Multispectral Pedestrian Detection: Benchmark Dataset and Baseline

Soonmin Hwang Jaesik Park Namil Kim Yukyung Choi In So Kweon
Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[smhwang,jspark,nikim,ykchoi]@rcv.kaist.ac.kr, iskweon77@kaist.ac.kr

Abstract

With the increasing interest in pedestrian detection, pedestrian datasets have also been the subject of research in the past decades. However, most existing datasets focus on a color channel, while a thermal channel is helpful for detection even in a dark environment. With this in mind, we propose a multispectral pedestrian dataset which provides well aligned color-thermal image pairs, captured by beam splitter-based special hardware. The color-thermal dataset is as large as previous color-based datasets and provides dense annotations including temporal correspondences. With this dataset, we introduce multispectral ACF, which is an extension of aggregated channel features (ACF) to simultaneously handle color-thermal image pairs. Multispectral ACF reduces the average miss rate of ACF by 15%, and achieves another breakthrough in the pedestrian detection task.

1. Introduction

Pedestrian detection is an active research area in the field of computer vision, since it is an essential and significant task for a surveillance or tracking system [6, 14, 24, 27, 28], as well as for pedestrian safety [9, 13, 15]. Although many researchers have studied various methods for a long time, pedestrian detection is still regarded as a challenging problem, limited by tiny and occluded appearances, cluttered backgrounds, and bad visibility at night. In particular, even though color cameras have difficulty getting useful information at night, most of the current pedestrian detectors are based on color images.

To address these challenges for automobile applications, generally two types of infrared sensors are used: near infrared ($0.75 \sim 1.3\mu$m) cameras or long-wavelength infrared ($7.5 \sim 13\mu$m, also known as the thermal band) cameras. Physically, pedestrians are more visible in thermal cameras than in near infrared cameras. This is because long-wavelength infrared cameras are more robust to the interferences produced by headlights and traffic signals. Even more importantly, a human body radiates in the long-wavelength infrared light wavelength of 9.3μm [25], which supports the suitability of thermal cameras for capturing humans.

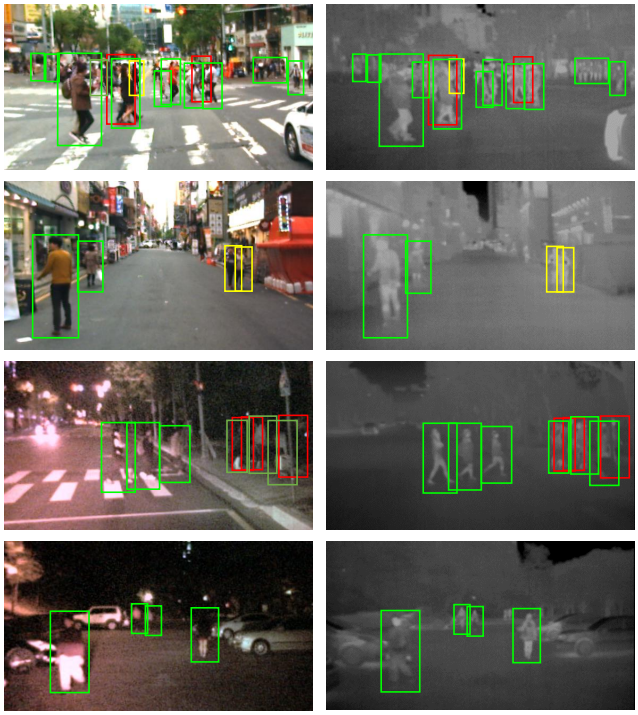


Figure 1. Examples of our *multispectral pedestrian dataset*. It has aligned pair of color (left column) and thermal (right column) images captured from day/night traffic scenes. The dense annotations provided with the dataset such as green, yellow, and red boxes indicate no-occlusion, partial occlusion, and heavy occlusion respectively. Images are cropped for better visualization.

Based on these facts, in this paper we introduce a *multispectral pedestrian dataset*[1] which provides thermal image sequences of regular traffic scenes as well as color image sequences. This work is motivated by other computer vision datasets such as Caltech 101 [19], Oxford buildings [23], Caltech pedestrian [10], and so on. These datasets have been contributed to stimulate their respective research fields. Likewise, our multispectral dataset designed to sup-

[1]Our multispectral pedestrian dataset is available online:
http://rcv.kaist.ac.kr/multispectral-pedestrian/

port the study of appropriate use of color-thermal images and to ultimately improve the accuracy of pedestrian detection.

Our contributions are threefold: (1) We introduce the multispectral pedestrian dataset, which provides aligned color and thermal image pairs. Our dataset has number of image frames as large as widely used pedestrian datasets [10, 15]. The dataset also contains nighttime traffic sequences which are rarely provided or discussed in previous datasets. (2) We analyze the complementary relationship between the color and thermal channels, and suggest how to combine the strong points of the two channels instead of using the color or thermal channel independently. (3) We propose several combinations of extended ACF with the thermal channel. One of our extensions reduces the average miss rate by 15% on the proposed multispectral pedestrian dataset.

In constrast to most previous datasets utilizing a color-thermal stereo setup, we use beam splitter-based hardware to physically align the two image domains. Therefore, our dataset is free from parallax and does not require an image alignment algorithm for post processing. To the best of our knowledge, this is the first work that provides aligned color and thermal image pairs captured in day and night. Examples of our dataset are shown in Fig. 1.

