Helmet use detection of tracked motorcycles using CNN-based multi-task learning.

LIN, H., DENG, J.D., ALBERS, D. and SIEBERT, F.W.

2020

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 10, 2020, accepted August 22, 2020, date of publication September 2, 2020, date of current version September 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021357

Helmet Use Detection of Tracked Motorcycles Using CNN-Based Multi-Task Learning

HANHE LIN[1], JEREMIAH D. DENG[2], (Member, IEEE), DEIKE ALBERS[3], AND FELIX WILHELM SIEBERT[4]

[1]Department of Computer and Information Science, Universität Konstanz, 78464 Konstanz, Germany
[2]Department of Information Science, University of Otago, Dunedin 9054, New Zealand
[3]Department of Mechanical Engineering, Technical University of Munich, 85748 Garching, Germany
[4]Department of Psychology, Friedrich-Schiller University of Jena, 07743 Jena, Germany

Corresponding author: Hanhe Lin (hanhe.lin@uni-konstanz.de)

This work is mainly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 251654672 - TRR 161 (Project A05), and partly supported by University of Otago Research Grant 2019.

ABSTRACT Automated detection of motorcycle helmet use through video surveillance can facilitate efficient education and enforcement campaigns that increase road safety. However, existing detection approaches have a number of shortcomings, such as the inabilities to track individual motorcycles through multiple frames, or to distinguish drivers from passengers in helmet use. Furthermore, datasets used to develop approaches are limited in terms of traffic environments and traffic density variations. In this paper, we propose a CNN-based multi-task learning (MTL) method for identifying and tracking individual motorcycles, and register rider specific helmet use. We further release the HELMET dataset, which includes 91,000 annotated frames of 10,006 individual motorcycles from 12 observation sites in Myanmar. Along with the dataset, we introduce an evaluation metric for helmet use and rider detection accuracy, which can be used as a benchmark for evaluating future detection approaches. We show that the use of MTL for concurrent visual similarity learning and helmet use classification improves the efficiency of our approach compared to earlier studies, allowing a processing speed of more than 8 FPS on consumer hardware, and a weighted average F-measure of 67.3% for detecting the number of riders and helmet use of tracked motorcycles. Our work demonstrates the capability of deep learning as a highly accurate and resource efficient approach to collect critical road safety related data.

INDEX TERMS Deep learning, traffic surveillance, motorcycle safety, helmet use detection, tracking.

I. INTRODUCTION

Nowadays, drivers' adherence to traffic laws is mainly monitored and enforced by traffic police officers through direct observation. Yet implementations of road surveillance infrastructure are increasingly being used to automatically identify safety related behaviors through traffic video analysis. Approaches have been developed to register relatively simple variables, such as traffic flow and density [1], [2], speed [3]–[5], traffic light violations [6], or collisions [7]. More recently, computer vision has been used to register more complex road user behaviors, such as driver mobile phone use [8] and unauthorized use of car-pooling lanes [9]. Since for many developing countries the main form of motorized transport consists of motorcycles, the detection of motorcycle helmet use of riders through machine learning has also been explored [10], [11]. The availability of exact and concurrent data about motorcycle helmet use on the street is crucial to injury prevention, as it can be used for targeted enforcement and effective education campaigns.

The registration of helmet use through human observers naturally consists of four basic elements, that any automated detection method must also possess to produce comparably detailed helmet use estimates. (1) Detection: Initially, active motorcycles need to be detected. (2) Tracking: Individual motorcycles need to be tracked through the road environment, to ensure that each motorcycle is only registered once, regardless of how long it is observed. (3) Rider differentiation: For an accurate calculation of motorcycle helmet use and to produce position-specific helmet use data, rider numbers and positions (i.e. distinguishing the driver and passenger(s)) per motorcycle need to be registered. (4) Site-diversity:

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

Helmet use numbers need to be accurately registered, independent of the road environment at an observation site. Hence, automated approaches need to show accuracy for more than one road environment. While these four basic elements of motorcycle helmet use observation come naturally to humans observers, existing automated detection approaches, either do not include all four elements or have low performance on some of them (see Section II). In particular the lack of rider differentiation is a crucial element for the application of automated helmet use detection in the field. Researchers repeatedly find evidence of an influence of rider position and rider number on helmet use on individual motorcycles [12]–[15]. Hence, the differentiation of rider helmet use for drivers and passengers is a crucial metric, that should not be omitted in automated detection approaches. The lack of broad applicability and robustness prevents the substitution of human observers through automated approaches in helmet use observation.

Hence, we present a deep learning based automatic detection approach that contains all four basic elements of human-observer helmet use registration, i.e. detection, tracking, rider differentiation, and site-diversity. The proposed work builds on and extends a previous approach for frame-based helmet use detection [10], which did not include tracking of motorcycles and in which the dataset was not made public. To encourage the development of diverse detection approaches, we make this dataset available with the publication of this article. In addition, we further propose a benchmark metric for the assessment of automated detection approaches.

In summary, our main contributions are twofold:

- We propose a comprehensive CNN-based approach for helmet use detection of tracked motorcycles, containing all basic elements utilized by human observers. A multi-task learning (MTL) framework is developed for both visual similarity learning and patch-based helmet use classification, which increases computational efficiency as well as detection accuracy. The source code and pre-trained model are available in [16].
- We publish a diverse, large-scale, annotated dataset for motorcycle detection, called HELMET. It contains 10,006 annotated motorcycles in 910 video clips, recorded throughout the country of Myanmar, containing 12 observation sites across 7 cities. To the best of our knowledge, it is the largest and most diverse motorcycle helmet use detection dataset. Based on the dataset, we propose a metric to evaluate the performance of helmet use detection algorithms, which takes account of both spatial and temporal detection. The dataset, together with the source code for performance evaluation, are available in [17].

II. RELATED WORK

To date, a number of approaches for the automated detection of motorcycle helmet use in recorded video data have been proposed [10], [11], [18]–[24], details of which can be found in Table 1. For the initial step of active motorcycle detection, approaches can be broadly categorized into conventional methods [18]–[21] and deep-learning-based methods [10], [11], [22]–[24].

For the detection of active motorcycles, most conventional methods follow similar procedures. First, a background subtraction method is used to extract moving objects/vehicles from the video data. After this, a binary classifier (e.g. a support vector machine (SVM)) is used to detect motorcycles. In another step, the head region of the motorcyclists is localized, and an additional classifier is used to distinguish helmet use from non-helmet use. To improve the performance of the binary classifier, hand-crafted features are used, a common one is to extract a histogram of oriented gradients (HOG) [25] from the detected head regions of riders. Such methods, however, do not work well when there are many motorcycles and/or there is more than one rider on a motorcycle. Instead of designing hand-crafted features, deep learning based methods strive to automatically develop representations from raw image data that are most suitable for the helmet use detection task. In [24], helmet use is classified in the detected head regions of riders using a convolutional neural network (CNN). In [22] and [11], two independent CNNs are trained, one is used to distinguish motorcycles from other vehicles, the other to classify helmet and non-helmet in the head region of riders. Since it is time-consuming to detect motorcycles and helmet use through two separate CNNs, [10] and [23] use one single CNN to detect motorcycles and helmet use simultaneously.

The tracking of individual motorcycles through single frames of a recorded video is only included in half of existing approaches presented in Table 1. While video data recorded with traffic surveillance infrastructure is inherently frame-based, helmet use data produced through automatic detection must be projected onto individual motorcycles, to allow a valid appraisal of helmet use. Hence, frame-based detection results for motorcycle and rider counts, as well as helmet use must be remapped to individual motorcycles which appear in multiple frames. This can only be achieved by approaches that link frame-based detection to cross-frame tracking. This tracking is missing in some approaches (e.g. [11]). To compensate for this lack of tracking, it is necessary to either use single frame detection at a fixed point/line in the frame to prevent the repeated detection of the same motorcycle) (e.g. [20]) or to collect helmet use data in every video frame without tracking, leading to the loss of information on the number of motorcycles registered at an observation site (e.g. [10]). Both of these shortcuts lead to a decrease in helmet use data quality and in addition prevent the use of multiple frames of an individual motorcycle for helmet use and rider detection.

