sensors MDPI

Article

Probabilistic Fusion for Pedestrian Detection from Thermal and Colour Images

Zuhaib Ahmed Shaikh *, David Van Hamme, Peter Veelaert and Wilfried Philips

TELIN-IPI, Ghent University–imec, 9000 Ghent, Belgium
* Correspondence: zuhaib.ahmed@ugent.be

Abstract: Pedestrian detection is an important research domain due to its relevance for autonomous and assisted driving, as well as its applications in security and industrial automation. Often, more than one type of sensor is used to cover a broader range of operating conditions than a single-sensor system would allow. However, it remains difficult to make pedestrian detection systems perform well in highly dynamic environments, often requiring extensive retraining of the algorithms for specific conditions to reach satisfactory accuracy, which, in turn, requires large, annotated datasets captured in these conditions. In this paper, we propose a probabilistic decision-level sensor fusion method based on naive Bayes to improve the efficiency of the system by combining the output of available pedestrian detectors for colour and thermal images without retraining. The results in this paper, obtained through long-term experiments, demonstrate the efficacy of our technique, its ability to work with non-registered images, and its adaptability to cope with situations when one of the sensors fails. The results also show that our proposed technique improves the overall accuracy of the system and could be very useful in several applications.

Keywords: sensor fusion; probabilistic fusion; naive Bayes; decision-level fusion

check for updates

Citation: Shaikh, Z.A.; Van Hamme, D.; Veelaert, P.; Philips, W. Probabilistic Fusion for Pedestrian Detection from Thermal and Colour Images. *Sensors* **2022**, *22*, 8637. https://doi.org/10.3390/s22228637

Academic Editor: Jean-Baptiste Thomas

Received: 11 October 2022
Accepted: 7 November 2022
Published: 9 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Pedestrian detection has been the focus of research in the field of computer vision over the last decades due to applications such as security, autonomous vehicles, and intelligent traffic management, to name a few [1]. For such systems, many heterogeneous sensors such as LiDAR, high-resolution cameras, thermal cameras, GPS, and others are used for several different purposes, such as pedestrian/cyclist/vehicle detection and route planning and estimation [1,2]. Visual sensors play a pivotal role in these systems as they provide much more information compared to other sensors. Nevertheless, no single visual sensor can deal with a dynamic environment, which includes varying lighting conditions due to time of day, varying weather conditions such as rain, fog, and snow, and temperature variation, necessitating the need for heterogeneous visual sensors. The variants of visual sensors come with their own merits and limitations. For instance, a thermal visual sensor can sense heat signatures of objects despite environmental complexity, but it can only provide limited information due to its inability to cover the entire visible spectrum of light. On the other hand, an RGB visual sensor provides much more information by covering a wider visible spectrum of light, but depends on several environmental conditions such as the amount of light, etc. In this regard, the American Automobile Association (AAA) evaluated several available solutions and concluded that none of them were good enough to detect pedestrians properly during difficult conditional environments, especially during the night [3]. Hence, the operation of these systems in a dynamic environment requires exploiting a combination of heterogeneous visual sensors to address the problems pertaining to changing environmental conditions [2,4]. This requires fusing data from heterogeneous visual sensors and addressing multiple related issues such as geometric alignment of visual data, different fields of view (FoVs), varying data-capture rates, and different resolutions.

The techniques of image data fusion specifically from heterogeneous visual sensors studied in the literature can be broadly classified into three main categories [5,6]. Pixel-level or early fusion combines several visual sensors' data at a pixel level. It is easy to interpret and better for visual appearance. However, the accuracy of fusion is degraded by a noisy sensor [7]. Furthermore, the fused data further requires feature extraction or classification. By contrast, feature-level or middle fusion is performed on the features extracted from visual sensor data. These methods can deal with a noisy sensor, as the rich features are available from another sensor(s). However, this method requires the design of a new classifier and a large training set [8]. Finally, fusion can be performed at the decision-level by considering the classified data from visual sensors; this is also known as late fusion. These methods can deal with the situation when a sensor is noisy or unavailable, as the classification is done independently, and classification errors are therefore uncorrelated. Furthermore, lower-level processing blocks can be optimised separately. A prerequisite of efficient pixel- and feature-level image fusion is the geometric alignment of images [9,10]. However, when using heterogeneous sensors with different positions, FoVs, and resolutions, the accuracy of the image registration process may be insufficient for these two types of fusion. Thus, there is a need for an efficient technique that can fuse the image data from heterogeneous visual sensors without the requirement of very accurate image alignment.

Decision-level fusion is more suitable in this respect, as it omits the need for accurate image alignment, can be more efficient in selecting information than a pixel fusion approach that treats all channels equally, has the capability to deal with noisy sensors, and allows independent optimization of the feature extraction and classification. Furthermore, there is no need to design a new classifier with the requirement of a large training set for multi-model input .

On the other hand, CNN-based fusion methods in recent years have shown notable progress in performance for multi-spectral pedestrian detection. Illumination-aware Faster R-CNN [11] adaptively merges colour and thermal sub-networks to obtain confidence scores defined over the illumination values via a gate function. Similarly, illumination-aware weights for fusion can be predicted using a gate function based on the illumination measure [12]. On the other hand, a confidence-aware fusion method [13] uses the confidence scores of the network to estimate the weights of each instance and effectively fuses the multi-modal information using those dynamic weights. The authors in [14] designed an illumination-aware feature alignment module to align two modality features and allow the network to be optimised adaptively according to illumination conditions.

Two-stream architectures with concatenated RGB-thermal feature maps used in recent studies have achieved significant improvements. Nevertheless, this dependency can cause a substantial loss in fusion performance if one of the inputs is unavailable or if a sensor fails. Moreover, the performance of the state-of-the-art fusion methods is strongly influenced by the quality of the registration between the thermal and colour images. Furthermore, illumination-aware methods usually consider only the variation of light in the colour images, while ignoring environmental changes in the thermal images.

Large changes in the images caused by illumination or other environmental factors clearly affect the performance of a detector. However, depending on the type of sensor, changes in appearance are caused by different environmental factors. For example, sufficient illumination is important for a colour camera, while it barely affects a thermal sensor, for which the temperature of a scene and object is paramount. This difference in conduct can be mitigated by carefully modelling the behaviour of the detectors in various environmental conditions, and taking this discrepancy into account in the fusion process.

In this regard, we introduce a probabilistic late-fusion method based on appearance models for colour and thermal images, which takes into account differences in light and temperature. We propose a way to take this context into account and choose an easy-to-measure method to evaluate its effectiveness. The ability of the detectors on colour images changes when the luminance is changed due to the amount of light present in the environment, and a similar effect can be observed in thermal images with varying temperatures. The proposed late fusion method is robust for less accurate registration, and continues to function even when the input from one of the sensors is unavailable.

For this, a naive Bayes-based fusion approach is proposed, which uses probability distributions estimated from a small annotated dataset. Moreover, a Monte Carlo sampling [15] variant is proposed to estimate the distributions required for the fusion process. This allows the use of off-the-shelf pre-trained detectors (e.g., CNNs) by modelling their output for any dataset without the need for re-training. Besides the convenient reusability of pre-trained detectors, the results in the paper also show a significant improvement in pedestrian detection due to sensor fusion, as compared to single detectors.

