The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking and Short/Long-term Re-Identification from Aerial Devices

S.V. Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, B.S. Harish and Hugo Proença, *Senior Member, IEEE*

***Abstract*—Over the years, unmanned aerial vehicles (UAVs) have been regarded as a potential solution to surveil public spaces, providing a cheap way for data collection, while covering large and difficult-to-reach areas. This kind of solutions can be particularly useful to detect, track and identify subjects of interest in crowds, for security/safety purposes. In this context, various datasets are publicly available, yet most of them are only suitable for evaluating detection, tracking and *short-term re-identification* techniques. This paper announces the free availability of the P-DESTRE dataset, the first of its kind to provide video/UAV-based data for *pedestrian long-term re-identification* research, with ID annotations consistent across data collected in different days. As a secondary contribution, we provide the results attained by the state-of-the-art pedestrian detection, tracking, short/long term re-identification techniques in well-known surveillance datasets, used as baselines for the corresponding effectiveness observed in the P-DESTRE data. This comparison highlights the discriminating characteristics of P-DESTRE with respect to similar sets. Finally, we identify the most problematic data degradation factors and co-variates for UAV-based automated data analysis, which should be considered in subsequent technologic/conceptual advances in this field. The dataset and the full specification of the empirical evaluation carried out are freely available at http://p-destre.di.ubi.pt/.**

***Index Terms*—Visual Surveillance, Aerial Data, Pedestrian Detection, Object Tracking, Pedestrian Re-identification, Pedestrian Search.**

I. Introduction

Video-based surveillance refers *the act of watching a person or a place, esp. a person believed to be involved with criminal activity or a place where criminals gather*[1]. Over the years, this technology has been used in far more applications than its roots in crime detection, such as traffic control and management of physical infrastructures. The first generation of video surveillance systems was based in closed-circuit television (CCTV) networks, being limited by the stationary nature of cameras. More recently, unmanned aerial vehicles (UAVs) have been regarded as a solution to overcome such limitations: UAVs provide a fast and cheap way for data collection, and can easily assess confined spaces, producing minimal noise while reducing the staff demands and cost. UAV-based surveillance of crowds can host crime prevention measures throughout the world, but it also raises a sensitive debate about faithful balances between security/privacy issues. In this context, it is important that legal authorities strictly define the cases where this kind of solutions can be used (e.g., missing child or disoriented elderly? Criminal seek?).

Being at the core of video surveillance, many efforts have been concentrated in the development of video-based pedestrian analysis methods that work in real-world conditions, which is seen as a *grand challenge*[2]. In particular, the problem of identifying pedestrians in crowds is especially difficult when the time elapsed between consecutive observations denies the use of clothing-based features (bottom row of Fig. 1).

Fig. 1. Key difference between the pedestrian *short-term re-identification* (upper row) and *long-term re-identification* problems (bottom row). In the former case, it is assumed that subjects keep the same clothes between consecutive observations, which does not happen in the *long-term* problem. Matching IDs across long-term observations is highly challenging, as the state-of-the-art re-identification techniques rely in clothing appearance-based features. **The P-DESTRE set is the first to supply video/UAV-based data for pedestrian *long-term re-identification*.**

To date, the research on pedestrian analysis has been mostly conducted on databases (e.g., [17], [30] and [11]) that provide data with short lapses of time between consecutive observations of each ID (typically within a single day), which allows to use clothing-based appearance features for identification (top row of Fig. 1). Also, datasets related to other problems are used (e.g., gait recognition [38]), where the data acquisition

A. Kumar, E. Yaghoubi and H. Proença are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Portugal, E-mail: arunkumarsv55@gmail.com, D2389@ubi.pt, hugomcp@di.ubi.pt

A. Das is with the India Statistical Institute, Kolkata, India, E-mail: abhijitdas2048@gmail.com

B. Harish is with the Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, India, E-mail: bsharish@jsstuniv.in

Manuscript received ? ?, 2020; revised ? ?, ?.

[1] https://dictionary.cambridge.org/dictionary/english/surveillance

[2] https://en.wikipedia.org/wiki/Grand_Challenges

conditions are evidently different of the seen in surveillance environments.

As a tool to support further advances in video/UAV-based pedestrian analysis, the P-DESTRE is a joint effort from research groups in two universities of Portugal and India. It is a multi-session set of videos, taken in outdoor crowded environments. "*DJI Phantom 4*"[3] drones controlled by human operators flew over various scenes of both universities *campi*, with the data acquired simulating the everyday conditions in surveillance environments. All subjects offered explicitly as volunteers and they were asked to act normally and ignore the UAVs. Moreover, the P-DESTRE set is fully annotated at the frame level by human experts, providing four families of meta-data:

- **Bounding boxes**. The position of each pedestrian at every frame is given as a bounding box, to support object detection, tracking and semantic segmentation experiments;
- **IDs**. Each pedestrian has a unique identifier that is kept consistent over all the data acquisition days/sessions. This is a singular characteristic that turns the P-DESTRE suitable for various kinds of identification problems. The *unknown* identities are also annotated, and can be used as distractors to increase the identification challenges;
- **Soft biometrics labels**. Each pedestrian is fully characterised by 16 labels: {'gender', 'age', 'height', 'body volume', 'ethnicity', 'hair colour', 'hairstyle', 'beard', 'moustache', 'glasses', 'head accessories', 'body accessories', 'action' and 'clothing information' (x3)}, which allows to perform soft biometrics and action recognition experiments.
- **Head pose**. 3D head pose angles are given in terms of *yaw*, *pitch* and *roll* values for all the bounding boxes, except backside views. This information was automatically obtained according to the Deep Head Pose [29] method.

As a consequence of its annotation, the P-DESTRE is the first suitable for evaluating video/UAV-based *long-term re-identification* methods. Using data collected over large periods of time (days/weeks), the re-identification techniques cannot rely in clothing-based features, which is the key characteristic that distinguishes between the *long-term* and the *short-term* re-identification problems (Fig. 1).

In summary, this paper offers the following contributions:

1) we announce the free availability of the P-DESTRE dataset, the first of its kind that is fully annotated at the frame level and was designed to support the research on video/UAV-based long-term re-identification. Moreover, the P-DESTRE set can be used in pedestrian detection, tracking, short-term re-identification and soft biometrics experiments;
2) we provide a systematic review of the related work in the scope of the P-DESTRE set, comparing its main discriminating features with respect to the related sets;
3) based in our own empirical evaluation, we report the results that state-of-the-art methods attain in the pedestrian detection, tracking and short-term re-identification tasks, when considering well-known surveillance datasets. The comparison between such results and those attained in P-DESTRE supports the originality of the novel dataset.

[3]https://www.dji.com/pt/phantom-4

The remainder of this paper is organized as follows: Section II summarizes the most relevant research in the scope of the novel dataset. Section III provides a detailed description of the P-DESTRE data. Section IV discusses the results observed in our empirical evaluation, and the conclusions are given in Section V.

II. Related Work

This section describes the most relevant UAV-based datasets and also pays special attention to datasets that focus the problems of pedestrian detection, tracking, re-identification and search.

A. UAV-Based Datasets

Various datasets of UAV-based data are available to the research community, most of them serving for object detection and tracking purposes. The 'Object deTection in Aerial images' [35] set supports research on multi-class object detection, and has 2,806 images, with 188K instances of 15 categories. The 'Stanford drone dataset' [28] provides video data for object tracking, containing 60 videos from 8 scenes, annotated for 6 classes. Similarly, the 'UAV123' [24] set provides 123 video sequences from aerial viewpoints, containing over 110K frames, annotated for object detection/tracking. The 'VisDrone' [40] consists of 288 videos/261,908 frames, with over 2.6M bounding boxes covering pedestrians, cars, bicycles, and tricycles. Finally, the largest freely available source is the 'Multidrone' [23], providing data for multiple category object detection and tracking. It contains videos of various actions, collected under various weather conditions and in different places, yet not all the data are annotated. The 'UAVDT' [9] is an image-based dataset that supports research on vehicle detection and tracking. It has 80K frames/ 841.5K bounding boxes, selected from 10 hours raw videos, that were manually annotated for 14 attributes (e.g., *weather condition*, *flying altitude*, *camera view*, *vehicle category* and *levels of occlusion*). Recently, to facilitate research on face recognition from video/UAV-based data, the 'DroneSURF' dataset [15] was released. This dataset is composed of 200 videos from 58 subjects, captured across 411K frames, and includes over 786K face annotations.

