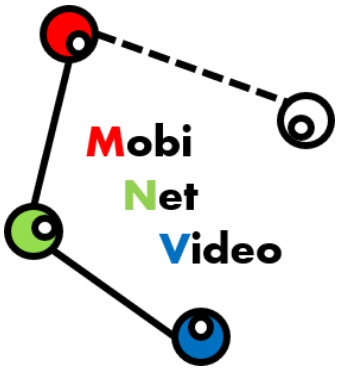


TEC2017-88169-R MobiNetVideo (2018-2020)

Visual Analysis for Practical Deployment of Cooperative Mobile Camera Networks

D1.3 v1

Evaluation datasets

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD

AUTHORS LIST

José M. Martínez	josem.martinez@uam.es
Álvaro García Martín	alvaro.garcia@uam.es
Marcos Escudero Viñolo	marcos.escudero@uam.es
Pablo Carballeira López	pablo.carballeira@uam.es
Juan Carlos San Mogiel Avedillo	juancarlos.sanmiguel@uam.es

HISTORY

Version	Date	Editor	Description
0.1	24/04/2019	José M. Martínez	Initial draft version
0.2	30/04/2019	Álvaro García-Martín	Contributions: object detection and re-identification: VIPer, Market and City
0.3	25/07/2019	Marcos Escudero, Pablo Carballeira	Contributions: Places365, VPU Generated Datasets for scene recognition.
0.4	26/07/2019	Juan Carlos San Miguel	Contributions: Multiple object tracking datasets
0.5	27/07/2019	José M. Martínez	Editorial checking
1.0	27/07/2019		First version

CONTENTS:

1. Introduction

1.1. Motivation

Work package 1 (WP1) aims at the initial establishment and maintenance of a development framework for the remaining work packages.

This deliverable describes the work related with the task T.1.3 Generation of datasets. Support to other tasks for generating test data and defining evaluation methodologies. It includes the selection of appropriate datasets (sequences and associated ground-truth) and their generation if required.

1.2. Document structure

This document contains the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: State of the art public datasets
- Chapter 3: VPU generated datasets
- Chapter 4: Evaluation methodologies
- Chapter 5: Conclusions

2. State of the art public datasets

2.1. Object detection

2.1.1. MOT2017Det

The MOT17 benchmark [1] includes a set of 14 (half for training and half for testing) sequences with crowded scenarios, different viewpoints, camera motions and weather conditions. The annotations for all sequences have been carried out by qualified researchers from scratch following a strict protocol, and finally double-checked to ensure highest annotation accuracy. Not only pedestrians are annotated, but also vehicles, sitting people, occluding objects, as well as other significant object classes. With this fine-grained level of annotation it is possible to accurately compute the degree of occlusion and cropping of all bounding boxes, which is also provided with the benchmark. The annotations of the testing sequences are not.

Sequences are very different from each other; we can classify them according to:

- Moving or static camera – the camera can be held by a person, placed on a stroller or on a car, or can be positioned fixed in the scene.
- Viewpoint – the camera can overlook the scene from a high position, a medium position (at pedestrian's height), or at a low position.
- Conditions – the weather conditions in which the sequence was taken are reported in order to obtain an estimate of the illumination conditions of the scene. Sunny sequences may contain shadows and saturated parts of the image, while the night sequence contains a lot of motion blur, making pedestrian detection and tracking rather challenging. Cloudy sequences on the other hand contain fewer of those artifacts.

Some examples from this dataset can be found in **Figure 1**. **Figure 2**, gives an overview of the training and testing sequence characteristics for the challenge, including the number of bounding boxes used.

Figure 1. MOT16 visual examples.

Training sequences										
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions	Source
MOT16-02	30	1920x1080	600 (00:20)	49	17,833	29.7	static	medium	cloudy	new
MOT16-04	30	1920x1080	1,050 (00:35)	80	47,557	45.3	static	high	night	new
MOT16-05	14	640x480	837 (01:00)	124	6,818	8.1	moving	medium	sunny	[13]
MOT16-09	30	1920x1080	525 (00:18)	25	5,257	10.0	static	low	indoor	new
MOT16-10	30	1920x1080	654 (00:22)	54	12,318	18.8	moving	medium	night	new
MOT16-11	30	1920x1080	900 (00:30)	67	9,174	10.2	moving	medium	indoor	new
MOT16-13	25	1920x1080	750 (00:30)	68	11,450	15.3	moving	high	sunny	new
Total training			**5,316 (03:35)**	**512**	**110,407**	**20.8**				

Testing sequences										
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions	Source
MOT16-01	30	1920x1080	450 (00:15)	23	6,395	14.2	static	medium	cloudy	new
MOT16-03	30	1920x1080	1,500 (00:50)	148	104,556	69.7	static	high	night	new
MOT16-06	14	640x480	1,194 (01:25)	217	11,538	9.7	moving	medium	sunny	[13]
MOT16-07	30	1920x1080	500 (00:17)	55	16,322	32.6	moving	medium	shadow	new
MOT16-08	30	1920x1080	625 (00:21)	63	16,737	26.8	static	medium	sunny	new
MOT16-12	30	1920x1080	900 (00:30)	94	8,295	9.2	moving	medium	indoor	new
MOT16-14	25	1920x1080	750 (00:30)	230	18,483	24.6	moving	high	sunny	new
Total testing			**5,919 (04:08)**	**830**	**182,326**	**30.8**				

Figure 2. Overview of the sequences currently included in the MOT16 benchmark.

Aside from pedestrians, the annotations also include other classes like vehicles, bicycles, etc. In **Figure 3**, we detail the types of annotations that can be found in each sequence of MOT17.

Annotation classes												
Sequence	Pedestrian	Person on vehicle	Car	Bicycle	Motorbike	Non motorized vehicle	Static person	Distractor	Occluder on the ground	Occluder full	Reflection	Total
MOT16-01	6,395	346	0	341	0	0	4,790	900	3,150	0	0	15,922
MOT16-02	17,833	1,549	0	1,559	0	0	5,271	1,200	1,781	0	0	29,193
MOT16-03	104,556	70	1,500	12,060	1,500	0	6,000	0	24,000	13,500	0	163,186
MOT16-04	47,557	0	1,050	11,550	1,050	0	4,798	0	23,100	18,900	0	108,005
MOT16-05	6,818	315	196	315	0	11	0	16	0	0	0	7,671
MOT16-06	11,538	150	0	118	0	0	269	238	109	0	0	12,422
MOT16-07	16,322	0	0	0	0	0	2,023	0	1,920	0	0	20,265
MOT16-08	16,737	0	0	0	0	0	1,715	2,719	6,875	0	0	28,046
MOT16-09	5,257	0	0	0	0	0	0	1,575	1,050	0	948	8,830
MOT16-10	12,318	0	25	0	0	0	1,376	470	2,740	0	0	16,929
MOT16-11	9,174	0	0	0	0	0	0	306	596	0	0	10,076
MOT16-12	8,295	0	0	0	0	0	1,012	765	1,394	0	0	11,464
MOT16-13	11,450	0	4,484	103	0	0	0	4	2,542	680	0	19,263
MOT16-14	18,483	0	1,563	0	0	0	712	47	4,062	393	0	25,260
Total	292,733	2,430	8,818	26,046	2,550	11	27,966	8,238	73,319	33,473	948	476,532

Figure 3. Overview of the types of annotations currently found in the MOT16 benchmark.

2.2. Object re-identification

2.2.1. VIPeR

VIPeR [2] consists of 632 people from two disjoint views. Each person has only one image per view. VIPeR suffers from subtantial viewpoint and illumination variations. Some examples from this dataset can be found in **Figure 4**.

