Published in IET Intelligent Transport Systems
Received on 8th August 2011
Revised on 8th March 2012
doi: 10.1049/iet-its.2011.0138

ISSN 1751-956X

# Helmet presence classification with motorcycle detection and tracking

*J. Chiverton*

*School of Information Technology, Mae Fah Luang University, 333 Moo 1, Thasud, Muang, Chiang Rai 57100, Thailand*
*E-mail: jpchiverton@theiet.org*

**Abstract:** Helmets are essential for the safety of a motorcycle rider, however, the enforcement of helmet wearing is a time-consuming labour intensive task. A system for the automatic classification and tracking of motorcycle riders with and without helmets is therefore described and tested. The system uses support vector machines trained on histograms derived from head region image data of motorcycle riders using both static photographs and individual image frames from video data. The trained classifier is incorporated into a tracking system where motorcycle riders are automatically segmented from video data using background subtraction. The heads of the riders are isolated and then classified using the trained classifier. Each motorcycle rider results in a sequence of regions in adjacent time frames called tracks. These tracks are then classified as a whole using a mean of the individual classifier results. Tests show that the classifier is able to accurately classify whether riders are wearing helmets or not on static photographs. Tests on the tracking system also demonstrate the validity and usefulness of the classification approach.

## 1 Introduction

Motorcycle helmets are essential for motorcycle safety [1]. Unfortunately their use is not always easily enforced, particularly where the culture is not accustomed to such practices (see e.g. [2–4]). The goal of this work is to provide an automated system approach to the detection of motorcycle riders not wearing a helmet.

A few systems have previously been proposed that include the detection of helmets as part of some other system goal. Chiu *et al.* [5] used helmet detection as an indicator of whether a motorcycle was present in a foreground region of the image data. Their technique relied on the use of a vertical histogram projection of the silhouette of the moving object to identify the location of the head of the rider. Then edges were detected and accumulated to determine if a circular object was present in the head region. The presence of a circular object was then used as an indicator of the presence of a helmet.

The security of retailers and automatic payment systems have prompted the investigation of techniques to automatically determine if individuals might be hiding their identity with motorcycle helmets [6, 7]. These techniques also rely on the assumption of the circular shape of the helmet and even the lack of presence of skin hues in the image data. Skin tones cannot be used for an automated traffic system as the system might be expected to detect motorcyclists' helmets from a rear view. Furthermore many styles of helmet do not obscure the face.

Circle and arc detection are relatively simple and well-defined techniques for the detection of circular objects. They are therefore perhaps one of the first techniques one might try for the detection of circular type objects, such as helmets. The assumption of a circular shape for a motorcycle helmet in [6, 7] results in a single numerical value representing the estimated size of the helmet which can be used, in combination with a threshold-type operation to classify the presence (or not) of a helmet. However, this single numerical value is highly dependent on noise and other artefacts including the fact that human heads and motorcycle helmets are both approximately circular, at least for part of their boundaries. These techniques are therefore not able to differentiate between a person wearing a helmet and a person who is not wearing a helmet. Nevertheless the detection of a head whether or not a helmet is being worn is still useful for the work described in [5] where the presence of a circular region is used as an indicator of the presence of a motorbike. For these reasons, both techniques result in far too many false positives for our application which we found after implementing both techniques. An example result is shown in Fig. 1. This has lead us to investigate alternative techniques for feature extraction and classification.

More robust object recognition techniques extract multiple features that are characteristic of the object as a whole (e.g. [8–10]). Multiple features are commonly set as the input to a classifier, such as the AdaBoost algorithm that has been used to detect faces in images [8] and more recently for the detection of cars in [9]. Another technique which has also been found to be fairly robust was initially demonstrated for person detection [10] using support vector machines (SVMs) and features based on localised histograms of oriented gradients (HOGs). Other similar techniques have also sought to describe some unique

**Fig. 1** *Illustration of potential false positives using a circle or arc feature extraction technique for helmet detection using a method similar to [6]*

characteristics of an object through histograms such as [11] incorporating spatial information for tracking. However, none of these techniques have been applied to the detection of people wearing helmets.

Automated traffic monitoring is a popular research topic primarily because of the demonstrable usefulness of working traffic monitoring systems (see e.g. [12]). Video-based traffic monitoring sensors are gaining increasing interest because of their potential to perform many useful traffic monitoring functions, including speed checking, traffic counting and general monitoring and control (see e.g. [13]). However, there are many potential problems that have yet to be solved to enable such systems to be fully exploited such as variable light and weather conditions [14] and shadow effects [15]. Traditional approaches to traffic monitoring have used various types of equipments such as radar, video and induction loops [12]. Some researchers have sought to combine the modalities to enhance the overall performance of more traditional devices [16], such as continuous wave (CW) Doppler radar speed measurement [17], which was combined with video data to enable more intelligent vehicle identification for individual speed measurements.

Computer vision systems for intelligent traffic monitoring are already industrially viable albeit with particularly special requirements in terms of the location of the video acquisition device. For example, many systems have to be positioned high above the road way to minimise the possibility of occlusions which still can result in occlusion-type errors (see e.g. [16]). Occlusions become inevitable for cameras set at a lower height, often referred to as low-angle video sensors. The detection of occlusions and/or tracking an object through an occlusion is the subject of many research works. For example [18] describes an approach for tracking vehicles over time at a low-angle combining feature tracking (in an estimated world coordinate system) and background subtraction.

Traffic monitoring sensors are often able to classify vehicles at a gross level using estimates of the length of a vehicle thereby differentiating between long vehicles such as trucks, medium-length vehicles such as cars and relatively short vehicles such as motorcycles (e.g. [12]). More recently, researchers have attempted to differentiate vehicles into eight categories including vehicle types such as sedans and pickups [19]; however, the performance assessment only included high-confidence measurements. Many of these computer vision techniques require some sort of camera calibration (see e.g. [20]), which can greatly affect the performance of a traffic monitoring sensor including, for example, the accuracy of motor vehicle speed estimation.

Motorcycles have been a special focus of computer vision-based traffic monitoring research [21]. Conventional magnetic-based counters are not able to accurately record motorcycle counts because of the potential freedom in location of a motorcycle in a comparatively wide road way where a magnetic counter requires the vehicle to drive over the device to enable a count to occur. Furthermore, for video-based sensors, motorcycles are often occluded by other larger vehicles because of the relatively small size of motorcycles and the high possibility of another vehicle passing the motorcycle. The work of Kanhere *et al.* [21] is mainly focused on generalised traffic monitoring tasks including various metrics of vehicle types, but they do not attempt to classify the presence of helmets on motorcycle riders which, however, is the primary focus of the work here. Similar to Kanhere *et al.*, we also identify motorcycles based on their relative dimensions and perform tracking rather than counting individual observations.

### 1.1 Our approach

We develop a framework that uses background subtraction to isolate moving vehicles and identify motorcycle riders by using their characteristic dimensions. The approximate region of the head of the riders is then isolated, which is used to calculate some image-based statistical information referred to here as features that are derived from the histograms of some operations applied to the region. The features selected here capture the important characteristics of helmets in comparison with heads without helmets. These characteristics are based on the reflective properties of the helmets where the tops of the helmets are found to be brighter than the bottom half of the helmet surface. Two example photographs of helmets illustrating this property are shown in Fig. 2. These features are then classified by an SVM with a linear kernel to classify whether the motorcycle rider is wearing a helmet or not. The classifier output for a single frame is then combined with the classifier output for other observed frames for the same motorcycle rider to produce a single decision on whether the motorcyclist is wearing a helmet. This information can then be more permanently recorded in combination with the relevant frame data or simply used for counting purposes.

## 2 Methodology

The system described here is divided into five components: Section 2.1 describes our motorcycle foreground object extraction technique; Section 2.2 outlines the helmet classification process; Section 2.3 describes motorcycle tracking using correspondence analysis and Kalman filters; Section 2.4 proposes a motorcycle track propagation and merging process that is used to join up or associate tracks that correspond to the same motorcycle that may have not been automatically associated because of poor segmentation or tracking in the preceding stages; and Section 2.5 defines three techniques to produce a single classification result from the multiple individual classification results produced for each frame where a single motorcycle has been tracked.




doi: 10.1049/iet-its.2011.0138

**Fig. 2** *Photographs of helmets taken at midday (left) and at dusk (right)*

Tops of the helmets are brighter than the bottoms of the helmets. This is typical of helmets in general because of their reflective properties which is further demonstrated by the mean feature vector illustrated in Fig. 6*b* where the last half of the feature vector contains histograms corresponding to histograms of the upper two quadrants of the helmets

### 2.1 Motorcycle foreground object extraction

*2.1.1 Background subtraction and connected component labelling:* An initial foreground mask is automatically extracted using the background subtraction technique described in [22]. The output of most background subtraction techniques are typically affected by noise and further mathematical morphology such as operations are often required to remove isolated noise pixels which is the approach taken here. Groups of pixels $\{\boldsymbol{x}\}$ that have been identified as potential foreground objects in the resulting binary mask $\boldsymbol{B}(x)$ will then result and are associated with sets of pixels $\mathcal{R}_i$ using connected component labelling

$$\mathcal{R}_i = \{x | (\boldsymbol{B}(\boldsymbol{x})) \wedge (\boldsymbol{x}' \in N(x), \boldsymbol{B}(\boldsymbol{x}') | \boldsymbol{x}' \in \mathcal{R}_i)\} \quad (1)$$

where $N(x)$ is a set of pixels adjacent to pixel $x$ so that if a neighbouring pixel $\boldsymbol{x}'$ is also labelled as part of the foreground, that is, $\boldsymbol{B}(\boldsymbol{x}')$ then pixel $\boldsymbol{x}'$ is also part of connected component $\mathcal{R}_i$.

*2.1.2 Possible motorcycle identification:* For an image frame at time instance $t$ there are $n^t \geq 0$ isolated foreground regions: $\boldsymbol{R}^t = \{\mathcal{R}_1^t, \ldots, \mathcal{R}_{n^t}^t\}$ a subset of these regions $\boldsymbol{R}_{\mathrm{M}} \subseteq \boldsymbol{R}$ may be motorcycle regions (dropping $t$) where we assume $\mathcal{R}_{\mathrm{m}} \in \boldsymbol{R}_{\mathrm{M}}$ contains all pixels of the motorcycle and the rider(s).

A number of steps were used to identify the members of the set $\boldsymbol{R}_{\mathrm{M}}$ from the superset $\boldsymbol{R}$. Initially, a (rotated minimum area) bounding box is found using a rotating calipers method around an isolated region $\mathcal{R}_i \in \boldsymbol{R}$. This bounding box then has height and width from which an aspect ratio can be calculated. The range of this aspect ratio was measured for motorcycles with their riders and found to be in the range [0.7, 2.3] (see Section 2.1.3). This aspect ratio was therefore used to initially reject regions with very different aspect ratios which were unlikely to correspond to motorcycles. Further regions were rejected from inclusion in $\boldsymbol{R}_{\mathrm{M}}$ by estimating the size of the head ($S_{\mathrm{H}}$, also described in Section 2.1.3). The head size estimate was used to reject regions where the head size estimate was found to be either too small or too large.

*2.1.3 Shape size estimation and localisation:* A technique was devised to estimate the size of the head of the motorcycle riders $S_{\mathrm{H}}$. An assumption was made regarding the constant proportion of the head area in comparison with the overall combined area of the motorcycle and rider for a particular (silhouette) view of a motorcycle and a rider.

Furthermore different views were assumed to result in different silhouette proportions. Manually extracted silhouettes of motorcycles and their riders from photographs were then generated and used to compare with the size of the heads of the riders in the same photographs (also using manually extracted silhouettes but of the heads). Individual views of motorcycles and their riders were estimated by fitting minimum bounding rotated rectangles around the motorcycles and riders. The ratio of the height to the width of these rectangles was then used as an estimate of the particular view of the rider. Data from this process can be seen in Fig. 3*a*.

Also shown in Fig. 3*a* is a straight line fit (by minimising the least-squares-error) to the data points. This straight line can be used to estimate the size of the head of a motorcycle rider $S_{\mathrm{H}}$ given the height $H_{\mathrm{M}}$ to width $W_{\mathrm{M}}$ ratio of a minimum bounding rotated rectangle ($H_{\mathrm{M}}/W_{\mathrm{M}}$) and the size of the motorcycle silhouette $S_{\mathrm{M}}$. First the relation between the head size to motorbike size ratio and the motorcycle height-to-width ratio is given in terms of an equation for a straight line

$$\frac{S_{\mathrm{H}}}{S_{\mathrm{M}}} = \mathcal{M}\frac{H_{\mathrm{M}}}{W_{\mathrm{M}}} + \mathcal{C} \quad (2)$$

where $\mathcal{M}$ and $\mathcal{C}$ are the slope and intercept of the fitted straight line. The size of the head of the motorcyclist can therefore be estimated with

$$S_{\mathrm{H}} = S_{\mathrm{M}}\left(\mathcal{M}\frac{H_{\mathrm{M}}}{W_{\mathrm{M}}} + \mathcal{C}\right) \quad (3)$$

To confirm the validity of this approach, the head to motorcycle area size ratios were estimated for the motorcycle silhouettes and the associated head silhouettes. The root mean square (RMS) difference errors between the estimated ratios and actual ratios were then calculated using the fitted straight line and also for fixed ratio values. Graphs of the errors for a range of fixed ratio values in comparison

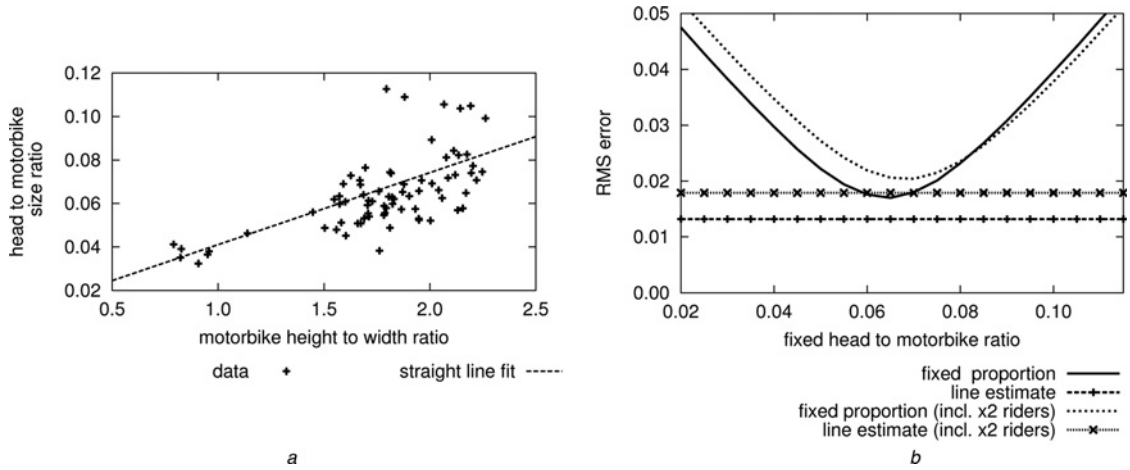


**Fig. 3** *Graphs depicting values (a) head to motor bike size ratio against motorbike height to width ratio and (b) RMS against fixed head to motorbike ratio*

*a* Relating motorcycle and rider view (height against width of minimum bounding rotated rectangles) with head to motorcycle area proportion (also shown is a straight line fit to the data)

*b* RMS errors between straight line estimated head to motorcycle area ratios and fixed head to motorcycle area ratios. Results shown are for motorcycles with a single rider and also for data including both one and two motorcycle riders. The straight line estimate produces the lowest RMS errors overall although there are close minimums for particular fixed head to motorcycle area ratios

with the errors obtained using the straight line can be seen in Fig. 3*b*.

The function defining the straight line fitted to the data does provide a minimum error approach to estimate the size of the head of a motorcycle rider. However, these results also demonstrate that a carefully selected fixed head to motorcycle area ratio may also provide reasonable head size estimates.

The head of the motorcycle riders is located by first assuming the motorcycle riders are upright and their head is not obscured or occluded. Then, using the estimated head size $S_{\mathrm{H}}$ the first $S_{\mathrm{H}}$ pixels in the motorcycle region $\mathcal{R}_i$ transversing in a row major order from top to bottom of the major axis of the motorcycle region are assigned to correspond as the head region $H_i$.

## 2.2 Classification

The main aim of this work is to identify whether a motorcycle rider is wearing a helmet or not from video data using a static camera. In the previous section, a technique was described to enable the identification of regions that are likely to contain motorcycle(s) and their rider(s) and then locating the likely helmet region for each rider. Features can be extracted from the head region information combined with the original image information to provide a feature vector for input to a classifier. The classifier used here is an SVM with a linear kernel.

A number of feature extraction processes were considered here: simple grey level histograms calculated globally over the estimated helmet region; grey level histograms calculated over four quadrants of the estimated helmet regions where the histograms were concatenated one after the other to form a single long feature vector; HOGs [10] where the gradients were estimated using a finite difference mask (i.e. of the form $(+1,\ 0,\ -1)$ and $(+1,\ 0,\ -1)^{\mathrm{T}}$) and also using $3 \times 3$ Sobel kernels; HOGs calculated using the maximum of the gradients from each of individual colour channels using both finite difference gradient masks and $3 \times 3$ Sobel gradient masks for each of the colour channels. All HOGs were calculated using just the four quadrants of the estimated helmet regions which differs from the original proposed HOG calculations in [10] because the estimated helmet regions are often relatively small regions, containing few pixels. Furthermore, the use of just four quadrants simplifies the computational requirements.

## 2.3 Object tracking

The background subtraction provides convenient binary masks of the foreground from which individual regions can be isolated and analysed to determine if the image data for a particular region contains a motorcycle or not $\mathcal{R}_i^t \in \boldsymbol{R}_{\mathrm{M}}^t$. For each new image frame at time $t$ from the video sequence correspondence analysis is performed to identify corresponding regions from the preceding time frame $t-1$, that is, region $\mathcal{R}_j^{t-1}$. Once corresponding regions at different time instances have been established the motion of each region is estimated using Kalman filters to determine the likely location of the region in a subsequent frame.

*2.3.1 Correspondence analysis:* The distance $d(i,j)$ between the centres of gravities (COGs) $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ of two regions $\mathcal{R}_i^t$ and $\mathcal{R}_j^{t'}$ in frames $t$ and $t'$ can be used as a correspondence function. It quantifies the relative correspondence between regions where the COG position from the previous time frame $\boldsymbol{c}_j$ can also be updated based on the Kalman filter motion model for the region $\mathcal{R}_j^{t'}$ (as done here). However, distance-based correspondence matching is not always perfect. For example, a distance-based correspondence function may fail if there are multiple tracked objects in close proximity to each other.

Therefore a number of possible correspondence functions are included and combined which attempt to measure the amount of agreement between two regions $\mathcal{R}_i^t$ and $\mathcal{R}_j^{t'}$ at two different time instances $t$ and $t'$. The two regions are referred to here by $(i,j)$. The correspondence functions are:

1. Distance between COGs $d(i,j) = \langle \boldsymbol{c}_i, \boldsymbol{c}_j \rangle$.
2. Comparison of photometric information $h(i,j)$, for example, by comparing histograms.
3. Difference in the size of the regions $s(i,j)$.




doi: 10.1049/iet-its.2011.0138

So that a correspondence function from the set of correspondence functions is $g_a \in \{h, d, s\}$.

Let $m(i, j) = 1$ be a class indicator variable that explicitly models the situation of when two regions $\mathcal{R}_i^t$ and $\mathcal{R}_j^{t'}$ correspond (otherwise $m(i, j) = 0$) and let $\boldsymbol{g}(i, j) = (h(i, j)d(i, j)s(i, j))^{\mathrm{T}}$ be a vector of similarity functions calculated for the two regions. Therefore we can consider a class conditional probability dependent on the class indicator variable where $P(\boldsymbol{g}(i, j)|m(i, j)) \rightarrow 1$ when $m(i, j) = 1$ and $P(\boldsymbol{g}(i, j)|m(i, j)) \rightarrow 0$ when $m(i, j) = 0$. The exponential of the negative mean square of the difference between the two regions for each of the correspondence functions are used to produce three functions, each with a range of $P(g_a|m(i, j)) \in [0, 1]$. Each of these probabilities are assumed to be independent (e.g. if two regions have similar colours then their distance is not affected), so that $P(\boldsymbol{g}(i, j)|m(i, j)) = P(h|m(i, j))P(d|m(i, j))P(s|m(i, j))$.