B. Pedestrian Analysis Datasets

As summarized in Table I, there are various datasets for supporting pedestrian analysis research. The pioneer set was the 'PRID-2011' [14], containing 400 image sequences of 200 pedestrians. Next, the 'CUHK03' [17] set aimed at providing enough data for deep learning-based solutions, and contains images collected from 5 cameras, comprising 1,467 identities and 13,164 bounding boxes. The 'iLIDS-VID' [32] set was the first to release video data, comprising 600 sequences of 300 individuals, with sequence lengths ranging from 23

TABLE I
COMPARISON BETWEEN THE P-DESTRE AND THE EXISTING DATASETS THAT SUPPORT THE RESEARCH IN PEDESTRIAN DETECTION, TRACKING AND SHORT/LONG-TERM RE-IDENTIFICATION (APPEARING IN CHRONOLOGICAL ORDER).

Dataset	Camera	Format	Task					Identities	Bound. Box	Environment	Height (m)
			Detection	Tracking	ReID	Search	Action Rec.				
PRID-2011 [14]	UAV	Still	✗	✗	✓	✗	✗	1,581	40K	Surveillance	[20, 60]
CUHK03 [17]	CCTV	Still	✗	✗	✓	✗	✗	1,467	13K	Surveillance	-
iLIDS-VID [32]	CCTV	Video	✗	✗	✓	✗	✗	300	42K	Surveillance	-
MRP [16]	UAV	Video	✓	✓	✓	✗	✗	28	4K	Surveillance	< 10
PRAI-1581 [32]	UAV	Still	✗	✗	✓	✗	✗	1,581	39K	Surveillance	[20, 60]
CSM [1]	(Various)	Video	✗	✗	✗	✓	✗	1,218	11M	TV	-
Market1501 [37]	CCTV	Still	✓	✓	✓	✗	✗	1,501	32,668	Surveillance	< 10
Mini-drone [6]	UAV	Videos	✓	✓	✗	✗	✓	-	> 27K	Surveillance	< 10
Mars [39]	CCTV	Video	✗	✗	✓	✗	✗	1,261	20K	Surveillance	-
AVI [30]	UAV	Still	✗	✗	✗	✗	✓	5,124	10K	Surveillance	[2, 8]
DukeMTMC-VideoReID [34]	CCTV	Video	✗	✗	✓	✗	✗	1,812	815K	Surveillance	-
iQIYI-VID [20]	(Various)	Video	✗	✗	✗	✓	✗	5,000	600K	TV	-
DRone HIT [11]	UAV	Still	✗	✗	✓	✗	✗	101	40K	Surveillance	25
LTCC [26]	CCTV	Still	✓	✗	✓	✓	✗	152	17K	Surveillance	-
P-DESTRE	UAV	Video	✓	✓	✓	✓	✓	269	> 14.8M	Surveillance	[5.5, 6.7]

to 192 frames. The 'MRP' [16] was the first UAV-based dataset specifically designed for the re-identification problem, containing a 28 identities and 4,000 bounding boxes. Roughly at the same time, the 'PRAI-1581' [32] data reproduces undoubtedly real surveillance conditions, but UAVs flew at too high altitude to enable re-identification experiments (up to 60 meters). This set has 39,461 images of 1,581 identities, and is mainly used for detection and tracking purposes. The 'Market-1501' [37] set was collected using 6 cameras in front of a supermarket, and contains 32,668 bounding boxes of 1,501 identities. Its extension ('MARS' [39]) was the first video-based set specifically devoted to pedestrian re-identification. Singularly, the 'Mini-drone' [6] set was created mostly to support abnormal event detection analysis, and has been also used for pedestrian detection, tracking and short-term re-identification purposes.

The 'DukeMTMC-VideoReID' [34] is a subset of the DukeMTMC [27] tracking dataset, used for pedestrian re-identification purposes. Authors also defined a performance evaluation protocol, enumerating the 702 identities used for training, the 702 testing identities, and the 408 distractor identities. Overall, this set comprises 369,656 frames of 2,196 sequences for training and 445,764 frames of 2,636 sequences for testing. The 'AVI' [30] set enables pose estimation/abnormal event detection experiments, with subjects in each frame annotated with 14 body keypoints. More recently, the 'DRoneHIT' [11] set supports image-based pedestrian re-identification experiments from aerial data, containing 101 identities, each one with about 459 images.

The 'CSM' [1] and 'iQIQI-VID' [20] sets were included in this summary because they previously released data for the long-term re-identification problem. However, their video sequences have notoriously different features from the acquired in surveillance environments: predominantly regard TV shows/movies. Similarly, the 'Long-Term Cloth-Changing (LTCC)' [26] set also supports long-term re-identification research and has 17,119 images from 152 identities, collected using CCTV footage and annotated across clothing-changes and different views.

Among the datasets analyzed, note that the Market1501, MARS, CUHK03, iLIDS-VID and DukeMTMC-VideoReID were collected using stationary cameras, and their data have notoriously different features of the resulting from UAV-based acquisition. Also, even though the PRAI-1581 and DRone HIT sets were collected using UAVs, they do not provide consistent identity information between acquisition sessions, and cannot be used in pedestrian search problem.

III. THE P-DESTRE DATASET

A. Data Acquisition Devices and Protocols

The P-DESTRE dataset is the result of a joint effort from researchers in two universities: the University of Beira Interior[4] (Portugal) and the JSS Science and Technology University[5] (India). In order to enable the research on pedestrian identification from UAV-based data, a set of *DJI® Phantom 4*[6] drones controlled by human operators flew over various scenes of both university campi, acquiring data that simulate the everyday conditions in outdoor urban environments.

All subjects in the dataset offered explicitly as volunteers and they were asked to completely ignore the UAVs (Fig. 2), that were flying at altitudes between 5.5 and 6.7 meters, with the camera pitch angles varying between 45° to 90°. Volunteers were students of both universities (mostly in the 18-24 age interval, $> 90\%$), $\approx 65/35\%$ males/females, and of predominantly two ethnicities ('white' and 'indian'). About

[4] http://www.ubi.pt
[5] https://jssstuniv.in
[6] https://www.dji.com/pt/phantom-4-pro-v2

28% of the volunteers were using glasses, 10% of them were using sunglasses. Data were recorded at 30fps, with 4K spatial resolution ($3,840 \times 2,160$), and stored in "mp4" format, with H.264 compression. The key features of the data acquisition settings are summarized in Table II, and additional details can be found at the corresponding webpage[7].

TABLE II
THE P-DESTRE DATA ACQUISITION MAIN FEATURES.

Image Acquisition Settings	
Camera Sensor: 1/2.3" CMOS, Effective pixels: 12.4 M	Frame Size: 3,840 × 2,160
Lens: FOV 94°, 20 mm (35 mm format equivalent) f/2.8 focus at ∞	ISO Range: 100-3200
Camera Pitch Angle: [45°, 90°]	Drone Altitude: [5.5, 6.7] meters
Format: MP4, 30 fps	Bit Depth: 24 bit
Volunteers	
Total IDs: 269	Gender: Male: 175 (65%); Female: 94 (35%)

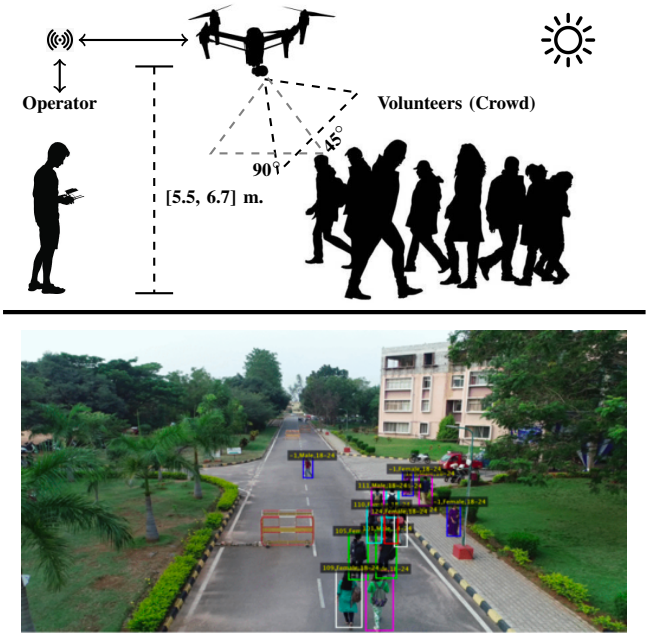

Fig. 2. At top: schema of the data acquisition protocol used. Human operators controlled DJI Phantom 4 aircrafts in various scenes of two university *campi*, flying at altitude between 5.5 and 6.7 meters, with gimbal pitch angles between 45° to 90°. The image at the bottom provides one example of a full scene of the P-DESTRE set.