The main contributions of this paper are as follows. (1) We present a luminance-change problem concerning colour and thermal due to changes in the environment and analyse its effect on the performance of detectors. (2) We propose a probabilistic late-fusion method conditioned on environmental conditions to address luminance and temperature changes in colour and thermal images. With this, the detector output is modelled without requiring retraining. In addition, the proposed method can work with accurately registered as well as poorly registered image pairs, and keeps working even when one of the sources is unavailable. (3) We also propose a method to compute likelihoods and priors using a variant of Monte Carlo sampling, which allows the computation of unbiased distributions swiftly. (4) The proposed method achieves state-of-the-art results on both the aligned FLIR dataset and our own captured dataset in terms of accuracy.

The rest of the paper is organised as follows. The overview of the proposed method is described in Section 2 with the fusion method and the methods for computing priors and likelihoods, the experimental setup, and the dataset formation. Section 3 contains the comparison results of the proposed method with different implementations for computing likelihood, different methods of random sample generation for computing likelihoods for the proposed technique, and the state-of-the-art fusion method for multi-spectral pedestrian detection. Section 4 concludes the paper with a description of potential future work.

2. Materials and Methods

2.1. Methodology

The proposed system uses images from two visual sources i.e., thermal and colour cameras. For thermal images, a pre-processing step is required to remove lens distortion, as mentioned in Section 2.5.

Existing pedestrian detectors are applied and their detection results are used for fusion and modelling, as shown in Figure 1.

Initially, during the modelling phase, a small annotated dataset is used to model the output of detectors conditioned on the environment variable that affects the sensor data, i.e., solar altitude representing day/evening/night for colour images and temperature for thermal images. This process uses the detection results to model the behaviour of detector(s) on the dataset with varying illumination in the thermal and colour images by computing relevant likelihood histograms, which are then used during the fusion process.

During the inference phase, the results of two detectors are fused using a naive Bayes-based algorithm (described in Section 2.2) with the environment information, i.e., temperature and solar altitude, and the likelihood histograms computed during the modelling phase, followed by greedy non-maximum suppression (NMS) [16] to remove duplicate detection from results.

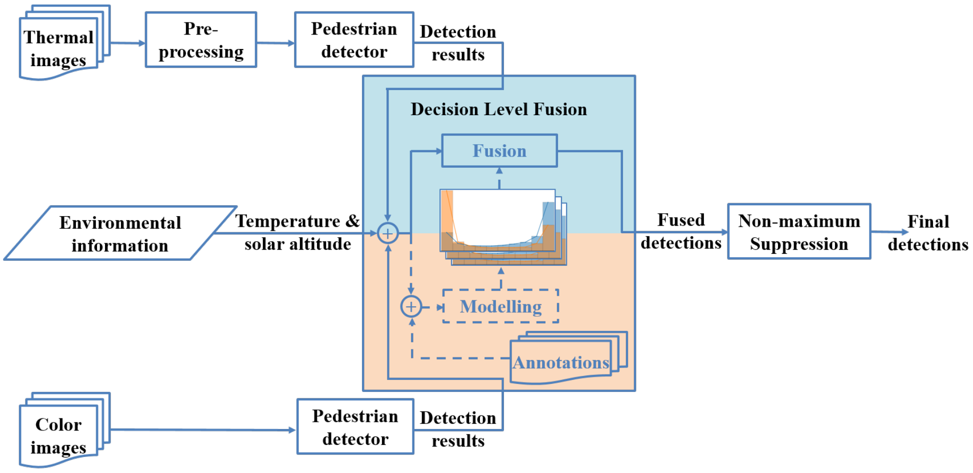


Figure 1. Proposed methodology.

For such a system, a generative adversarial network (GAN) [17] could be useful to generate more datasets from existing images to perform more robust modelling; however, due to common problems such as non-convergence of model parameters, vanishing parameters, and mode collapse, it is hard to train a sturdy GAN [17]. For this, naive Bayes for late fusion is considered due to its benefits such as a few parameters to set, simplified design process, computational speed, easy scalability, and not requiring a large amount of data [18]. Additionally, it is a well-established technique for modelling detection and classification problems [19].

2.2. *Naive Bayes-Based Fusion*

The task of the fusion process is to estimate the probability that a person is present at each possible location x in the field of view covered by the sensors, denoted as $p_x(ped)$. It is estimated as a conditional probability given the scores of the detectors' output for bounding boxes near that location, denoted as $p_x(ped|s_I, s_R)$, where s_I and s_R are the detection scores from detectors applied on thermal and colour images, respectively.

The fusion process computes the probabilities for all the detections from two detectors. It uses the modelled output of the detectors as an approximated likelihood/histogram, as discussed in Section 2.4, using a Naive Bayes approach.

The equation for computing probabilities using the proposed fusion is given as follows:

$$posterior = \frac{likelihood \times prior}{marginal probability} \tag{1}$$

In Equation (1), the posterior is estimated by the fusion process, i.e., $p_x(ped|s_I, s_R)$ conditioned on the output scores of the detectors at a given location x in the registered images. Using the chain rule, the Equation (1) becomes:

$$p_x(ped|s_I, s_R) = \frac{p_x(s_I, s_R|ped)p_x(s_R|ped)p_x(ped)}{p_x(s_I, s_R|ped)p_x(s_R|ped)p_x(ped) + p_x(s_I, s_R| \sim ped)p_x(s_R| \sim ped)p_x(\sim ped)} \tag{2}$$

Since the variables s_I and s_R in Equation (2) are assumed to be independent when conditioned on the variable ped, Equation (2) can be written as:

$$p_x(ped|s_I, s_R) = \frac{p_x(s_I|ped)p_x(s_R|ped)p_x(ped)}{p_x(s_I|ped)p_x(s_R|ped)p_x(ped) + p_x(s_I| \sim ped)p_x(s_R| \sim ped)p_x(\sim ped)} \tag{3}$$

where $p_x(s|ped)$ and $p_x(s| \sim ped)$ are the likelihoods for a correct (true positive) and incorrect detection (false positive), respectively, while $p_x(ped)$ is the prior.

Although, for simplicity, we will assume that $p_x(ped|s_I, s_R)$ only depends on the scores of the detector, and not on the actual position in the image, we still maintain the subscript x to emphasise that the posterior refers to the occurrence of a pedestrian at a certain position in the image.

The likelihood functions in Equation (3) express the general case where the probability $p_x(s|ped)$ describes the distribution of detector scores across all conditions. However, detector performance can vary strongly in the function of specific circumstances, for example, night vs. day, for detection in RGB camera images. As a result, the posterior probability in Equation (3) may poorly represent real pedestrian presence in specific conditions. To obtain more accurate posterior probabilities, likelihood functions are further conditioned on the luminance category G, because the luminance factor strongly affects the confidence scores in experiments produced by detectors on RGB and thermal images.