We introduce our new dataset and analyze its statistics in Sec. 2. With our new dataset, in Sec. 3 we discuss the extensions of ACF [9] to handle additional information from the aligned thermal image. In Sec. 4, we evaluate the effectiveness of the additional channel in various conditions by means of pedestrian detection performances. Lastly, we summarize our findings and suggest future directions with our dataset.

2. Multispectral Pedestrian Dataset

This section introduces our imaging hardware and the calibration procedure for capturing the aligned multispectral images. An analysis of the dataset is described in the next section.

2.1. Imaging Hardware

Hardware specification. As shown in Fig. 2, we developed imaging hardware consisting of a color camera, a thermal camera, a beam splitter, and a three-axis camera jig. The beam splitter in the system transmits the thermal band of the incident light and reflects the visible band. It also helps the optical center of the two cameras to be coincident. The beam splitter is made of Zinc coated Silicon wafer for this optical purpose. We used the PointGrey Flea3, a global shuttered color camera and the FLIR-A35 thermal camera. The color camera has 640×480 pixels of spatial resolution with a 103.6° vertical field of view. The thermal camera has 320×256 pixels of spatial resolution with a 39° vertical field of view. Note that the color camera has a larger field of view than the thermal camera. We intended to use the original thermal image in the aligned image domain by sacrificing the border area of the color image. The frame rate of the two cameras is equal to 20 fps.

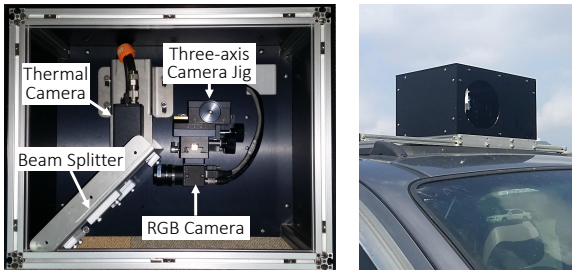


Figure 2. Our hardware configuration for capturing multispectral images. (Left) Top view. (Right) Mounted on the rooftop of a car.

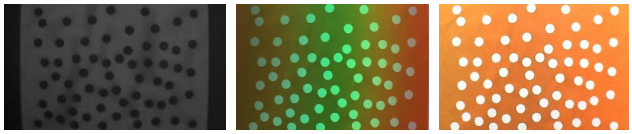


Figure 3. A hole pattern used for the color-thermal camera calibration. (Left) thermal image. (Middle) color image having color distortion due to beam splitter. (Right) after color correction.

Camera calibration. The concept for our hardware was previously introduced by Bienkowski *et al*. [3] for nondestructive evaluation purposes. Since the calibration method for aligning two image domains is not mentioned in [3], we briefly introduce our calibration approach here. First, we compute a translation between the two cameras mounted on the hardware using stereo camera calibration. Here, we can regard that the optical axes of the two camera beyond the beam splitter are parallel due to the hardwired arrangement. Therefore, there is only translation between the two image domains, and we only adjust the camera positions using the three-axis jig until the translation becomes zero. After the adjustment, the two image domains are rectified to have the same virtual focal length. After these procedures, the two image domains share the same focal length and the same principal point and there is no baseline. The virtually aligned image domain has 640×512 pixels of spatial resolution, and has a 39° vertical field of view, which is as similar to human vision. As a conventional checker board pattern is not observable in a thermal camera, we used a special calibration board [16, 17] having a number of holes. When it is heated, there is a temperature difference between the board and holes, which are therefore observable in the thermal camera. Examples of the hole pattern images are shown in Fig. 3.

Color correction. The captured color image shows color

Figure 4. Image pairs captured by our hardware. (Left) thermal image. (Middle) color image. (Right) blending of the two images.

distortion because the reflection ratios of the visible band from the beam splitter are uneven depending on the incident light directions (shown as Fig. 3). To handle this problem, we capture a reference which is an image of a white plane but showing the color distortion. As our color camera has a linear camera response function, the reference image is equivalent to the per-pixel reflection coefficient of the visible band. Therefore, we alleviate the color distortion by dividing the intensity level of captured images with these reflection coefficients.

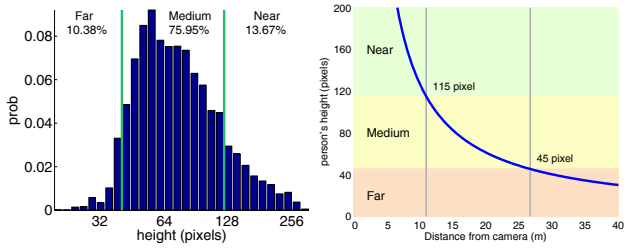
2.2. Data Collection and Ground truth

Data capture. The hardware is mounted on the roof of a car, and used for capturing ego-centric images of the traffic scenes. In particular, we captured various scenes at day and night time to consider changes in light conditions. An example of proposed dataset is shown in Fig. 4.

