data

MDPI

Review

A Systematic Survey of ML Datasets for Prime CV Research Areas—Media and Metadata

Helder F. Castro [1,*], