
Dataset Supplementary Material

Ivan Nikolov Mark P. Philipsen Jinsong Liu Jacob V. Dueholm
Anders S. Johansen Kamal Nasrollahi Thomas B. Moeslund

Visual Analysis and Perception Lab

Aalborg University, Denmark

{iani, mpph, jili, jvdu, asjo, kn, tbm}@create.aau.dk

1 Datasheets for Datasets

1.1 Motivation

For what purpose was the dataset created? The Long-term Thermal Drift (LTD) dataset was created to provide a large dataset for studying the effects of concept drift on deep learning models, using the naturally occurring seasonal and day and night drifts. The dataset contains changing conditions from winter to summer, together with weather phenomena like rain, snow, fog, bright sunshine, etc.

Who created this dataset (e.g., which team, research group) and behalf of which entity(e.g., company, institution, organization)? The dataset is part of the work of the Visual Analysis and Perception (VAP) Lab, part of Aalborg University, Denmark. The dataset was created by Ivan Nikolov, Mark Philipsen, Jinsong Liu, Anders Johansen and Jacob Dueholm, under the supervision of Thomas Moeslund and Kamal Nasrollahi.

Who funded the creation of the dataset? The dataset was created as a part of the project "Preventing drowning accidents", with a funding body of TrygFonden <https://www.respektforvand.dk/english>

1.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The dataset comprises of two parts. First, a large number of 2-minute long video clips, captured at equal intervals every 30 minutes for an 8 month period. Second, a table of metadata for each clip, containing folder information, timestamps and weather data.

How many instances are there in total (of each type, if appropriate)? The dataset contains 8,940 2-minute clips or around 298 hours of video. The metadata contains information for each clip instance.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The dataset contains equally spaced 2-minute clips from a period of 8 month. The initial raw data contained holes and not all months were fully covered from beginning to end. The captured data was equally sampled so a full coverage was obtained, while keeping the overall size manageable.

What data does each instance consist of? Each instance consists of a 2-minute clip in a .mp4 format of black and white thermal data with a resolution of 288×384 . Each instance also has a representative row from the metadata containing:

- Folder structure consisting of - main date folder and video clip name with a number and time

33 • A full timestamp
34 • Weather information at the time of the clip - temperature, humidity, precipitation, dew point,
35 wind direction, wind speed, sun radiation, minutes of sunshine

36 **Is there a label or target associated with each instance?** Each video is associated with a timestamp
37 and weather data. In addition, a small subset of images are extracted from the video data from the
38 months of February, April and August and the pedestrians in these images are annotated with bounding
39 boxes.

40 **Is any information missing from individual instances?** Nothing is missing from individual in-
41 stances.

42 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social
43 network links)?** The only relationships between the video clips are the timestamps showing which
44 clip comes after which at which date and time. These are made explicit in the metadata file.

45 **Are there recommended data splits (e.g., training, development/validation, testing)?** The
46 dataset comes with scripts for direct video selection through the use of the metadata, as well as
47 the separation of the clips into frames for easier computer vision analysis. As the video dataset is
48 directly connected to the metadata, the metadata is used for querying and selecting parts of the dataset.
49 Depending on the requirements of the user parts of the dataset can be divided for training, validation
50 or testing. This selection can be done on the basis of the timestamps, as well as more coarse grained
51 selections made from them. Such selections can be - based on day of the week, weekday or weekend,
52 day or night, working hours or after work hours, etc. The selection of subsets for training, testing,
53 validation can also be done based on the weather conditions. Researchers interested only in specific
54 weather conditions, or in the change of model performance based on weather phenomena can opt to
55 use data only from specific weather, depending on for example temperature, humidity, wind speed or
56 sun radiation. Finally, splitting the data based on a season is also possible - training on winter months
57 in Denmark like January, February, March and testing on summer months like July and August. In its
58 current state the dataset can be a useful tool for unsupervised learning algorithms, with additional
59 video and frame level annotations being added for more fine grained supervised learning use-cases.

60 **Are there any errors, sources of noise, or redundancies in the dataset?** Possible errors and
61 sources of noise can be found in some of parts of the video clips, as the used camera could contain
62 frame-level noise artifacts. In addition, due to the varied nature of the weather conditions the data
63 was captured, some noise and sudden brightness and contract changes are possible.

64 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,
65 websites, tweets, other datasets)?** The data is self-contained, as all additional information has been
66 gathered before-hand.

67 **Does the dataset contain data that might be considered confidential (e.g., data that is protected
68 by legal privilege or by doctorpatient confidentiality, data that includes the content of individ-
69 uals non-public communications)?** No. All data is captured in a open public space and by the
70 nature of thermal data preserves more the anonymity of the captured people.

71 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening,
72 or might otherwise cause anxiety?** No. The dataset consists of everyday life occurrences of people
73 walking, running, driving, talking, etc.

74 **Does the dataset relate to people?** Yes. The dataset contains many instances of people.