Ground truth annotation. Among the grabbed frames, 95,328 color-thermal pairs[2] were manually annotated for the total of 103,128 dense annotations and 1,182 unique pedestrians. To annotate the ground truth, we used Piotr's Computer Vision Toolbox[8], but it was modified to display color and thermal images simultaneously. The modification helps with annotation because a distant pedestrian at nighttime is rarely observable in the color channel. We also modified the toolbox to give occlusion tags instead of occlusion regions for each bounding box. Similar to Dollár *et al*. [10], the object has one of four labels. Obviously an individual pedestrian was labeled as a *person*. Not distinguishable individuals were labeled as *people*. People riding a two-wheeled vehicle were labeled as *cyclist*. In a highly cluttered scene, even human annotators sometimes cannot clearly determine whether a human shaped object is a pedestrian or not. This object is labeled as *person?* and it is ignored in the evaluation[3]. After the annotation, the bounding boxes also have temporal correspondences indicating the person index over the frames. In our dataset, a person appears 74.80 frames on average (corresponds to 3.74 seconds).

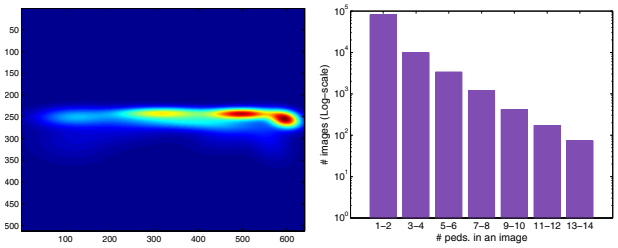
[2]Note that the frame number was not enlarged by horizontal mirroring. However, in the training stage of the baseline algorithm, we mirrored the positive samples to make more general examples.

[3]In our dataset, the number of *person?* is only 1,434 (1.66%) compared to a total of 86,152 *person* annotations. Therefore, it does not significantly affect the reliability of the evaluation result.



(a) Distribution of peds. height (b) Distance vs height

Figure 5. We define evaluation conditions with respect to the scale. Considering the general driving speed and the braking distances, we set 45 pixel and 115 pixel to the criteria of scale conditions. More than 75 % of pedestrians belong to medium scale. It means the detection algorithms try to focus on this condition.



(a) Distribution of peds. centers (b) Number of peds. per frame

Figure 6. (a) Due to the right-handed traffic condition, most pedestrians appear at the right side of the image. (b) The proposed dataset contains a lot of crowded scenes.

Train and Test sets. To divide annotated color-thermal image pairs into train and test datasets, we used the following criterion. First, the numbers of pedestrians appearing in the two sets were similar. Second, the frame numbers of day/night images in the two sets were similar. Third, the two sets were not overlapped. Compared to random division, this scheme helps to avoid data bias and over-fitting on a certain scene.

2.3. Properties of Dataset

Scale. Since the key application of pedestrian detection is accident avoidance, we classified the size of annotated bounding boxes based on the braking distance of the vehicles. In urban areas where pedestrians usually appear, we regarded the general driving speed as $30 \sim 55$ km/h. The expected braking distances under this driving condition are $11 \sim 28$ m (including braking delay due to the reaction of drivers) [7]. That corresponds to $45 \sim 115$ pixels of height in our aligned image domain if the height of the pedestrians is around 1.7m. We classified the annotations within these sizes as *medium*. As shown in Fig. 5(a), *near* and *far*, which are smaller or larger than *medium*, were also determined. Figure 5(b) shows the relation between the pedestrian's height in pixel units and its corresponding distance in meters.

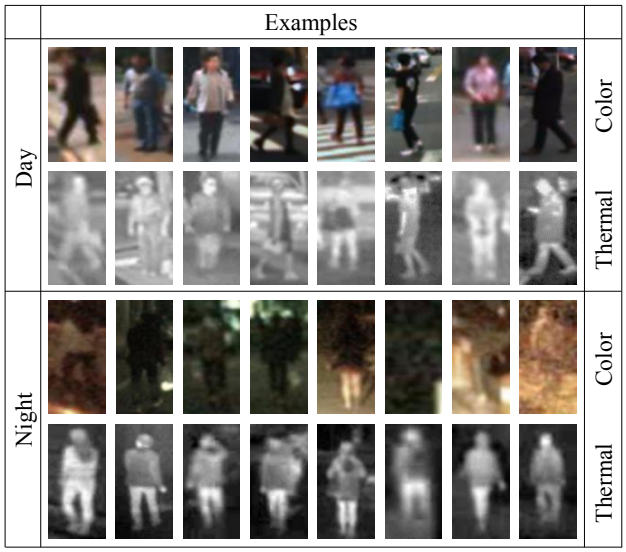


Figure 7. Examples of annotated pedestrians with *no occlusion* tag. It shows color and thermal image pairs at day and night times.

Occlusion. If a pedestrian is suddenly occluded by other pedestrians or objects in the scene, we annotated it with one of the three occlusion tags. Pedestrians who were never occluded were tagged as *no occlusion*; those occluded to some extent up to one half were tagged as *partial occlusion*; and those whose contour was mostly occluded were tagged as *heavy occlusion*. Among the total annotations, over 75% of pedestrians were tagged not occluded (78.6%), and the remainder were partial occlusion (12.6%) and heavy occlusion (8.8%).

Position. Figure 6 (a) shows the center of annotated pedestrians represented as the distribution of a Gaussian mixture model. Our hardware was set up to cover the view of a general driver. This setup constrains the appearance of pedestrians in certain regions. Therefore, pedestrians were distributed in a narrow band across the center of the image. Pedestrians mostly appear at the right side of the image, because the car drives under the right-handed traffic condition. We also show the number of pedestrians per frame in log-normalized scale in Fig. 6 (b).