B. Annotation Data

The P-DESTRE set is fully annotated at the frame level, by human experts. For each video, we provide one text file with the same filename (plus the ".txt" extension), containing all the corresponding meta-information in comma-separated file format. In these files, each row provides the information for one bounding box in a frame (total of 25 numeric values). The annotation process was divided into four phases: 1) pedestrian detection; 2) tracking; 3) identification and soft biometrics characterisation; and 4) 3D head pose estimation.

At first, the well-known Mask R-CNN [13] method was used to provide an initial estimate of the position of every pedestrian in the scene, with the resulting data subjected to human verification and correction. Next, the deep sort method [33] provided the preliminary tracking information, which again was corrected manually. As result of these two initial steps, we obtained the rectangular bounding boxes providing the regions-of-interest (ROI) of every pedestrian in each frame/video. The next phase of the annotation process was carried out manually, with human annotators that knew personally the volunteers of each university setting the ID information and characterising the samples according to the soft labels. Finally, we used the Deep Head Pose [29] method to obtain the 3D head pose angles for all elements (except backside views), expressed in terms of *yaw*, *pitch* and *roll* values.

Table III provides the details of the labels annotated for every instance (pedestrian/frame) in the dataset, along with the ID information, the bounding box that defines the ROI and the frame information. For every label, we also provide a list of its possible values.

C. Typical Data Degradation Factors

As expected, the acquisition of video/UAV-based data in crowded outdoor environments, from at-a-distance and simulating covert protocols, has led to extremely heterogeneous samples, degraded in multiple perspectives. Under visual inspection, we identified the six major factors that the most frequently reduced the quality data, and augment the challenges of automated image analysis:

1) **Poor resolution/blur**. As illustrated in the top row of Fig. 3, some subjects were acquired from large distances (over 40 m.), with the corresponding ROIs having very poor resolution. Also, some parts of the scenes laid outside the cameras depth-of-field, in result of a large range in objects depth. This led to blurred samples. In both cases, the amount of information available per bounding box is reduced;
2) **Motion blur**. This factor yielded from the non-stationary nature of the cameras and the subjects' movements. In practice, for some bounding boxes, an apparent streaking of the body silhouettes is observed;
3) **Partial occlusions**. As a result of the scene dynamics and due to the multiple objects in the scenes, partial occlusions of subjects were particularly frequent. According to our perception, this might be the most concerning factor of UAV-based data, as illustrated in the third row of Fig. 3;
4) **Pose**. Under covert data acquisition protocols and without accounting for subjects cooperation, many samples regard profile and backside views, in which identification and soft biometric characterisation are particularly difficult;
5) **Lighting/shadows**. As a consequence of the outdoor conditions, many samples are over/under-illuminated,

[7] http://p-destre.di.ubi.pt/download.html

TABLE III
THE P-DESTRE DATASET ANNOTATION PROTOCOL. FOR EACH VIDEO, A TEXT FILE PROVIDES THE ANNOTATION AT FRAME LEVEL, WITH THE ROI OF EACH PEDESTRIAN IN THE SCENE, TOGETHER WITH THE ID INFORMATION AND 16 OTHER SOFT BIOMETRIC LABELS

Attributes	Values
Frame	**1**, **2**, ...
ID	**-1**: 'Unknown', **1**, **2**, ...
Bounding Box	$[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}, \boldsymbol{w}]$ (Top left column, top left row, height, width)
Head Pose	[**flag**, ***yaw***, ***pitch***, ***roll***] (flag: -1=not-available, 1=avaliable)
Age	**0**: 0-11, **1**: 12-17, **2**: 18-24, **3**: 25-34, **4**: 35-44, **5**: 45-54, **6**: 55-64, **7**: $>$ 65, **8**: 'Unknown'
Height	**0**: 'Child', **1**: 'Short', **2**: 'Medium', **3**: 'Tall', **4**: 'Unknown'
Body Volume	**0**: 'Thin', **1**: 'Medium', **2**: 'Fat', **3**: 'Unknown'
Ethnicity	**0**: 'White', **1**: 'Black', **2**: 'Asian', **3**: 'Indian', **4**: 'Unknown'
Hair Color	**0**: 'Black', **1**: 'Brown', **2**: 'White', **3**: 'Red', **4**: Gray', **5**: 'Occluded', **6**: 'Unknown'
Hairstyle	**0**: 'Bald', **1**: 'Short', **2**: 'Medium', **3**: 'Long', **4**: Horse Tail,' **5**: 'Unknown'
Beard	**0**: 'Yes', **1**: 'No', **2**: 'Unknown'
Moustache	**0**: 'Yes', **1**: 'No', **2**: 'Unknown'
Glasses	**0**: 'Yes', **1**: 'Sunglass', **2**: 'No', **3**: 'Unknown'
Head Accessories	**0**: 'Hat', **1**: 'Scarf', **2**: 'Neckless', **3**: 'Occluded', **4**: 'Unknown'
Upper Body Clothing	**0**: 'T-shirt', **1**: 'Blouse', **2**: 'Sweater', **3**: 'Coat', **4**: 'Bikini', **5**: 'Naked', **6**: 'Dress', **7**: 'Uniform', **8**: 'Shirt', **9**: 'Suit', **10**: 'Hoodie', **11**: 'Cardigan'
Lower Body Clothing	**0**: 'Jeans', **1**: 'Leggins', **2**: 'Pants', **3**: 'Shorts', **4**: 'Skirt', **5**: 'Bikini', **6**: 'Dress', **7**: 'Uniform', **8**: 'Suit', **9**: 'Unknown '
Feet	**0**: 'Sport', **1**: 'Classic', **2**: 'High Heels', **3**: 'Boots', **4**: 'Sandals, **5**: 'Nothing', **6**: Unknown'
Accessories	**0**: 'Bag', **1**: 'Backpack', **2**: 'Rolling', **3**: 'Umbrella', **4**: 'Sportif', **5**: 'Market', **6**: 'Nothing', **7**: 'Unknown'
Action	**0**: 'Walk', **1**: 'Run', **2**: 'Stand', **3**: 'Sit', **4**: 'Cycle', **5**: 'Exercise', **6**: 'Pet', **7**: 'Phone', '**8**: 'Leave Bag', **9**: 'Fall', **10**: 'Fight', **11**: 'Date', **12**: 'Offend', **13**: 'Trade'

Fig. 3. Examples of the six factors that - under visual inspection and in a qualitative analysis - constitute the major challenges to automated image analysis in video/UAV-based data. These are the predominant data degradation factors in the P-DESTRE set and the most important co-variates for the responses of automated systems.

with shadowed regions due to the remaining objects in the scene (e.g., buildings, cars, trees, traffic signs...);

6) **UAV elevation angle**. When using gimbal pitch angles close to 90°, the longest axis of the subjects body is almost parallel to the camera axis. In such cases, images contain exclusively a top-view perspective of the subjects, with reduced amount of discriminating information (bottom row of Fig. 3).

When comparing the major features of CCTV and UAV-based data, the *pitch* factor of images is particularly evident. Due to the UAVs altitude, subjects appear almost invariably with negative pitch angles (over 95% of the P-DESTRE images have pitch angles between -10° and 50°), which - according to the results reported in Section IV - appears to be a relevant data degradation factor. Also, the non-stationary feature of UAVs increases the heterogeneity of the resulting data, which again augments the challenges in performing reliable automated image analysis.

D. P-DESTRE Statistical Significance

Let α be a confidence interval. Let p be the error rate of a classifier and $\hat{p}$ be the estimated error rate over a finite number of test patterns. At an α-confidence level, we want that the true error rate does not exceed $\hat{p}$ by an amount larger than $\varepsilon(n,\ \alpha)$. Guyon et al. [12] defined $\varepsilon(n,\ \alpha) = \beta p$ as a fraction of p. Assuming that recognition errors are Bernoulli trials, authors concluded that the number of required trials n to achieve (1-α) confidence in the error rate estimate is given by:

$$n = -ln(\alpha)/(\beta^2 p). \quad (1)$$

Using typical values $\alpha = 0.05$ and $\beta = 0.2$, authors recommend a simpler form, given by: $n \approx \frac{100}{p}$.