For rider number and position detection, only one of the approaches listed in (Table 1) generates detailed information on this [10]. And while other approaches (e.g. [20]) use head counts on the motorcycle as a substitute for rider numbers, this information is not mapped on rider positions

TABLE 1. Existing helmet use detection studies and related datasets.

Reference	Year	Detection	Tracking	Rider Differentiation	Observation sites	Annotated hours	Annotated frames	Annotated motorcycles	Frame rate	Resolution
[18]	2012	✓	✓	✗	–	–	>230	–	–	–
[19]	2013	✓	✓	✗	1	2	3,245	669	25	1280×720
[20]	2013	✓	✗	✗	1	–	440	–	30	640×480
[21]	2016	✓	✗	✗	1	2	–	105	30	–
[22]	2017	✓	✗	✗	5	3.5	–	3,678	25/30	–
[23]	2018	✓	✗	✗	1	4.2	–	493	–	–
[11]	2019	✓	✗	✗	6	–	5,000	–	24/25/30	360×238 to 1920×1080
[24]	2020	✓	✓	✗	1	2	–	1,645	25	1250×720
[10]	2020	✓	✗	✓	12	2.5	91,000	10,006	10	1920×1080
Ours	2020	✓	✓	✓						

– No information provided.

(i.e. driver vs. passenger). As the specific position and number of riders on a motorcycle directly relates to their helmet use [12]–[15], the lack of this critical information presents a clear barrier for the application of automated helmet use detection approaches in the field.

On the element of site-diversity, the existing datasets used to develop automated motorcycle helmet use detection approaches (Table 1) show a critical lack of diverse observation sites and a general lack of detailed information on road environments used. Five of the datasets [19]–[21], [23], [24] only contain data from one recording site, prohibiting robust evaluation of the developed solutions in diverse traffic environments. The two datasets which contain more recording sites do not distinguish helmet use between motorcycle drivers and passengers [11], [22]. This lack of data diversity and level of annotation detail in existing datasets hinders the development of widely applicable detection solutions.

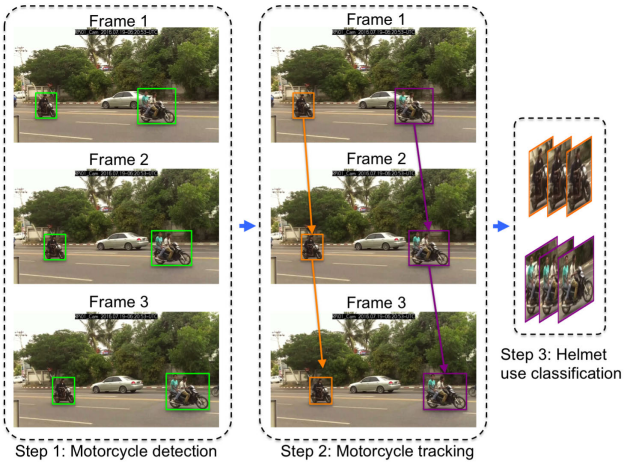


FIGURE 1. An illustration of the proposed approach for helmet use detection of tracked motorcycles.

III. THE PROPOSED APPROACH

Our proposed approach for helmet use detection of tracked motorcycles consists of the three steps, which are visualized in Fig. 1. In the first step of our approach, we use a fine-tuned pre-trained RetinaNet [26] for the detection of active motorcycles, i.e. motorcycles with at least one rider on them, on a single frame level. In the second step, each detected active motorcycle is tracked through adjacent frames, using both the motion state of the motorcycle as well as the visual similarity between detected active motorcycles. In the last step, when a track terminates, i.e. an individual motorcycle leaves the view of the video camera, the helmet use class of the tracked motorcycle is predicted, i.e. rider number, their position, and their helmet use are identified. All three steps are described in detail in the following sections.

A. MOTORCYCLE DETECTION

Detecting a motorcycle in a single frame is a classic object detection task. To this end we trained a state-of-the-art object detection algorithm to detect motorcycles in the dataset. Today's prevalent algorithms for object detection can be subdivided in two broad approaches: one-stage and two-stage. While the two-stage algorithms have overall higher accuracies in object detection, they are comparably slower, as frames are processed twice, once for identifying potential object locations in a frame, and once more for detecting the actual objects. Single-stage methods combine the steps of localizing potential objects and object detection into a single processing stage, which results in a small decrease in accuracy, but a large decrease in the processing time. A relatively new single-stage method is RetinaNet [26], which uses a multi-scale feature pyramid combined with focal loss to successfully overcome detection accuracy limitations. RetinaNet achieves faster detection than two-stage methods, while having a higher detection accuracy than comparable single-stage methods such as YOLO [27]. We therefore applied a RetinaNet model for detecting motorcycles.

Since motorcycle detection is very similar to other object detection tasks, instead of training from scratch, we fine-tuned a RetinaNet model with pre-trained weights obtained by the COCO dataset [28].

B. MULTIPLE MOTORCYCLE TRACKING

To clarify the procedure of motorcycle tracking, let $\mathcal{V} = \{v^{(1)}, \ldots, v^{(k)}\}$ be the set of existing tracks at time t. Using the notations in Table 2, a track is denoted by $v^{(i)} = (B^{(i)}, s^{(i)}_{t-1}, P^{(i)}_{t-1})$, where $B^{(i)}$ is a buffer to store the measurements that are assigned to the track $v^{(i)}$, $s^{(i)}_{t-1}$ and $P^{(i)}_{t-1}$

TABLE 2. Notations for multiple motorcycle tracking.

Symbol	Description
$m_t^{(j)}$	Measurement j at time t
$b_t^{(j)}$	Predicted bounding box j at time t
$x_t^{(j)}$	Cropped image from predicted bounding box j at time t
$B^{(i)}$	Buffer to the measurements that are assigned to track i
$s_t^{(i)}$	Estimated state of track i at time t
$\hat{s}_t^{(i)}$	Predicted state of track i at time t
$P_t^{(i)}$	Estimated state covariance of track i at time t
$\hat{P}_t^{(i)}$	Predicted state covariance of track i at time t
K	Kalman gain
A	State transition matrix, constant
H	Measurement matrix, constant
Q	Process noise covariance, constant
R	Measurement noise covariance, constant
I	Identity matrix

denote the state vector and the state covariance matrix of a Kalman filter [29]. Meanwhile, let $\mathcal{M} = \{m_t^{(1)}, \ldots, m_t^{(n)}\}$ be newly arrived measurements at time t. Each measurement is denoted as $m_t^{(j)} = (b_t^{(j)}, x_t^{(j)})$, where $b_t^{(j)}$ is the predicted bounding box, and $x_t^{(j)}$ is the cropped image patch from the predicted bounding box. Each image patch is re-scaled to 192×192. Furthermore, we normalize the bounding box by the frame width and height so that all the numbers fall between 0 and 1. Given a bounding box (l, u, w, h), its centroid z is computed as $z = (l + w/2, u + h/2)$.

