outdoor use. The thermal camera also provides Bluetooth wireless control and video data recording via its on-board microSD card as well as an analog video output.

Both stereo vision and thermal cameras must remain fixed relative to each other for consistency of data collection. A threaded rod is custom cut to length and each end is threaded into the respective cameras tripod mounting hole. This provides a rigid connection between the color and thermal cameras. An electrical junction box is utilized as an appropriately sized, water proof box that provides high impact resistance. The top lid is replaced with an impact resistant clear acrylic sheet such that the stereo vision cameras can be situated safely behind it. A circular hole is cut into the top lid to accommodate for the thermal camera lens to fit through and mounted via the lens barrel. This is essential, as even clear acrylic would block most, if not all the IR spectrum that is used by the thermal camera.

The mounting system is designed, modeled, and built utilizing aluminum extrusions. The entire structure is completely portable and can be mounted to any vehicle with a ski rack. The aluminum extrusions can sit between the front and back ski rack hold-downs. On the other hand, cable management is crucial in our design as long cables are needed for communication between the laptop inside the vehicle and the cameras on the roof. To avoid interference and safety issues, the cables must run down the back of the vehicle, through the trunk and into the vehicle cabin, which needs approximately 20 feet of cable. This creates an issue for the ZED stereo vision cameras, as it operates on high speed USB 3.0 protocol that allows for a 10 feet maximum length due to signal degradation and loss. To resolve this issue, an active USB extension cable is used. A total of four cables terminated from the camera setup are wrapped together with braided cable sleeves to prevent tangling and ensure robustness.

An analog frame grabber is employed to capture the realtime analog output of the IR camera instead of directly recording to the on-board microSD card. It is to ensure proper synchronization between the thermal camera and stereo vision cameras. With analog frame grabber, we are able to precisely capture at 30 FPS. AVI files are generated using software provided along with the frame grabber. These AVI files are then converted into image sequences. The thermal images are synchronized with color images by software timestamps, and manual correction during post processing. We use this imperfect approach due to the limitation of our instruments and/or APIs provided by the SDK. A more reliable approach is by using trigger-based synchronization, as described in [30].

### B. Data Collection and Experimental Setup

Our dataset is made available online at http://computing. wpi.edu/dataset.html. The data are collected while driving on city roads. Highway driving data are not collected since pedestrians are hardly seen on highways. A total number of 58 data sequences are extracted from approximately three hours of driving on city roads across multiple days and lighting conditions. There are 4330 frames in total, in which a person or multiple people are in clear view and un-occluded. Figure 2 shows the histogram of our sample height, which indicates that more than half of the pedestrian samples in our dataset are less than 50 pixels in height, and that makes our dataset challenging. Each frame contains the stereo color images, thermal image and disparity map. Since cameras have different angle of view and field of view, the 58 usable sequences are rather short, ensuring the pedestrians are within the view of all cameras. Furthermore, video frames without any pedestrians are not included in our dataset. Currently, our dataset does not contain categorized scenarios such as sunny days, foggy days, etc. However, as more data becoming available, the dataset can be expanded to include those scenarios, and further analysis can be performed on it.



Figure 2: Histogram of Sample Height.

## IV. PROPOSED METHOD

## A. Overview

Figure 3 shows the flowchart of our proposed pedestrian detection method. Disparity data are generated from stereo color data. Thermal data are obtained from the thermal cameras and reconstructed according to the point registration using trifocal

4

tensor. By aligning data from multi-cameras, features can be extracted from each sensor using the same window or region of interests, which corresponds to the same real-world area or object. Instead of concatenating the features of different data sources and training a single classifier, feature extraction and classification are performed independently for each data source before the decision fusion stage. The decision fusion stage uses the confidence scores of the classifiers, along with some additional constraints to make the final decision. The proposed detector system can be reconfigured using different feature extraction and classification methods, such as HOG with SVM or CCF with AdaBoost. The decision fusion stage can utilize information from one or multiple classifiers. The performance of different configurations can be evaluated and compared.



Figure 3: Framework of the proposed pedestrian detection method.