Considering the statistics of the P-DESTRE set (Fig. 4), in terms of the number of data acquisition sessions/days per volunteer and the number of bounding boxes per volunteer/session, it is possible to obtain the lower bounds for the statistical confidence in experiments related with identity

verification at the frame level, assuming the 1) short-term re-identification; and 2) long-term re-identification problems.

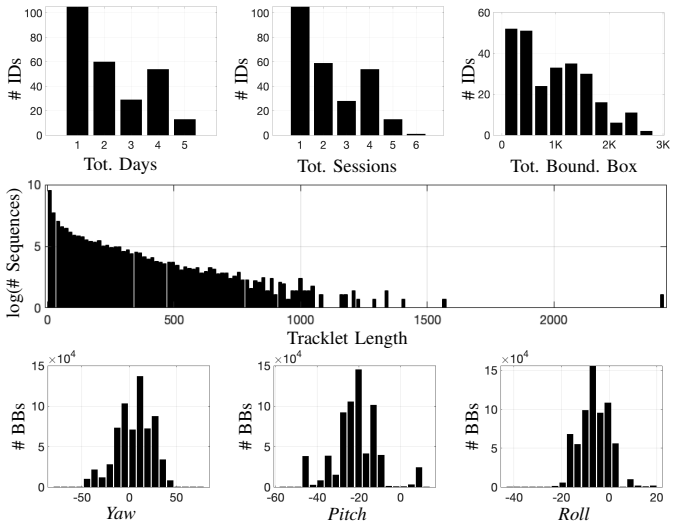

Fig. 4. P-DESTRE statistics. Top row: number of days with data per volunteer (at left), number of data acquisition sessions per volunteer (at center), and number of bounding boxes per volunteer (at right). The histogram at the middle row provides the summary statistics for the length of the tracklet sequences. Finally, the bottom row provides the total of bounding boxes (BBs) per 3D head pose angle, expressed in terms of *yaw*, *pitch* and *roll* values.

In the short-term re-identification setting, considering that each frame (bounding box) with a valid ID ($\geq$ 1) generates a valid template, that all frames of the same ID acquired in different sessions of the same day can be used to generate *genuine* pairs and that frames with different IDs (including '*unknown*') compose the *impostors* set, the P-DESTRE dataset enables to perform 1,246,587,154 (genuine) + 605,599,676,264 (impostor) comparisons, leading to a $\hat{p}$ value with a lower bound of approximately 1.647×10^{-10}. Regarding the pedestrian long-term re-identification problem, where the *genuine* pairs must have been acquired in different days, the dataset enables to perform 2,160,586,581 (genuine) + 605,599,676,264 (impostor) comparisons, leading to a $\hat{p}$ value with a lower bound of approximately 1.645×10^{-10}. Note that these are lower bounds, that do not take into account the portions of data used for learning purposes. Also, these values will increase if we do not assume the independence between images and error correlations are taken into account.

IV. Experiments and Results

In this section we report the results obtained by methods that represent the state-of-the-art in four tasks: pedestrian 1) detection; 2) tracking; 3) short-term re-identification; and 4) long-term re-identification. For contextualisation, we report not only the performance obtained in the P-DESTRE set, but also provide baseline results attained by the same techniques in well-known datasets. Also, for each problem, we illustrate the typical failure cases that we have subjectively perceived during our experiments.

A. Pedestrian Detection

The RetinaNet [19], R-FCN [7] methods were initially considered to represent the state-of-the-art in pedestrian detection, as both outperformed in the PASCAL VOC 2007/2012 [10] challenge ('Person Detection' category). Then, the well-known SSD [21] method was also chosen as baseline, as it is the most widely detector reported in the literature, and its results can be easily contextualised. Accordingly, this section reports a comparison between the performance of the three object detectors in the P-DESTRE/PASCAL sets.

In summary, RetinaNet is composed of a backbone network and two task specific subnetworks. It uses a feature pyramid network as backbone model, to obtain a convolutional feature map over the entire input image. Two sub-networks use this feature representation: the first one classifies the anchor boxes and the second model performs the bounding box regression, to refine the localization of the detected objects. R-FCN uses a fully convolutional architecture, where the translation invariance is obtained by position-sensitive score maps that use specialized convolutional layers to encode the deviations with respect to default positions. A position-sensitive ROI pooling layer is appended on top of the fully connected layers. The SSD model eliminates the proposal generation and feature resampling steps by encapsulating all the processing into a single network. It discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. In our experiments, in a data augmentation setting, the sizes of the learning patches were randomly sampled by [0.1, 1] factor, and horizontally flipped with probability 0.5.

For the PASCAL VOC 2007/2012 set, the official development kit[8] was used to evaluate the methods on the 'Person' category, using 10-fold cross validation. Regarding the P-DESTRE set, a 10-fold cross validation scheme was used, with the data in each split randomly divided into 60% for learning, 20% for validation and 20% for test, i.e., 45 videos were used for learning, 15 for validation and 15 videos for test purposes. The full specification of the samples used in each split and the scores returned by each method is provided in[9].

TABLE IV
Comparison between the average precision (AP) obtained by three methods considered to represent the state-of-the-art in pedestrian detection, in the P-DESTRE and PASCAL VOC 2007/2012 sets.

Method	Backbone	PASCAL VOC	P-DESTRE
RetinaNet [19]	ResNet-50	86.44 $\pm$ 1.03	63.10 $\pm$ 1.64
R-FCN [7]	ResNet-101	84.43 $\pm$ 1.85	59.29 $\pm$ 1.31
SSD [21]	Inception-V2	74.70 $\pm$ 2.69	55.63 $\pm$ 2.93

The results are summarized in Table IV for all datasets/methods, in terms of the average precision obtained at intersection-of-union values equal to 0.5 (i.e., AP@IoU=0.5). Also, Fig. 5 provides the precision/recall curves for both data sets and all detection methods, with the P-DESTRE

[8] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/#devkit
[9] http://p-destre.di.ubi.pt/pedestrian_detection_splits.zip

values being represented by red lines and the PASCAL VOC 2007/2012 results represented by green lines. The shadowed regions denote the standard deviation performance in the 10 splits, at each operating point. Overall, all methods decreased notoriously their effectiveness from the PASCAL VOC set to the P-DESTRE set, in some cases with error rates increasing over 160%. In the case of the R-FCN method, in a small region of the performance space (recall ≈ 0.2), the levels of performance for P-DESTRE and PASCAL VOC were approximately equal, yet the precision values then remain stable for much higher recall values in the PASCAL VOC set.

When comparing the performance of the three techniques tested, we observed that RetinaNet slightly outperformed the competitors in both datasets, in all cases with the R-FCN being the runner-up. The SSD algorithm not only got evidently the lowest average performance among all methods, but also its variance was the largest, which points for the lower robustness of this technique to most of the data co-variates in both the PASCAL VOC and P-DESTRE sets. The observed ranks among the three methods not only accord previous object evaluation initiatives [10], but also the substancial lower performance observed in P-DESTRE than in PASCAL VOC supports the hypothesis claimed in this paper: the P-DESTRE set has evidently different features with respect to previous similar sets.

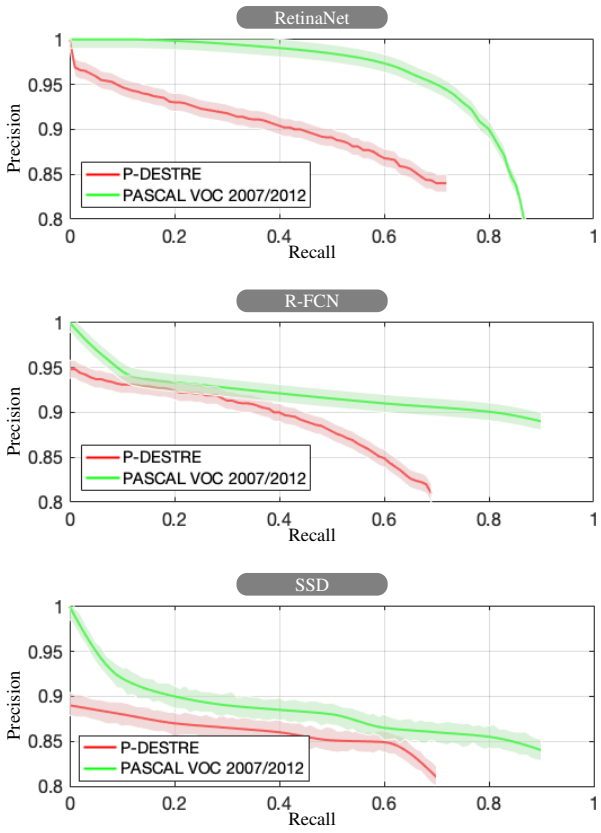

Fig. 5. Comparison between the precision/recall curves observed in the PASCAL VOC 2007/2012 (green lines) and P-DESTRE (red lines) sets. Results are given for the RetinaNet (top plot), R-FCN (middle plot) and SSD (bottom plot) object detection methods.