To address this, Equation (1) is reformed by considering the posterior as the probability of the presence of a pedestrian *ped* given the detection score, and global luminance category from colour (RGB) and thermal (infrared—IR) detection s_R, G_R and s_I, G_I, respectively, as $p_x(ped|s_I, G_I, s_R, G_R)$ for possible locations x. Thus, the likelihood according to the chain rule would be as in Equation (4).

$$p_x(s_I, G_I, s_R, G_R|ped) = p_x(s_I|ped, G_I, s_R, G_R)p_x(G_I|ped, s_R, G_R)p_x(s_R|ped, G_R)p_x(G_R|ped) \tag{4}$$

However, some of the variables in Equation (4) are independent, and some are dependent on other variables as well. For example, the detection score s_I is only dependent on the presence of pedestrian *ped* and the environmental variable G_I. G_I is a completely independent factor itself. Therefore, by considering conditional independence among these variables, the Equation (4) can be written in a simplified form as:

$$p_x(s_I, G_I, s_R, G_R|ped) = p_x(s_I|ped, G_I)p_x(G_I)p_x(s_R|ped, G_R)p_x(G_R) \tag{5}$$

Similarly, the marginal probability $p_x(s_I, G_I, s_R, G_R)$ for Equation (1) can be derived as:

$$\begin{aligned} p_x(s_I, G_I, s_R, G_R) = p_x(s_I|ped, G_I)p_x(G_I)p_x(s_R|ped, G_R)p_x(G_R)p_x(ped) + \\ p_x(s_I| \sim ped, G_I)p_x(G_I)p_x(s_R| \sim ped, G_R)p_x(G_R)p_x(\sim ped) \end{aligned} \tag{6}$$

By substituting $p_x(ped)$ as prior, likelihood as Equation (5), and marginal probability as Equation (6) in Equation (1), the simplified equation for naive Bayes-based fusion can be written as:

$$\begin{aligned} &p_x(ped|s_I, G_I, s_R, G_R) = \\ &\frac{p_x(s_I|ped, G_I)p_x(s_R|ped, G_R)p_x(ped)}{p_x(s_I|ped, G_I)p_x(s_R|ped, G_R)p_x(ped) + p_x(s_I| \sim ped, G_I)p_x(s_R| \sim ped, G_R)p_x(\sim ped)} \end{aligned} \tag{7}$$

where $p_x(G_I)$ and $p_x(G_R)$ are cancelled out from the numerator and denominator, and $p_x(s|ped, G)$ and $p_x(s| \sim ped, G)$ are the likelihoods for correct and incorrect detections of thermal and colour images.

The methods to compute the prior $p_x(ped)$, i.e., the probability of a pedestrian being present in the field of view at location x, and the likelihoods $p_x(s|ped, G)$ and $p_x(s| \sim ped, G)$, i.e., the likelihood of a detection at position x being either a true positive or false positive conditioned on the environmental variable G, are further discussed in Sections 2.3 and 2.4. After computing the posterior, detections that belong to the same pedestrian are then associated, as discussed in Section 2.2.1.

2.2.1. Association Method

Each possible combination of detections from different sensors is evaluated using the posterior probabilities of Equation (7) and the amount of overlap measured by intersection over union (IoU) [20] between the detections. Moreover, when a detection is only available from a single sensor, a likelihood is computed for an imaginary detection with score equal to zero for the other sensor. This helps the fusion process to handle omissions.

Bounding box association between different detectors is solved by using a cost matrix that involves the scores of both detectors as well as the IoU of the two bounding boxes, where the matches are selected with minimal cost. To improve numerical stability when likelihoods are very small, we apply a log function:

$$cost(I_i, R_j) = -log(p_x(ped|s_{I_i}, G_{I_i}, s_{R_j}, G_{R_j})) - log(IoU(bb_{I_i}, bb_{R_j})) \tag{8}$$

where I_i, R_j are the detections from the colour and thermal images, respectively, $p_x(ped|s_{I_i}, G_{I_i}, s_{R_j}, G_{R_j})$ is their posterior, and $IoU(bb_{I_i}, bb_{R_j})$ is the IoU of their bounding boxes.

Furthermore, multiple detections for the same pedestrian may occur due to less overlapping detections, i.e., less than the IoU threshold th; therefore, greedy non-maximum suppression (NMS) is used to suppress duplication among the select matches in colour and thermal image detections.

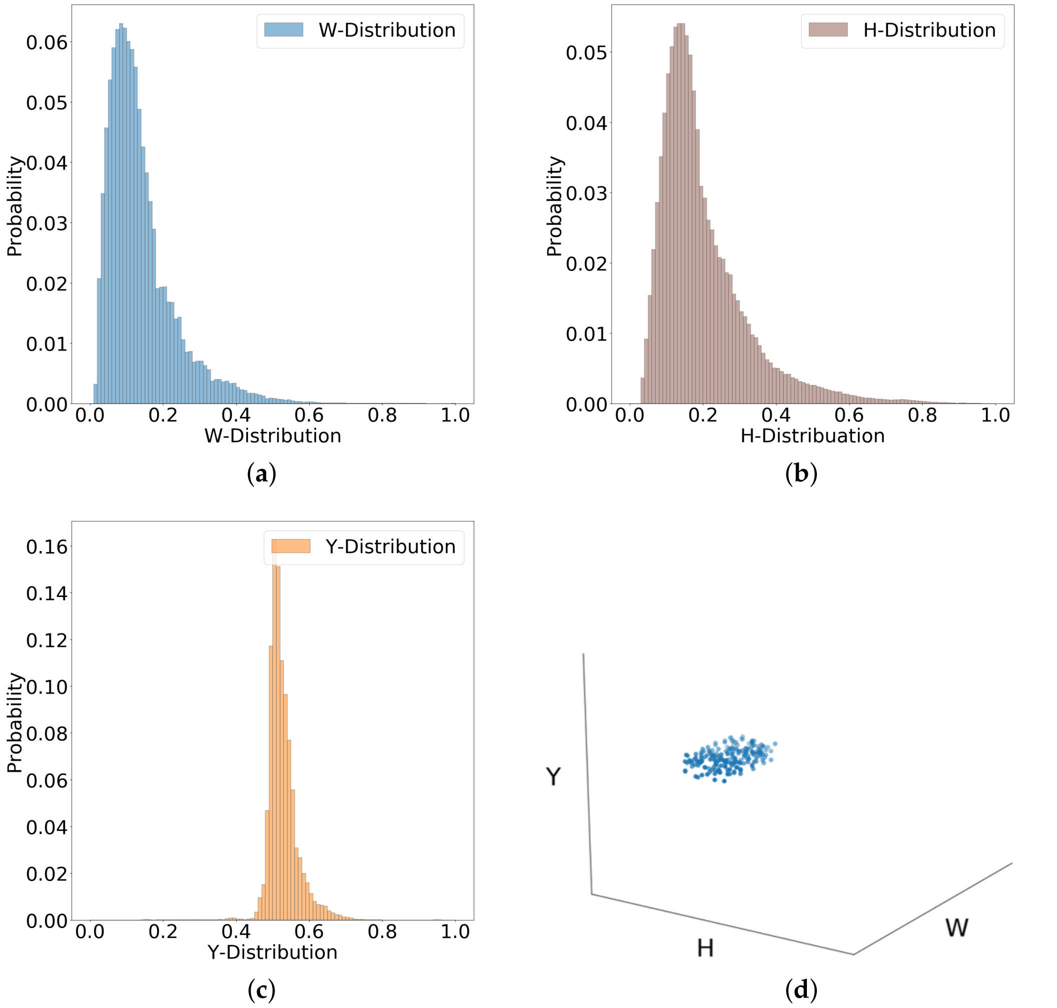
2.3. Prior

The prior is the probability of an event to occur, which, is in this case the presence of pedestrian(s) in the scene at all possible locations. Moreover, the prior is also used to compute the value of $p_x(\sim ped)$ as $1 - p_x(ped)$, which describes the probability of a pedestrian not being present at a certain location in the scene.