Probabilities for each pair of regions in the two time frames $t$ and $t$' can therefore be used to form a proximity matrix $\mathcal{X}$ for the pairs of regions in the image frames with elements given by

$$\mathcal{X}_{i,j} = P(\boldsymbol{g}(i, j)|m(i, j)) \tag{4}$$

This matrix may possess multiple modes for a particular region where region $\mathcal{R}_i$ from time frame $t$ may have high-correspondence probability values for two different regions at time frame $t'$, for example, $\mathcal{R}_{j_1}$ and $\mathcal{R}_{j_2}$. A simple (but not globally optimal) way of selecting a correspondence is by choosing the correspondence with the greatest value, for example, $(i, j_2)$. However, a best match of a region from time frame $t$ with a region from time frame $t'$ may prevent a better match overall from being found going in the opposite direction, that is, a best match of a region from $t$' to a region from $t$.

Therefore correspondences between regions are calculated here from $\mathcal{X}$ using a global optimisation technique that operates over the matrix as a whole similar to the spectral technique proposed in [23].

*2.3.2 Sustained correspondence analysis and motion estimation:* A motion model can be built up for a sequence of regions that have been found to correspond over a number of time frames. The motion model used here is a Kalman filter-based approach. Accurate motion estimation is important because it can assist in distance-based correspondence analysis, where the distance between the COGs will be reduced for corresponding regions. A region might be successfully tracked over a small number frames. However, such regions will often be because of irrelevant motions such as trees blowing in the wind that may have confused the initial background subtraction process. A sequence of corresponding regions is therefore rejected from further processing if it has not been successfully tracked for longer than a fixed number of frames, for example, 10 frames. This helps to eliminate tracks that are because of irrelevant motion.

### 2.4 Track propagation and merging

Occasionally the tracking system may have errors where the tracked object may not be successfully tracked in a number of frames. Tracking may fail for a number of reasons such as a poor segmentation, result in the background subtraction stage or over segmentation of the image as a whole because of irrelevant motion confusing the background subtraction process such as, as already mentioned trees blowing in the wind. These failures are likely to be reasonably sporadic or perhaps intermittent where only a few or individual frames at a time might have been affected. The tracking system has been designed to overcome such limitations through a technique referred to here as 'track propagation' where each successful disjoint track is included in a set of disjoint object tracks. These object tracks are then propagated and tests are performed to determine if any of the disjoint tracks track objects where their geometric tracks can be successfully intersected in image space using the track propagation data. Intersected disjoint tracks are then merged to form a new single disjoint track that includes the original track information and the track propagation information.

### 2.5 Multi-inference process

A single classification result $\Theta_i$ is needed for each track $\mathcal{T}_i$, but each track can contain regions from many frames and there is a classification result for a tracked region at each time instance. The segmented regions can be considered to be time varying spatial information with segmentation and/or classification errors at each time instance. Therefore some additional information is needed to determine which classification results are likely to be correct and then all the results need to be combined into a single result taking into account this confidence information.

A simple approach to inference across multiple decisions is to calculate the arithmetic mean of the decisions. This process becomes a bit complicated for discrete class sets without an obvious ordering; however, the problem here is a binary classification process so that the order of the classes is not important. Therefore we can associate a numerical value with each class from a set of class indicator variables, that is, $C = \{0, 1\}$, where, for example, 0 indicates 'without helmet' and 1 indicates 'with helmet'. Thus, the classification for track $\mathcal{T}_i$ can be calculated as the mean of the individual classification results $\theta_{i_a}^{\tau+a}$ weighted by a time varying confidence function $\pi_{i_{t-\tau}}^t$

$$\Theta_i = \frac{\sum_{t=\tau}^{\tau+n} (\pi_{i_{t-\tau}}^t \theta_{i_{t-\tau}}^t)}{\sum_{t=\tau}^{\tau+n} \pi_{i_{t-\tau}}} \tag{5}$$

Three different confidence functions are considered here. Firstly, as a control, the weighting function is set to a constant so that (5) is just an unweighted mean. Secondly, lighting in video data may vary and the segmentation and/or classification processes may be adversely affected by extreme lighting conditions such as too dark or too light areas. Therefore the weighting function is defined to be higher when the background immediately surrounding the segmented head region is neither too dark nor too light. Lastly, if the motorcycle is far away from the camera then the size of the observed head in the image data will be very small which will make accurate segmentation and classification more difficult. Therefore a weighting function that is proportional to the size of the segmented head region is also considered.

## 3 Experimental methodology

### 3.1 Classifier training

An SVM with a linear kernel was used here to classify whether a motorcycle rider was wearing a helmet or not as

**Fig. 4** *Example training and or test images from frames of two high-definition (HD) video sequences and a variety of static photographs*

Sequences of videos and photographs were acquired at different times of day, different days and different times of year. Furthermore a wide range of different views of motorcycles were also obtained, both from position of the camera and because some locations resulted in views of the motorcycles when they were turning around a corner

discussed in Section 2. Classifiers such as an SVM require supervised training. The SVM was trained here by calculating feature vectors from 230 static images (photographs) and frames from video sequences containing motorcycle riders with instances of helmet and non-helmet wearing individuals. These images were captured at various different times of day, on a number of different days and from a variety of different views, including variations in elevation. A number of example training images can be seen in Fig. 4. This variety in acquisition conditions helped to ensure the data fully represents the potential variability that might occur in the field. The helmet region (or head region for the non-helmet instance) was manually outlined for each of the images to provide accurate image information for the SVM training.

### 3.2 Feature comparison

A number of different sources of features were investigated using the described framework to determine the best feature space, defined here as the classifier using the feature vector that provided the highest classification accuracy. Classification accuracy is calculated with

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

where true positives (TP) is the number of head regions with helmets classified correctly; true negatives (TN) is the number of head regions without helmets classified correctly; false positives (FP) is the number head regions without helmets classified incorrectly; and false negatives (FN) is the number of head regions with helmets classified incorrectly.

Investigated features included: a single grey level histogram for the entire head region; four grey level histograms for four regions of the head split into approximately equal sized quadrants; and HOGs using gradients calculated with a number of different gradient masks. These masks were used to extract the image gradients for both grey level and colour image data (thus producing a combination of two different results) where the masks included 'simple' (i.e. central finite difference) masks ($1 \times 3$ and $3 \times 1$) and Sobel-based masks ($3 \times 3$). Different amounts of smoothing using a Gaussian low-pass filter mask were also applied as part of the feature vector calculation process to determine whether a reduction in high-frequency image information may assist with the classification process.