For an existing track $v^{(i)}$, we first predict its new state $\hat{s}_t^{(i)}$ and new state covariance $\hat{P}_t^{(i)}$ using the estimated state $s_{t-1}^{(i)}$ and estimated state covariance $P_{t-1}^{(i)}$ at time $t-1$:

$$\begin{aligned} \hat{s}_t^{(i)} &= A s_{t-1}^{(i)}, \\ \hat{P}_t^{(i)} &= A P_{t-1}^{(i)} A^T + Q. \end{aligned} \tag{1}$$

Next, we compute the distance between all tracks and measurements, yielding a distance matrix D, where D_{ij} denotes the distance between track $v^{(i)}$ and measurement $m_t^{(j)}$. With matrix D, the measurement-to-track association is solved by the Munkres assignment algorithm [30].

To measure the distance D_{ij}, the conventional way is to compute the motion distance, namely, the squared Mahalanobis distance [31] between the predicted Kalman state of track $v^{(i)}$ and the centroid of the predicted bounding box $z_t^{(j)}$:

$$D_{ij}^{\mathrm{M}} = (z_t^{(j)} - H\hat{s}_t^{(i)})^T (H\hat{P}_t^{(i)} H^T + R)^{-1} (z_t^{(j)} - H\hat{s}_t^{(i)}). \tag{2}$$

While the motion distance is a suitable association metric when moving objects are sparse, the density of motorcycles in our dataset is very high, which results in a poor measurement-to-track association when multiple motorcycles are very close to each other. To address this limitation, in addition to the motion distance, we compute visual dissimilarity:

$$D_{ij}^{\mathrm{V}} = \frac{1}{N} \sum_{n=1}^{N} \|\phi(x^{(n)}; \theta) - \phi(x_t^{(j)}; \theta)\|_2, \tag{3}$$

where $x^{(n)}$ denotes the N cropped image patches that are assigned to track $v^{(i)}$, and $\phi(\cdot; \theta)$ corresponds to the feature vector learned by a InceptionV3 deep neural network model, to be defined later in Section III-D.

Hence, we have a combined distance as the product of the motion distance and the visual dissimilarity:

$$D_{ij} = D_{ij}^{\mathrm{M}} \cdot D_{ij}^{\mathrm{V}}. \tag{4}$$

To sum up, Eq. (4) indicates any track and measurement are similar only if they have similar visual appearances *and* similar motions.

Applying the Munkres assignment algorithm to the distance matrix D, any new measurement can either be assigned to an existing track or initiate a new track. If a measurement $m_t^{(j)}$ is assigned to an existing track $v^{(i)}$, the track is updated as:

$$\begin{aligned} K &= \hat{P}_t^{(i)} H^T (H\hat{P}_t^{(i)} H^T + R)^{-1}, \\ s_t^{(i)} &= \hat{s}_t^{(i)} + K(z_t^{(j)} - H\hat{s}_t^{(i)}), \\ P_t^{(i)} &= (I - KH)\hat{P}_t^{(i)}, \\ B^{(i)} &= B^{(i)} \cup m_t^{(j)}, \end{aligned} \tag{5}$$

where K is Kalman gain; otherwise it initiates a new track $v^{(i+1)}$, with track information updated by:

$$\begin{aligned} s_t^{(i+1)} &= \begin{bmatrix} l_t^{(j)} + w_t^{(j)}/2 \\ 0 \\ u_t^{(j)} + h_t^{(j)}/2 \\ 0 \end{bmatrix}, \\ P_t^{(i+1)} &= \begin{bmatrix} 0.100 & 0 & 0 & 0 \\ 0 & 0.025 & 0 & 0 \\ 0 & 0 & 0.100 & 0 \\ 0 & 0 & 0 & 0.025 \end{bmatrix}, \\ B^{(i+1)} &= m_t^{(j)}. \end{aligned} \tag{6}$$

If no measurement is assigned to an existing track, the track is updated as:

$$\begin{aligned} s_t^{(i)} &= \hat{s}_t^{(i)}, \\ P_t^{(i)} &= \hat{P}_t^{(i)}, \end{aligned} \tag{7}$$

which allows temporary occlusion or missing detection.

We close an existing track if no new measurement is assigned to it for more than 8 consecutive frames. We only keep closed tracks with a duration greater than 5 frames and a proportion of visible frames in a track greater than 60%. While the lack of new information after 8 consecutive frames reliably closes tracks of motorcycles that drive out of the camera's view, it also helps to close tracks that were incorrectly started by a false positive detection. The rule for a 5 frame minimum to keep a track then reliably leads to the deletion of these false positive tracks.

C. HELMET USE CLASSIFICATION

For a closed track of sufficient length, its helmet use is estimated by pooling the helmet use prediction of cropped

image patches within the track. More specifically, let $(x^{(n)})_{n=1}^{N}$ be the cropped image patches that are assigned to a tracked motorcycle, then the track's helmet use class is estimated as:

$$\hat{y} = \underset{y\in\{1,2,\ldots,C\}}{\arg\max} \frac{1}{N}\sum_{n=1}^{N} g(x^{(n)}; W), \tag{8}$$

where $g(\cdot; W)$ is a deep convolutional neural network (CNN), parameterized by W.

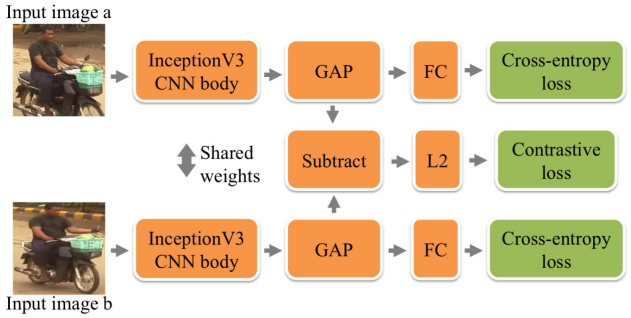


FIGURE 2. The proposed architecture for patch-based helmet use classification and visual similarity learning.

D. MTL FOR PATCH-BASED HELMET USE CLASSIFICATION AND VISUAL SIMILARITY LEARNING

In our approach, apart from fine-tuning RetinaNet for motorcycle detection, we need to train CNNs for two purposes. One is visual similarity learning. That is, we want the distance between two image patches to be small if they are in the same track, but large if they belong to different tracks. The other purpose is patch-based helmet use classification, i.e. we want to predict the helmet use class (rider number, position, and helmet use) given a cropped image patch. In our approach, instead of training two separate CNNs which is time-consuming, we apply multi-task learning (MTL) to learn both tasks simultaneously.

The architecture of the proposed deep learning model is illustrated in Fig. 2. A given pair of image patches $x^{(a)}$ and $x^{(b)}$ with 192 × 192 resolution, is feed into a Siamese network [32] that uses an InceptionV3 [33] CNN body with shared weights θ. The network body is truncated, such that the global average pooling layer (GAP) and the final fully-connected (FC) layer are removed. Each image patch x is transformed into a 2048-dimensional feature vector $\phi(x; \theta)$ after passing the output of InceptionV3 CNN body to a GAP layer.

With these two 2048-dimensional feature vectors, the model has three tasks to learn:

1) Given the feature vector $\phi(x^{(a)}; \theta)$, predict the helmet use class $p^{(a)} = f\left(\phi(x^{(a)}; \theta); w^{(a)}\right)$, where $f(\cdot)$ is a softmax regression model, parameterized by weight $w^{(a)}$.
2) Compute the Euclidean distance between the transformed feature vectors of image patch $x^{(a)}$ and image patch $x^{(b)}$:

$$d(x^{(a)}, x^{(b)}) = \|\phi(x^{(a)}; \theta) - \phi(x^{(b)}; \theta)\|_2, \tag{9}$$

3) Given $\phi(x^{(b)}; \theta)$, predict helmet use class $p^{(b)} = f\left(\phi(x^{(b)}; \theta); w^{(b)}\right)$, with the softmax regression model $f(\cdot)$ parameterized by weight $w^{(b)}$.