The estimation of the prior $p_x(ped)$ at a given location x in the scene depends on the prevalence of pedestrians in the dataset. However, since the prior will be used to compute the posterior $p_x(ped|s_I, G_I, s_R, G_R)$ for scores produced by a detector, we also have to model the behaviour of this detector. More specifically, we have to take into account the role of the gating function used by the detector to determine whether a pedestrian is present at a certain location. Furthermore, we also have to take into account that the positions where a pedestrian may occur are not evenly distributed over the image, and that the height and width of a bounding box also have their own typical distributions. To accomplish this, we apply the Monte Carlo sampling method to the ground truth set to obtain reliable estimates $\tilde{p}_x(ped)$ for the prior.

For estimation, the generation of the realistic random sample, i.e., random bounding boxes, is difficult due to different distributions of width, height, and location of bounding boxes. Therefore, multivariate distribution is computed based on the ground truth set, as shown in Figure 2.

The distributions shown in Figure 2 are computed with 100 bins based on the resolution of the images. The height H and width W of the bounding boxes in pixels are smaller and, therefore, are multiplied twice and four times for better visualization, respectively. Furthermore, the pedestrians and their size in the image, i.e., W and H, in the captured dataset appear about evenly on the x-axis, and only the y-axis Y is taken into account for this multivariate distribution. Additionally, we have also investigated other methods for generating random samples and compared the final fusion results accordingly in Section 3.



(**a**) (**b**)

(**c**) (**d**)

Figure 2. Multivariate distribution for generating random samples. (**a**) W-distribution. (**b**) H-distribution. (**c**) Y-distribution. (**d**) Random samples generation.

The Algorithm 1 for estimation of the prior works as follows. After initializing *pr*, random samples, i.e., random bounding boxes, are drawn from multivariate distribution in Step 2, by obtaining the Y-position and corresponding size of the bounding box, i.e., width and height values, as shown in Figure 2d. Steps 3 to 10 compare these random samples with the annotations of a randomly selected image and increment *pr* if there is a matched annotation, i.e., if the IoU between the random sample and the annotation is greater than the threshold.

Algorithm 1: Estimating $p_x(ped)$.

```
Data: Ground truth set $D$ of bounding boxes where pedestrians appear, sample
      size $N$ and IoU threshold $th$
Result: Estimate $\tilde{p}_x(ped)$ for the prior
1.  $pr \leftarrow 0$
2.  $R \leftarrow$ DRAW_SAMPLES($N$)
3.  for $r \in R$
4.      Choose a random image $i$ from the dataset
5.      Let $D_i \subset D$ be all ground truth annotations for image $i$
6.      $M_i = \{d_i \in D_i | \text{IoU}(d_i, r) > \text{th}\}$
7.      if $|M_i| > 0$
8.          $pr \leftarrow pr + 1$
9.      end
10. end
11. $\tilde{p}_x(ped) \leftarrow \frac{pr}{N}$
```

Drawing random samples from the ground truth (i.e., bounding boxes) ensures that the prior takes into account gating (i.e., the IoU criterion), and uses appropriate distributions for the position and dimensions of the bounding boxes. By repeating the above resampling multiple times, we also obtain a value for the variance of $\tilde{p}_x(ped)$.

Annotations from the ground truth that match the random sample (i.e., gating with IoU > 0.25) only contribute to the prior. In last step, the number of samples matched with ground annotations is normalised. For our dataset, the estimated value for the prior found was $p_x(ped) = 0.3$.

2.4. Likelihood

The likelihood function is the probability of the observation, i.e., the detection score of the detector over the parameters of the model, which are the presence or absence of the pedestrian and the global luminance category G. Thus, the likelihood functions $p_x(s|ped, G)$ and $p_x(s| \sim ped, G)$ are to be determined for each value of G. Therefore, the dataset categories (mentioned as in Tables 1 and 2) are further used to compute likelihoods to model the output of a detector. Two likelihoods are considered; the first is the probability of detection being correct, i.e., the probability of detection score for all possible locations x conditioned on the event (pedestrian presence) and global luminance category. Similarly, the second likelihood is for the detection being wrong, where the event is the pedestrian not being present.

Initially, the detector is applied to the images of dataset categories, and detection results per image are used to classify correct (true positive) and incorrect (false positive) detections using IoU by comparing the detections with the ground truth annotations.

The detection is considered correct if the overlap area between detection and ground truth annotation is more than the threshold value; otherwise, it is wrong. The ground truth annotations are computed using a semi-auto method and, therefore, are not pixel-level accurate. Due to this, an IoU threshold of 0.25 is used in the whole fusion process rather than the standard IoU threshold of 0.5. Moreover, during the classification of detection results, priority is given to the detections with the highest detection score. Therefore, true positives and false positives are formed from the detection results of all the dataset categories.

2.4.1. Classical Method

The likelihood histograms are computed for each bin based on the correct (true positive) and incorrect (false positive) detections. Furthermore, a Savitzky–Golay filter [21] is used to smooth the histogram data. The computed histogram for a single category from the dataset is shown in Figure 3.

Figure 3 shows the computed histogram is based on 10 bins, where the likelihood of correct detection is higher when the detection score is also higher. On the other hand, the likelihood of detection being wrong is higher when the detection score is lower. This method provides an easy way to compute likelihood histograms; however, the likelihood values are sensitive to the number of bins used.

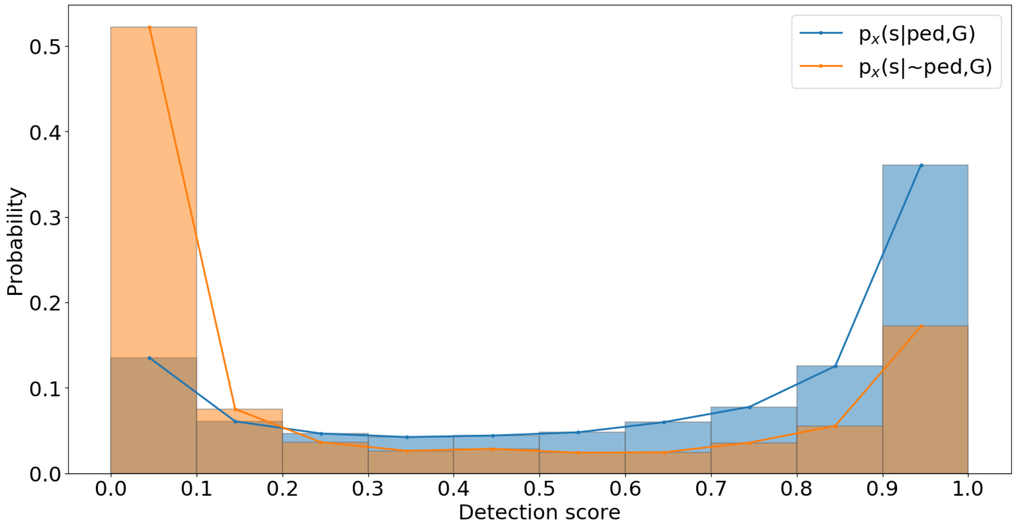
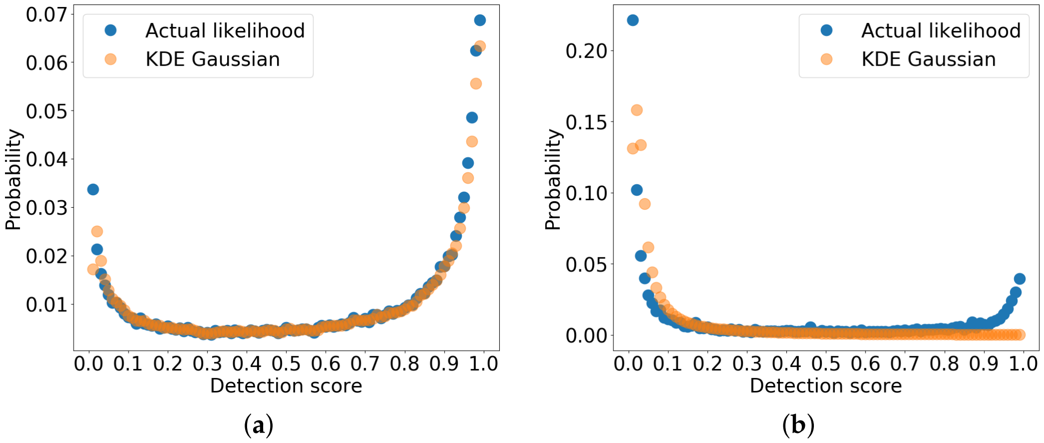