Figure 4. VIPeR visual examples.

2.2.2. Market1501

During the dataset collection of the Market1501 [3], a total of six cameras were placed in front of a campus supermarket, including five 1280×1080 HD cameras, and one 720×576 SD camera. Overlapping exists among these cameras. This dataset contains 32,668 bboxes of 1,501 identities. Due to the open environment, images of each identity are captured by at most six cameras. Each annotated identity is captured by at least two cameras, so that cross-camera search can be performed. Some examples from this dataset can be found in **Figure 5**.

Figure 5. Market1501 visual examples.

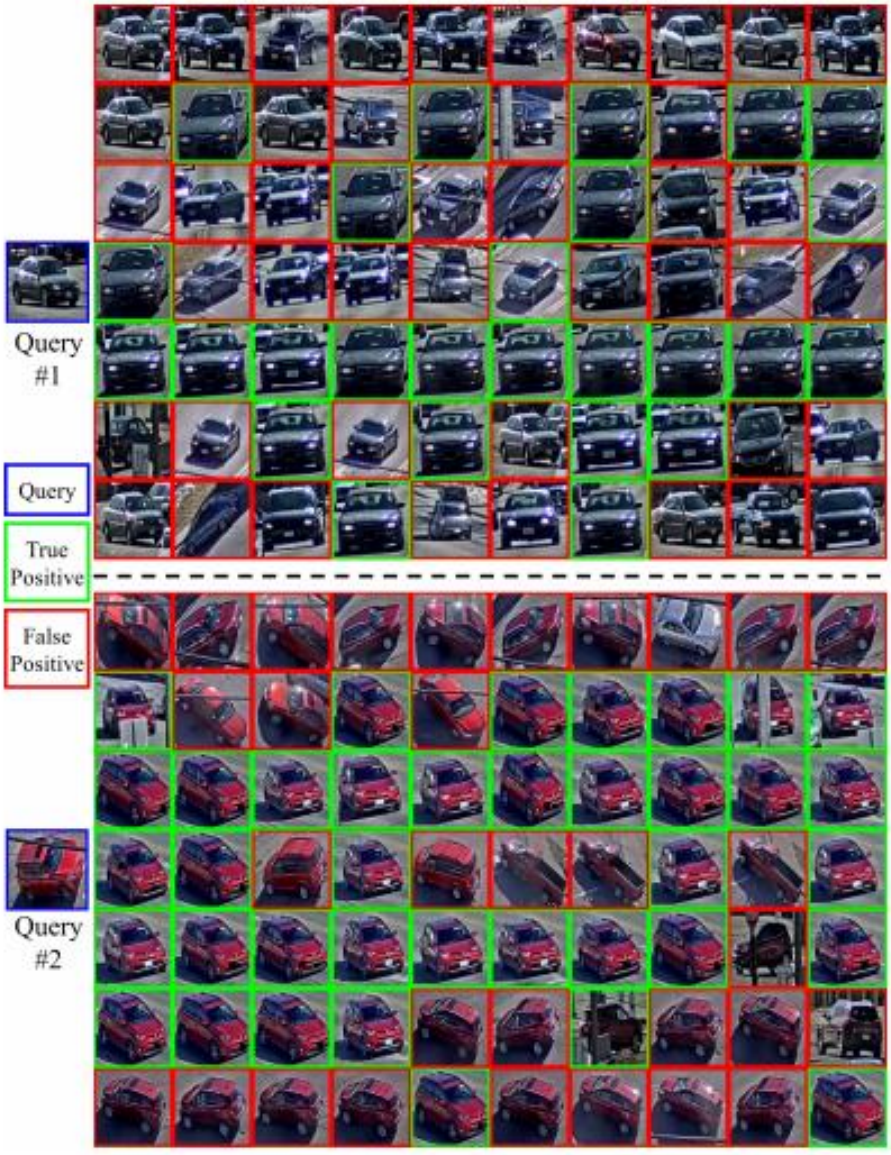


Figure 6. CityFlow-ReID visual examples.

2.2.3. CityFlow-ReID

The CityFlow-ReID dataset [4] contains 3.25 hours of videos collected from 40 cameras spanning across 10 intersections in a mid-sized U.S. city. The distance between the two furthest simultaneous cameras is 2.5 km, which is the longest among all the existing benchmarks. The dataset covers a diverse set of location types, including intersections, stretches of roadways, and

highways. With the largest spatial coverage and diverse scenes and traffic conditions, it is the first benchmark that enables city-scale video analytics.

Cameras at the same intersection sometimes share overlapping field of views (FOVs) and some cameras use fisheye lens, leading to strong radial distortion of their captured footage. Besides, because of the relatively fast vehicle speed, motion blur may lead to failures in object detection and data association.

A sampled subset from CityFlow, noted as CityFlowReID, is dedicated for the task of image-based ReID. CityFlow-ReID contains 56,277 bounding boxes in total, where 36,935 of them from 333 object identities form the training set, and the test set consists of 18,290 bounding boxes from the other 333 identities. The rest of the 1,052 images are the queries. On average, each vehicle has 84.50 image signatures from 4.55 camera views. Some examples from this dataset can be found in **Figure 6**.

2.3. Scene Recognition

2.3.1. Places365

Places365 dataset [7] is explicitly intended for scene recognition. It composed of 10 million of images comprising 434 scene classes. There are two versions of the dataset: Places365-Standard with 1.8 million train and 36000 validation images from 365 scene classes, and Places365-Challenge-2016, in which the size of the training set is increased up to 6.2 million extra images, including 69 new scene classes—leading to a total of 8 million train images from 434 scene classes—. Our experiments are carried out using the Places365-Standard dataset. **Figure 7** presents example images and scene classes of this dataset.

The examples in **Figure 7** illustrate the challenges of this dataset which combines a broad number of non-homogeneous classes with blurry *boundaries* between classes.

Field

Forest

Lagoon

Coast

Kitchen

Kitchenette

Figure 7. Places365 visual examples. Each row contains visual examples of different scene classes sharing common objects.

2.4. Multiple object tracking

2.4.1. Generic datasets for Multiple object tracking

2.4.1.1. MOT – Multiple Object Tracking

The MOT dataset [8] consists of 14 challenging video sequences (7 for training - 5316 frames, 7 for testing - 5919 frames). It focuses on tracking pedestrians. Version MOT17 (there is no paper for this dataset, only for the previous one) is provided with 3 sets of detections (DPM, Faster-RCNN and SDP) and a ground truth.

Figure 8. An overview of the MOT16 dataset. Top: Training sequences; bottom: test sequences

The latest version CVPR2019 contains 8 sequences (4 train, 4 test) in unconstrained environments.

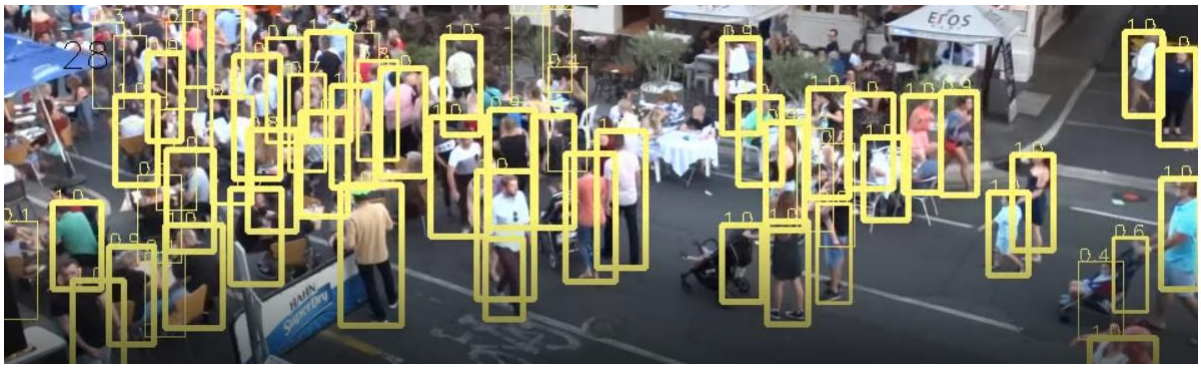


Figure 9. An example of the CVPR19 dataset.