#### B. Trifocal tensor

These three cameras have different angle of view and field of view, making the point registration (pixel level alignment) essential to windowed detection method cross multi-spectral images. Simple overlay with fixed pixel offsets does not work because every object has its own offset values depending on the distance to camera. Therefore, trifocal tensor [4], [31] is used for pixel level alignment over the color and thermal images. The trifocal tensor  $\mathcal{T}$  is a set of three  $3 \times 3$  matrices that can be denoted as  $\{T_1, T_2, T_3\}$  in matrix notation, or  $\mathcal{T}_i^{jk}$  in tensor notation [31] with two contravariant and one covariant indices. The idea of the trifocal tensor is that given a view point correspondence  $\mathbf{x} \leftrightarrow \mathbf{x}' \leftrightarrow \mathbf{x}''$ , there is a relation

$$[\boldsymbol{x}']_{\times} \left(\sum_{i} x^{i} \mathbf{T}_{i}\right) [\boldsymbol{x}'']_{\times} = \mathbf{0}_{3 \times 3}.$$
 (1)

One method to compute the trifocal tensor  $\mathcal{T}$  is by using the normalized linear algorithm. Given a point-point-point correspondence  $x \leftrightarrow x' \leftrightarrow x''$ , there is a relation

$$x^i x'^j x''^k \epsilon_{jqs} \epsilon_{krt} \mathcal{T}_i^{qr} = 0_{st}$$

where 4 out of 9 equations are linearly independent for all choices of s and t. The tensor  $\epsilon_{ijk}$  is defined for  $i, j, k = 1, \ldots, 3$  as follows:

0	unless $i, j, k$ are distinct
<b>{</b> +1	if $ijk$ is an even permutation of 123
-1	if $ijk$ is an odd permutation of 123

Therefore at least 7 point-point-point correspondences are needed to compute the 27 elements of the trifocal tensor. The trifocal tensor can be computed from a set of equations in the form of At = 0, using the algorithm for least-squares solution of a homogeneous system of linear equations.

Given the correct correspondence  $x \leftrightarrow x'$ , it is possible to determine the corresponding point x'' in the third view without reference to image content. It can be denoted as  $x''^{k} = x^{i} l'_{i} T^{jk}_{i}$  and can be obtained by using the trifocal tensor and fundamental matrix  $F_{21}$ . The line l' goes through x' and is perpendicular to  $\mathbf{l}'_e = \mathbf{F}_{21} \boldsymbol{x}$ . Both the trifocal tensor and fundamental matrix F<sub>21</sub> can be pre-computed and only need to be computed once as long as the placement of the cameras remains unchanged. An alternative method is epipolar transfer  $\mathbf{x}'' = (\mathbf{F}_{31}\mathbf{x}) \times (\mathbf{F}_{32}\mathbf{x}')$ . However, this method has a serious problem that it fails for all points lying on the trifocal plane. Therefore, trifocal tensor is a practical solution for point registration. In our experiment, trifocal tensor is estimated using a checkerboard. The pattern is made of different materials, making it visible in both color and thermal camera. Figure 4 shows the usage of trifocal tensor in aligning color and thermal images, including reconstructed thermal camera frames using trifocal tensor, aligned to left.



(a) Color image.

(b) Thermal image.



(c) Reconstructed thermal im- (d) Red-cyan anaglyph of color age using trifocal tensor and and reconstructed thermal imdisparity information. ages.

Figure 4: Proper alignment of color and thermal images using trifocal tensor.

#### C. Sliding windows vs. region of interest

There are two main methods to locate a pedestrian: sliding window detection and Region of Interest (ROI) extraction. In sliding window detection, it applies a small sliding window over the entire image, often in different scales, to perform an exhaustive search. Each window is classified followed by some post-processing, such as bounding box grouping. The ROI extraction finds out the potential candidates first by some pre-processing techniques such as color and pixel intensity to filter out negatives from these candidates by using a classifier or some other constraints. It is often more efficient, as the number of candidates is much less than the amount of sliding windows.

For pedestrian detection, both ROI extraction and sliding window detection have been employed in the literature. The sliding window detection method is an universal approach but is computationally expensive. On the other hand, ROI extraction is often used for thermal images, because pedestrians are often hotter than the surrounding environment. The ROIs are segmented based on the pixel intensity values. However, we find that the ROI extraction on thermal images does not always work well. The assumption that the pedestrians are hotter is not always true for various reasons. For instance, a pedestrian wearing heavy layers of clothing does not appear with distinctively high pixel intensity values in a thermal image, and thus a pedestrian can not be located by simple morphological operations. As another example, the temperature of the road surface exposed to intense sunlight has higher temperature than the human bodies. Although false positives introduced by hot objects such as vehicle engines can be filtered in later steps, the losses of true positives become a serious problem. As a result, we feel the sliding window detection method is more reliable in case of these complex scenarios. The classifier can analyze the windowed samples thoroughly and make an accurate decision. Figure 5 shows some examples of our pedestrian samples in color images and corresponding thermal images, where row 1 and 3 are color samples and corresponds to thermal samples in row 2 and 4, respectively.