Figure 3. Likelihood histograms from a dataset category: likelihood for true positives-$p_x(s|ped, G)$ and false positives-$p_x(s| \sim ped, G)$.

2.4.2. Kernel Density Estimation

Kernel density estimation [22] can be used to overcome the number of bins problem in the likelihood histograms. This is done by estimating the probability density using the Gaussian kernel. However, this non-parametric estimation method is sensitive to bandwidth and, therefore, the estimation is performed with several bandwidth values and compared to actual data distribution using the Kolmogorov–Smirnov test [23] to find the best fit estimation. The best-selected estimation compared with the actual distribution of 100 bins for a single category from the dataset is shown in Figure 4.



(**a**) (**b**)

Figure 4. Estimated likelihood from a dataset category. (**a**) True positives likelihood-$p_x(s|ped, G)$. (**b**) False positives likelihood-$p_x(s| \sim ped, G)$.

The estimations are made separately for correct (true positive) and incorrect (false positive) detections for each dataset category for each modality with different bandwidth values, depending upon the best fit with actual data. From Figure 4, it can be seen that the estimation fits the actual data without combining several data samples in the bins, although these estimation methods are biased near the boundaries and flatten the peaks of the distribution [24], as seen in the figure as well. However, the bias of density estimation is better than histogram estimation [25].

2.4.3. Monte Carlo Sampling

The likelihoods computed in Section 2.4.1 and estimated at a higher bin resolution in Section 2.4.2 are based on the detector-provided bounding boxes per image and, hence, only consider correct and incorrect detections, disregarding the possible locations. Moreover, this creates a bias in the rich contrast regions of the images.

For this purpose, Monte Carlo sampling can be used to approximate the distribution of data by considering all bounding box samples in one space. Moreover, random samples are generated and gated for all possible locations in the space, which also produces an unbiased sample generation. However, due to the limitations of drawing samples directly, as mentioned earlier, a similar variation of Monte Carlo sampling is used (as proposed in Section 2.3), where random bounding boxes are drawn from the multivariate distribution shown in Figure 2.

The Algorithm 2 for likelihood estimation requires correct (true positive) detections based on the classification performed on detector output using IoU, along with user-specified parameters such as size for random samples, number of bins for histograms, and gating function parameters, i.e., IoU threshold for considering all the matching detections with the random samples.

After initializing the histograms, random samples are drawn from multivariate distribution (as shown in Figure 2) in Step 2. Steps 3 to 16 compare each bounding box r with the detections of a detector to determine how likely the detection score $s(m_i)$ is for correct or incorrect detections. A random image is selected from the dataset and only those detections that are matched with the sample r (which has a higher IoU than the threshold) in Steps 4 to 6 are selected. If there are not any matched detections with the random sample, then another random sample r is selected and compared with the detections of another random image i in Steps 7 to 9. Otherwise, in Step 10, only the detection with the highest score in that region of an image is considered to overcome the overlapping detections with a lower score.

Algorithm 2: Likelihood estimation.

```
Data: Bounding boxes of all true-positive detections (TP) with their detection scores,
      sample size $N$, IoU threshold (th), number of bins (b), and dataset images I
Result: Two normalised histograms (TPD and FPD) for the density estimation of
      $p_x(s_R|ped)$ and $p_x(s_R|\sim ped)$
1.  Initialise TPD and FPD as histograms with all $b$ bin counts set to zero
2.  $R \leftarrow$ DRAW_SAMPLES($N$)
3.  for $r \in R$
4.      Choose a random image $i$ from the dataset
5.      Let $D_i \subset D$ be all detections for image $i$
6.      $M_i = \{d_i \in D_i | \text{IoU}(d_i, r) > \text{th}\}$
7.      if $M_i = \emptyset$
8.          goto step 3
9.      end
10.     Let $m_i \in M_i$ be the matched detection with the highest score $s(m_i)$
11.     if $m_i \in TP_i$
12.         increase the $j$th bin count of TPD by one, where $j = \lfloor s(m_i) \times b \rfloor$
13.     else
14.         increase the $j$th bin count of FPD by one, where $j = \lfloor s(m_i) \times b \rfloor$
15.     end
16. end
17. Normalise both histograms
```

The matched detection with the highest score is compared with TP for the correct or incorrect histogram Steps 11 to 15. The detection score $s(m_i)$ of the bounding box m_i is multiplied by the number of bins to find the right bin in Steps 12 and 14. In the last step, both histograms are normalised with their sums. By imposing a threshold on the IoU with random sample r, we take into account the gating function of the detector.

The likelihood estimation from approximated histogram for correct (true positive) detection can be represented by the conditional probabilistic term $p_x(s|ped, G)$, where s represents the detection score, ped represents the presence of a pedestrian as correct detection, and G represents the global luminance. Similarly, likelihood estimation from histogram of incorrect (false positive) detection is $p_x(s| \sim ped, G)$, with $\sim ped$ representing a detection without pedestrian.

In Figure 5, the two shown likelihood histograms are approximated from a single category of the dataset from images of a single modality; similarly, the distributions are computed for all categories of the dataset for both colour and thermal images. In the likelihood histogram for true positives, the detections with the highest score are the most correct ones, and in false positives, the incorrect detections usually have the lowest scores. These likelihoods are used to model the output of the detectors and are also used in the fusion process, as discussed in Section 2.2.

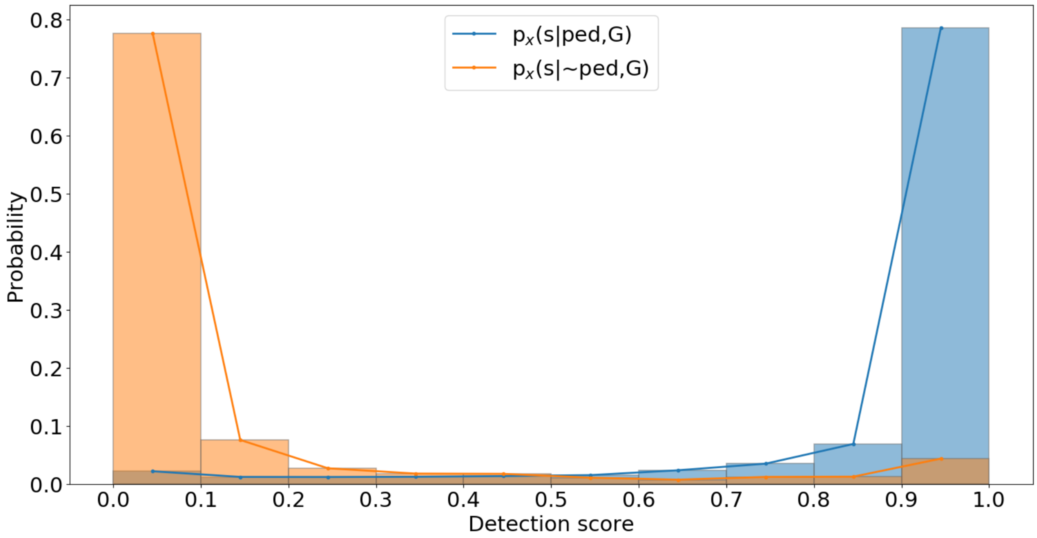


Figure 5. Approximated likelihood histograms from a dataset category: likelihood for true positives—$p_x(s|ped, G)$, and false positives—$p_x(s| \sim ped, G)$.