Appearance change. Figure 7 shows several examples of pedestrians in the day and night time. The color image in daytime shows a distinct human shape due to the strong sunlight. On the other hand, the shape in the color image at nighttime is not distinguishable due to the dark environment. However, the thermal image shows a distinct shape at nighttime, because the temperature difference is greater when the air temperature is cooler, so the pedestrians having a fixed temperature, can be clearly captured in the nighttime. In the daytime, the strong sun radiation causes background clutters. For these reasons, we can expect to obtain better performance by using the strong points of the color and the thermal images throughout the day.

	Training		Testing		Properties								
	# pedestrians	# images	# pedestrians	# images	# total frames	occ. labels	color	thermal	moving cam.	video seqs.	temporal corr.	aligned channels	publication
INRIA [4]	1.2k	1.2k	566	741	2.5k		✓			✓			‘05
ETHZ [14]	2.4k	499	12k	1.8k			✓			✓			‘08
Daimler [13]	15.6k	6.7k	56.4k	21.8k	28.5k		✓		✓	✓			‘09
Caltech [10]	192k	128k	155k	121k	250k	✓	✓		✓	✓	✓		‘09
KITTI [15]	12k	1.6k	–	–	80k	✓	✓		✓	✓			‘12
OSU-T [5]	984	1.9k	–	–	0.2k			✓		✓			‘05
LSI [21]	10.2k	6.2k	5.9k	9.1k	15.2k			✓	✓	✓			‘13
ASL-TID [24]	–	5.6k	–	1.3k	4.3k			✓		✓			‘14
TIV [28]	–	–	–	–	63k			✓		✓			‘14
OSU-CT [6]	–	–	–	–	17k		✓	✓		✓		✓	‘07
LITIV [27]	–	–	16.1k	5.4k	4.3k		✓	✓		✓		✓	‘12
Ours	41.5k	50.2k	44.7k	45.1k	95k	✓	✓	✓	✓	✓	✓	✓	‘15

Table 1. Comparison of several pedestrian datasets. The horizontal lines divide the image types of the dataset (color, thermal, and color-thermal) based on the image types. The first four columns indicate number of pedestrian and images in the training and testing dataset ($k = 10^3$). Properties column summarizes additional characteristics of the datasets. Note that our dataset is largest color-thermal dataset providing occlusion labels and temporal correspondences captured in a non-static traffic scene.

2.4. Comparison to Existing Datasets

Table 1 provides a summary of existing pedestrian datasets. According to the image type, the datasets are classified into: color, thermal, and color-thermal.

Most of the existing color datasets [4, 10, 13, 14, 15] provide color image sequences captured in daytime under fine weather conditions. Caltech [10] and KITTI [15] in particular are the most widely used datasets having various real driving scenarios. Caltech [10] has the largest number of frames in the video format. They also have temporal correspondences of the bounding boxes, which give an identification index over the frame in the same target. KITTI [15] is used for validating various computer vision applications such as stereo vision, optical flow, visual SLAM, and object detection using color images only.

Thermal datasets [5, 21, 24, 28] are usually designed for object detection and tracking. The OSU-T dataset [5] is made for benchmarking tracking algorithms, and some datasets provide a trajectory instead of a bounding box [24, 28]. Olmeda *et al.* [21] provides a pedestrian detection dataset captured by thermal camera on a moving vehicle. Notable benchmark, referred to as TIV [28], provides multi-view or multi-resolution image sequences, and have annotated labels such as person, bat, runner, bicycle, motorcycle,

and car. In addition, TIV [28] provides a high resolution thermal image (up to 1024×1024) and provides the largest number of frames among the thermal datasets.

Our approach is classified as color-thermal dataset as it provides aligned color and thermal images. Compared to [6, 27], our dataset has an ego-centric moving view of the traffic scene, and provides a much larger number of annotated frames. In addition, our approach provides temporal correspondences and occlusion labels, which are useful information for pedestrian detection, identification, and tracking. Our setup is also related to the pedestrian detection system [18] which consists of a pair of color cameras and a infrared camera. However, compared to our system, it requires additional stereo matching of the color images and aligns color-thermal image pairs using a trifocal tensor.

3. Baseline Approaches

To handle the color and thermal channels effectively, our baseline algorithm is built upon the aggregated channel features (ACF) pedestrian detector [9]. This is natural choice because the algorithm can accommodate multiple channels showing different modalities. For instance, it uses chromatic and gradient channels augmented from a single color image. In this manner, the thermal channel can be regarded as another augmented channel in this algorithm. Here, we benefit from our capturing hardware because the alignment problem between color and thermal channels is removed. In addition, the ACF pedestrian detector [9] is widely used as a basis algorithm for the concurrent state-of-the-art pedestrian detectors [20, 22, 29].

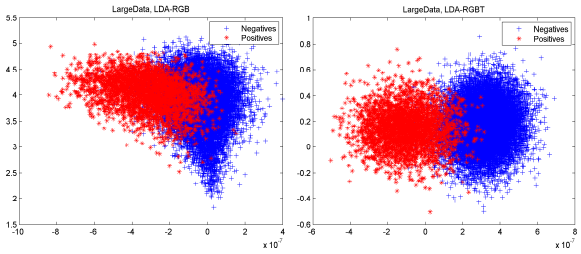
With this idea, we first review standard ACF designed for color images and introduce our extension to additionally handle the thermal channel.