A brief overview of the CVPR19 dataset is presented in the Table 1.

Name	FPS	Resolution	Length	Tracks	Boxes	Description
CVPR19-05	25	1920x1080	3315 (02:13)	1251	815068	Crowded square by night time.
CVPR19-03	25	1920x880	2405 (01:36)	754	414734	People leaving entrance of stadium by night time, elevated viewpoint.
CVPR19-02	25	1920x1080	2782 (01:51)	295	199752	Crowded indoor train station.
CVPR19-01	25	1920x1080	429 (00:17)	90	26219	Crowded indoor train station.

Table 1. CVPR19 dataset

2.4.1.2. KITTI-MOTS

This is the MOTS (Multiple Object Tracking and Segmentation) version of the famous KITTI dataset [9], captured from a moving vehicle in the city environment. The dataset combines bounding boxes for tracking with segmentation masks for the MOTS task. Currently the dataset only includes the training data, the test and evaluation are under construction.

Figure 10. An example of the KITTI-MOTS dataset.

The dataset contains 12 sequences and 5027 frames and considers both pedestrians and cars. The total number of tracks is 530. The total number of annotated segmentation masks is 65213. The resolution of video sequences is 1392 x 512 pixels. Annotations are provided in .txt and

.png formats. Those in .txt format include the object ID, class ID, and the mask can be decoded using *cocotools*.

2.4.1.3. CAVIAR

The CAVIAR project [10] (*Context Aware Vision using Imagebased Active Recognition*), from INRIA Labs, started in 2002 and ended in 2005. It was dedicated to the development of algorithms to richly describe and understand video scenes. In CAVIAR Test Case Scenario, two sets of data were provided. A first set was filmed in the entrance lobby of INRIA Labs (indoor), and the second one in a hallway in a shopping center in Lisbon (also indoor). For the first set, six different scenarios were considered (Walking, Browsing, Resting, slumping or fainting, Leaving bags behind, People/groups walking together and splitting up and Two people fighting), and for each scenario, a various number of sequences were recorded (from three to six, for a total of 28 video sequences). For the second set, two views (corridor view and front view) of twenty-two different scenes were given (hence providing forty-four sequences). Annotations contain object ID.

2.4.1.4. LTB35

Benchmark [11] for visual single object tracking. Consists of 50 HD videos from real world scenarios, encompassing a duration of over 400 minutes (676K frames).

2.4.2. VisDrone

VisDrone [12] is a large-scale video object detection and tracking dataset, including 79 video clips with approximate 1.5 million annotated bounding boxes in 33,366 frames. Some other useful annotations, such as object category, occlusion, and truncation ratios, are also provided for better data usage. The dataset is collected with several drones, in various scenarios, which are taken at different locations, but share similar environments and attributes.

The evaluation protocol used taken from E. Park, W. Liu, O. Russakovsky, J. Deng, F.-F. Li, and A. Berg, "Large Scale Visual Recognition Challenge 2017," [1] is used to evaluate the tracking performance. Specifically, each algorithm is required to output a list of bounding box with confidence scores and the corresponding identities. We sort the tracklets (formed by the bounding box detections with the same identity) according to the average confidence of their bounding box detections. A tracklet is considered correct if the intersection over union (IoU) overlap with ground truth tracklet is larger than a threshold. Similar to [1], three thresholds are used in evaluation, i.e., 0.25, 0.50, and 0.75. The performance of an algorithm is evaluated by

averaging the mean average precision (mAP) across object classes over different thresholds. The evaluation code for Task 4 is available on the VisDrone github.

Figure 11. An overview of the VisDrone dataset.

Number of snippets			
Dataset	Training	Validation	Test-challenge
Multiple object tracking	56 clips 24,201 frames	7 clips 2,819 frames	16 clips 6,333 frames

Table 1. VisDrone MOT challenge dataset 2019

2.4.2.1. Campus - Stanford Drone Dataset

This dataset [13] contains video sequences taken with drones in different campuses. It consists of eight unique scenes and includes objects such as pedestrians, bikes, skateboarders, cars, buses, and golf carts (19K targets in total). There are 929.5k frames in total. Target trajectories along with their target IDs are annotated. The videos are collected with drones and have a resolution of 1400 x 1904 pixels.

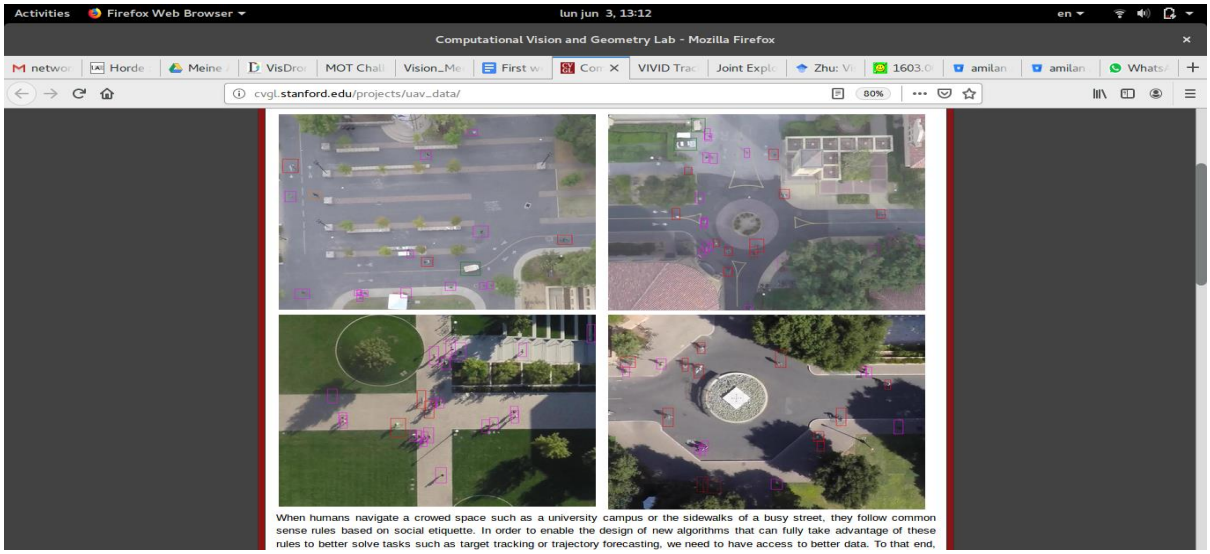


Figure 11. An overview of the Campus dataset..