2.5. Experimental Setup

For the dataset, video frames from a FLIR Thermicam-390 LWIR thermal camera and Intel Realsense are captured. The traffic during the recordings was persons, persons in groups, cycles, bikes, and cars.

Recordings were made during the sunny and cloudy days without rain/snow because of the permeable recording setup. The recording scenario of the dataset is given in Table 1.

Table 1. Recording scenario for dataset.

Date & Time	Condition	Solar Altitude	Temperature
15-09-2020 @ 21:26–21:54	Warm night	[−14°, −18°]	24 °C
16-09-2020 @ 19:12–19:36	Warm afternoon-evening	[07 °, 03°]	23 °C
05-11-2020 @ 16:53–17:25	Cold evening	[02°, −02°]	12 °C
18-11-2020 @ 16:59–17:48	Cold evening-night	[−01°, −09°]	13 °C
20-11-2020 @ 10:09–11:06	Chilly morning	[13°, 17°]	04 °C
10-12-2020 @ 18:12–19:02	Chilly night	[−14°, −22°]	03 °C

The lens parameters for the thermal camera are calculated with the help of a halogen lamp, chessboard, and the Caltech toolbox [26], as described in [27], to remove lens distortion. The colour images are free from lens distortion.

The colour images are annotated semi-automatically using an auto-labelling tool [28]. The annotations for thermal images are formed with the help of colour images' annotations with negligible parallax, using the stereo projection method [29,30].

The bounding box coordinates for thermal images are computed from the normalised bounding box coordinates of colour (RGB) images and then thermal (infrared—IR) image coordinates. This operation performs scaling and translation on annotations to transfer them from RGB to IR images considering the same orientation for both of the cameras, as in Equation (9).

$$(x_I\ ,\ \ y_I) = \left(\frac{x_R - o_{x_R}}{f_{x_R}} f_{x_I} + o_{x_I}\ ,\ \ \frac{y_R - o_{y_R}}{f_{y_R}} f_{y_I} + o_{y_I}\right) \tag{9}$$

where (f_{x_R}, f_{y_R}) and (f_{x_I}, f_{y_I}) are the focal length, (o_{x_R}, o_{y_R}) and (o_{x_I}, o_{y_I}) are principal points, and (x_R, y_R) and (x_I, y_I) are the bounding box points of colour and thermal camera images, respectively.

For the experiments, we have used YOLO version 3 [31] as the pedestrian detector for both colour and thermal images. Although YOLO is only trained on colour images, applying it on inverted thermal images produced useful results, as shown in Section 3.

2.6. Dataset Division

The luminance in colour and thermal images is affected by different environmental factors due to different modalities. For colour images, the luminance in the images is dependent on the position of the natural light source. For thermal images, the environmental temperature is important.

Therefore, the dataset is divided into categories based on the selected environmental variable, which affects the global luminance category denoted as *G*, presented in Table 2. The distributions of these categorised datasets are computed and used as likelihood histograms in the fusion process, which results in an improvement for the fusion, as described in Section 3.

Table 2. Dataset division.

Dataset Category	**Environmental Variable**	**Global Luminance Category *G***	**No. of Images**
Colour			
Day	Solar altitude $> 6°$	1	26,573
Evening	Solar altitude = $[6°\ , -6°]$	2	24,144
Night	Solar altitude $< -6°$	3	13,365
Thermal			
Cold	Temperature < 20 °C	1	34,328
Warm	Temperature ≥ 20 °C	2	29,754

The dataset is divided into the mentioned factors for colour and thermal images with the help of weather information [32,33], presented in Table 1. The colour images are divided with three different ranges for the day, evening, and night, depending on the solar altitude [34], as shown in Figure 6.

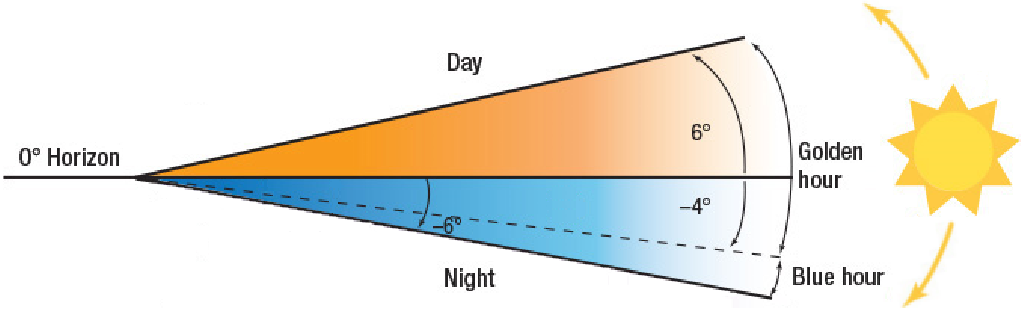


Figure 6. Solar altitude between Sun and local horizon.

The altitude ranges for the colour dataset categories considered are shown in Figure 6 for day and night, while the blue and golden hours are considered as evening. The solar altitude data is obtained from weather information and the colour dataset is divided accordingly. The thermal images are divided with temperature ranges, also using the weather information.

3. Results

The proposed fusion method is implemented with different approaches to compute the prior and likelihoods. The results of these implementations are shown in Figure 7.

The fusion is performed on different dataset categories; mAP (mean average precision) [35] is considered as a performance factor. It can be seen in Figure 7 that the implementation of the proposed method with the proposed Monte Carlo sampling variant performs better than other implementations.

On the other hand, the fusion with KDE is slightly better than the classical method in most cases, but it sometimes degrades when estimation becomes inappropriate with KDE due to multiple and high peaks in the actual distribution. Moreover, detections of YOLO for RGB and IR images are also compared, where it is observed that the mAP of the proposed fusion results are much better, especially with the difficult conditions, than applying a detector on a single sensor.

For comparison, we have used precision, recall, and mAP factors. These factors as computed as:

$$mAP = \frac{\sum_{n=1}^{N} AP_n}{N} \tag{10}$$

where

$$AP = \sum_{r \epsilon \{0.0, 0.1, \ldots, 1.0\}} (R_r - R_{r-1}) P_r \tag{11}$$

and,

$$P = \frac{tp}{tp + fp}, \; R = \frac{tp}{tp + fn} \tag{12}$$

tp, fp and fn are true positives, false positives, and false negatives. N is total number of images with ground truths; P_r and R_r are the precision and recall at the r threshold, respectively.