3.1. Standard ACF

For color image input, the standard ACF [9] have 10 augmented channels (LUV+M+O): LUV denotes 3 channels of CIELUV color space, M denotes 1 channel of gradient magnitude, and O denotes 6 channels of gradient histogram which is a simplified version of histogram of oriented gradients (HOG) [4]. In ACF [9], they utilize the bootstrapping procedure which is to mine hard negatives among a tremendous number of negatives, and re-train the AdaBoost classifier [1] several times. Finally, they apply an efficient rejection method called soft cascade to boost detection time. In this manner, a powerful pedestrian detection framework is constructed.

3.2. Multispectral ACF

We utilized the ACF pedestrian detector [9] as our baseline and extended it to encode the thermal intensity channel. For the extension, we suggest three baselines as follows: (1) ACF+T (2) ACF+T+TM+TO (3) ACF+T+THOG. Here,



(a) ACF (b) Multispectral ACF

Figure 8. Linear discriminant analysis (LDA) on the ACF and multispectral ACF features. The high dimensional features are projected into 2D domain using LDA. In the above two figures, red and blue dots indicate positive and negative samples respectively. By adding thermal channels, the positive and negative samples become more distinctive.

ACF is the aforementioned feature defined for the color channel. T, T+TM+TO, and T+THOG indicates the additional channel features augmented from the thermal channel. The individual explanations follow.

T. This channel feature uses the thermal intensity directly. To improve the detection performance, we enhanced the contrast of the image using histogram equalization.

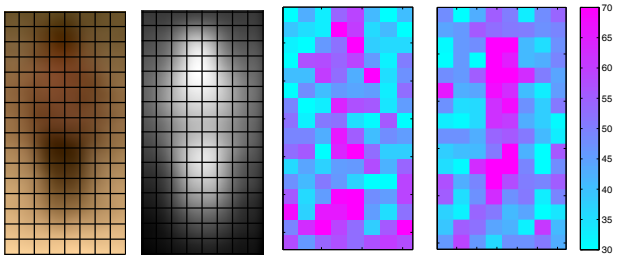
T+TM+TO. This extension consists of three channels: T, TM and TO. T is the aforementioned thermal channel, TM is the normalized gradient magnitude of the thermal images, TO is the histogram of oriented gradients of the thermal images. The TM and TO are acquired from the same method as standard ACF.

T+THOG. This extension uses the T and HOG feature [4] of the thermal image (denoted as THOG). Compared to TO which computes 6 directions of histograms, THOG computes more gradient orientations and has additional normalization steps on the local histograms.

Note that the three extensions utilize the intensity and gradient information of the thermal channel. We were motivated by recent work [26] which utilized gradients of a thermal image as an important cue. We self-evaluated these extensions on various conditions: different scales, occlusion tags, and capturing time (day or night). The result in Fig. 10 indicates that the three extensions outperforms ACF, and ACF+T+THOG shows the best performance. This is because ACF+T+THOG has most elaborate representation of the human shape. Based on this observation, we selected ACF+T+THOG as a desirable combination for the channel feature, and we name it *multispectral ACF* for the ramainder of this paper.

3.3. Analysis of Multispectral ACF

We compared the multispectral ACF to standard ACF to observe the benefits that resulted from the thermal channel.



(a) Avg. using color images (b) Avg. using thermal images (c) Voting map (ACF) (d) Voting map (multispectral ACF)

Figure 9. Average images of annotated pedestrians using (a) color channel and (b) thermal channel. (c, d) voting maps indicating frequently used feature grid. The voting map of multispectral ACF is more concentrated around the upper part of human body.

Distinctiveness of multispectral ACF. For the qualitative analysis, we trained the AdaBoost classifier [1] using ACF and multispectral ACF, respectively. After the training, each classifier had two groups of features belonging to either the hard negative class or the positive class. Here, we applied linear discriminant analysis (LDA) [11] to the two groups of features to visualize their distribution. LDA finds an optimal projection vector which minimizes a variance of the same class and maximizes a variance of the different classes. In this manner, features are projected into 2D spaces, and one of the results are shown in Fig. 8.

For the quantitative analysis, we introduce the following steps to measure the distance between positive and negative features. The analysis was separately applied to the two feature groups obtained from ACF and multispectral ACF. First, for each of the feature groups, we applied k-means clustering [12]. Second, we made histograms of positive and negative features by binning the corresponding cluster labels. Third, we measured the distance between the positive and negative histograms. As a result, the Bhattacharyya distance [2] using multispectral ACF was found to be larger (0.65) than the distance using the ACF (0.43). This implies that the multispectral ACF shows more distinctiveness than the ACF.

Frequently used local features. Figure 9 (a-b) show the average images of the positive samples. In our baseline algorithm [9], weak classifiers select few cells in the regular grid of the bounding box and classify positive and negative samples using the most discriminant cell. Based on the average image, the classifier can be regarded as well-trained if the features around the human shape regions are frequently used. To observe locations of frequently used features, we made voting maps in the regular grid as shown in Fig. 9 (c-d). In Fig. 9 (c), a learned model with color images (using ACF) has many features located outside of the human shape. This is caused by the significant background clutter, which is common in color images. On the other hand, the voting map of the multispectral ACF, shown in Fig. 9 (d), displays gathered votes in the upper part of the human shape. This visualization implies that the multispectral ACF can build a more spatially concentrated feature set.