Scenes	Videos	Bicyclist	Pedestrian	Skateboarder	Cart	Car	Bus
gates	9	51.94	43.36	2.55	0.29	1.08	0.78
little	4	56.04	42.46	0.67	0	0.17	0.67
nexus	12	4.22	64.02	0.60	0.40	29.51	1.25
coupa	4	18.89	80.61	0.17	0.17	0.17	0
bookstore	7	32.89	63.94	1.63	0.34	0.83	0.37
deathCircle	5	56.30	33.13	2.33	3.10	4.71	0.42
quad	4	12.50	87.50	0	0	0	0
hyang	15	27.68	70.01	1.29	0.43	0.50	0.09

2.4.2.2. UA-DETRAC

Object tracking benchmark [14] consisting of 60 videos for training and 40 for testing. It contains in total 140.1k frames and 4 object categories. The AVSS2019 Challenge consists of 10 hours of videos at 24 different locations at Beijing and Tianjin in China. The videos are recorded at 25 frames per seconds (fps), with resolution of 960 × 540 pixels. There are 8,250 vehicles that are manually annotated, leading to a total of 1.21 million labeled bounding boxes of objects.

Figure 12. An overview of the UA-DETRAC dataset.

2.4.2.3. UAVDT Benchmark

UAV benchmark [15] focusing on complex scenarios that include new challenges, e.g., high density, small object, and camera motion. Contains about 80k representative frames are fully annotated with bounding boxes as well as up to 14 kinds of attributes (e.g., weather condition, flying altitude, camera view, vehicle category, and occlusion) for three fundamental computer vision tasks: object detection, single object tracking, and multiple object tracking.

Figure 12. An overview of the VisDrone dataset.

2.4.2.4. UAV123

UAV123 [16] is made for single object tracking. It contains video captured from low-altitude UAVs. There are 123 aerial video sequences with a total of 110k frames.

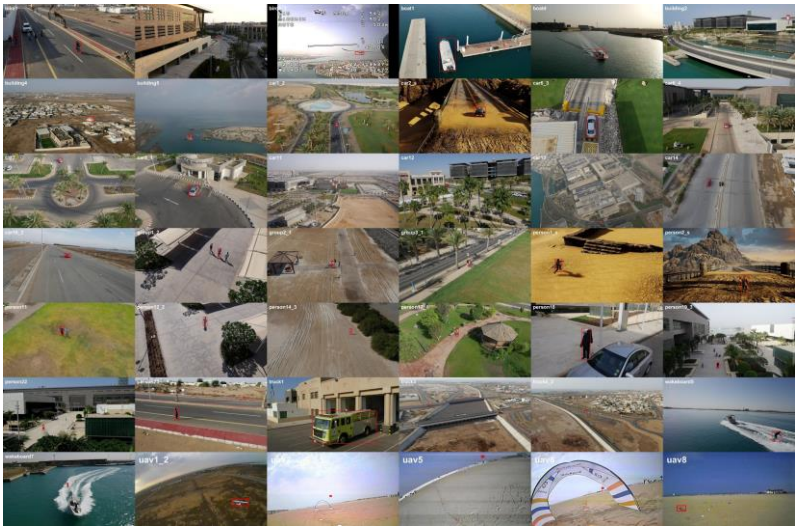


Figure 13. An overview of the UAV123 dataset.

2.4.2.5. VIVID

Contains [17] five RGB and three thermal videos for single target tracking. The targets are mostly vehicles. Moreover, videos present changes in scale, different levels of occlusion and discontinuities.

Figure 14. Sample frames of RGB and thermal sequences of the VIVID dataset.

2.4.2.6. LAMOD

Dataset [18] for moving object detection, which is an extension of VIVID and UAV123 datasets. Sequences are hand-annotated for moving objects in every frame. Ground-truth annotation includes the values for top left x, top left y, width and height, but do not include the object ID. In the first batch, annotations for 14 sequences are released.

2.4.2.7. Visual object tracking for UAVs

Dataset [19] for single object tracking of 70 sequences with high diversity captured by drone cameras. The sequences focus mostly on tracking people and cars. Other videos are added from YouTube. Videos cover different types of camera motion, including rotation and translation, also simple and complex backgrounds, and contain occlusions. Moreover, all bounding boxes are manually annotated in all video frames. Each row in the ground-truth files represents the bounding box of the target in that frame, (x, y, box-width, box-height).

Figure 15. Sample frames of the dataset

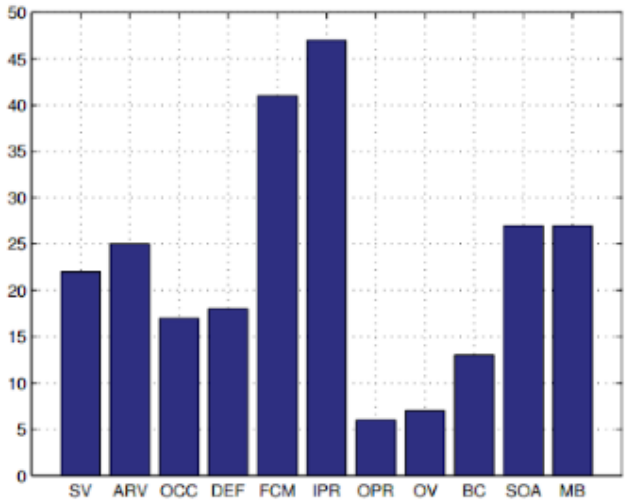


Figure 16.. Attribute distribution of the entire dataset. Each subset of sequences corresponds to one of the attributes, namely: scale variation (SV), aspect ratio variation (ARV), occlusion (OCC), deformation (DEF), fast camera motion (FCM), in-plane rotation (IPR), out-of-plane rotation (OPR), out- of-view (OV), background cluttered (BC), similar objects around (SOA), and motion blur (MB).

2.4.2.8. Okutama-action (multi-human tracking under development)

A dataset [20] for aerial-view human action detection. Contains 43 sequences with a resolution of 3840x2160 and an average duration of one minute with abrupt camera movement. Annotated with track ID, although it belongs to the same person only for 180 consecutive frames, then the person gets a new ID for the next 180 frames.

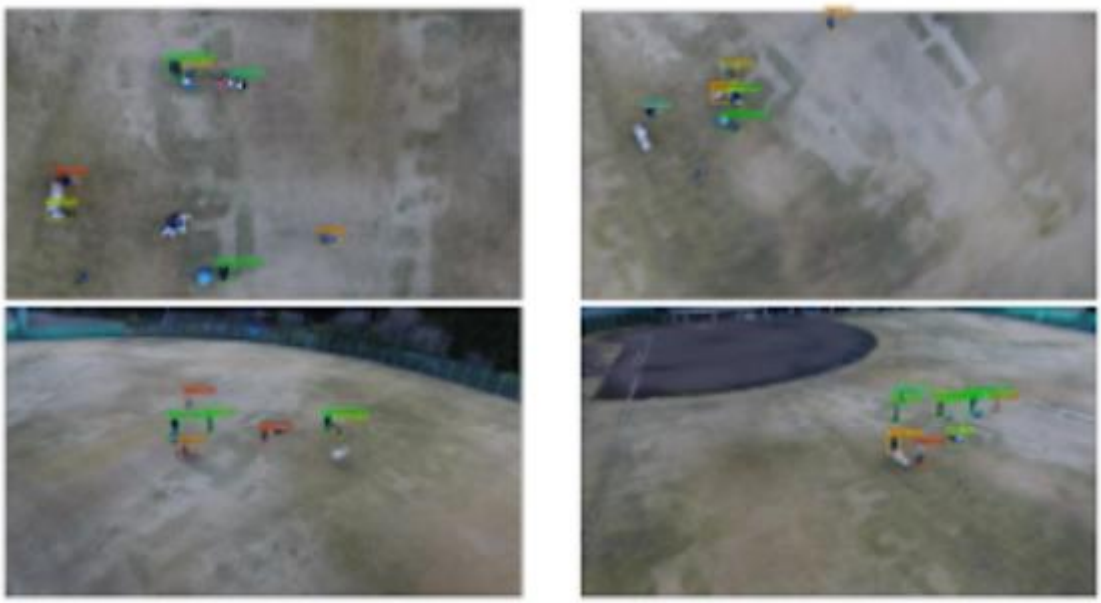


Figure 17. An overview of the Okutama-action dataset.