Figure 5: Examples of pedestrians in color and thermal images.

However, sliding window detection method also has its own drawbacks, besides much higher computational cost. The total number of windows in an images often reaches  $10^5$  or more. Even a fair classifier with False Positives Per Window (FPPW)

of  $10^{-4}$  would still result 10 False Positives Per Image (FPPI). Since 2009, the evaluation metric has been changed from FPPW to FPPI [7]. To solve this problem, many state-of-the-art CNN-based classifier have been proposed in recent years. An alternative approach is to combine information from additional sensors. Our proposed approach of multi-spectral cameras is along this line.

# D. Detection

In this paper, we only compare the HOG and CCF methods for the task of pedestrian detection. The reason is explained in Section VI-A. The HOG method used in this paper is based on [1].

The HOG features have been widely used in object detection. It defines overlapped blocks in a windowed sample, and cells within blocks. The histogram of the unsigned gradients of several different directions are computed in all blocks, and are concatenated as features. The HOG features are often combined with SVM and sliding window method for detection on different scaling levels.

At the training stage, the positive samples are manually labeled. The initial negative samples are randomly selected on training images as long as they do not overlap with the positive samples. All samples are scaled to a standard window size of  $20 \times 40$  for training. The size of the minimum sample in our data is  $11 \times 22$ . After the initial training, the detector is tested on the training set and more false positives are added back to the negative samples set. These false positives are often called hard negatives and this procedure is often called hard negatives mining. This procedure can be repeated for a few times until the performance improvement becomes marginal.

Once the detector is trained, it is ready to perform detection on the test dataset and give a decision score for each window. Each frame with original size of  $640 \times 480$  is scaled into different sizes. The detector with a fixed size of  $20 \times 40$  is then applied to the scaled images to find pedestrians of various sizes at different locations in a frame.

CCF uses low level features from a pre-trained CNN model, cascaded with a boosting forest model such as Real AdaBoost [32] as a classifier. The lower level features from the first few CNN layers are considered generic descriptors for objects, which contain richer information than channel features. Meanwhile, the boosting forest model replaces the remaining parts of CNN. Thus we avoid training a complete end-to-end CNN model for a specific object detection application which would require large resources of computation, storage and time. In our experiment, we apply similar settings as described in [12], except for the parameters of the scales and number of octaves, in order to detect pedestrians far away that are as small as  $20 \times 40$  pixels. The conv3-3 layer in VGG-16 model are used as feature extraction. The windowed sample size is  $128 \times 64$ instead of  $20 \times 40$ . The feature dimension of the  $20 \times 40$  sample is 1296. The training samples of CCF are from the training stage of HOG, similar to the method described in [12] which use aggregated channel features (ACF) [33] to select training samples for CCF. Caffe [34] is used for feature extraction of CCF on a GPU-based computer platform. At the test stage,

6

CCF method runs on the GPU platform is considerably faster than the HOG method, but it requires more memory and disk space for data storage.

## E. Information fusion

The idea of combing the information from color image, disparity map and thermal data for decision making is referred as information fusion. One approach is to concatenate these features together [4]. A single classifier can be trained on the concatenated features and the final decisions of the test instances can be obtained from the classifier. This approach has an disadvantage that the classifier training becomes a challenge as the dimension of features increases. Furthermore, if a new type of feature needs to be added or an existing feature needs to be removed, the classifier need to be re-trained, which is time consuming.

An alternative approach of information fusion is to employ multiple classifiers and an example can be found in [35]. Each classifier makes decision on a certain type or subset of features and the final result is obtained by using a decision fusion technique such as majority voting or sum rule [36]. This approach has an advantage that the structure of a system is reconfigurable. Without re-training the classifiers, adding or removing different types of features becomes very convenient. Therefore, we choose the later approach to make our system reconfigurable so that it evaluates various settings and methods. Specifically, an SVM is used at the decision fusion stage and its inputs are confidence scores from classifiers in the previous stage, which is more appropriate than commonly used statistical decision fusion method in the case of multi-source data [37], [38]. The data from different sources are often not equally reliable, and so are the classifiers. The confidence scores must be weighted when obtaining the final decision from information fusion.