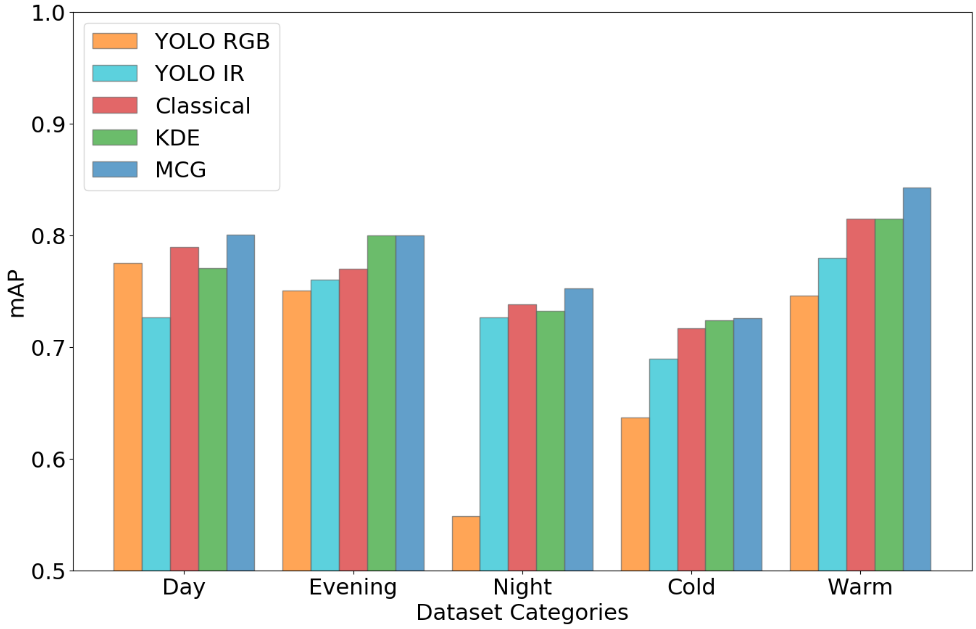


Figure 7. Comparison of different implementations for the proposed fusion and YOLO detections.

Furthermore, the results of the proposed method are compared with a similar method, but without being conditioned on the environmental variable *G*, as presented in Table 3.

Table 3. Comparison of proposed fusion with and without considering the conditional variable.

Dataset Category	Proposed Fusion			Fusion without G		
	P	**R**	**mAP**	**P**	**R**	**mAP**
Day	0.51	0.70	<u>0.80</u>	0.50	0.70	0.79
Evening	0.41	0.76	<u>0.80</u>	0.47	0.73	0.79
Night	0.41	0.61	<u>0.75</u>	0.55	0.55	0.74
Cold	0.40	0.65	<u>0.73</u>	0.46	0.62	0.73
Warm	0.50	0.73	<u>0.84</u>	0.54	0.70	0.83

It can be seen from the results that by just considering a single conditional variable *G*, the overall results of the method improve. Therefore, the results prove that the proposed fusion process improves the detection accuracy of the system by a considering conditional variable, especially in challenging dynamic environments where the accuracy of a single sensor detector starts degrading.

To investigate the relative importance of the prior and likelihood estimations, different strategies to draw random samples are compared, from the most realistic to a naive method, using multivariate distribution as discussed in Section 2.3, polynomial regression, normal distribution, and uniform distribution. The final results are shown in Figure 8.

Polynomial regression is used to compute the polynomial relation between the Y-axis and the corresponding width and height of the bounding box from the annotated dataset. The independent normal distributions are used for another experiment by considering mean and variance from the annotated dataset; finally, uniform distribution is used for random sample generation, where the sample ranges are obtained from the annotated dataset.

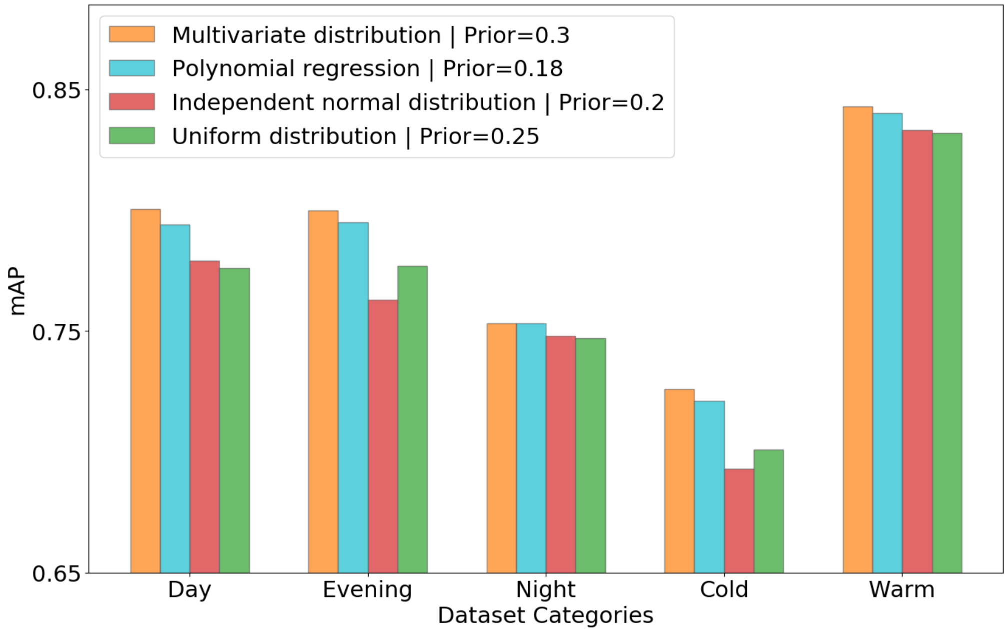


Figure 8. Comparison of different random sample generation methods for the proposed fusion.

Figure 8 shows the difference in mAP after fusion when using different methods to generate samples to compute prior and likelihood histograms. This experiment clearly illustrates the importance of obtaining a realistic estimate and the role of sample distribution in its estimation, as described in Section 2.3.

Additionally, the proposed method is also compared with the state-of-the-art fusion methos, as shown in Table 4. Most of the current state-of-the-art methods require fine image registration between thermal and colour images; therefore, the aligned FLIR dataset [36,37] is used for this comparison.

Table 4. Comparison of proposed fusion with the state-of-the-art method on aligned FLIR dataset.

FLIR Dataset		Proposed Fusion			MBNet [14]		
RGB	**IR**	**P**	**R**	**mAP**	**P**	**R**	**mAP**
	✓	0.86	0.55	<u>0.69</u>	0.45	0.14	0.27
✓		0.63	0.45	<u>0.61</u>	0.83	0.01	0.01
✓	✓	0.66	0.65	<u>0.73</u>	0.40	0.55	0.69

The results in Table 4 show that the performance of the current state-of-the-art method drastically decreases if the input from one of the sensors is unavailable. On the other hand, the proposed method performs better than the state-of-the-art-method when the input images from both of the sensors are available, and it is also able to cope if one of the sensors is unavailable.

Furthermore, the current state-of-the-art methods cannot be applied to the non-registered dataset without altering and retraining the fusion technique, while the proposed method performs effectively on both non-registered and registered datasets.