75 **Does the dataset identify any subpopulations (e.g., by age, gender)?** The dataset contains people
76 of different ages and genders, but those are not identified in any particular way. The fact the data is
77 thermal, makes this identification especially hard.

78 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indi-
79 rectly (i.e., in combination with other data) from the dataset?** It is possible to identify individuals
80 based on the limited features that can be gathered from the thermal data, as well as from their habits,
81 work commute and positions in time. As the dataset contains a long-term video surveillance from a
82 single location and additional timestamps in the metadata, a person can be potentially identified and
83 tracked for a period of time every time they pass through the field of view of the camera.

84 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that re-
85 veals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union**

Table 1: An overview of the dataset statistics for the thermal videos

Property	Value
Dataset release year	2021
Number of video clips	8940
Length of the dataset	298 hours
Resolution	288×384
Clip format	mp4
Frame rate	25 fps
Encoding	x224
Bit depth	8 bits

Table 2: An overview of the dataset weather metadata statistics

Features	Min	Max
Temperature [$^{\circ}C$]	-13.0	29.1
Humidity [%]	22	100
Precipitation [kg/m^2]	0	4.4
Dew Point [$^{\circ}C$]	-14.5	19.8
Wind Direction [<i>degrees</i>]	0	360
Wind Speed [m/s]	0	11.6
Sun Radiation [W/m^2]	0	1084
Minutes of Sunshine [<i>min</i>]	0	10

memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? Even with the limited identifying features available from the thermal images, from a long distance and a limited resolution, the gender, ethnicity of people passing through the field-of-view can be possibly identified and tracked for longer periods of time. No confidential information, is contained in the dataset.

In Table 1 the statistics of the thermal images part of the dataset are given, while Table 2 presents the weather metadata statistics. Figure 1 shows examples of people captured as part of the dataset, in either day or night, from different months. Variation of clothing and heat signatures are evident.

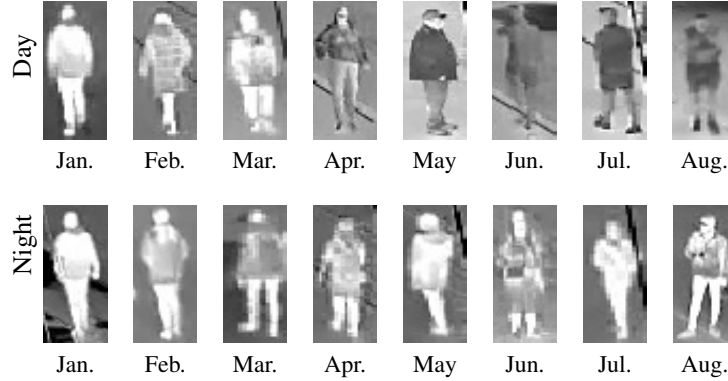


Figure 1: Example captured people at day and night for each of the months present in the dataset.

1.3 Collection Process

How was the data associated with each instance acquired? The dataset video clip timestamps were captured directly from the raw data stream. The weather data was associated to each clip using these timestamps, using the open source Danish Meteorological Institute (DMI) weather API.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? The raw data stream was captured using a Hikvision DS-2TD2235D-25/50 long wavelength infrared (LWIR) thermal camera, capturing wavelengths between 8 and 14 μm . The data stream was then automatically separated into days and sampled using the provided timestamps. The weather data was queried from the DMI weather API. The provided weather data was with a refresh rate of 10 minutes.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The dataset was captured from a data stream of raw data. The sampling was done deterministically, by first separating the raw data into days based on the timestamps. From these daily data batches, a sampling is done of a 2-minute clip every 30

108 minutes. As the raw dataset contains discontinuities, these holes are also present between the sampled
109 clips. This missing data varies between hourly holes and fully missing days.

110 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors)**
111 **and how were they compensated (e.g., how much were crowdworkers paid)?** Only researchers
112 from Aalborg University were involved in capturing, processing, sampling and analysing the data
113 from dataset.

114 **Over what timeframe was the data collected?** The data was gathered in the period of 2020 and
115 2021. The months of May, June, July and August were captured in 2020, while January, February,
116 March and April were captured from 2021. The provided timestamps match with these periods.

117 **Were any ethical review processes conducted (e.g., by an institutional review board)?** No.

118 **Does the dataset relate to people?** Yes. People are present in many of the clips in different activities
119 - walking, running, driving, talking, sitting, etc.

120 **Did you collect the data from the individuals in question directly, or obtain it via third parties**
121 **or other sources (e.g., websites)?** The data was gathered from the harbor surveillance system present
122 in Aalborg, Denmark.

123 **Were the individuals in question notified about the data collection?** No, as the data was captured
124 from a public space, over a long time period, making notifying people impossible.

125 **Did the individuals in question consent to the collection and use of their data?** No, as the
126 individuals were not necessary aware of being part of the observation.

127 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke**
128 **their consent in the future or for certain uses?** No, as the data was gathered from a public space
129 and filmed individuals were not aware.