Using the MTL model, the helmet use classification model in Eq. (8) can be rewritten as $g(\cdot; W) = f(\phi(\cdot; \theta); w)$, where the visual similarity learning task and the helmet use classification task shares the weights θ in the training and predicting process, which not only significantly decreases the computational cost, but also improves generalization by using the domain information contained in the related tasks [34].

For the first and third tasks, we use the cross-entropy loss for optimization:

$$\begin{aligned} L_1(x^{(a)}, y^{(a)}) &= -\sum_{i=1}^{K} y_i^{(a)} log(p_i^{(a)}) \\ L_3(x^{(b)}, y^{(b)}) &= -\sum_{i=1}^{K} y_i^{(b)} log(p_i^{(b)}) \end{aligned} \tag{10}$$

where y is a one-hot vector that encodes the ground truth helmet use class, and K is the number of annotated helmet use classes. In the HELMET dataset, $K = 36$.

For the second task, we consider the contrastive loss [35]:

$$\begin{aligned} L_2(x^{(a)}, x^{(b)}) = &\sum_{(a,b)\in S} \max\left(d(x^{(a)}, x^{(b)}) - \tau_1, 0\right) \\ &+ \sum_{(a,b)\in D} \max\left(\tau_2 - d(x^{(a)}, x^{(b)}), 0\right), \end{aligned} \tag{11}$$

where S is an index set consisting of image pairs that come from the same track, and D is an index set consisting of image pairs that come from different tracks.

By minimizing the contrastive loss function, we expect the distance $d(x^{(a)}, x^{(b)})$ of an image pair in the same track is less than a threshold τ_1 and that of an image pair in different tracks is larger than a threshold τ_2. In this work, τ_1 and τ_2 are set as 1 and 5 empirically.

The loss function of the MTL framework is defined as:

$$L = \sum_{k=1}^{3} \gamma_k L_k, \tag{12}$$

where γ_k corresponds to the weight of task k. In our work, we assume each task has equal weight, namely $\gamma_k = 1/3$.

IV. HELMET DATASET

The HELMET dataset is an extension of our previous work [10], [12]. Here we give a brief introduction how we create and annotate the dataset, as well as how to evaluate the performance of helmet use detection approaches based on the dataset.

A. DATASET CREATION AND ANNOTATION

The source data for the HELMET dataset consists of 385 hours of traffic video, recorded in 2016 over a two month period in the country of Myanmar. Video data collection was planned and conducted in close consultation with the Myanmar Traffic Police Force, to ensure adherence to local laws and regulation. Using two video-cameras built from a Raspberry Pi 3 mini-computer and a Raspberry Pi camera module, 13 observation sites around seven cities in Myanmar were recorded at a rate of 10 frames per second, with a resolution of 1920 × 1080 pixels. The recorded data include diverse road environments, various traffic densities and different weather conditions. Before selecting video data for the HELMET dataset, the underlying source data was cleaned up, and video sections were removed when they contained motion blur due to cloudy weather or rain, which would have prohibited their detailed annotation. After this pre-processing, video clips of 242 hours length taken at 12 observation sites remained.

To most efficiently utilize the available annotation resources, it was decided to preferentially annotate video sections that contain a high number of motorcycles. After splitting up the video data from each observation site into 10 second video clips (100 frames per clip), the pre-trained YOLO9000 [27] object detection algorithm was applied to identify the number of motorcycles in each clip. Broadly maintaining the share of individual observation sites in the source data, 910 video clips with the highest number of motorcycles (identified through YOLO9000) were chosen for annotation. The resulting distribution of the 910 sampled video clips (91,000 frames) is presented in Table 3.

TABLE 3. Properties of the HELMET Dataset.

Observation site	Duration (minutes)	Annotated frames	Annotated motorcycles	Annotated motorcyclists
Bago_highway	5.8	3,500	225	322
Bago_rural	11.2	6,700	163	232
Bago_urban	10.5	6,300	475	694
Mandalay_1	38	22,800	2,224	3,197
Mandalay_2	31.7	19,000	4,387	7,106
Naypyitaw_1	8.5	5,100	167	247
Naypyitaw_2	7.2	4,300	213	310
NyaungU_rural	13.8	8,300	312	516
NyaungU_urban	11.2	6,700	579	782
Pakokku_urban	12.5	7,500	955	1,389
Pathein_rural	2	1,200	71	111
Yangon_II	6.5	3,900	235	387
Overall	151.7	91,000	10,006	15,293

Data in the HELMET dataset was annotated by drawing a rectangular bounding box around motorcycles, and adding information on the number of riders, their positions, and rider specific helmet use. The structure for rider position annotation is shown in Figure 3, distinguishing between the driver (D), multiple passengers (P1-P3), and a child passenger (P0) standing on the floorboard of the motorcycle in front of the driver. Bounding boxes of individual motorcycles are linked over subsequent frames in the annotation process, i.e. the identification of bounding boxes belonging to individual motorcycles is possible in the HELMET dataset,

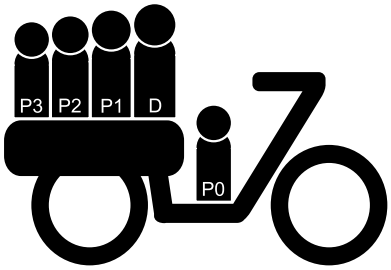


FIGURE 3. Rider position encoding.

forming the basis of the developed tracking approach. This *track* of motorcycle bounding boxes belonging to a single motorcycle, imprinted with information on rider number, rider position, and helmet use, is defined as a continuous helmet use event (CHUE). The number of CHUEs is identical with the number of individual motorcycles observed (Table 3).

All annotation was conducted using the program BeaverDam [36] and each annotation was verified by a second annotator. The 91,000 annotated frames from 12 observation sites form the HELMET dataset, which can be accessed by researchers free of charge with the publication of this article [17].

B. EVALUATION METRIC

To facilitate a consistent evaluation of the performance of approaches for helmet use detection of tracked motorcycles, we adapt a metric for continuous visual event recognition proposed in [37] for use with CHUE.

Let a CHUE be a tuple $E = (L, T)$, where L is helmet use class, and T is motorcycle track. The motorcycle track T is defined as $T = (f_i, b_i)_{i=1}^{N}$, where f_i is the frame number, and $b_i = (l_i, u_i, w_i, h_i)$ is the bounding box defined by location information within the frame: left (l_i), upper (u_i), width (w_i), and height (h_i), and N is the duration of the track.

A detected helmet use event E_{Detect} is regarded as a correct detection w.r.t. a ground truth event E_{GT} only if it satisfies the following conditions:

- Given a bounding box pair from E_{Detect} and E_{GT} in an individual frame f_i, a correct *frame* detection is defined as an intersection over union (IoU) of above 50% between E_{Detect} and E_{GT} inf_i.
- Given a number of correct frame detections, a correct *track* detection is registered if the ratio of correct individual detections in E_{Detect} in relation to the track duration N of E_{GT} is above 50%.
- Given a predicted helmet use class L_{Detect}, a correct detection is registered if L_{Detect} is identical to the labeled class L_{GT}.

Following these criteria, we are able to measure the performance of an approach by the following metrics: precision, recall, and weighted aggregate F-measure.

- Precision is the ratio of the number of correct E_{Detect} to the total number of E_{Detect} (correct and incorrect) in the i-th class.
- Recall is the ratio of the number of correct E_{Detect} to the number of correct E_{Detect} combined with missed E_{GT} in i-th class.
- For i-th class, the F-measure is the harmonic mean of precision and recall:

$$F_i = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

- Since samples across all helmet use classes L_{GT} are imbalanced, we use a weighted aggregate F-measure in the dataset, defined as:

$$F_{\text{weighted}} = \frac{1}{w_i \times \sum_{i=1}^{C} \frac{1}{F_i}} \tag{14}$$

where C is the number of helmet use classes, and the weight on the i-th class w_i is proportional to the number of samples in the i-th class; w_i's sum to one.