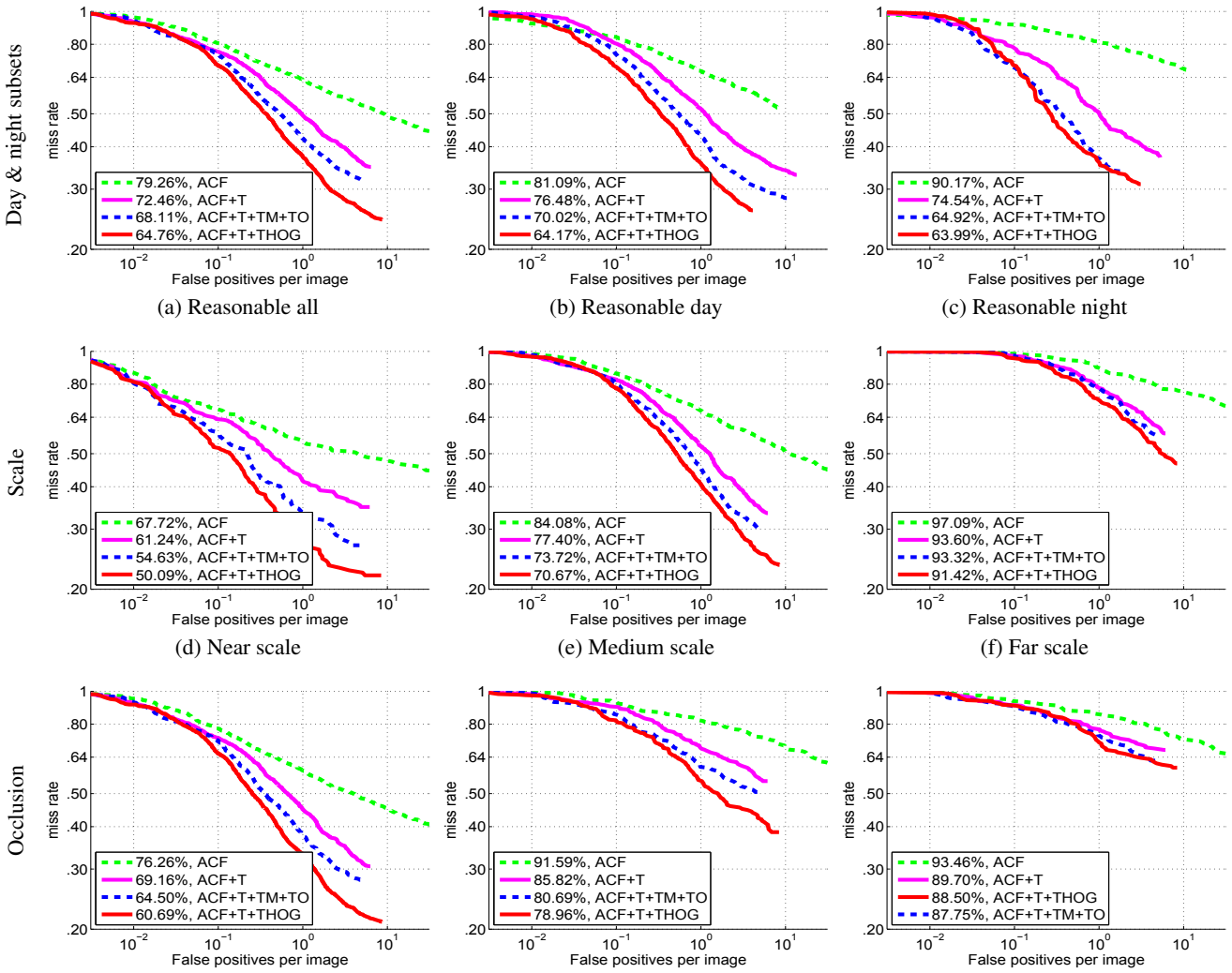
4. Experimental Results

To measure the effectiveness of the thermal channel in various conditions, we evaluated the ACF and its extended candidates as described in Sec. 3.2. For all these experiments, the detectors were trained and tested on the proposed dataset using the public ACF implementation[8]. Since our focus was on evaluating the effect of the thermal channel, the parameters were fixed in all experiments. We plotted the miss rate using a per-image evaluation scheme (FPPI) and summarized the performance with a single value by using log-average miss rate over the range of $[10^{-2}, 10^{0}]$ as suggested by Dollar et al.[10] Figure 10 shows the evaluation results for the various subsets of the test set described below.

Day and night. For this experiment, we used a subset named *reasonable* which is a representative subset of the proposed dataset. The *reasonable* subset consists of not/partially occluded pedestrians which are larger than 55 pixels. The dataset is divided into *reasonable day* and *reasonable night* based on the capturing time. In Fig. 10 (a), all three extensions using a color-thermal channel performed better than ACF using only the color channel. This is valid regardless of daytime or nighttime as shown in Fig. 10 (b), (c). Apparently, in case of the nighttime when the pedestrian is hardly distinguishable in the color image, the thermal channel seems to be dominant at detecting pedestrians.