# F. Additional constraints

1) Disparity-size: Besides the extracted features from an image frame, additional constraints can be incorporated into the decision fusion stage to further improve the detector performance. An example is the disparity-size relationship. Figure 6 shows the disparity and height relationship of the positive samples in the form of a linear regression line  $d = \begin{bmatrix} h & 1 \end{bmatrix} \times B$ , where d is mean disparity, h is the height of the sample, and B is a  $2 \times 1$  coefficient matrix. Given a pair of mean disparity  $\hat{d}$  and height  $\hat{h}$  of a sample, the residual  $r = |\hat{d} - [\hat{h} \quad 1] \times B|$  can be used to estimate whether this sample is possibly a pedestrian or not.

From Figure 6 we can see a number of samples have very small mean disparity and are far below the regression line. This is because the disparity information is not accurate when an object is far away from camera. In fact, the stereo vision camera we use automatically clamps the disparity value at certain distance. Object beyond that distance results zero disparity, which makes the estimation for small size samples inaccurate.



Figure 6: The relationship between the mean disparity and the height of an object.

2) Road horizon: During detection, a few reasonable assumptions can be made to filter out more false positives while retaining the true positives. The assumptions vary depending on the application, including color, shape, position, etc. One assumption here is that pedestrians stand on the road, i.e., the lower bound of a pedestrian must below the road horizon. The road horizon can be automatically detected in an image. This kind of simple constraint may or may not improve the detector performance, and experiments should be carried out to determine its effectiveness.

## V. PERFORMANCE EVALUATION

There are a total of 58 labeled video sequences in our dataset. We use 39 of them for training and the remaining 19 for test. Figure 7 shows the performance of different settings, including disparity map, color image, thermal data, and their combinations, all based on HOG features. Generally, the more types of information are used, the better performance is achieved. The disparity-only setup performs the worst. The color image only is better, followed by the combination of color and disparity. Note that the thermal-only setup outperforms the combination of color and disparity. The heat signature of pedestrians seems more recognizable in thermal images. The combination of color, thermal and disparity information achieves the best performance, with about 36% log-average miss rate.

Figure 8 shows the performance of the HOG features, added with disparity-size information and road horizon constraint. The road horizon improves the log-average MR by about 5%. Despite little improvement provided by adding the disparity-size information alone, the combination both provides nearly 7% improvement in log-average MR.

Figure 9 shows the performance of different settings using CCF. Performance of disparity only is the worst. Thermal image performs very well. However, it is interesting to see the disparity does not provide any improvement when combined with color or thermal. In fact, combing with disparity results lower performance. This is due to the fact that CCF implementation accepts 8-bit image as input, thus the precision of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIV.2019.2904389, IEEE Transactions on Intelligent Vehicles



Figure 7: Performance of different input data combinations, all using HOG features.



Figure 8: Performance improvement by adding disparity-size and road horizon constraints.

the disparity is not accurate. In comparison, CCF outperforms HOG almost on all settings except for disparity. The best performance comes from CCF with the combination of color and thermal, which achieves 9% log-average MR. Similarly, we also attempt to add disparity-size information and road horizon constraint to the CCF method, but the performance changes are negligible, possibly because the CCF method is already performing very well at 9% log-average MR.

#### VI. DISCUSSION

#### A. Why HOG and CCF?

While there are more advanced deep learning networks that have better performance on Caltech-USA dataset, we only compare the HOG and CCF methods for the task of pedestrian detection for the follow reasons:

- The HOG method was always included as a baseline in Caltech-USA dataset. Among 44 methods reported on the Caltech-USA dataset [7], 30 of them employed HOG or HOG-like features.
- The CCF achieved good performance on Caltech-USA dataset. The idea of combining low level CNN feature



Figure 9: Performance of different input data combinations, all using CCF.

and a boosting forest model avoids training a CNN from end to end, which requires huge amount of data and is time consuming. The advantage of CCF is especially obvious in this paper, when our dataset is relatively small, and different combinations of features are used as input data. Training different versions of CNN to find the best combination can be done when more data become available in the future.