TABLE 4. Segmentation of video clips in the dataset for performance evaluation, the training-validation-test split ratio is 70%, 10%, and 20%.

Observation site	Training	Validation	Test	Overall
Bago_highway	24	4	7	35
Bago_rural	41	6	11	58
Bago_urban	44	6	13	63
Mandalay_1	159	23	45	227
Mandalay_2	111	16	31	158
Naypyitaw_1	36	5	10	51
Naypyitaw_2	30	4	9	43
NyaungU_rural	57	8	17	82
NyaungU_urban	47	7	13	67
Pakokku_urban	52	8	15	75
Pathein_rural	8	1	3	12
Yangon_II	27	4	8	39
Overall	636	92	182	910

V. EXPERIMENTAL RESULTS

A. TRAINING SETUP

To evaluate our proposed method, the 910 annotated video clips were randomly divided into three non-overlapping subsets: a training set (70%), a validation set (10%), and a test set (20%) according to each individual site, as shown in Table 4. We used the training set to train our proposed method, and used the validation set to find the best generalizing model. Given the best generalizing model, we report the final model performance on the test set.

For multiple motorcycle tracking, the parameters A, H, Q, and R of Kalman filter are predefined and given by:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$Q = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.25 \end{bmatrix}$$

$$R = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}$$

To train and evaluate the MTL model, we first generated all pairs of image patches in each video clip. Next we randomly sampled 2,000,000, 100,000, and 200,000 pairs from training, validation, and test sets respectively. In each subset, 50% image pairs come from the same tracks and 50% image pairs come from different tracks. In our work, the CNN body was initialized with the pre-trained weights on ImageNet [38] and all FC layers were initialized with random weights.

For both deep learning models, i.e. the motorcycle detection model and the MTL model, the Adam optimizer [39] was used with the default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a custom learning rate α. In our experiments, we tried $\alpha = 10^{-2}, 10^{-3}, \ldots, 10^{-5}$ and chose the value that gives the best validation result. Considering the large size of data, we only trained for 10 epochs with a batch size of 2 for the motorcycle detection model and 128 for the MTL model. In the training process, we saved the best model that gave the minimum loss on the validation set and reported its final performance on the test set.

B. RESULTS AND ANALYSIS FOR MOTORCYCLE DETECTION

The first step in our approach is the detection of active motorcycles (Step 1 in Fig. 1). As described in section III-A, a fine-tuned RetinaNet was used for this task.[1] To evaluate the performance of our model, we use the average precision (AP) metric [40]. The AP on the test set is very high, achieving 95.3% for the detection of motorcycles. Fig. 4 visualizes the motorcycle detection results for four sampled frames. It can be observed that the motorcycle detection through RetinaNet is very close to the human annotation despite the occlusions occurring to some motorcycles and riders.

C. RESULTS AND ANALYSIS FOR MULTI-TASK LEARNING

As helmet use detection and visual similarity are two tasks in the MTL procedure, their results are presented separately.

For helmet use classification, as there are two such tasks in MTL, we selected the output with a lower loss on the validation set. Using this model resulted in an 80.6% accuracy for the detection of motorcycle helmet use classes on all 54,529 annotated bounding boxes in the test set. In other words, in 80.6% of all detected active motorcycles in the test set, our approach correctly classified the number of riders, their position, and their helmet use.

For visual similarity learning, two types of errors can occur, different motorcycles can be falsely classified as belonging to the same track, or the same motorcycle can be falsely classified as belonging to different tracks. To evaluate

[1] Implemented using https://github.com/fizyr/keras-retinanet

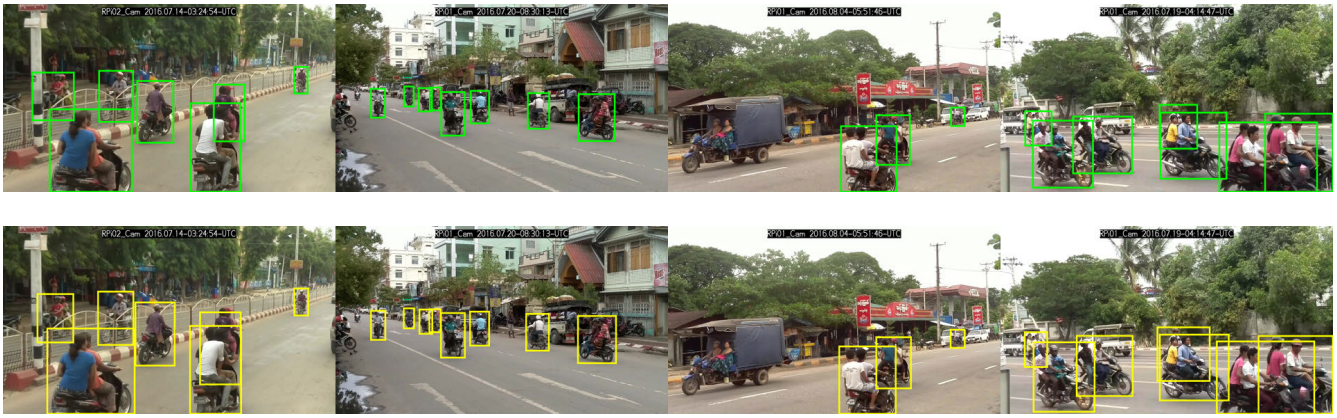


FIGURE 4. Motorcycle detection results on sampled frames, where the top row corresponds to the ground truth human annotation (green bounding boxes), and the bottom row corresponds to the results (yellow bounding boxes), predicted by fine-tuned RetinaNet.

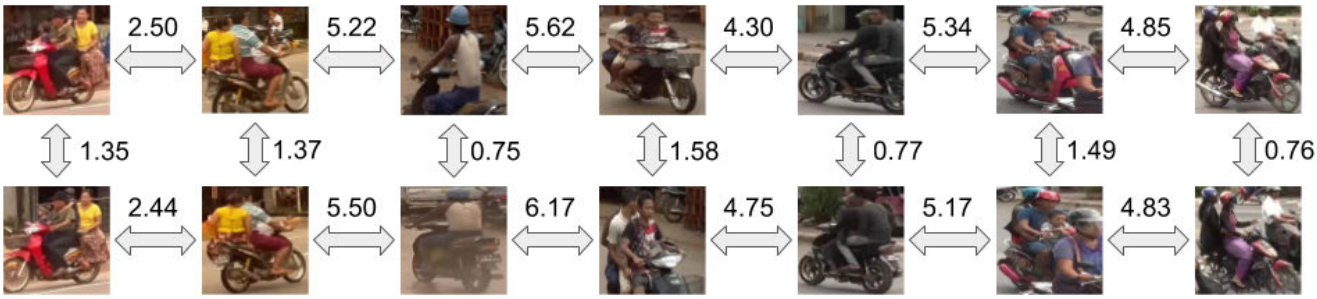


FIGURE 5. Examples of the output distances of the MTL model between image pairs of the same track (vertical comparison) and different tracks (horizontal comparison), with smaller numbers showing higher visual similarity.

the performance of visual similarity learning we draw the receiver operating characteristic (ROC) curve and compute the area under the ROC curve (AUC), as shown in Fig. 6. We tune the threshold on the result of visual similarity learning of 200,000 pairs, where axis x and y correspond to false positive rate (FPR) and true positive rate (TPR), defined as:

$$\begin{aligned} \text{FPR} &= \frac{\text{false positive}}{\text{false positive} + \text{true negative}}, \\ \text{TPR} &= \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \end{aligned} \tag{15}$$

where "true positive" and "true negative" corresponds to the number of correctly detected pairs from the same motorcycle track and correctly detected pairs from different tracks respectively. "False negative" describes the number of pairs from the same track which were detected as belonging to a different track, and "false positive" describes the number of pairs that are falsely detected as belonging to the same track. The proposed model applied to visual similarity learning resulted in an AUC of 0.967. Fig. 5 shows the output visual distances of the MTL model between sampled image pairs of the same and different tracks. For these examples, a threshold of 2.0 would classify every image pair correctly.