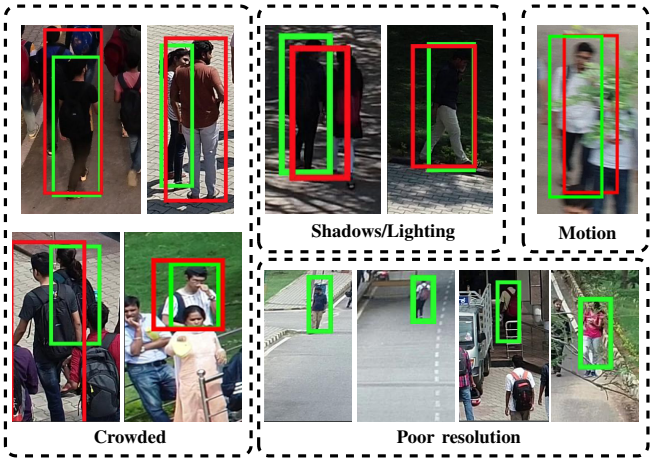

Fig. 6. Typical cases where the object detectors returned the worst scores, i.e., failed to appropriately detect the pedestrians. The green boxes represent the ground-truth, while the red colour denotes the detected boxes.

In a qualitative perspective, we observed that all methods faced particular difficulties in crowded scenes, when only a small part of the subjects silhouette is unoccluded, as illustrated in Fig. 6. Considering that RetinaNet is anchor-based, and that the predefined anchor boxes have a set of handcrafted aspect ratios and scales that are data dependent, performance might have been seriously affected. Even though RetinaNet has clearly outperformed its competitors, the challenging conditions in the P-DESTRE set have still notoriously degraded its effectiveness, when compared to the PASCAL VOC baseline. By analysing the instances in both sets, we observed that the P-DESTRE set has notoriously more *hard* cases than PASCAL VOC, with a significant portion of severely degraded samples (i.e., with severe occlusions, extreme poor resolution and strong local lighting variations/shadows).

In summary, our experiments point for the requirement of novel strategies to handle the specific problems that yield from UAV-based data acquisition. Not only the state-of-the-art solutions provide levels of performance that are still far from the demanded to deploy this kind of solutions in real-environments, but most methods are also sensitive to particularly frequent co-variates in UAV-based imaging (e.g., motion-blur and shadows). Another concerning point is the density of subjects in the scenes, with crowded environments easily providing severe occlusions that constraint the effectiveness of the object detection phase.

B. Pedestrian Tracking

For the tracking task, the TracktorCV [2] and V-IOU [5] methods were initially selected to represent the state-of-the-art, according to: 1) their performance in the MOT challenge[10]; and 2) the fact that both provide freely available implementations, which is important to guarantee that we obtain a fair evaluation between datasets. Moreover, we considered additionally one method (IOU [4]) that is among the most widely reported in the literature. We compared the effectiveness attained by the three techniques in the P-DESTRE and

[10] https://motchallenge.net

MOT challenge sets, in order to perceive the relative hardness of tracking pedestrians in UAV-based data in comparison to a stationary camera setting. In terms of evaluation protocols, the rules provided for the MOT challenges were rigorously met for the MOT evaluation. For the P-DESTRE set, a 10-fold cross validation scheme was used, with the data in each split randomly divided into 60% for learning, 20% for validation and 20% for test, i.e., 45 videos were used for learning, 15 for validation and 15 videos for test purposes. The full details of each split are available at[11].

The TracktorCV method comprises two steps: 1) a regression module, that uses the input of the object detection step to update the position of the bounding box at a subsequent frame; and 2) an object detector that provides the set of bounding boxes for the next frames. The IOU method was developed based on two assumptions: i) the detection step returns a detection per frame for every object to be tracked; and ii) the objects in consecutive frames have high overlap (according to an Intersection-over-Union perspective). Based on these two assumptions, IOU tracks objects without considering image information, which is a key point that contributes for its computational effectiveness. Further, the short tracks are eliminated according to an acceptance threshold. The V-IOU algorithm is an extension of the IOU algorithm that attenuates the problem of false negatives, by associating the detections in consecutive frames according to spatial overlap information. For all three methods, the hyper-parameters were tuned according to the way authors suggested, and are given in[12].

In terms of performance measures, our analysis was based in the Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP) and F1 values, as described in [3]. The summary results attained by both algorithms and datasets are given in Table V. Once again, a consistent degradation in performance from the MOT-17 to the P-DESTRE set was observed, even though the deterioration was in absolute terms far less than the observed for the detection task (here, an decrease in the F1 values of around 10% was observed). It is interesting to observe the larger variance values obtained for tracking methods with respect to the values provided for the detection step. This was justified by the smaller number of learning/test instances available for tracking (working at *sequence*/video level) than for detection (that works at frame level).

When comparing the results of all methods, the TracktorCv outperformed its competitors (V-IOU as runner-up) both in non-aerial and aerial data, decreasing the error rates around 9% with respect to the second best techniques. As expected, the IOU technique obtained invariably the worst performance among all methods tested, which also accords previous tracking performance evaluation initiatives carried out. In all cases, we observed a positive correlation between their typical failure cases, which were invariably related to crowded scenes, and two particularly concerning cases: 1) scenes where, due to extreme pedestrian density, subjects' trajectories cross others at every moment; and 2) when severe occlusions of the body silhouettes occur. Both factors augment the likelihood of observing *fragmentations*, i.e., with the trackers erroneously switching identities of two trajectories in the scene, and wrong *merge* cases, with the trackers erroneously merging two ground truth identities into a single one.

When subjectively comparing the data in MOT-17 and P-DESTRE datasets, it is evident that P-DESTRE contains more complex scenarios, more cluttered backgrounds (e.g., many scenes have 'grass' grounds and tree branches) and more poor resolution subjects. Also, we noted that the trackability of pedestrians also depends on the tracklet length (i.e., the number of consecutive frames where an object appears), with the values in MOT-17 varying from 1 to 1,050 (average 304) and in P-DESTRE varying from 4 to 2,476 (average 63.7 $\pm$ 128.8), as illustrated in Fig. 4.

TABLE V
COMPARISON BETWEEN THE TRACKING PERFORMANCE ATTAINED BY THREE ALGORITHMS CONSIDERED TO REPRESENT THE STATE-OF-THE-ART IN THE P-DESTRE AND MOT-17 DATA SETS.

Method	Dataset	MOTA	MOTP	F-1
TracktorCv [2]	**MOT-17**	65.20 $\pm$ 9.60	62.30 $\pm$ 11.00	89.60 $\pm$ 2.80
	P-DESTRE	56.00 $\pm$ 3.70	55.90 $\pm$ 2.60	87.40 $\pm$ 2.00
V-IOU [5]	**MOT-17**	52.50 $\pm$ 8.80	57.50 $\pm$ 9.50	86.50 $\pm$ 1.90
	P-DESTRE	47.90 $\pm$ 5.10	51.10 $\pm$ 5.80	83.30 $\pm$ 8.40
IOU [4]	**MOT-17**	45.51 $\pm$13.61	46.02 $\pm$ 12.40	78.21 $\pm$ 3.12
	P-DESTRE	38.27 $\pm$ 8.42	39.68 $\pm$ 4.92	74.29 $\pm$ 6.87

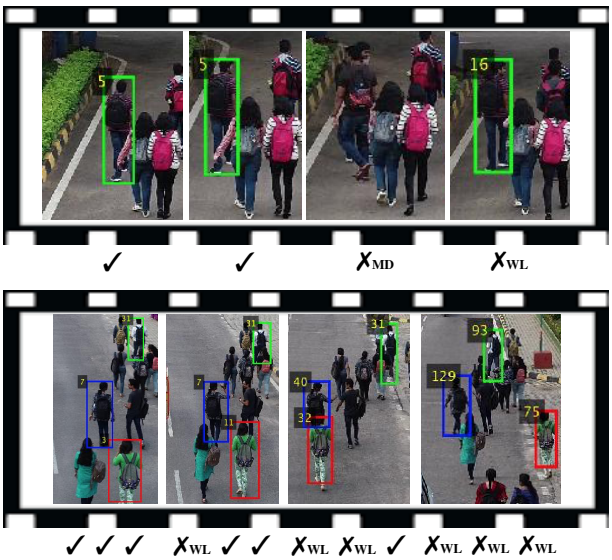

Fig. 7. Examples of sequences where the tracking methods faced difficulties, either missing the ground-truth targets at some point or producing a *fragmentation* that resulted in a wrong label assignment. *MD* stands for "missed detection" and *WL* represents "wrong label" assignment.