2.4.2.9. MultiDrone Public Dataset

The public MultiDrone Dataset has been assembled using both pre-existing audiovisual material and newly filmed UAV shots. A large subset of these data has been annotated for facilitating scientific research, in tasks such as visual detection and tracking of bicycles, football players, human crowds, etc. The following list gives an overview over the available data and respective details of the full dataset:

1. More than 10 hours of UAV footage depicting football, rowing and cycling (DW)
2. More than 115 GBs of professional aerial/UAV shots depicting bicycle races (RAI, DW, AUTH)

3. More than 2.5 GBs of general, professional and semi-professional UAV shots in varied environments (AUTH)
4. More than 32 GBs of aerial/UAV shots depicting human crowds (AUTH)
5. More than 2 GBs of aerial/UAV shots depicting boat races (AUTH, RAI)
6. More than 6 GBs of ground shots depicting a UAV flying (AUTH)
7. More than 2.5 GBs of UAV shots depicting buildings-of-interest (AUTH)
8. Additional UAV datasets for face de-identification, person detection, potential landing site detection (AUTH)
9. More than 25 minutes of simulation videos using a state-of-the-art real-time 3D graphics engine, used to characterise optimal drone parameters for specific scenarios and shot types in terms of viewing experience. (UoB)
10. Skeletal joints of all the detected people (including cyclists and spectators) corresponding to more than 2 hours of footage (IST)

The permission is needed in order to access the dataset. The resolution of sequences in the dataset varies from sequence to sequence from 768 x 432 to 4096 x 2160 pixels. The following categories of objects are presented for tracking: boats, bicycles, people, UAV's, faces.

Figure 18. An overview of the Multidrone dataset.

2.4.2.10. CARPK

Car parking lot dataset [21] for object counting. The images are collected with the drone-view at approximate 40 meters height. The image set is annotated by bounding box per car. All labeled bounding boxes have been well recorded with the top-left points and the bottom-right points. It is supporting object counting, object localizing, and further investigations with the annotation format in bounding boxes.

Figure 19. An overview of the CARPK dataset

In contrast to the PUCPR dataset, this dataset supports a counting task with bounding box annotations for all cars in a single scene. Most important of all, compared to other car datasets, CARPK is the only dataset in drone-based scenes and also has a large enough number in order to provide sufficient training samples for deep learning models.

The dataset does not provide object ID information. The annotation files only contain the positions of the bounding boxes.

```
1   843 629 946 710 1
2   825 555 929 639 1
3   836 499 927 569 1
4   817 430 926 510 1
5   820 372 914 448 1
6   617 324 721 367 1
7   624 376 735 426 1
8   627 435 737 484 1
9   623 483 736 530 1
10  630 527 743 586 1
11  633 585 754 641 1
```

Figure 20. CARPK annotations format

Dataset	Sensor	Multi Scenes	Resolution	Annotation Format	Car Numbers	Counting Support
OIRDS	satellite	✓	low	bounding box	180	✓
VEDAI	satellite	✓	low	bounding box	2,950	✓
COWC	aerial	✓	low	car center point	32,716	✓
PUCPR	camera	×	high	bounding box	192,216	×
CARPK	**drone**	✓	**high**	**bounding box**	**89,777**	✓

Table 2. Comparison of aerial view car-related datasets

2.4.2.11. Multi-Target Detection and Tracking from a Single Camera in Unmanned Aerial Vehicles (UAVs)

The dataset [22] focuses on tracking of UAV's in the video sequences taken from another UAV. The dataset comprises 50 video sequences of 70250 frames with 30 fps frame rate. They are recorded by a GoPro 3 camera (HD resolution: 1920x1080 or 1280x760) mounted on a custom delta-wing airframe. For each video, there are multiple target UAVs (up to 8) which have various appearances and shapes. The targets in the dataset are manually annotated in the videos by using VATIC software to generate ground-truth dataset for performance evaluation.

Figure 21. An overview of the Multi-target dataset.

3. VPU generated datasets

3.1. Scene Recognition

3.1.1. Places365 Lifelogging version

The task of scene recognition has been classically evaluated using still images representing scenes. In the context of the MobiNet video project, we have created a new dataset that extrapolates Places365's classes to lifelogging/egocentric videos. The dataset is made up of 450 videos recorded with smartphones, go-pro and handheld cameras. Videos have been obtained by downloading YouTube videos licensed as Creative Commons.

For each scene class in Places365, we include between one (90% of the classes) and four videos. The average length of the videos is 638 frames—around twenty-one seconds—, and the median length is 600 frames per video, —around twenty seconds. In overall, the dataset is approximately 34.1 GB large.

Examples of the dataset are depicted in **Figure 21**. The dataset is available at http://www-vpu.eps.uam.es/webvpu/en/recursos-publicos/datasets/.

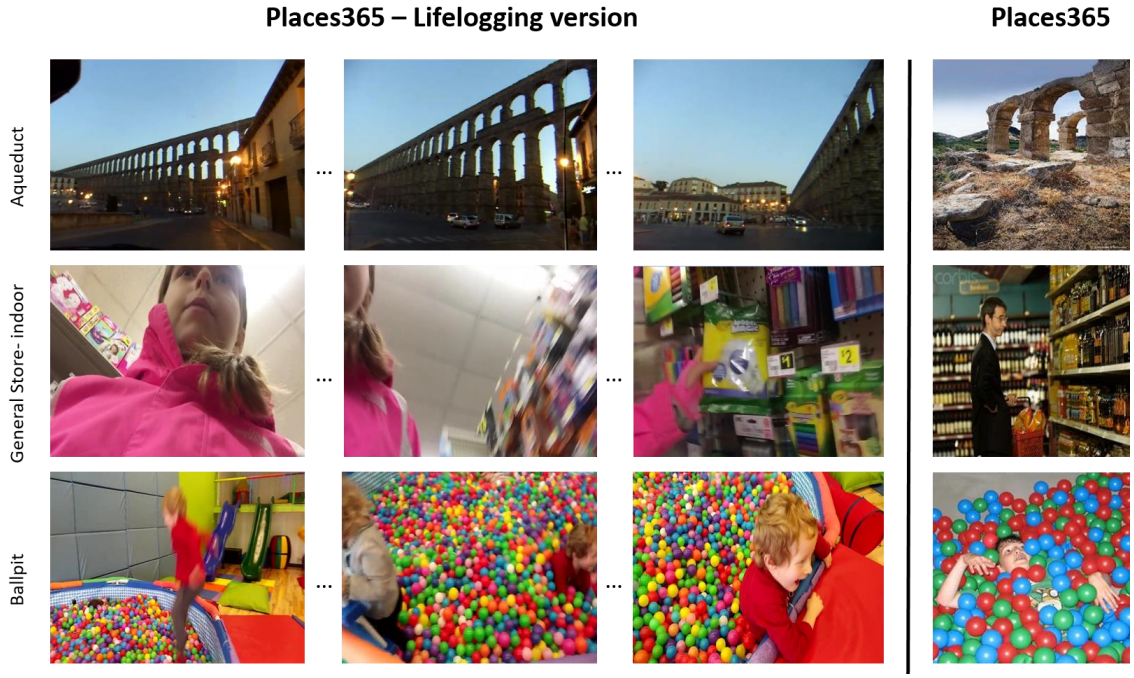


Figure 21. Places365- lifelogging version selected frames of three classes (right) with examples of corresponding classes in Places365 (left).