### 3.3 Classifier performance assessment

The performance of the developed algorithms were tested on the image data using cross validation where the classifier stage was trained using some of the data in combination with the manually defined ground truth and then tested with the remaining data. Leave-one-out cross validation was used to maximise the number of training data (230 images). Some of the data were taken with a still camera in multi-shot mode where multiple shots (4 frames − 2 frames/s) were acquired of the same individual albeit at different time




doi: 10.1049/iet-its.2011.0138

instances. Therefore a further classifier assessment scheme was undertaken where eight images (adjacent in time) were left out of the training set (to ensure all shots of the same individual were excluded except the training image) and then the middle image was used in the test stage, referred to here as leave-eight-out cross validation. This process was repeated across all the images in the collection (230 images), similar to the leave-one-out schema.

Classifier performance was also assessed in comparison with: (i) time of day and (ii) angle of view of the motorcycle. The time of day information was automatically extracted through the meta-data encoded with the images. Angle of view of the motorcycle was estimated using the height-to-width ratio of the minimum bounded rotated rectangle surrounding the motorcycle silhouette which was extracted manually from a subset of the image data.

Assessment was then also performed for the tracking and classification system as a whole using video data (or any individual frames), which had not been used in the training process.

## 4 Results

### 4.1 Comparison of features for identifying helmetsagainst non-helmets

Classification accuracy results for the SVM classifier with various features on the static image data sets using leave-one-out and leave-eight-out cross validation can be seen in Fig. 5. These results clearly demonstrate the superiority of the multiple region grey level histograms where classification accuracy is almost 100% for the case of zero smoothing using the leave-one-out cross validation assessment. The classification accuracy is reduced for all feature types using the leave-eight-out cross validation assessment. Nevertheless the classifier using the multiple region grey level histograms still out performs the other classifiers with a relatively high accuracy of 96% for zero smoothing. The smoothing or low-pass filtering of the image data improves some of the results using the HOG features for some low-pass filter kernel sizes, but the increase in performance is still not sufficient (86%).

A further investigation was then carried out to observe the major differences between the various feature vectors. The mean feature vector for the two classes of helmet wearing and non-helmet wearing motorcycle riders were calculated for the four feature vectors. These mean feature vectors are shown in Fig. 6. The mean vector is not particularly useful for visualising a feature space, but it does help to illustrate major differences between the statistics of the classes. Little difference can be observed in the single grey level histograms. However, dividing the head area into four regions reveals some very useful differences between the histograms of the four regions that can be used by a suitably trained classifier to easily separate the two classes. This can be compared with the HOGs which are quite noisy and it is difficult to observe any particular difference between the two mean feature vectors. The grey level histograms are particularly useful in identifying helmets because the top of the helmet typically reflects some light and the bottom part of the helmet is relatively much darker. This property is true even when observing a helmet at midday or at dusk because the sky is typically much brighter than the ground, the sun is rarely below the rider and the sun lights up the entire sky. This is demonstrated, in part, by the last half of the feature vector corresponding to the upper two quadrants of the helmet. These two upper quadrants show many more pixels with lighter grey level intensities in comparison with the bottom two quadrants corresponding to the first half of the feature vector. This can be further illustrated by the two example photographs of helmets captured at midday and at dusk as shown earlier, in Fig. 2.

### 4.2 Imaging conditions: time of day and view of motorcycle

Classification accuracy with respect to variations in time of day and variations in view of the motorcycle estimated through the motorcycle height-to-width ratio can be seen in Fig. 7.

Also shown in these two figures are the frequency of the number of motorcycle riders for a particular time of day or view of the motorcycle. The results show that classification accuracy remains relatively high irrespective of the time of day, the particular view of the motorcycle rider or the number of image data available for training. These results therefore help to demonstrate that a traffic monitoring

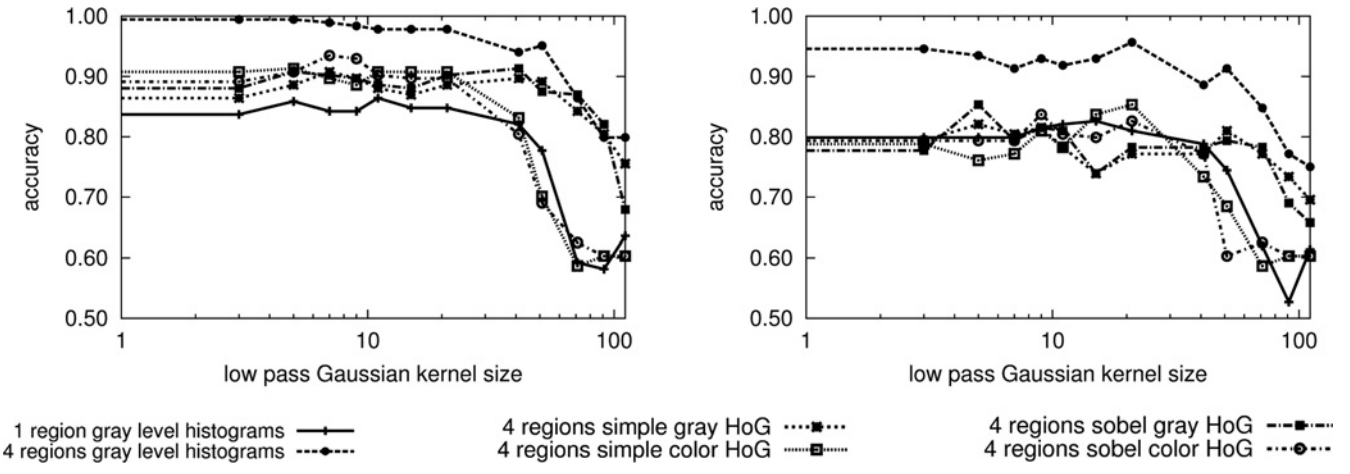


**Fig. 5** *Comparison of classification accuracy (helmetagainst no helmet) for a variety of different features and different amounts of smoothing using two different classifier assessment schemes*

On the left leave-one-out cross validation is used. Leave-eight-out cross validation results are shown on the right, demonstrating reduced performance for classifiers using all types of feature vectors. However, the best classifier over all the classifiers is the same for both cross validation schemes