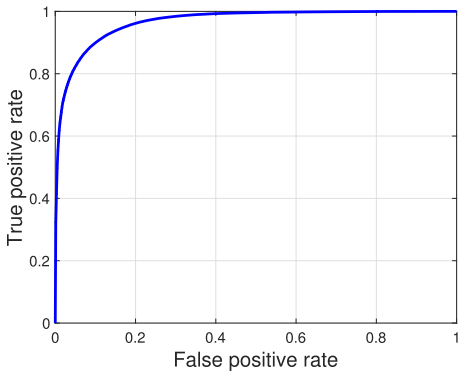


FIGURE 6. The ROC curve of visual similarity learning on the test set (20,000 image pairs).

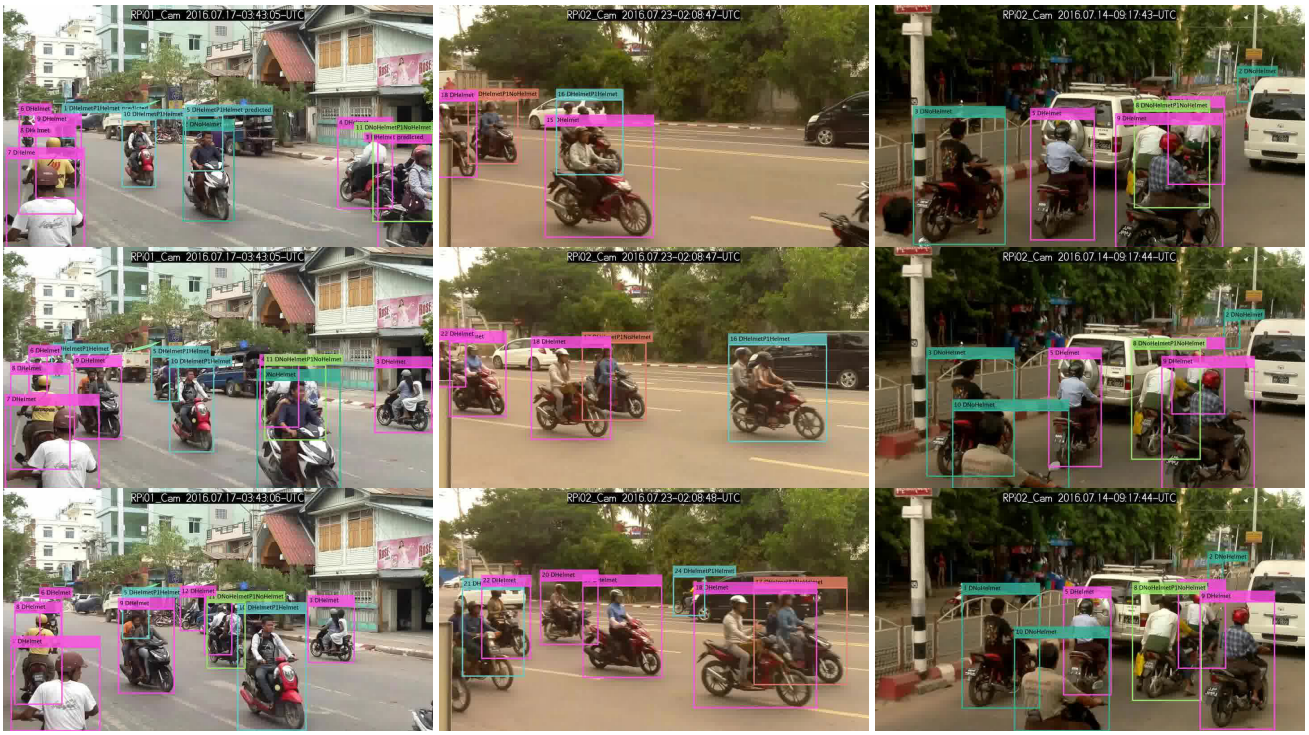
D. ABLATION STUDY

To investigate the effectiveness of the key components of our approach, we conduct six ablation experiments, in which different components of our approach are removed or replaced to learn more about their contribution to detection accuracy and computational efficiency.

As shown in Table 5, for the first two ablation experiments, we use frame-based motorcycle and helmet use detection from a single CNN (YOLOv2 or RetinaNet), as the basis for detection, adding motorcycle tracking through motion distance from our approach. The motorcycle class is determined by the majority of predicted helmet use classes of a track. For the additional four ablation experiments, we used the approach presented in this paper, but use different networks, i.e., YOLOv2 or RetinaNet, for motorcycle detection and

TABLE 5. Results of ablation study on the test set of HELMET dataset.

	Frame-based detection	Tracking	Track-based detection	Weighted F-measure (%)	Time (FPS)
1	YOLOv2 (motorcycle and helmet use)	Motion-based	Voting	32.4	25.12
2	RetinaNet (motorcycle and helmet use)	Motion-based	Voting	44.5	13.42
3	YOLOv2 (motorcycle)	Motion-based	Helmet use	56.1	9.51
4	YOLOv2 (motorcycle)	Hybrid	Helmet use	60.0	9.44
5	RetinaNet (motorcycle)	Motion-based	Helmet use	61.5	8.41
6	RetinaNet (motorcycle)	Hybrid	Helmet use	67.3	8.32



(a) Mandalay_1 (b) Mandalay_2 (c) Pakokku_urban

FIGURE 7. Helmet use detection of tracked motorcycles on sampled frames in three observation sites. Each column corresponds to the frames sampled (every 5th frame) from a video clip. where different bounding box colors corresponds to different predicted classes. The predicted helmet use class and track id are labelled at the top of bounding boxes.

different schemes, motion-based or hybrid, for motorcycle tracking.

For computational speed, it can be observed that the first two approaches are achieving the highest number of processed frames per second (25.12 and 13.42 FPS), as they only use one network to simultaneously predict motorcycle and helmet use class. However, it can be observed that this high speed is achieved at the expense of detection accuracy, as the two approaches are prone to produce missing detections due to the imbalanced helmet use classes. This in turn decreases tracking performance. Comparing RetinaNet and YOLOv2, the advantage of the multi-scale feature pyramid with focal loss is apparent, as RetinaNet has a higher accuracy than YOLOv2, at the expense of processing speed. This difference is present for detecting motorcycle and helmet use simultaneously or separately. Finally, combining motion similarity with visual similarity during tracking (our *hybrid tracking* approach) improves the F-measure value with little additional computational cost.

E. RESULTS AND ANALYSIS OF HELMET USE DETECTION

Detailed results of our proposed approach for helmet use detection of tracked motorcycles in each individual class are presented in Table 6. We achieved a 67.3% weighted F-measure on the test set of the HELMET dataset. Our approach works well on common classes of up to two riders per motorcycle. Considering only these common classes, the weighted F-measure improves to 70.6%. Fig. 7 shows detection results on some sampled frames. For more detailed results, we have attached video samples of our approach as *supplementary files*.

In addition we present the location-wise performance in Fig. 8, which shows the weighted F-measure for each observation site in the HELMET dataset. Video clips from the test set for all sites can be found in the *supplementary files* of this article. It can be observed that the approach works well for most locations, with nine of the observation sites showing an F-measure of around 70% and above. However, comparatively low accuracy can be observed for Bago_urban

TABLE 6. Performance evaluation in each individual class.