3.1.2. A unified dataset for semantic segmentation

Recently, several datasets for the training and evaluation of semantic segmentation have been proposed. Reference benchmarks provide annotations of large sets of data encompassing hundreds of semantic classes but usually using a scarcely and unbalanced class sampling (e.g. ADE20K, COCO Stuff). However, some of them provide dense and exhaustive annotations for a small subset of classes in specific scenarios (e.g. Cityscapes, Mapillary, Taskonomy). Together, they provide a diverse, rich and varied set of semantic classes, images and scenarios. Unfortunately, semantic classes are not aligned along datasets and different labels are used for identifying equivalent semantic classes (e.g. *people* / *person*).

In the context of the MobiNet Video project, we target the creation of a complete and diverse dataset for training semantic segmentation methods able to cope with the high variety and instability of lifelogging scenarios. To this aim, we propose to integrate top relevant semantic segmentation datasets into one: a unified dataset for semantic segmentation.

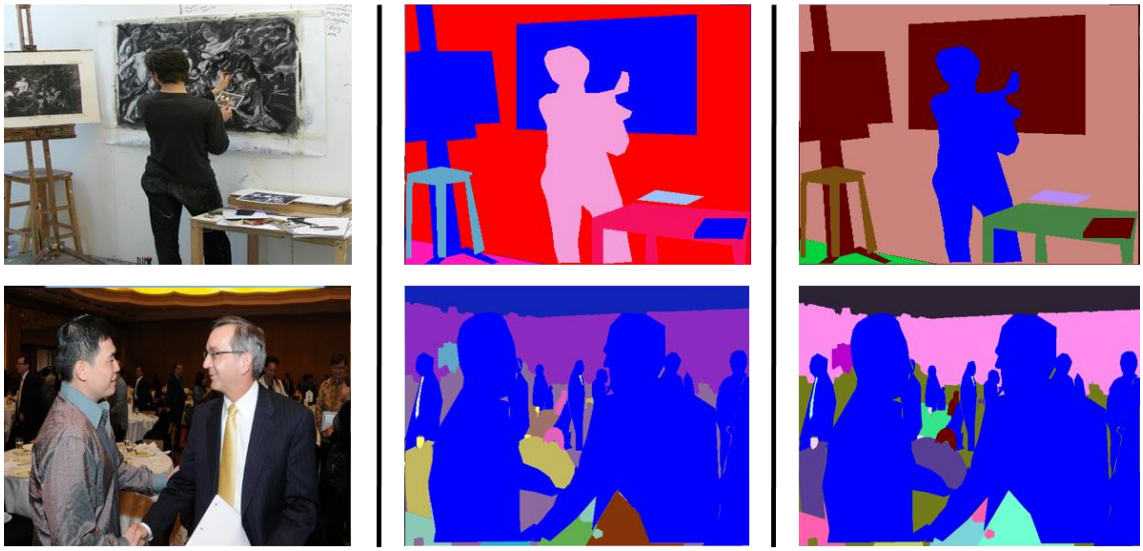


Figure 22. Two visual examples of a unified dataset for semantic segmentation. From left to right: colour image, semantic segmentation provided in the source dataset and relabelled semantic segmentation in the proposed dataset. Note how equivalent conceptual objects (*persons*/*people*) share the same label in the proposed unified dataset.

As a starting point, we start by merging analogue classes along datasets, identifying them with a single label (see **Figure 22**). This simple strategy allows us to obtain a larger dataset (see class distribution in **Figure 23**). However, there are situations which require further disambiguation as the existence of several subclasses of the same class (e.g. see wall and its

subclasses in **Figure 23**). Currently, we are working in designing automatic strategies to disambiguate these classes.

The current version of the dataset is available at http://www-vpu.eps.uam.es/webvpu/en/recursos-publicos/datasets/.

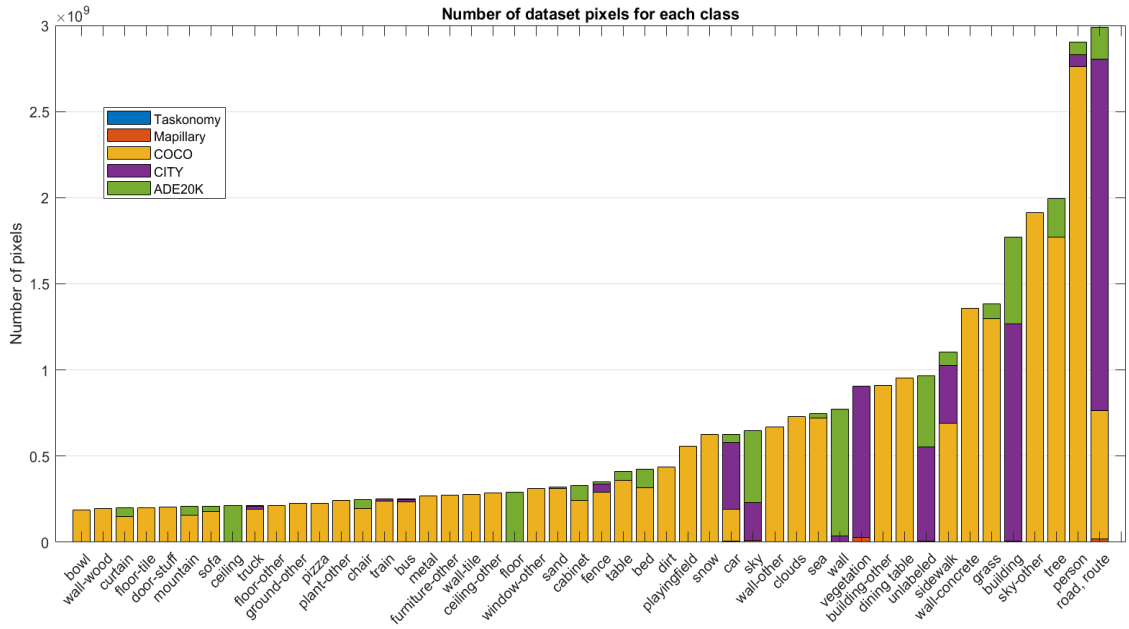


Figure 23. Pixel-wise distribution of semantic classes in the proposed dataset (selection). Contributions of source dataset are indicated in colours.

3.2. Google Street view dataset for lifelogging applications

Recent works [28] have demonstrated the potential of combining deep learning architectures with the availability of large urban image databases to develop innovative applications, based on computer vision for wearable cameras; in this case navigation in urban environments without the need of using GPS location coordinates. This example demonstrates the potential of making use of available large image databases, such as Google Street View, to enable the development of lifelogging applications.

The work in [29] develops a preliminary tool to extract images from the Google Street View database through predefined routes, using the Google Directions API [30] and Google Street View API [31]. This tool allows to define parameters in the extraction of the images such as the transport mode, horizontal and vertical angle (heading and pitch), or camera field of view (fov). Such tools allow to extract images that can be used to develop computer vision algorithms for lifelogging applications. Figure 24 shows examples of images extracted using the developed

tool, for different locations and scenarios. Figure 25 shows examples of images extracted from the same location but corresponding to different parameter values (in this case horizontal fov).