Scale. In this experiment, we examined trained detectors using three subsets of the dataset which were defined based on the size of the bounding box. As shown in Fig. 6, these were classified into near ($\sim$28 m, 115 pixels$\sim$), medium (11$\sim$28 m, 45$\sim$115 pixels) and far (28 m$\sim$, $\sim$45 pixels). These subsets contain non-occluded pedestrians captured in daytime and nighttime. As shown in Fig. 10 (d)-(f), the multispectral ACF generally outperforms ACF on the three scales. In general, as the height of a pedestrian gets smaller, the miss rate gets larger. Our detector also follows this tendency (*near*: 50.09%, *medium*: 70.67% and *far*: 91.42%). Interestingly, the performance gap between ACF and multispectral ACF gets larger if the scale increases (*near*: 17.63%, *medium*: 13.41% and *far*: 5.67%). We believe this is due to the low-resolution of the thermal camera, which can capture a human shape better if the pedestrian is not too distant.

Occlusion. For this experiment, we made three subsets based on the occlusion tags: no-occlusion, partial-occlusion ($\sim$50% of area occluded) and heavy occlusion (50%$\sim$ of



(a) Reasonable all (b) Reasonable day (c) Reasonable night

(d) Near scale (e) Medium scale (f) Far scale

(g) No occlusion (h) Partial occlusion (i) Heavy occlusion

Figure 10. False positive per image (FPPI) versus miss rate in various conditions. Our multispectral ACF is denoted as ACF+T+THOG.

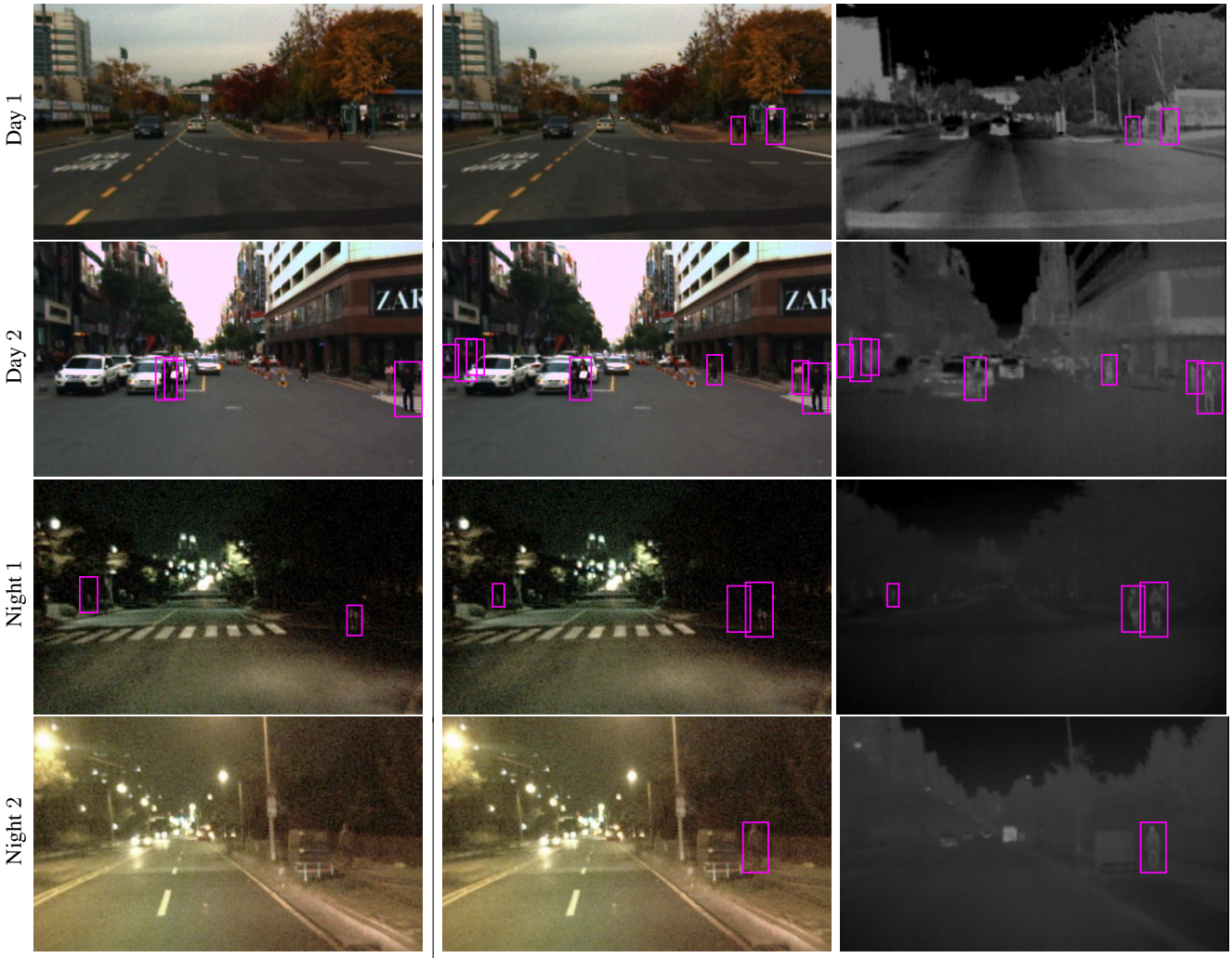
the area occluded). The dataset contains daytime and nighttime images with various scales. The evaluation results are shown in Fig. 10 (g)-(i). The performance of our multispectral ACF acceptably degrades as the occlusion level increases. However, the performance of stadard ACF drops significantly even under partial occlusion. This implies that the additional thermal channel is helpful in case of partial occlusion as well.