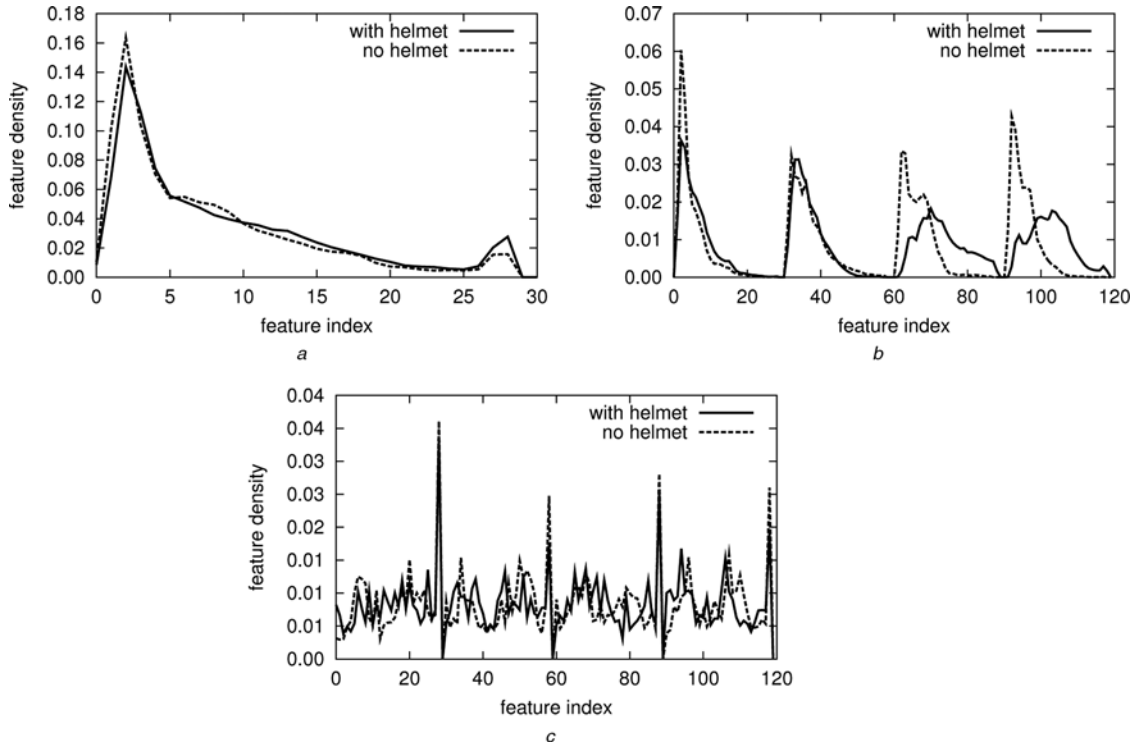


**Fig. 6** *Comparison of the mean feature vectors calculated for various features extracted from the two classes of helmet wearing and non-helmet wearing motorcycle riders*

*a* Feature vectors based on a single grey level histogram calculated over the head area of the riders
*b* Mean feature vector for four grey level histograms calculated over four sub-regions of the head area of the riders
*c* Feature vectors calculated using HOGs with simple finite difference kernels on grey scale image data
Feature vectors in *b* and *c* were created using the bottom two image quadrants for the first 60 features and the upper two image quadrants for the remaining 60–120 features. *b* illustrates significant differences in the upper two quadrants, showing brighter intensities in the upper helmet region, illustrated photographically in Fig. 2

system could use prior training data and not require on-site training.

### 4.3 Application to video footage of motorcycle riders

After discovering the superior performance that was achievable using the localised grey-scale histograms, the trained classifier was combined with the traffic object segmentation and tracking framework and applied to a series of six video sequences containing motorcycle riders with and without helmets. The frames for these video data sequences were not used to train the original classifier. Some illustrative example frames for simultaneous tracking and helmet classification can be seen in Fig. 8 and the classifications for the detected tracks for this video sequence can be seen in Fig. 9.

The classification of each region for each time frame results in a sequence of (possibly differing) classifications for a single-tracked object as shown in Fig. 9. A single classification result was therefore calculated for each track using the weighted means calculated by (5). Mainly for illustrative purposes the weighted mean is calculated continuously, demonstrating the possible changes in the final classification result. The final classification result was found to be correct for all the tracks using the head size weighting, but the unweighted and the too dark/too light weighted means each resulted in one mis-classification of a

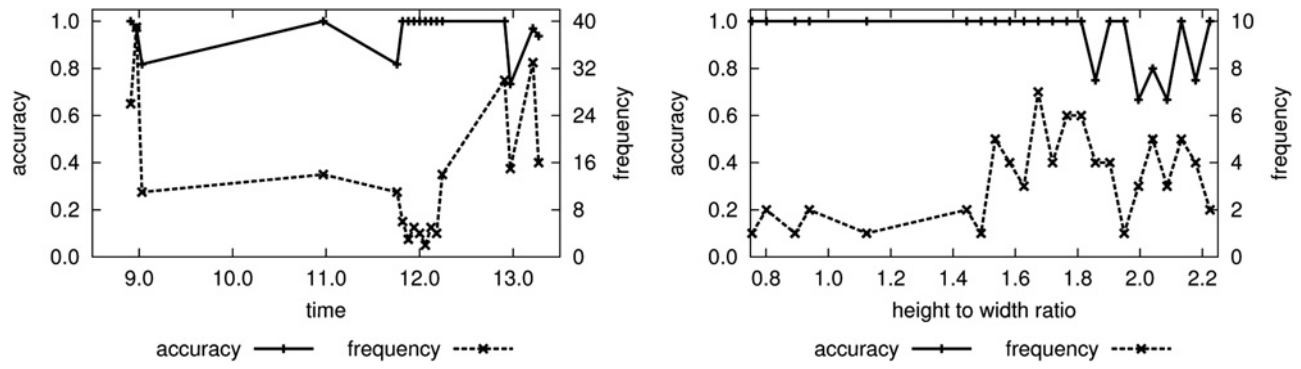


**Fig. 7** *Classification accuracy with respect to (left) variations in time of day and (right) variations in view of the motorcycle estimated through the motorcycle height-to-width ratio*




doi: 10.1049/iet-its.2011.0138


**Fig. 8** *Example image frames with colour coded classification outlines where light grey is indicative of no helmet and middle grey indicates the presence of a helmet*

*a* $t = 1746$
*b* $t = 1775$
*c* $t = 1790$
*d* $t = 1805$

final track. This suggests that the head size is a useful indicator of the correctness of the individual frame-based classifications. This was confirmed by observing the actual size of the head region to be very small (60 pixels or; 3 pixel radius) for quite a few instances where the classification for individual frames was found to be wrong.

The overall motorcycle track detection and helmet classification accuracies for correctly classifying individual tracks of motorcyclists wearing or not wearing helmets in the series of six videos were then calculated. The results can be seen in Table 1.

The first five columns of Table 1 show the results for the tracking system's ability to correctly identify motorcycles with an overall motorcycle track detection accuracy of 83%. The last column shows the helmet classification accuracy for the detected motorcycle tracks with an overall helmet classification accuracy of 85%.

As might be expected, these classification performance values are lower than the classification accuracy obtained with the static photographs partly because of difficulties in obtaining accurate segmentations of the heads of the motorcyclists and errors in the tracking system as a whole.

### 4.4 Track propagation and merging

An important component of the tracking system is the ability to continue to track objects even if they are temporarily no longer visible, referred to as undergoing an occlusion. Occluding objects can be static objects such as sign posts or moving objects such as other motorcyclists. The results shown in Figs. 8 and 9 demonstrate the tracking continues despite a number of occlusions. This is possible because of the track propagation and merging process. The tracking system stops classifying the tracked object during occlusions because there is no data available for classification. This is indicated by series of data points (circles) in the middle region of the plots in Fig. 9. However, tracking continues when a motorcyclist becomes visible once more. A single track is obtained for a single motorcyclist by merging two (or more) individual tracks together through the track propagation and merging process.