3) The goal of this paper is to investigate the combination of multi-spectral cameras and its improvement on pedestrian detection. We publicize our dataset, so other researchers can continue this study to discover many better solutions in the future.

#### B. How to interpret the results?

As shown in Figure 9 and explained in Section V, the best performance is achieved when combining color and thermal data, and introducing disparity as additional feature does not improve the performance. However, this does not mean the disparity information is useless, nor stereo vision is unnecessary. As described in Section IV-B, trifocal tensor must be employed to align the thermal and color data, which requires disparity information. It is impossible to align the color data with the thermal data using a single color camera and a single thermal camera, because the entire image cannot be transformed using point matching techniques due to the difference of color and thermal data in nature.

On the other hand, the performance is still highly dependent on the instrument. Our thermal camera has a resolution of  $640 \times 480$ , which is relatively low. To accommodate the resolution and FOV of the thermal camera, the color cameras have to be set to the same resolution. In addition, color cameras are sensitive to the lighting condition, therefore the quality of the image sometimes cannot be guaranteed. Figure 10 shows an example, with bounding box drawn on the detected pedestrian in both color and thermal images. It is obvious that the thermal image provide much better information about the presence of the pedestrian, while it is hardly identifiable in the color image due to the shadow.



Figure 10: A pedestrian is embedded in the shadow of a color image.

Although thermal images seem to be dominant in our experiment, its reliability still needs improvement. Figure 11 shows a thermal image taken on a hot sunny day. Two pedestrians circled are not bright enough compared to the surroundings, which is contradictory to the assumption of distinct thermal intensity in many existing research works. In this case, the methods or operations based on pixel intensity values become unreliable, such as intensity thresholding, head recognition using hot spot, etc. On the contrary, some shape or gradient based methods may still perform well, such as HOG and CCF described in this paper.



Figure 11: An example thermal image with two pedestrians.

Finally, it is worth noting that possible camera parameters estimation errors may have an impact on the performance. The feature extraction across images requires accurate image point registration, which can't be done without accurate camera parameters. Therefore, the possible camera parameters estimation errors should be minimized during the calibration stage, possibly by using more point pairs over a set of images.

### VII. CONCLUSIONS

In this paper, a novel pedestrian detection instrumentation is designed using both thermal and RGB-D stereo cameras. Data are collected from on-road driving and an experimental dataset is built with pedestrians labeled as ground truth. A reconfigurable multi-stage detection framework is proposed. Trifocal tensor is used to align data from multiple cameras. It is then possible to combine features from different sources and compare their performance. Both HOG and CCF based detection methods are evaluated using the multi-spectral dataset with various combinations of thermal, color, and disparity information. The experimental results show that CCF significantly outperforms the HOG features. The combination of color and thermal images using CCF method results the best performance of 9% log-average miss rate. For the future work, other advanced feature extraction and classification methods will be considered to further improve the pedestrian detector performance.

#### REFERENCES

- N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, June 2005, pp. 886–893 vol. 1.
- [2] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 304–311.
- [3] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications*, vol. 25, no. 1, pp. 245–262, 2014. [Online]. Available: http://dx.doi.org/10.1007/s00138-013-0570-5
- [4] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodalstereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 4, pp. 619–629, Dec 2007.
- [5] S. Zhang, R. Benenson, M. Omran, J. H. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" *CoRR*, vol. abs/1602.01237, 2016. [Online]. Available: http://arxiv.org/abs/1602. 01237
- [6] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb
- [7] R. Benenson, M. Omran, J. H. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" *CoRR*, vol. abs/1411.4304, 2014. [Online]. Available: http://arxiv.org/abs/1411.4304
- [8] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," pp. 91.1–91.11, 2009, doi:10.5244/C.23.91.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/ 4824-imagenet-classification-with-deep-convolutional-neural-networks. pdf
- [10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556
- [11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), June 2015.
- [12] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features for pedestrian, face and edge detection," *CoRR*, vol. abs/1504.07339, 2015. [Online]. Available: http://arxiv.org/abs/1504.07339
- [13] W. Li, D. Zheng, T. Zhao, and M. Yang, "An effective approach to pedestrian detection in thermal imagery," in 2012 Eighth International Conference on Natural Computation (ICNC), May 2012, pp. 325–329.
- [14] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in 2006 IEEE Intelligent Vehicles Symposium, 2006, pp. 206–212.
- [15] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 288 – 299, 2007, special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum. [Online]. Available: http://www.sciencedirect.com/ science/article/pii/S1077314206001925
- [16] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," *IEEE Workshop on Applications* of Computer Vision and the IEEE Workshop on Motion and Video Computing, vol. 1, pp. 364–369, 2005.
- [17] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63–71, March 2005.
- [18] D. Olmeda, A. de la Escalera, and J. M. Armingol, "Contrast invariant features for human detection in far infrared images," in 2012 IEEE Intelligent Vehicles Symposium (IV), June 2012, pp. 117–122.