130 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data**
131 **protection impact analysis) been conducted?** No.

132 **1.4 Preprocessing/cleaning/labeling**

133 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing,**
134 **tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing**
135 **of missing values)?**

136 The data gathering pipeline contains the following steps:

- 137 • The raw video stream was gathered directly from the harbor surveillance server. The server
138 provided video files of varying sizes, encompassing single or multiple days. Each file was
139 timestamped with the starting date and time.
- 140 • The raw files were separated into smaller files encompassing a single day, using the starting
141 timestamps and the length of the videos
- 142 • The raw daily files were further sub-sampled into 2-minute long clips, from every 30 minutes
143 of raw footage. Timestamps for each clip were calculated from the daily timestamps
- 144 • The videos were recoded into black and white format to lower the storage requirements
- 145 • Using the timestamps for each clip and the known geographical location of the camera, the
146 weather data was gathered into a metadata file with a row for each clip
- 147 • As part of the research done using the dataset, a small subset of frames were extracted
148 from clips from the months January, February, April and August. In total 700 frames were
149 extracted and subsequently annotated with bounding boxes around people using the open
150 source program Labellmg.

151 **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support**
152 **unanticipated future uses)?** The raw data was saved and will be provided in a second repository at
153 a later date.

154 **Is the software used to preprocess/clean/label the instances available?** The script pipeline for
155 processing the raw data is available on github. For additional dependencies and libraries, check the
156 provided repository.

157 1.5 Uses

158 **Has the dataset been used for any tasks already?** As of the time of publication, one paper has
159 subsets of the dataset Liu et al. [2021].

160 **Is there a repository that links to any or all papers or systems that use the dataset?** There is no
161 repository for papers using the dataset.

162 **What (other) tasks could the dataset be used for?** The dataset can be used for training and
163 verification of models on data from different seasons. As cars, bicycles, motorcycles and other
164 vehicles are present in the dataset, it can be used for detecting and tracking vehicle movement. The
165 dataset can be used for tracking and re-identification of pedestrians and groups of people. The dataset
166 contains both single people, as well as large crowds of people, which can be used for training models
167 in crowded scenarios. Finally, the dataset can be used for anomaly detection and identification of
168 crowds, vehicles, animals, among other.

169 **Is there anything about the composition of the dataset or the way it was collected and prepro-**
170 **cessed/cleaned/labeled that might impact future uses?** The dataset does not provide any large
171 risks for future users, as it is hard to identify a person from the thermal video and the camera does
172 not focus on one particular person, but groups of people.

173 **Are there tasks for which the dataset should not be used?** The dataset should not be used for
174 verifying people’s identities or tracking habits and patterns of individuals.

175 1.6 Distribution

176 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution,**
177 **organization) on behalf of which the dataset was created?** The dataset will be publicly available
178 to researchers.

179 **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** The processed
180 dataset will be available on Kaggle in the link and on a second repository at a later date, together with
181 the raw videos. The code for processing and cleaning the dataset is available on Github, together
182 with the code for training and testing deep learning models and analysing their results, provided here.

183 **When will the dataset be distributed?** The dataset is available online from August, 2021.

184 **Will the dataset be distributed under a copyright or other intellectual property (IP) license,**
185 **and/or under applicable terms of use (ToU)?** The dataset will not be under any prohibitive license,
186 but there will be a request to cite the paper the dataset is introduced in.

187 **Have any third parties imposed IP-based or other restrictions on the data associated with the**
188 **instances?** There are no restrictions on access to the dataset.

189 **Do any export controls or other regulatory restrictions apply to the dataset or to individual**
190 **instances?** No.

191 1.7 Maintenance

192 **Who will be supporting/hosting/maintaining the dataset?** The dataset will be hosted on Kaggle,
193 with continues necessary support provided as needed.

194 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** The
195 owner of the dataset will be the Visual Analysis and Perception (VAP) lab. As such all questions and
196 comments should be addressed through their website.

197 **Is there an erratum?** All changes and additional data will be announced through the VAP lab
198 website and through social media.

199 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete in-**
200 **stances)?** Yes. If required new improved versions of the dataset will be provided.

201 **If the dataset relates to people, are there applicable limits on the retention of the data associ-**
202 **ated with the instances (e.g., were individuals in question told that their data would be retained**
203 **for a fixed period of time and then deleted)?** No.

204 **Will older versions of the dataset continue to be supported/hosted/maintained?** Datasets on
205 Kaggle have a versioning history, which will be maintained until otherwise specified in news on the
206 VAP lab website.

207 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for**
208 **them to do so?** For getting in contact and discussions on collaboration and extending the dataset,
209 please get in touch with the VAP lab at the specified website

210 **References**

211 Jinsong Liu, Mark Philip Philipsen, and Thomas B Moeslund. Supervised versus self-supervised
212 assistant for surveillance of harbor fronts. In *16th International Conference on Computer Vision*
213 *Theory and Applications (VISAPP)*, 2021.