Summary. Through our experiments, we validated the effectiveness of the joint use of color-thermal images. Our trained detectors showed consistent improvements for all conditions compared to the color image based detector. The thermal image was helpful even when visual information was lacking, in far scale (Fig. 10 (f)) or occluded cases (Fig. 10 (h), (i)).

5. Conclusion

We introduced an multispectral pedestrian dataset of real traffic scenes. The proposed dataset has rich information involving thermal images, various real traffic scenes, lots of annotations with occlusion tags, and temporal correlations of the annotations. In addition, we thoughtfully defined the subsets of the dataset in terms of day/nighttime, braking distances, and occlusion levels.

We analyzed the effects of the thermal channel with regard to the distinctiveness of channel features. We also verified our extension of ACF, called multispectral ACF, in various conditions. Through the experiments, we determined that the aligned multispectral images are very helpful for resolving pedestrian detection problems in various conditions. We expect that the proposed dataset can encourage the development of better pedestrian detection methods.



(a) Detection using ACF [9] (trained by color images)

(b) Detection using our multispectral ACF (trained by both color (left) and thermal (right) images)

Figure 11. Examples of detection results at FPPI 1. (a) Detection results of ACF [9] trained by color images only. (b, c) Detection results of the multispectral ACF (described in Sec. 3.2) trained by both color and thermal images. The ACF misses some pedestrians which are hard to distinguish from background and produce some false positives. On the contrary, the multispectral ACF can detect pedestrian correctly even in the challenging nighttime images. Images are cropped for better visualization.

Acknowledgement: We thank anonymous reviewers giving constructive comments to our work. We also appreciate KAIST-RCV labmates who help to finish the tedious annotation task. This work was supported by the Development of Autonomous Emergency Braking System for Pedestrian Protection project funded by the Ministry of Trade, Industry and Energy of Korea. (MOTIE)(No.10044775)

References

[1] R. Appel, T. Fuchs, P. Dollar, and P. Perona. Quickly boosting decision trees – pruning underachieving features early. In *International Conference on Machine Learning (ICML)*, 2013.

[2] A. Bhattacharyya. *On a measure of divergence between two statistical populations defined by their probability distributions*. Bulletin of the Calcutta Mathematical Society, 1943.

[3] L. Bienkowski, C. Homma, K. Eisler, and C. Boller. Hybrid camera and real-view thermography for nondestructive evaluation. *Quantitativ InfraRed Thermography*, 254, 2012.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[5] J. Davis and M. Keck. A two-stage approach to person detection in thermal imagery. In *Proceeding of Workshop on Applications of Computer Vision (WACV)*, 2005.

[6] J. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Com-*

puter Vision and Image Understanding, 106(2–3):162–182, 2007.

[7] Department for Transport (DfT) Driver and Vehicle Standards Agency (DVSA). *The Official Highway Code*. TSO (The Stationery Office), United Kingdom, 2007.

[8] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(8):1532–1545, 2014.

[10] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] R. O. Duda, P. E. Hart, and D. H. Stork. *Pattern Classification (2nd ed.)*. Wiley Interscience, 2000.

[12] C. Elkan. Using the triangle inequality to accelerate k-means. In *International Conference on Machine Learning (ICML)*, 2003.

[13] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(12):2179–2195, 2009.

[14] A. Ess, B. Leibe, K. Schindler, and L. van Gool. A mobile vision system for robust multi-person tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] J. Jung, Y. Jeong, J. Park, H. Ha, J. D. Kim, and I.-S. Kweon. A novel 2.5d pattern for extrinsic calibration of tof and camera fusion system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011.

[17] J. Jung, J.-Y. Lee, Y. Jeong, and I. S. Kweon. Time-of-flight sensor calibration for a color and depth camera pair. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Accepted.

[18] S. J. Krotosky and M. M. Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8:619–629, 2007.

[19] F.-F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28:594–611, 2006.

[20] W. Nam, P. Dollár, , and J. H. Han. Local decorrelation for improved pedestrian detection. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[21] D. Olmeda, C. Premebida, U. Nunes, J. Armingol, and A. de la Escalera. Pedestrian classification and detection in far infrared images. *Integrated Computer-Aided Engineering*, 20:347–360, 2013.

[22] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[24] J. Portmann, S. Lynen, M. Chli, and R. Siegwart. People detection and tracking from aerial thermal views. In *Proceeding of IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[25] L. St-Laurent, X. Maldague, and D. Prévost. Combination of colour and thermal sensors for enhanced object detection. In *Information Fusion, 2007 10th International Conference on*, pages 1–8. IEEE, 2007.

[26] M. Teutsch, T. Mller, M. Huber, and J. Beyerer. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In *Proceedings of IEEE International Conference of Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014.

[27] A. Torabi, G. Mass, and G.-A. Bilodeau. An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116:210–221, 2012.

[28] Z. Wu, N. Fuller, D. Theriault, and M. Betke. A thermal infrared video benchmark for visual analysis. In *Proceeding of 10th IEEE Workshop on Perception Beyond the Visible Spectrum (PBVS)*, 2014.

[29] S. Zhang, C. Bauckhage, and A. B. Cremers. Informed haar-like features improve pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.