[11] http://p-destre.di.ubi.pt/pedestrian_tracking_splits.zip
[12] http://p-destre.di.ubi.pt/parameters_tracking.zip

C. Pedestrian Short-term Re-Identification

We selected three well known re-identification algorithms to represent the state-of-the-art and assessed their performance. The MARS [39] dataset was selected to represent the stationary datasets, as it is currently the largest video-based source that is freely available.

According to the results reported on a challenge [36], the GLTR [18], COSAM [31] and NVAN [22] methods were selected. The GLTR exploits multi-scale temporal cues in video sequences, by modelling separately short- and long-term features. Short-term components capture the appearance and motion of pedestrians, using parallel dilated convolutions with varying rates. Long-term information is extracted by a temporal self-attention model. The key in COSAM is to capture intra video attention using a co-segmentation module, extracting task-specific regions-of-interest that typically correspond to pedestrians and their accessories. This module is plugged between convolution blocks to induce the notion of co-segmentation, and enables to obtain representations of both the spatial and temporal domains. Finally, the Non-local Video Attention Network (NVAN) exploits both spatial and temporal cues by introducing a non-local attention operation into the backbone CNN at multiple feature levels. Further, it reduces the computational complexity of the inference step by exploring the spatial and temporal redundancy that is observed in the learning data.

In a 5-fold setting, both datasets were divided into random splits, each one containing the learning, query and gallery sets, in proportions 50:10:40. For the MARS dataset, the evaluation protocol described in[13] was used. For the P-DESTRE dataset, we considered 1,894 tracklets of 608 IDs, with an average number of frames per tracklet of 67,4. The full specification of the samples used for learning/validation/test purposes in each split is given in[14].

Regarding the GLTR method, the ResNet50 was used as backbone model, with the learning rate set to 0.01. In the COSAM method, the Se-ResNet50 architecture was used as backbone model. The COSAM layer was plugged between the forth and fifth convolution layers, with the learning rate set to 0.0001 and the reduction dimension size set to 256. For the NVAN method, we also used ResNet50 architecture as backbone network and plugged two non-local attention layers (after *Conv3_3* and *Conv3_4*) and three non-local layers (after *Conv4_4*, *Conv4_5*, and *Conv4_6*). The input frames were resized into 256×128. The model was trained using the Adam algorithm, with 300 epochs and learning rate set to 0.0001.

The summary results are provided in Table VI. In opposition to the detection and tracking problems, it is interesting to note that no significant decreases in performance were observed from the MARS to the P-DESTRE data, which points for the suitability of the existing short-term re-identification solutions for UAV-based data. Fig. 8 provides the cumulative rank-n curves for all algorithms/datasets. The red lines represent the P-DESTRE results and the green series denote the MARS values. Results are given in terms of the identification rate with respect to the proportion of gallery identities retrieved (i.e., hit/penetration plot). Apart the outperforming results of NVAN, it is particularly interesting to note the apparently contradictory results of the GLTR and COSAM algorithms in the MARS and P-DESTRE sets. In all cases, in terms of the top-20 performance, the P-DESTRE results were far worse than the corresponding MARS values. However, for larger ranks (starting at 5% of the enrolled identities), the P-DESTRE values were solidly better than the ranks observed for MARS. Also, in case of heavily degraded MARS instances, algorithms returned almost random results, which was not observed for the P-DESTRE. This might be justified by the fact that P-DESTRE contains more *poor quality* data than MARS, yet it does not provide *extremely degraded* (i.e., almost *impossible*) instances that turn the identification into a quasi-random process.

TABLE VI
COMPARISON BETWEEN THE RE-IDENTIFICATION PERFORMANCE ATTAINED BY THREE STATE-OF-THE-ART METHODS IN THE P-DESTRE AND MARS DATA SETS.

Method	Dataset	mAP	Rank-1	Rank-20
GLTR [18]	**MARS**	77.74 ± 1.07	84.72 ± 2.61	95.80 ± 2.34
	P-DESTRE	77.68 ± 9.46	75.96 ± 11.77	95.48 ± 3.17
COSAM [31]	**MARS**	78.35 ± 1.66	84.03 ± 0.91	96.97 ± 0.98
	P-DESTRE	80.64 ± 9.91	79.14 ± 12.43	97.10 ± 1.85
NVAN [22]	**MARS**	81.13 ± 1.35	85.94 ± 0.94	97.20 ± 0.97
	P-DESTRE	82.78 ± 10.35	80.42 ± 12.38	98.34 ± 1.93

Based in these experiments, Fig. 9 highlights some notorious cases for re-identification purposes. The upper row represents the particularly hazardous cases in terms of *convenience*, where different IDs were erroneously perceived as the same. This was mostly due to similarities in clothing, together with shared soft biometric labels between different IDs. The bottom row provides the particularly dangerous cases for *security* purposes, where methods had difficulties in identifying a known ID. Here, errors often yielded from notorious differences in pose and scale between the query/gallery data. Along with the background clutter, these factors were observed to decrease the effectiveness of the feature representations, and were among the most concerning for re-identification performance.

D. Long-term Pedestrian Re-identification

As stated above, the pedestrian video-based long-term re-identification problem was the main motivation for the development of the P-DESTRE dataset. Here, there is not any guarantee about the clothing appearance of subjects, nor about the time elapsed between consecutive observations of one ID. In such circumstances, the analysis of alternative features should be considered (e.g., face, gait or soft-biometrics based).

Considering that there are not yet methods in the literature specifically designed for this kind of task, we have chosen

[13] http://www.liangzheng.com.cn/Project/project_mars.html
[14] http://p-destre.di.ubi.pt/pedestrian_reid_splits.zip

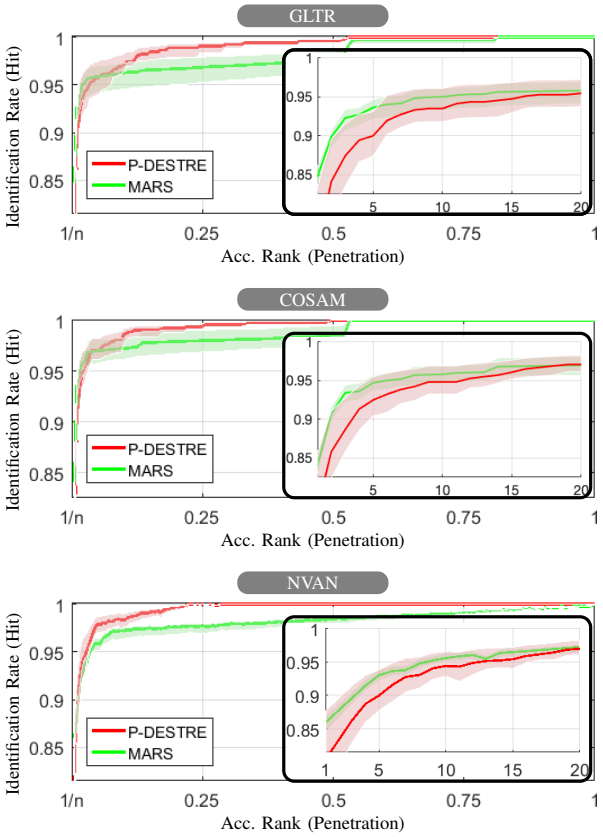

Fig. 8. Comparison between the closed-set identification (CMC) curves observed in the MARS (green lines) and P-DESTRE (red lines) sets for the GLTR, COSAM and NVAN re-identification techniques. Zoomed-in regions with the top-1 to 20 results are shown in the inner plots.