For example, consider the plot in Fig. 9*c* in relation to the motorcyclist travelling towards the camera in the frames shown in Fig. 8. This motorcyclist passes and is occluded by two other motorcyclists. The first occlusion occurs before $t = 1746$, as indicated by Fig. 8*a* and the first series of data points in the plot. The second occlusion occurs between frames $t = 1775$ and 1790, as indicated by Figs. 8*b* and *c* and the second series of data points in the plot. The entire track for this motorcyclist will therefore have been created from (at least) three individual tracks. The individual tracks will have been propagated (using the motion estimation model) both forwards and backwards in time and consequently merged where they have been found to intersect to form a single continuous track. Classification of whether this single instance of a motorcyclist is wearing a helmet or not is then possible.

### 4.5 Discussion

The motorcycle with the track illustrated by the plot in Fig. 9*f* appears to have slightly improved overall classification when using the dark/light background weighted mean. This motorcyclist is not wearing a helmet and the background for the last few frames is particularly dark and without texture. The background subtraction stage is consequently unable to accurately segment the entire head of the rider but the relative size of the head of the motorcyclist is increasing because the

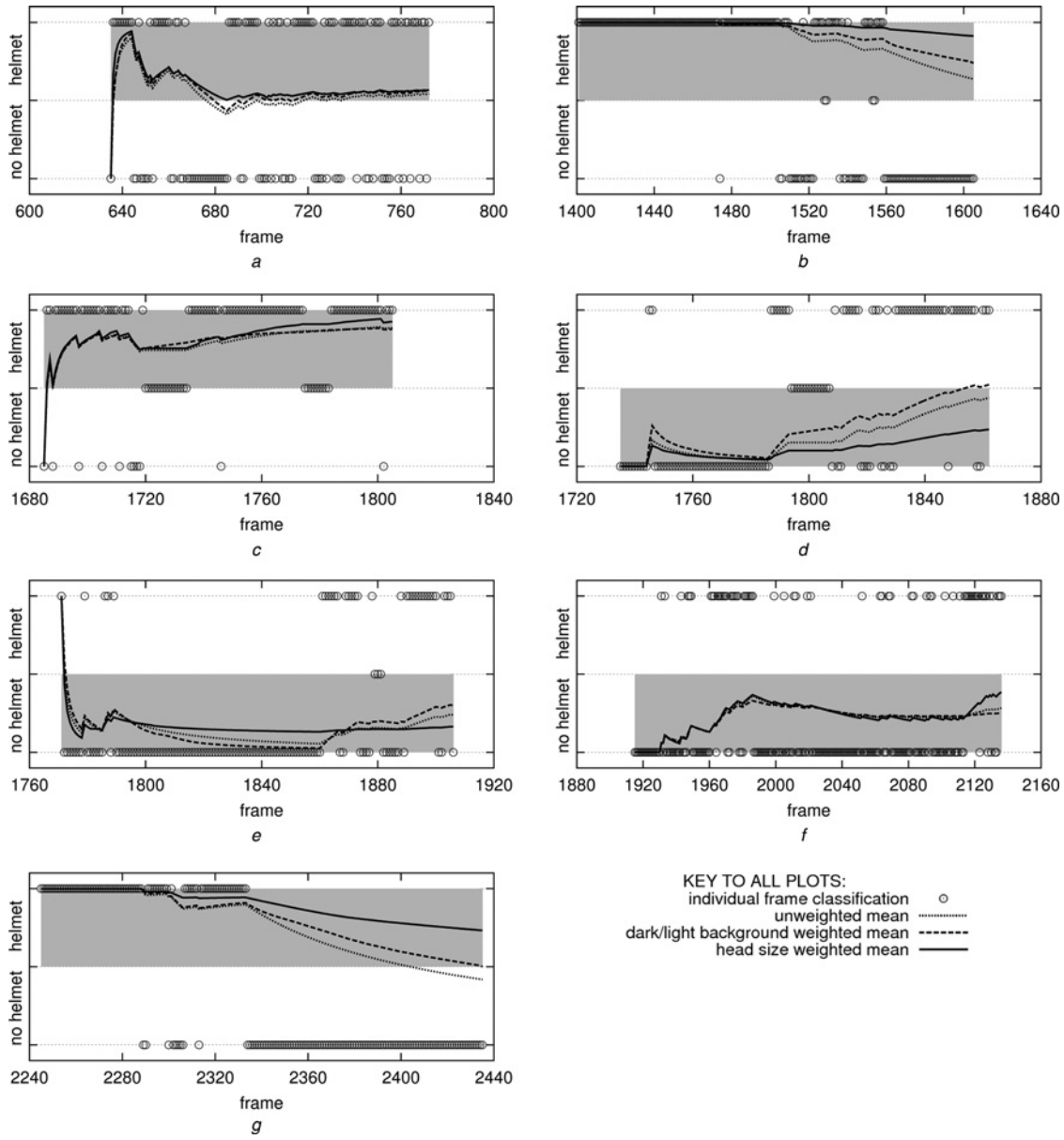


**Fig. 9** *Video sequence classification results*

*a* $t \in [635, 0772]$
*b* $t \in [1401, 1605]$
*c* $t \in [1685, 1805]$
*d* $t \in [1735, 1862]$
*e* $t \in [1771, 1906]$
*f* $t \in [1915, 2136]$
*g* $t \in [2245, 2435]$

Data points (circles) represent individual region classification on a per frame basis. The shaded region represents the true classification result region. Data points on centre lines are unclassified because of the track propagation process. The continuous lines represent cumulative classification results taken by three different averaging operations (unweighted mean, inverse weighting for saturated lighting conditions and head size weight) over the classification results to provide a combined classification result overall. The most correct estimation overall is given by the head size weighted mean

rider is travelling towards the camera. Therefore the input to the classifier is a corrupted feature vector that does not accurately represent the data that the classifier expects and the head size weighting incorrectly weights the classifier results because the head size is increasing. This information could be used to develop more advanced classification weighting schemes. Alternatively, additional shape-based information could be included in the system to provide more accurate head and/or helmet location and size estimates.