Class	Position					No. of Tracks				Helmet use detection
	D	P1	P2	P3	P0	Training	Validation	Test	Overall	F-measure (%)
1	✓	–	–	–	–	3,054	446	786	4,286	74.8
2	✓	✓	–	–	–	1,623	205	412	2,240	73.8
3	✗	–	–	–	–	878	118	234	1,230	66.4
4	✗	✗	–	–	–	672	99	145	916	60.9
5	✓	✗	–	–	–	323	41	67	431	48.7
6	✗	✗	✗	–	–	168	11	31	210	25.5
7	✗	✓	–	–	–	99	12	17	128	25.0
8	✓	✗	✓	–	–	88	23	16	127	25.8
9	✓	✗	✗	–	–	55	8	12	75	34.8
10	✗	✗	–	–	✗	38	5	11	54	11.8
11	✓	✓	–	–	✗	29	4	16	49	36.4
12	✗	✗	✗	–	✗	25	4	6	35	25.0
13	✓	–	–	–	✗	20	4	11	35	30.8
14	✗	–	–	–	✗	20	4	5	29	0
15	✓	✗	–	–	✗	21	4	1	26	0
16	✓	✓	✓	–	–	17	1	4	22	0
17	✓	✓	–	–	✓	15	1	6	22	28.6
18	✓	✗	✓	–	✗	13	5	2	20	22.2
19	✓	–	–	–	✓	11	0	0	11	n/a
20	✓	✗	✗	–	✗	10	0	0	10	n/a
21	✗	✗	✓	–	–	6	3	0	9	n/a
22	✗	✗	✗	✗	–	5	1	0	6	n/a
23	✓	✓	✗	–	–	6	0	0	6	n/a
24	✓	✗	✗	✓	–	3	2	1	6	0
25	✓	✗	✓	–	✓	3	0	1	4	0
26	✓	✓	✓	–	✗	2	1	0	3	n/a
27	✓	✗	✗	✗	–	3	0	0	3	n/a
28	✗	✓	✓	–	–	3	0	0	3	n/a
29	✗	✓	–	–	✗	2	0	0	2	n/a
30	✓	✗	✗	–	✓	2	0	0	2	n/a
31	✗	✗	–	–	✓	1	0	0	1	n/a
32	✗	✗	✗	✗	✗	1	0	0	1	n/a
33	✓	✓	✓	–	✓	0	0	1	1	0
34	✗	✗	✓	–	✗	0	0	1	1	0
35	✓	✗	✗	✓	✗	0	0	1	1	0
36	✓	✗	✗	✗	✗	0	0	1	1	0
						7,216	1,002	1,788	10,006	weighted avg: 67.3

✓ Motorcyclist in corresponding position wear a helmet
✗ Motorcyclist in corresponding position does not wear a helmet
– There is no motorcyclist in corresponding position
n/a Not applicable since there is no test data in the corresponding class

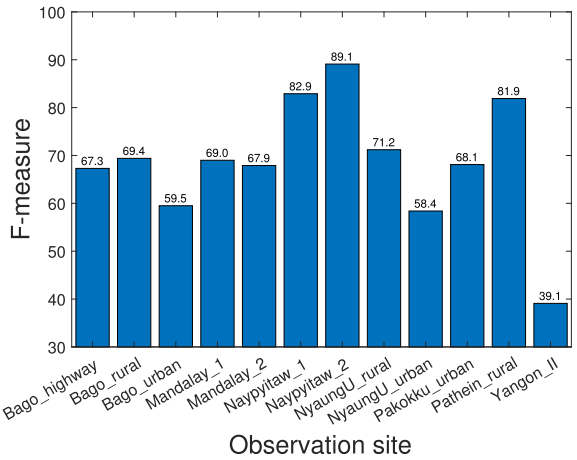


FIGURE 8. Performance evaluation in each observation site.

and NyaungU_urban, with F-measures slightly below 60%, and Yangon_II with an F-measure of only 39.1%. Looking at the video clips in the test dataset (see *supplementary files*), it becomes apparent, that the three sites with the lowest F-measures have properties that can be linked to the low accuracy of our approach in these environments. Bago_urban and Yangon_II contain a crossroad, i.e. motorcycles appear in the observation camera's view in unusual angles compared to the other sites. The observation site NyaungU_urban contains a large number of parked motorcycles, on which riders rest. Our approach detects and registers these motorcycles, leading to an inaccurate detection, since parked motorcycles were not annotated in the annotation process, as they are not actively used.

F. COMPUTATIONAL COST

Our approach was implemented using the Python Keras library with Tensorflow as a backend and ran on two NVIDIA Titan Xp GPUs. In our implementation, instead of keeping every cropped image patch in a track, we retain its visual feature and helmet use prediction output only, which reduces both computational space and time.

The overall processing speed of our method is 8.32 FPS. More specifically, the computational time for motorcycle

detection is 0.059 seconds per frame; the computational time for visual feature extraction and patch-based helmet use classification is 0.058 seconds per frame; and the computational time for tracking is negligible, merely 0.003 seconds per frame.

VI. CONCLUSION

In this paper, we have proposed a deep learning based method to automatically perform three elements of human observer motorcycle helmet use registration, i.e. detection and tracking of active motorcycles, as well as identification of rider number per motorcycle, rider position, and rider specific helmet use. In addition, we have applied our approach to video data from diverse road environments, which included adverse factors such as occlusion, differences in camera angle, an imbalanced number of coded classes, as well as differing rider numbers per motorcycle and varying traffic densities. All of these elements make our approach more comprehensive than earlier approaches for the automated detection of motorcycle helmet use (see Table 1). Our results show a generally high accuracy of our approach. For the element of frame-based detection of motorcycles, we achieve an average precision of 95.3%. The visual similarity element of motorcycle tracking of our approach achieves 0.967 AUC, in this first application of CNN-based tracking of active motorcycles. For the element of detection of helmet use class, i.e. the registration of rider number, position, and rider specific helmet use, we achieve an accuracy of 80.6% on a frame based level. Especially the imbalanced number of classes in the HELMET dataset contribute to wrong classifications. For the comprehensive application of our approach, all its elements are combined, i.e. motorcycle detection, tracking, and helmet use class prediction are jointly applied. Our results show a weighted F-measure of 67.3% for the helmet use detection of tracked motorcycles, showing that our approach can be used to generate reliable motorcycle, rider number, and position specific helmet use estimates. The results of our ablation study show that our approach achieves a comparatively high accuracy against ablation experiments. While this high accuracy comes at the expense of computational efficiency, our approach can process more than 8 FPS on consumer hardware, which is close to real-time speed for 10 FPS video data. Overall, our work shows that all four basic elements of helmet use registration through human observers can be implemented in a CNN-based approach that is computationally efficient on consumer hardware. Furthermore, the inclusion of detailed rider differentiation is an enhancement of existing approaches. In addition to presenting our helmet use detection approach, we publish the HELMET dataset with this paper, which includes diverse traffic video data that can be used to train and evaluate similar approaches. Since existing datasets have a number of shortcomings and are not readily available to researchers, we hope that the publication of the HELMET dataset will advance the development and evaluation of detection approaches similar to the one in this paper.

There are some limitations to our work. The detection accuracy can be much compromised when dealing with uncommon traffic environments, or street scenes with parked motorcycles. Hence, the current approach is partly constrained in real-world applicability, as observation site specific elements could decrease detection accuracy. And while the HELMET dataset is a first step towards using more diverse datasets for the development of automated helmet detection approaches, further data needs to be collected to make approaches universally applicable.