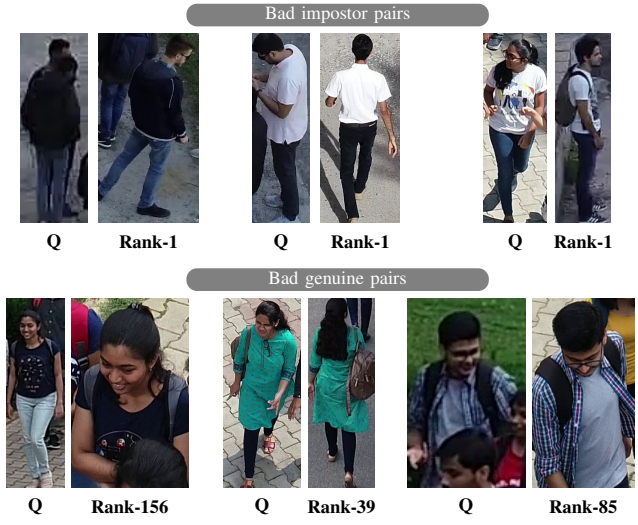

Fig. 9. Examples of the instances that got the worst re-identification performance. The upper row illustrates typical false matches, almost invariably related with clothing styles and colours. The bottom row provides some examples of cases where (due to differences in pose and scale), the true identities could not be retrieved among the top positions. "'Q" represents the query image and "Rank-i" provides the rank of the corresponding gallery image.

a combination of two well-known re-identification techniques that combine face and body features. Similarly to the previous tasks, the goal was to obtain an approximation for the effectiveness attained by the existing solutions in UAV-based data. Such levels of performance constitute a baseline for this problem and can be used as basis for further developments.

The facial regions-of-interest were detected by the SSH method [25] (acceptance threshold=0.7), from where a feature representation was obtained using the ArcFace [8] model. For the body-based analysis, the COSAM [31] model provided the feature representation. Both models were trained *from scratch*. The data were sampled into 5 trials, each one containing learning/gallery/query instances in proportions 50:10:40. As for the previous tasks, the full specification of the samples used in each split is given in[15].

For the ArcFace method, the MobileNetV2 was used as backbone model, and the learning rate set to 0.01. Regarding COSAM, the Se-ResNet50 was used as backbone model, and the COSAM layer was plugged into the forth and fifth convolutional layers, with learning rate equal to $1e^{-4}$ and dimension size equal to 256. Each model was trained reparately, and during the test phase, the mean value of the ArcFace facial features in the tracklet were appended to the body-based representation yielding from COSAM. The Euclidean norm was used as distance function between such concatenated representations.

Fig. 10 provides the cumulative rank-n curves obtained, in terms of the successfull identification rates with respect to the proportion of gallery identities (i.e., hit/penetration plot). As expected, when compared to the short-term re-identification setting, performance was substantially lower (rank-1 $\approx$ 79.14% for re-identification $\rightarrow \approx$ 49.88% for search), which accords the human perception for the additional difficulty of *search* with respect to *re-identify*.

TABLE VII
BASELINE LONG-TERM PEDESTRIAN RE-IDENTIFICATION PERFORMANCE OBTAINED BY AN ENSEMBLE OF ARCFACE [8] + COSAM [31] IN THE P-DESTRE DATA SET.

Method	mAP	Rank-1	Rank-20
ArcFace [8] + COSAM [31]	34.90 $\pm$ 6.43	49.88 $\pm$ 8.01	70.10 $\pm$ 11.25

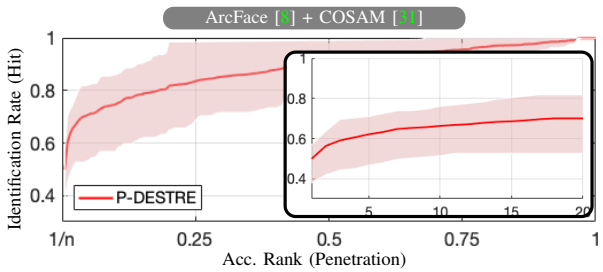

Fig. 10. Closed-set identification (CMC) curves obtained for the long-term re-identification problem in the P-DESTRE dataset. The inner plot provides the top-20 results as a zoomed-in region.

[15]http://p-destre.di.ubi.pt/pedestrian_search_splits.zip

Based in our qualitative analysis of the results, Fig. 11 provides three types of examples: the upper row shows some successful identification cases, in which the model retrieved the true identity in the first position. In most cases, we noted that subjects kept *some* piece of clothing/accessories between observations (e.g., glasses or backpack) and the same hairstyle. The remaining rows illustrate the failure cases: the second row provides examples of the hazardous cases for *convenience* purposes, in which due to similarities in pose, accessories and soft biometric labels between the query and gallery images, false matches have occurred. Finally, the bottom row provides examples of *security sensitive* cases, where the IDs of the queries were retrieved in high positions (ranks 56, 73 and 98), i.e., the system failed to detect a subject of interest in a crowd.

Fig. 11. Examples of the instances where good/poor pedestrian search performance was observed. The upper row illustrates particularly successful cases, while the bottom rows show pairs of images where the used algorithm had notorious difficulties to retrieve the correct identity. "'Q" represents the query image and "Rank-i" provides the rank of the retrieved gallery image.

The challenges of long-term re-identification are illustrated in Fig. 12, providing the differences between the probabilities of obtaining a top-i correct identification (hit), $\forall i \in \{1, \ldots, n\}$, i.e., retrieve the identity corresponding to a query up to the ith position, for the search and re-identification problems. Here, $\mathrm{P}_s(i)$ and $\mathrm{P}_r(i)$ denote the probabilities of observing a *hit* in the search P_s and re-identification P_r tasks, i.e., negative $\big(\mathrm{P}_s(i)$ - $\mathrm{P}_r(i)\big)$ denote higher probabilities for re-identification success than for search success. The zoomed-in region given at the right part of the Figure shows the additional difficulty (of almost 40 percentual points) in retrieving the true identity in a single shot (difference between top-1 values). Then, the gap between the accumulated values of P_s and P_r decreases in a monotonous way, and only approaches 0 near the full penetration rate, i.e., when all the known identities are retrieved for a query. In summary, it is much more difficult to identify pedestrians when no clothing information can be used, which paves the way for further developments in this kind of technology. According to our goals in developing this data source, the P-DESTRE set is a tool to support such advances in the state-of-the-art.

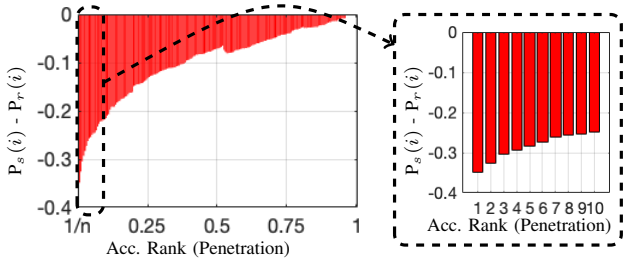

Fig. 12. Differences between the probability of retrieving the true identity of a query among the top-i positions, $\forall i \in \{1, \ldots, 100\}$, for the pedestrian long-term re-identification (P_s) and short-term re-identification (P_r) problems.

V. Conclusions

This paper announced the availability of the P-DESTRE dataset, which provides video sequences of pedestrians taken from UAVs in outdoor environments. The key point of the P-DESTRE set is to provide full annotations that enable the research on *long-term pedestrian re-identification*, where the time elapsed between consecutive observations of IDs forbids the use of clothing-based features. Apart this, the P-DESTRE set is also suitable for research on UAV/video-based pedestrian detection, tracking, short-term re-identification and soft biometrics analysis.

Additionally, as a secondary contribution, we offered the results of our own evaluation of the state-of-the-art in the pedestrian detection, tracking and short-term re-identification problems, comparing the performance attained in data acquired from stationary (CCTV) and from moving/UAV devices. Such results point for a particular hardness of the existing solutions to detect and track subjects UAV-based data. In opposition, the existing short-term re-identification techniques appear to be relatively robust to the features typical of UAV-based data.

Overall, the decreases in performance observed from CCTV to UAV-based data support the originality and usefulness of P-DESTRE. hence, potential directions for further developments of long-term UAV-based re-identification include the use of *attention-based* networks that disregard portions of the input data known to be ineffective for long-term re-identification (e.g., clothes or hairstyles). Another important field will be the development of *domain adaptation* techniques robust to changes in the UAV-acquisition settings and environments heterogeneity.

Acknowledgements

This work is funded by FCT/MEC through national funds and co-funded by the FEDER - PT2020 partnership agreement

under the projects UID/50008/2019, POCI-01-0247-FEDER-033395 and C4: Cloud Computing Competence Centre.