Figure 24 Examples of Google Street View images extracted from several different locations

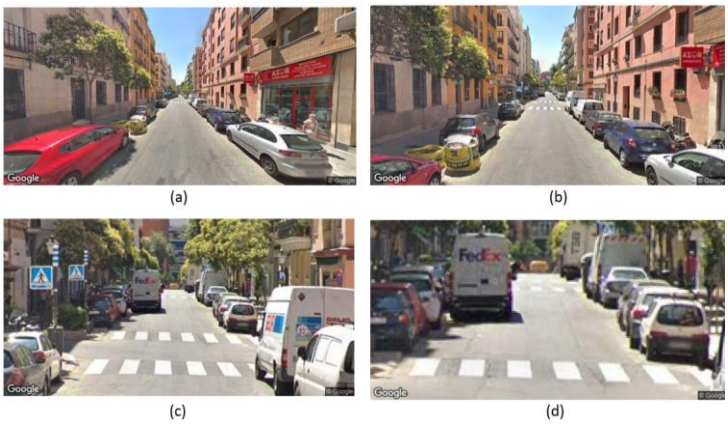


Figure 25 Examples of Google Street View images extracted in the same location with different values of horizontal fov: (a) 120º, (b) 80º (c) 40º (d) 20º.

Extending the work in [28], this preliminary tool for the creation of Google Street View datasets for lifelogging applications has been updated to generate structured datasets that collect the useful metadata that apply to the images of a given route. A summary of the structure and metadata of the database can be found in Table 3. During the development of the Mobinet video project, this tool will be used to generate urban image datasets, as required for the specific needs of each task.

File System Structure:

Every route (origin and destination) generates a route folder with name ' **A,B_Y,Z**', where:

- A and B are the latitude and longitude of the origin respectively; and
- Y and Z are the latitude and longitude of the destination respectively.

Each route folder can contain different versions of the route (different parameters). Each version is contained in a folder with name '**H={HD}_P={P}_FOV={F}_M={M}_S={W}x{H}-jpegs**' where:

- **{HD}**: is the heading value
- **{P}**: is the pitch value
- **{F}**: is the field of view value
- **{M}**: is the mode of the direction
- **{W}**: is the width of the images
- **{H}**: is the height of the images

Metadata:

Additionally, the metadata of each route version is saved in a JSON file with a root *JSON Object* stores all parameters that are common to the route/single-location as *key-value pairs*. It also contains a *JSON Array*, with key *'images'* that contains as many *JSON Objects* as the images/waypoints. Each image *JSON Object* stores the value of the parameters that are specific to each image/waypoint:

Possible parameters of the root *JSON Object* and *JSON Array* are:

- *from:* string with origin latitude and longitude, concatenated by a comma.
- *to:* string with destination latitude and longitude, concatenated by comma.
- *width:* integer value of the width of the images.
- *height:* integer value of the height of the images.
- *fpx:* real value of the frames per meter or frames per second.
- *heading:* integer value of heading used on the query of the Google Street View API.
- *pitch:* integer value of the pitch used on the query of the Google Street View API.
- *fov:* field of view integer value used on the query of the Google Street View API.
- *seqNumber:* integer value of from 00000 to 99999 that identies the waypoint/image.
- *lat:* real value of the latitude of the location.
- *lng:* real value of the longitude of the location.

Table 3 Structure and metadata of the data obtained using the tool to extract images from the Google Street View database.

4. Evaluation methodologies

4.1. Object detection

There are two common prerequisites for quantifying the performance of a detector. One is to determine for each hypothesized output, whether it is a true positive (TP) that describes an actual (annotated) object, or whether the output is a false alarm (or false positive, FP). This decision is typically made by thresholding based on a defined distance (or dissimilarity) measured. An object that is missed by any hypothesis is a false negative (FN). A good result is expected to have as few FPs and FNs as possible. Next to the absolute numbers, we also show the false positive ratio measured by the number of false alarms per frame (FAF), sometimes also referred to as false positives per image (FPPI) in the object detection literature [1].

In the most general case, the relationship between ground truth objects and a tracker output is established using bounding boxes on the image plane [5]. The intersection over union (a.k.a. the Jaccard index) is usually employed as the similarity criterion, while the threshold td is set to 0.5 or 50%.

Obviously, it may happen that the same object is covered by multiple outputs. The second prerequisite before computing the numbers is then to establish the correspondence between all annotated and hypothesized objects under the constraint that a true object should be recovered at most once, and that one hypothesis cannot account for more than one object. By measuring the intersection over union of bounding boxes and matching those from ground truth annotations and results, measures of recall and precision can be computed, obtaining the Precision-Recall curves.

4.2. Object re-identification

For the evaluation of image based ReID, the results are usually represented by a matrix mapping each query to the test images ranked by distance. Following [6], two metrics are used to evaluate the accuracy of algorithms: mean Average Precision (mAP), which measures the mean of all queries' average precision (the area under the Precision Recall curve), and the rank-K hit rate, denoting the possibility that at least one true positive is ranked within the top K positions. In our experiments, the top 1, 5, 10 and 20 and the mAP measured by the top 100 matches for each query is adopted for comparison.

4.3. Scene Recognition

Scene recognition benchmarks are generally evaluated via the *Top@k* accuracy metric with $1 \leq k \leq K$ and K being the number of classes. The *Top@1* accuracy measures the percentage of validation/testing images whose top-scored class coincides with the ground-truth label. Generally, *Top@k* accuracy, represents the percentage of validation/testing images whose ground-truth label corresponds to any of the k top-scored classes.

The *Top@k* accuracy metrics are biased to classes over-represented in the validation set; or, in other words, under-represented classes barely affect these metrics. In order to cope with unbalanced validation sets, we propose to use an additional performance metric, the Mean Class Accuracy (MCA):

$$\text{MCA} = \frac{\sum_{i=1}^{K} Top_i@1}{K}$$

, where the value inside the summation is the *Top@1* metric for scene class i. Note that MCA equals *Top@1* for perfectly balanced datasets.

4.4. Multiple object tracking

Unlike traditional single-target tracking, multi-target tracking is normally evaluated with more metrics. These metrics can be grouped into those having into account individual trajectories, and those which do not. The following metrics contain trajectory information:

1. MOTA - Multiple Object Tracking Accuracy. This measure combines three error sources: false positives, missed targets and identity switches in the following manner.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}$$

where t is the frame index and GT is the number of ground truth objects.

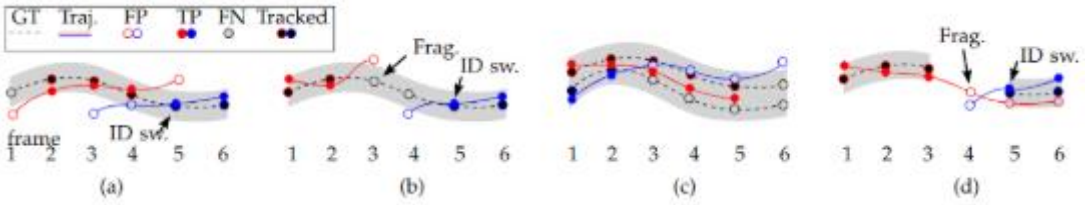
2. IDF1 - The ratio of correctly identified detections over the average number of ground-truth and computed detections.

3. MT - Mostly tracked targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span.

4. ML - Mostly lost targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span.

5. ID switches - An ID switch is when a track j is matched to a ground truth target i, and at some point a different track k is assigned to the target i. A relative value is provided as: #ID switches/Recall. This can be seen in the figure, when a ground truth trajectory (black) is estimated first by an ID represented in red, and then switches to another ID represented by blue.