As discussed earlier in the introduction, two previously published techniques [5, 7] were implemented and assessment of their performance was undertaken. The results of these tests were unsatisfactory for our system producing too many false positives and false negatives. These techniques were based on the idea that the helmet can be detected as a circular object, which is a valid assumption except image gradient information was used to detect edges in or around the head region to cumulatively estimate circular boundaries. There are typically many sources of high-image gradient magnitude in or around the head region not necessarily because of the boundary of the head or helmet. Furthermore, the motorcycle riders are




doi: 10.1049/iet-its.2011.0138

**Table 1** Overall classification accuracy for the tracking system including motorcycle track detection performance and helmet classification performance

| Sequence | Motorcycle detection | | | | | Helmet classification |
|---|---|---|---|---|---|---|
| | TN | FN | FP | TP | Accuracy | Accuracy |
| 1 | 3 | 1 | 1 | 7 | 0.83 | 1.00 |
| 2 | 3 | 1 | 1 | 7 | 0.83 | 0.86 |
| 3 | 9 | 0 | 4 | 4 | 0.76 | 0.75 |
| 4 | 0 | 1 | 0 | 6 | 0.86 | 0.67 |
| 5 | 4 | 1 | 1 | 8 | 0.86 | 0.88 |
| 6 | 0 | 0 | 0 | 1 | 1.00 | 1.00 |
| overall accuracy | | | | | 0.83 | 0.85 |

moving and as a result the image data may be corrupted by motion blur, which more than likely reduces the prominence of useful edges in the image data. This can be compared with the results presented here which demonstrate that the performance of the classifier remains approximately constant even when the image data has been low-pass filtered using relatively large low-pass Gaussian kernel sizes. Image gradient techniques or an assumption of a circular region at a particular location of the detected motorcycle region could be used in the system proposed here to possibly improve the segmentation process of the heads of the motorcyclists. This might then improve the input to the helmet classifier and result in more accurate classification of the presence of a helmet on a motorcyclist.

## 5 Conclusions

A new technique for the classification of the presence of a helmet on motorcycle riders has been presented. Tests show very accurate classification results for static photographs ($>95\%$). Tests on videos of motorcycle riders also confirm the validity and usefulness of the classification technique.

## 6 Acknowledgments

The author is very grateful to Mitsui Sumitomo Insurance Welfare Foundation (MSIWF) for providing a grant on traffic safety. He would also like to thank Surapong Uttama for his help with talking with students and liasing with various people at Mae Fah Luang University. All the software was programmed in C++ using, where possible, the Open source Computer Vision (OpenCV) library, version 2.3.1.

## 7 References

1 Bayly, M., Regan, M., Hosking, S.: 'Intelligent transport systems and motorcycle safety' (Monash University, Accident Research Centre, 2006), p. 260
2 Bianco, A., Trani, F., Santoro, G., Angelillo, I.F.: 'Adolescents' attitudes and behaviour towards motorcycle helmet use in Italy', *Eur. J. Pediatr.*, 2005, **164**, (4), pp. 207–211
3 Pitaktong, U., Manopaiboon, C., Kilmarx, P.H., *et al.*: 'Motorcycle helmet use and related risk behaviors among adolescents and young adults in Northern Thailand', *Southeast Asian J. Trop. Med. Public Health*, 2004, **35**, (1), pp. 232–241
4 Hung, D.V., Stevenson, M.R., Ivers, R.Q.: 'Prevalence of helmet use among motorcycle riders in Vietnam', *Inj. Prev.*, 2006, **12**, (6), pp. 409–413
5 Chiu, C.C., Ku, M.Y., Chen, H.T.: 'Motorcycle detection and tracking system with occlusion segmentation'. IEEE CS Eighth Int. Workshop on WIAMIS'07, 2007, pp. 32–32
6 Wen, C.Y., Chiu, S.H., Liaw, J.J., Lu, C.P.: 'The safety helmet detection for ATM's surveillance system via the modified Hough transform'. IEEE 37th Annual ICCST, 2003, pp. 364–369
7 Liu, C.C., Liao, J.S., Chen, W.Y., Chen, J.H.: 'The full motorcycle helmet detection scheme using canny detection'. IPPR 18th Conf., CVGIP, 2005, pp. 1104–1110
8 Viola, P., Jones, M.: 'Robust real-time face detection', *Int. J. Comp. Vis.*, 2004, **57**, (2), pp. 137–154
9 Stojmenovic, M.: 'Algorithms for real-time object detection in images', in Nayak, A, Stojmenevic, I. (Eds.): 'Handbook of applied algorithms' (Wiley, 2008), pp. 317–346
10 Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. Int. Conf. on Computer Vision Pattern Recognition, IEEE, 2005, pp. 886–893
11 Birchfield, S., Rangarajan, S.: 'Spatiograms versus histograms for region-based tracking'. Int. Conf. Computer Vision Pattern Recognition, IEEE, 2005, pp. 1158–1163
12 Minge, E.: 'Evaluation of non-intrusive technologies for traffic detection' (Minnesota Department of Transportation, Office of Policy Analysis, Research and Innovation, SRF Consulting Group, US Department of Transportation, Federal Highway Administration, 2010), pp. 2010–2036
13 Semertzidis, T., Dimitropoulos, K., Koutsia, A., Grammalidis, N.: 'Video sensor network for real-time traffic monitoring and surveillance', *IET Intell. Transp. Syst.*, 2010, **4**, (2), pp. 103–112
14 Buch, N., Velastin, S.A., Orwell, J.: 'A review of computer vision techniques for the analysis of urban traffic', *IEEE Trans. Intell. Transp. Syst.*, 2011, **PP**, (99), pp. 1–20
15 Song, K.T., Tai, J.C.: 'Image-based traffic monitoring with shadow suppression', *Proc. IEEE*, 2007, **92**, (2), pp. 413–426
16 Middleton, D., Longmire, R., Turner, S.: 'State of the art evaluation of traffic detection and monitoring systems', Arizona Department of Transportation, US Department of Transportation, Federal Highway Administration, Texas Transportation Institute, Texas A and M University, 2007, FHWA-AZ-07-627(1)
17 Roy, A., Gale, N., Hong, L.: 'Automated traffic surveillance using fusion of Doppler radar and video information', *Math. Comput. Modeling*, 2011, **54**, pp. 531–543
18 Kanhere, N.K., Birchfield, S.T., Sarasua, W.A.: 'Vehicle segmentation and tracking in the presence of occlusions'. Transport Research Board Annual Meeting, November 2006, vol. 1944, pp. 89–97
19 Morris, B.T., Trivedi, M.M.: 'Learning, modeling, and classification of vehicle track patterns from live video', *IEEE Trans. Intell. Transp. Syst.*, 2008, **9**, (3), pp. 425–437
20 Kanhere, N.K., Birchfield, S.T.: 'A taxonomy and analysis of camera calibration methods for traffic monitoring applications', *IEEE Trans. Intell. Transp. Syst.*, 2010, **11**, (2), pp. 441–452
21 Kanhere, N.K., Birchfield, S.T., Sarasua, W.A., Khoeini, S.: 'Traffic monitoring of motorcycles during special events using video detection', *Transp. Res. Record: J. Transp. Res. Board*, 2010, **2160**, (10–3933), pp. 69–76
22 Kaewtrakulpong, P., Bowden, R.: 'An improved adaptive background mixture model for real-time tracking with shadow detection', in Jones, G.A., Paragios, N., Carlos, S. (Eds.): 'Video-based surveillance systems' (Kluwer Academic Publishers, 2002), pp. 135–144
23 Scott, G., Longuet-Higgens, H.: 'An algorithm for associating the features of two images', *Proc. R. Soc. B: Biol. Sci.*, 1991, **244**, (1309), pp. 21–26