REFERENCES

[1] M. Ahmed, M. Jahangir, H. Afzal, A. Majeed and I. Siddiqi. Using Crowd-source based features from social media and Conventional features to predict the movies popularity. In proceedings of the *IEEE International Conference on Smart Cities, Social Communication and Sustained Communication (SmartCity)*, pag. 273–278 2015. 3

[2] P. Bergmann, T. Meinhardt and L. Leal-Taixe. Tracking without bells and whistles. *ArXiv*, https://arxiv.org/abs/1903.05625v3, 2019. 7, 8

[3] K. Barnardin and R. Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, 10.1155/2008/246309, 2008. 8

[4] E. Bochinski, V. Eiselein and T. Sikora. High-Speed tracking-by-detection without using image information. In Proceedings of the *IEEE International Conference on Advanced Video and Signal Based Surveillance*, doi: 10.1109/AVSS.2017.8078516, 2017. 7, 8

[5] E. Bochinski, T. Senst and T. Sikora. Extending IOU based multi-object tracking by visual information. In Proceedings of the *IEEE International Conference on Advanced Video and Signal Based Surveillance*, doi: 10.1109/AVSS.2018.8639144, 2018. 7, 8

[6] M. Bonetto, P. Korshunov, G. Ramponi and T. Ebrahimi. Privacy in Mini-drone Based Video Surveillance. in Proceedings of the *Workshop on De-identification for privacy protection in multimedia*, doi: 10.13140/RG.2.1.4078.5445, 2015. 3

[7] J. Dai, Y. Li, K. He and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In proceedings of the *International Conference on Neural Information Processing Systems*, pag. 379–387, 2016. 6

[8] J. Deng, J. Guo, N. Xue and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In proceedings of the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*,doi: 10.1109/CVPR.2019.00482, 2019. 10

[9] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang and Q. Tian The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. In proceedings of the *European Conference on Computer Vision*, pag. 370–386, 2018. 2

[10] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn and A. Zisserman. The PASCALVisual Object Classes Challenge: A Retrospective. *International Journal Computer Vision* , 111, pag. 318-327, 2015. 6, 7

[11] A. Grigorev, Z. Tian, S.R̃ho, J. Xiong, S. Liu and F. Jiang. Deep person re-identification in UAV images. *EURASIP Journal on Advanvced Signal Processing*, 54, doi: 10.1186/s13634-019-0647-z, 2019. 1, 3

[12] I. Guyon, J. Makhoula, R. Schwartz and V. Vapnik. What size test set gives good error rate estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1*, pages 52–64, February 1998. 5

[13] K. He, G. Gkioxari, P. Dollar and R. Girshick. Person Re-Identification by Descriptive and Discriminative Classification. *ArXiv*, https://arxiv.org/abs/1703.06870v3, 2018. 4

[14] M. Hirzer, C. Beleznai, P. Roth and H. Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In proceedings of the *Scandinavian Conference on Image Analysis*, pag. 91–102, 2011. 2, 3

[15] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa and P. B. Sujit. Dronesurf: Benchmark dataset for drone-based face recognition. In proceedings of the 14^{th} *IEEE International Conference on Automatic Face and Gesture Recognition*, pag. 1–7, 2019. 2

[16] R. Layne, T. Hospedales and S. Gong. Investigating open-world person re-identification using a drone. In proceedings of the *European Conference on Computer Vision*, pag. 225–240, 2014. 3

[17] W. Li, R. Zhao, T. Xiao and X. Wang. DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification. In proceedings of the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*,doi: 10.1109/CVPR.2014.27, 2014. 1, 2, 3

[18] J. Li, J. Wang1, Q. Tian, W. Gao and S. Zhang. Global-Local Temporal Representations For Video Person Re-Identification *ArXiv*, https://arxiv.org/abs/1908.10049v1, 2019. 9

[19] T-Y Lin, P. Goyal, R. Girshick, K. He and P. Dollar. Focal Loss for Dense Object Detection *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), pag. 318–327, 2020. 6

[20] Y. Liu, B. Peng, P. Shi, H. Yan, Y. Zhou, B. Han, Y. Zheng, C. Lin, J. Jiang, Y. Fan, T. Gao, G. Wang, J. Liu, X. Lu and D. Xie. iQIYI-VID: A Large Dataset for Multi-modal Person Identification. *ArXiv*, https://arxiv.org/abs/1811.07548v2, 2019. 3

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu and A.C. Berg. SSD: Single shot multibox detector. In proceedings of the *European Conference on Computer Vision*, doi: 10.1007/978-3-319-46448-0_2, 2016. 6

[22] C.T. Liu, C.W. Wu, Y.C.F. Wang and S.Y. Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In proceedings of the *British Machine Vision Conference*, https://arxiv.org/abs/1908.01683, 2019. 9

[23] I. Mademlis, V. Mygdalis, N.Nikolaidis, M. Montagnuolo, F. Negro, A. Messina and I.Pitas. High-Level Multiple-UAV Cinematography Tools for Covering Outdoor Events. *IEEE Transactions on Broadcasting*, 65(3), pag. 627–635, 2019. 2

[24] M. Mueller, N. Smith and B. Ghanem. A Benchmark and Simulator for UAV Tracking In proceedings of the *European Conference on Computer Vision*, doi: 10.1007/978-3-319-46448-0_27, 2016. 2

[25] M. Najibi, P. Samangouei, R. Chellappa and L. Davis. SSH: single stage headless face detector. In proceedings of the *International Conference on Computer Vision*, doi: 10.1109/ICCV.2017.522, 2017. 10

[26] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y. Jiang and X. Xue. Long-Term Cloth-Changing Person Re-identification. *ArXiv*, https://arxiv.org/abs/2005.12633, 2020. 3

[27] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In Proceedings of the *European Conference on Computer Vision Workshops*, arXiv:1609.01775v2, 2016. 3

[28] A. Robicquet, A. Sadeghian, A. Alahi and S. Savarese. Learning Social Etiquette: Human Trajectory Prediction In Crowded Scenes. In Proceedings of the *European Conference on Computer Vision*, doi: 10.1007/978-3-319-46484-8_33, 2016. 2

[29] N. Ruiz, E. Chong and J. Rehg. Fine-Grained Head Pose Estimation Without Keypoints. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, doi: 10.1109/CVPRW.2018.00281, 2018. 2, 4

[30] A. Singh, D. Patil and S. Omkar. Eye in the Sky: Real-time Drone Surveillance System (DSS) for Violent Individuals Identification using ScatterNet Hybrid Deep Learning Network. In proceedings of the *IEEE Computer Vision and Pattern Recognition Workshops 2018*, doi: 10.1109/CVPRW.2018.00214, 2018. 1, 3

[31] A. Subramaniam, A. Nambiar and A. Mittal. Co-segmentation Inspired Attention Networks for Video-based Person Re-identification. In proceedings of the *International Conference on Computer Vision*, pag. 562-572, 2019. 9, 10

[32] X. Wang and R. Zhao. Person re-identification: System design and evaluation overview. *Person Re-Identification*, Springer, doi: 10.1007/978-1-4471-6296-4_17, 2014. 2, 3

[33] N. Wojke, A. Bewley and D. Paulus. Simple online and realtime tracking with a deep association metric. In proceedings of the *IEEE International Conference on Image Processing*, pag. 3645–3649, 2017. 4

[34] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In proceedings of the emphIEEE Computer Vision and Pattern Recognition Workshops 2018, doi: 10.1109/CVPR.2018.00543, 2018. 3

[35] G-S. Xia, X. Bai, J. Ding, Z, Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo and L. Zhang. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In proceedings of the emphIEEE Computer Vision and Pattern Recognition, doi: 10.1109/CVPR.2018.00418, 2018 2

[36] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao and S. Hoi. Deep Learning for Person Re-identification: A Survey and Outlook. *ArXiv*, https://arxiv.org/abs/2001.04193v1, 2020. 9

[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian. Scalable Person Re-identification: A Benchmark. In proceedings of the *IEEE International Conference on Computer Vision*, doi: 10.1109/ICCV.2015.133, 2015. 3

[38] S. Zheng, J. Zhang, K. Huang, R. He and T. Tan. Robust View Transformation Model for Gait Recognition. In proceedings of the *IEEE International Conference on Image Processing*, doi: 10.1109/ICIP.2011.6115889, 2011. 1

[39] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang and Q. Tian. MARS: A video benchmark for large-scale person re-identification. In proceedings of the *European Conference on Computer Vision*, Lecture Notes in Computer Science, vol 9910, pag. 868–884, 2016. 3, 9

[40] P. Zhu, L. Wen, X. Bian, H. Ling and Q. Hu. Vision Meets Drones: A Challenge. *ArXiv*, https://arxiv.org/abs/1804.07437, 2018. 2