6. Fragmentation - Number of track fragmentations. It counts the number of times a ground truth trajectory is interrupted, that is, when a trajectory changes status from tracked to untracked, and tracking is resumed at a later point. A relative value is also provided as: #fragmentations/Recall. A fragmentation can be seen in figures (b) and (c), when the ground truth target (black) stops being estimated by the tracker (red or blue).



Moreover, the metrics below do not have into account individual trajectories:

7. MOTP - Multiple Object Tracking Precision. The misalignment between the annotated and the predicted bounding boxes.

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_{t} c_t}$$

where ct denotes the number of matches in frame t and dt,i is the bounding box overlap of target i with its assigned ground truth object. If computed in 3D, the definition changes slightly to:

$$MOTP_{3D} = 1 - \frac{\sum_{t,i} d_{t,i}}{t_d \bullet \sum_t c_t}$$

FP - The total number of false positives.

FN - The total number of false negatives.

Hz - Processing speed (in frames per second excluding the detector) on the benchmark.

Better performance of a tracking algorithm corresponds to the higher values of MOTA, MOTP, IDF1, MT and Hz metrics, and the lower values of ML, FP, FN, IDSW and Fragmentation metrics.

None of the aforementioned metrics can fully reflect the quality of a given detector, but MOTA seems to best correlate with the average ranking across all 10 metrics.

Some benchmarks provide their own detections in the datasets. Therefore, when the results of some tracker are submitted for the challenge, the authors indicate whether they used a publically available detections or not. For example, the MOT Challenge has a special field 'Detector' which can be either public or private.

5. Conclusions

This deliverable complies the different datasets (sequences and associated ground-truth), both selected from the state of the art:

- Object detection: MOT2017Det
- Object re-identification: VIPeR, Market150, CityFlow-ReID
- Scene Recognition: Places365
- Multiple Object tracking: MOT – Multiple Object Tracking, KITTI-MOTS, CAVIAR, LTB35 11, VisDrone challenge datasers

and generated within the project:

- Scene Recognition: Places365 Lifelogging version, A unified dataset for semantic segmentation, Google Street View Dataset for lifelogging applications

Also the different evaluation methodologies to be used within the project are described.

References

[1] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, Konrad Schindler, MOT16: A Benchmark for Multi-Object Tracking, proceedings of the IEEE Computer Vision and Pattern Recognition conference, 2016.

[2] D. Gray and H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, proceedings of the IEEE European Conference on Computer Vision, 2008.

[3] L. Zheng et al, Scalable person re-identification: A benchmark, proceedings of the IEEE International Conference on Computer Vision, 2015.

[4] Zheng Tang et al, CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, proceedings of the IEEE Computer Vision and Pattern Recognition conference, 2019.

[5] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, (http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html).

[6] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark, proceedings of the IEEE International Conference on Computer Vision, 2015.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, no. 99, pp. 1–1, 2017.

[8] Milan, A., Leal-Taixé, L., Reid, I., Roth, S. & Schindler, K. MOT16: A Benchmark for Multi-Object Tracking. arXiv:1603.00831 [cs], 2016., (arXiv: 1603.00831).

[9] Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., & Leibe, B. (2019). MOTS: Multi-Object Tracking and Segmentation. arXiv preprint arXiv:1902.03604.
https://arxiv.org/pdf/1902.03604.pdf

[10] Fisher, R., Santos-Victor, J., & Crowley, J. (2005). CAVIAR: Context aware vision using image-based active recognition. 2011-11-01].
http://homepages.inf.ed.ae.uk/rbf/CAVIAR/eaviar.htm

[11] Moudgil, A., & Gandhi, V. (2018, December). Long-term visual object tracking benchmark. In Asian Conference on Computer Vision (pp. 629-645). Springer, Cham.
https://link.springer.com/content/pdf/10.1007%2F978-3-030-20890-5.pdf

[12] Zhu, P., Wen, L., Bian, X., Ling, H., & Hu, Q. (2018). Vision meets drones: a challenge. arXiv preprint arXiv:1804.07437.

[13] Robicquet, A., Sadeghian, A., Alahi, A., & Savarese, S. (2016, October). Learning social etiquette: Human trajectory understanding in crowded scenes. In European conference on computer vision (pp. 549-565). Springer, Cham.
https://link.springer.com/chapter/10.1007/978-3-319-46484-8_33.

[14] Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M. C., Qi, H., ... & Lyu, S. (2015). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. arXiv preprint arXiv:1511.04136.
https://arxiv.org/abs/1511.04136.

[15] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., ... & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 370-386). https://arxiv.org/abs/1511.04136.

[16] Mueller, M., Smith, N., & Ghanem, B. (2016, October). A benchmark and simulator for uav tracking. In European conference on computer vision (pp. 445-461). Springer, Cham. https://ivul.kaust.edu.sa/Pages/pub-benchmark-simulator-uav.aspx.

[17] Collins, R., Zhou, X., & Teh, S. K. (2005, January). An open source tracking testbed and evaluation web site. In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (Vol. 2, No. 6, p. 35) https://www.ri.cmu.edu/publications/an-open-source-tracking-testbed-and-evaluation-web-site/

[18] Berker Logoglu, K., Lezki, H., Kerim Yucel, M., Ozturk, A., Kucukkomurler, A., Karagoz, B., ... & Erdem, A. (2017). Feature-based efficient moving object detection for low-altitude aerial platforms. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2119-2128). https://www.ri.cmu.edu/publications/an-open-source-tracking-testbed-and-evaluation-web-site/

[19] Li, S., & Yeung, D. Y. (2017, February). Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In Thirty-First AAAI Conference on Artificial Intelligence. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14338/14292

[20] Barekatain, M., Martí, M., Shih, H. F., Murray, S., Nakayama, K., Matsuo, Y., & Prendinger, H. (2017). Okutama-action: an aerial view video dataset for concurrent human action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 28-35). https://www.ri.cmu.edu/publications/an-open-source-tracking-testbed-and-evaluation-web-site/

[21] Hsieh, M. R., Lin, Y. L., & Hsu, W. H. (2017). Drone-based object counting by spatially regularized regional proposal network. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4145-4153). https://www.ri.cmu.edu/publications/an-open-source-tracking-testbed-and-evaluation-web-site/

[22] Li, J., Ye, D. H., Chung, T., Kolsch, M., Wachs, J., & Bouman, C. (2016, October). Multi-target detection and tracking from a single camera in Unmanned Aerial Vehicles (UAVs). In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(pp. 4992-4997). IEEE. https://engineering.purdue.edu/~bouman/UAV_Dataset/pubs/IROS-2016.pdf

[23] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision, 127(3), 302-321. 2019.

[24] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223). 2016.

[25] Caesar, H., Uijlings, J., & Ferrari, V. Coco-stuff: Thing and stuff classes in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1209-1218). 2018.

[26] Neuhold, G., Ollmann, T., Rota Bulo, S., & Kontschieder, P. The mapillary vistas dataset for semantic understanding of street scenes. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4990-4999). 2017.

[27] Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3712-3722). 2018.

[28] P. Mirowski, M. K. Grimes, M. Malinowski, K. M. Hermann,K. Anderson, D. Teplyashin, K. Simonyan, K. Kavukcuoglu,A. Zisserman, and R. Hadsell. "Learning to navigate in cities without a map" in Proc. NIPS 2018.

[29] Detección de objetos en imágenes urbanas de Google Street View (Object detection in Google Street Vieew urban images), Paula Guerra Toni, (advisor: Pablo Carballeira López), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2019.

[30] Google Street View Static API: (accesed Jul. 2019) https://developers.google.com/maps/documentation/streetview/intro

[31] Google Directions API: (accesed Jul. 2019) https://developers.google.com/maps/documentation/directions/intro