- [19] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in 2010 IEEE International Conference on Image Processing, Sept 2010, pp. 2313–2316.
- [20] M. Bertozzi, A. Broggi, C. H. Gomez, R. I. Fedriga, G. Vezzoni, and M. DelRose, "Pedestrian detection in far infrared images based on the use of probabilistic templates," in 2007 IEEE Intelligent Vehicles Symposium, June 2007, pp. 327–332.
- [21] T. T. Zin, H. Takahashi, and H. Hama, "Robust person detection using far infrared camera for image fusion," in 2007 Second International Conference on Innovative Computing, Information and Control, Sept 2007, pp. 310–310.
- [22] L. Spinello and K. O. Arras, "People detection in rgb-d data," in 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, Sept 2011, pp. 3838–3843.
- [23] M. Munaro, F. Basso, and E. Menegatti, "Tracking people within groups with rgb-d data," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Oct 2012, pp. 2101–2107.
- [24] M. Bertozzi, A. Broggi, A. Lasagni, and M. D. Rose, "Infrared stereo vision-based pedestrian detection," in *IEEE Proceedings. Intelligent Vehicles Symposium*, 2005., June 2005, pp. 24–29.
- [25] I. R. Spremolla, M. Antunes, D. Aouada, and B. E. Ottersten, "Rgb-d and thermal sensor fusion-application in person tracking." in VISIGRAPP (3: VISAPP), 2016, pp. 612–619.
- [26] 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, July 2010.
- [27] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 1037–1045.
- [28] K. H. Lee and J. N. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1429–1438, Sept 2015.
- [29] W. Liu, R. W. H. Lau, X. Wang, and D. Manocha, "Exemplar-amms: Recognizing crowd movements from pedestrian trajectories," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2398–2406, Dec 2016.
- [30] R. Brehar, C. Vancea, T. Mariţa, I. Giosan, and S. Nedevschi, "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), Sept 2015, pp. 385–392.
- [31] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [32] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999. [Online]. Available: http://dx.doi.org/10.1023/A: 1007614523901
- [33] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, Aug 2014.
- [34] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," arXiv preprint arXiv:1408.5093, 2014.
- [35] M. Rohrbach, M. Enzweiler, and D. M. Gavrila, "High-level fusion of depth and intensity for pedestrian classification," in *Joint Pattern Recognition Symposium*. Springer, 2009, pp. 101–110.
- [36] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar 1998.
- [37] B. Waske and J. A. Benediktsson, "Fusion of support vector machines for classification of multisensor data," *IEEE Transactions on Geoscience* and Remote Sensing, vol. 45, no. 12, pp. 3858–3866, Dec 2007.
- [38] R. Pouteau, B. Stoll, and S. Chabrier, "Support vector machine fusion of multisensor imagery in tropical ecosystems," in 2010 2nd International Conference on Image Processing Theory Tools and Applications (IPTA), July 2010, pp. 325–329.



Zhilu Chen received the B.E. degree in microelectronics from Xi'an Jiaotong University, Xi'an, China, in 2011, and the M.S. degree in electrical and computer engineering from Worcester Polytechnic Institute (WPI), Worcester, MA in 2013. He is currently pursuing the Ph.D. degree from the Electrical and Computer Engineering Department, Worcester Polytechnic Institute, Worcester, MA. His research interests are computer vision, machine learning and GPU acceleration for Advanced Driver Assistance Systems.



Xinming Huang (M'01–SM'09) received the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA in 2001. He is currently a Professor in the Department of Electrical and Computer Engineering at Worcester Polytechnic Institute (WPI), Worcester, MA. Previously he was a Member of Technical Staff with Bell Labs of Lucent Technologies, from 2001 to 2003. His research interests are in the areas of integrated circuits and embedded systems, with emphasis on reconfigurable computing, wireless communications, information security,

computer vision, and machine learning.