For future research, we intend to enhance the HELMET dataset by incorporating scenes with more diverse traffic infrastructure, e.g. crossroads, to ensure more robust application of the approach. Also, more training data that contains parked motorcycles will be acquired and used for training, so that these objects will not be detected as false positive.

REFERENCES

[1] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," *Transp. Res. C, Emerg. Technol.*, vol. 6, no. 4, pp. 271–288, Aug. 1998.

[2] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[3] S. Gupte, O. Masoud, R. F. K. Martin, and N. P. Papanikolopoulos, "Detection and classification of vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 3, no. 1, pp. 37–47, Mar. 2002.

[4] J. Wu, Z. Liu, J. Li, C. Gu, M. Si, and F. Tan, "An algorithm for automatic vehicle speed detection using video camera," in *Proc. 4th Int. Conf. Comput. Sci. Edu.*, Jul. 2009, pp. 193–196.

[5] T. N. Schoepflin and D. J. Dailey, "Dynamic camera calibration of roadside traffic management cameras for vehicle speed estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 2, pp. 90–98, Jun. 2003.

[6] D.-W. Lim, S.-H. Choi, and J.-S. Jun, "Automated detection of all kinds of violations at a street intersection using real time individual vehicle tracking," in *Proc. 5th IEEE Southwest Symp. Image Anal. Interpretation*, Apr. 2002, pp. 126–129.

[7] H. Veeraraghavan, O. Masoud, and N. P. Papanikolopoulos, "Computer vision algorithms for intersection monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 2, pp. 78–89, Jun. 2003.

[8] Y. Artan, O. Bulan, R. P. Loce, and P. Paul, "Driver cell phone usage detection from HOV/HOT NIR images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 225–230.

[9] Y. Artan, O. Bulan, R. P. Loce, and P. Paul, "Passenger compartment violation detection in HOV/HOT lanes," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 395–405, Feb. 2016.

[10] F. W. Siebert and H. Lin, "Detecting motorcycle helmet use with deep learning," *Accident Anal. Prevention*, vol. 134, Jan. 2020, Art. no. 105319.

[11] B. Yogameena, K. Menaka, and S. Saravana Perumaal, "Deep learning-based helmet wear analysis of a motorcycle rider for intelligent surveillance system," *IET Intell. Transp. Syst.*, vol. 13, no. 7, pp. 1190–1198, Jul. 2019.

[12] F. W. Siebert, D. Albers, U. Aung Naing, P. Perego, and S. Chamaiparn, "Patterns of motorcycle helmet use–A naturalistic observation study in Myanmar," *Accident Anal. Prevention*, vol. 124, pp. 146–150, May 2019.

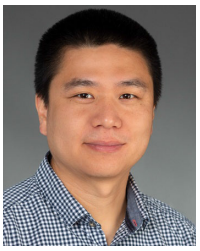
[13] R. D. Ledesma, S. S. López, J. Tosi, and F. M. Poó, "Motorcycle helmet use in mar del plata, argentina: Prevalence and associated factors," *Int. J. Injury Control Saf. Promotion*, vol. 22, no. 2, pp. 172–176, Apr. 2015.

[14] D. V. Hung, M. R. Stevenson, and R. Q. Ivers, "Prevalence of helmet use among motorcycle riders in Vietnam," *Injury Prevention*, vol. 12, no. 6, pp. 409–413, Dec. 2006.

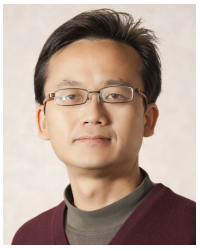
[15] A. M. Bachani, N. T. Tran, S. Sann, M. F. Ballesteros, C. Gnim, A. Ou, P. Sem, X. Nie, and A. A. Hyder, "Helmet use among motorcyclists in cambodia: A survey of use, knowledge, attitudes, and practices," *Traffic Injury Prevention*, vol. 13, no. sup1, pp. 31–36, Mar. 2012.

[16] H. Lin. (2020). *Helmet Use Detection Source Code*. [Online]. Available: https://github.com/LinHanhe/Helmet_use_detection

[17] H. Lin and F. W. Siebert. (2020). *The HELMET Dataset*. [Online]. Available: https://osf.io/4pwj8/
[18] J. Chiverton, "Helmet presence classification with motorcycle detection and tracking," *IET Intell. Transport Syst.*, vol. 6, no. 3, pp. 259–269, 2012.
[19] R. Silva, K. Aires, T. Santos, K. Abdala, R. Veras, and A. Soares, "Automatic detection of motorcyclists without helmet," in *Proc. XXXIX Latin Amer. Comput. Conf. (CLEI)*, Oct. 2013, pp. 1–7.
[20] R. Waranusast, N. Bundon, V. Timtong, C. Tangnoi, and P. Pattanathaburt, "Machine vision techniques for motorcycle safety helmet detection," in *Proc. 28th Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2013, pp. 35–40.
[21] K. Dahiya, D. Singh, and C. K. Mohan, "Automatic detection of bike-riders without helmet using surveillance videos in real-time," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3046–3051.
[22] C. Vishnu, D. Singh, C. K. Mohan, and S. Babu, "Detection of motorcyclists without helmet in videos using convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 3036–3041.
[23] N. Boonsirisumpun, W. Puarungroj, and P. Wairotchanaphuttha, "Automatic detector for bikers with no helmet using deep learning," in *Proc. 22nd Int. Comput. Sci. Eng. Conf. (ICSEC)*, Nov. 2018, pp. 1–4.
[24] L. Shine and C. V. Jiji, "Automated detection of helmet on motorcyclists from traffic surveillance videos: A comparative analysis using hand-crafted features and CNN," *Multimedia Tools Appl.*, vol. 79, pp. 14179–14199, Feb. 2020.
[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland. Springer, 2014, pp. 740–755.
[29] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
[30] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, Mar. 1957.
[31] G. J. McLachlan, "Mahalanobis distance," *Resonance*, vol. 4, no. 6, pp. 20–26, 1999.
[32] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.
[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[34] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[35] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.
[36] A. Shen, "Beaverdam: Video annotation tool for computer vision training labels," M.S. thesis, EECS Dept., Univ. California, Berkeley, CA, USA, Dec. 2016. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-193.html
[37] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, and C. C. Chen, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. CVPR*, Jun. 2011, pp. 3153–3160.
[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[40] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.



HANHE LIN received the Ph.D. degree from the Department of Information Science, University of Otago, New Zealand, in 2016. He is currently a Postdoctoral Researcher with the Department of Computer and Information Science, University of Konstanz, Germany. His research interests include machine learning and deep learning-based application, visual quality assessment, and crowdsourcing.



JEREMIAH D. DENG (Member, IEEE) received the B.Eng. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 1989, the M.Eng. degree from the South China University of Technology (SCUT), Guangzhou, China, in 1992, and the Ph.D. degree under the co-supervision of SCUT and The University of Hong Kong, Hong Kong, in 1995. He joined the University of Otago, Dunedin, New Zealand, as a Research Fellow, in 1999, where he is currently an Associate Professor with the Department of Information Science. He has published over 110 technical articles in machine learning, signal processing, and mobile computing.

DEIKE ALBERS received the M.Sc. degree in human factors engineering from the Technical University of Munich (TUM), where she is currently pursuing the Ph.D. degree with the Chair of Ergonomics. She is also a Researcher. Her research interests include traffic safety and validity of usability assessments under different testing conditions.

FELIX WILHELM SIEBERT received the M.Sc. degree in human factors from the Technische Universität Berlin, Germany, and the Ph.D. degree in psychology from the Leuphana University of Lüeneburg, Germany. He holds a postdoctoral position with the Department of Psychology, Friedrich-Schiller University of Jena, Germany. His research interests include safety of vulnerable road users and impact of new forms of mobility on transport systems.