4. Discussion

The results presented in this paper show that our proposed fusion technique, combined with the proposed variant of Monte Carlo sampling to compute prior and likelihood, improves the detection accuracy of the system, especially during difficult dynamic situations and even when one of the sensors is unavailable, as compared to the state-of-the-art methods. Additionally, the proposed method can easily be implemented without retraining the detector on a huge annotated dataset, with minimal changes in the parameters. Furthermore, adding a single relatively uninformative measurement such as global luminance is shown to have the potential to improve the accuracy of simple naive Bayes fusion, and it can further be improved by investigating object-based variables such as skin complexion of pedestrians [38] detected or local contrast measures for detection, in order to acquire balanced modelling for several sub-populations [39] from the existing datasets, which would be conditioned similarly using the proposed fusion method.

In the future, we aim to use our proposed method with a particle filter tracker, which will improve the accuracy further by using successive detections from previous frames and will also be useful for systems such as autonomous driving and traffic management, to name a few.

Author Contributions: Conceptualization, methodology, validation, investigation, writing—original draft preparation, Z.A.S.; validation, investigation, writing—review and editing, resources, D.V.H.; investigation, writing—review, supervision, P.V.; formal analysis, writing—review, supervision, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brunetti, A.; Buongiorno, D.; Trotta, G.F.; Bevilacqua, V. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing* **2018**, *300*, 17–33. [CrossRef]
2. Chakraborty, S.; Laware, H.; Castanon, D.; Zekavat, R. High precision localization for autonomous vehicles via multiple sensors, data fusion and novel wireless technologies. In Proceedings of the Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), New York, NY, USA, 20–22 October 2016; pp. 1–9.
3. Research Report: Automatic Emergency Braking with Pedestrian Detection. Available online: https://www.aaa.com/AAA/commo\o/aar/files/Research-Report-Pedestrian-Detection.pdf?cjevent=91c7ad8b070311ea813900930a180513&utm_medium=affiliate&cm\p=AFC_membership_na_prospecting_affiliate_MWG (accessed on 10 August 2021).
4. Zanchin, B.C.; Adamshuk, R.; Santos, M.M.; Collazos, K.S. On the instrumentation and classification of autonomous cars. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 2631–2636.
5. Goshtasby, A.A.; Nikolov, S. Image fusion: Advances in the state of the art. *Inf. Fusion* **2007**, *8*, 114–118. [CrossRef]
6. Ghassemian, H. A review of remote sensing image fusion methods. *Inf. Fusion* **2016**, *32*, 75–89. [CrossRef]
7. Zhao, W.; Xu, Z.; Zhao, J. Gradient entropy metric and p-Laplace diffusion constraint-based algorithm for noisy multispectral image fusion. *Inf. Fusion* **2016**, *27*, 138–149. [CrossRef]
8. Ross, A. *Encyclopedia of Biometrics*; Springer: Boston, MA, USA, 2009; Chapter: Fusion-Feature level, pp. 597–602.
9. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A. *Remote Sensing Image Fusion*; CRC Press: Boca Raton, FL, USA, 2015; pp. 150–190.
10. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [CrossRef]
11. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition* **2019**, *85*, 161–171. [CrossRef]
12. Guan, D.; Cao, Y.; Yang, J.; Cao, Y.; Yang, M.Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Inf. Fusion* **2019**, *50*, 148–157. [CrossRef]
13. Li, Q.; Zhang, C.; Hu, Q.; Fu, H.; Zhu, P. Confidence-aware Fusion using Dempster-Shafer Theory for Multispectral Pedestrian Detection. *IEEE Trans. Multimed.* **2022**. [CrossRef]
14. Zhou, K.; Chen, L.; Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 787–803.
15. Mooney, C.Z. *Monte Carlo Simulation*; Sage Publications: Thousand Oaks, CA, USA, 1997.
16. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef] [PubMed]
17. Alqahtani, H.; Kavakli-Thorne, M.; Kumar, G. Applications of generative adversarial networks (gans): An updated review. *Arch. Comput. Methods Eng.* **2021**, *28*, 525–552. [CrossRef]
18. Stern, M.; Beck, J.; Woolf, B.P. *Naive Bayes Classifiers for User Modeling*; Center for Knowledge Communication, Computer Science Department, University of Massachusetts: Amherst, MA, USA, 1999.
19. Murphy, K.P. *Lecture Notes on Naive Bayes Classifiers*; University of British Columbia: Vancouver, BC, Canada, 2006.

20. Tanimoto, T.T. *An Elementary Mathematical Theory of Classification and Prediction*; Internal Report IBM Corp.: New York, NY, USA, 1958.
21. Schafer, R.W. What is a Savitzky-Golay filter? *IEEE Signal Process. Mag.* **2011**, *28*, 111–117. [CrossRef]
22. Węglarczyk, S. Kernel density estimation and its application. In Proceedings of the ITM Web of Conferences 23, EDP Sciences, Poznan, Poland, 23–24 April 2018; pp. 1–8.
23. Massey, F.J., Jr. The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78. [CrossRef]
24. Zambom, A.Z.; Ronaldo, D. A review of kernel density estimation with applications to econometrics. *Int. Econom. Rev.* **2013**, *5*, 20–42.
25. Sheather, S.J. Density estimation. *Stat. Sci.* **2004**, *19*, 588–597. [CrossRef]
26. Camera Calibration Toolbox for Matlab. Available online: https://data.caltech.edu/records/20164 (accessed on 13 July 2021).
27. Shaikh, Z.A.; Allebosch, G.; Veelaert, P.; Wilfried, P. Automatic annotation of pedestrians in thermal images using background/foreground segmentation for training deep neural networks. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, 1–4 December 2020; pp. 1444–1451.
28. Dimitrievski, M.; Shopovska, I.; Hamme, D.V.; Veelaert, P.; Philips, W. Automatic labeling of vulnerable road users in multisensor data. In Proceedings of the IEEE International Conference on Intelligent Transportation-ITSC, Indianapolis, IN, USA, 19–22 September 2021; pp. 1–10.
29. Collins, R. (Penn state University, Penn State, United States). Lecture Notes on Camera Projection—CSE-486, 2007. Available online: http://www.cse.psu.edu/~rtc12/CSE486/ (accessed on 15 October 2021).
30. Trucco, E.; Verri, A. *Introductory Techniques for 3-D Computer Vision*; 1st ed.; Prentice Hall: Hoboken, NJ, USA, 1998; pp. 34–40.
31. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
32. Weather in Belgium. Available online: https://www.timeanddate.com/weather/belgium (accessed on 6 October 2021).
33. Sun rise and Sunset in Belgium. Available online: https://www.timeanddate.com/astronomy/belgium (accessed on 6 October 2021).
34. Kalogirou, S.A. *Solar Energy Engineering Processes and Systems*, 2nd ed.; Academic Press: Cambridge, MA, USA, 2014; Chapter: Environmental Characteristics, pp. 51–63.
35. Schütze, H.; Manning, C.D.; Raghavan, P. *An Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2009; pp. 151–175.
36. Zhang, H.; Fromont, E.; Lefèvre, S.; Avignon, B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AL, USA, 19–22 September 2020; pp. 276–280.
37. Aligned FLIR Dataset. Available online: https://drive.google.com/file/d/1xHDMGl6HJZwtarNWkEV3T4O9X4ZQYz2Y/view (accessed on 20 April 2022).
38. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
39. Datta, A.; Swamidass, S.J. Fair-Net: A Network Architecture For Reducing Performance Disparity Between Identifiable Sub-Populations. *arXiv* **2021**, arXiv:2106.00720.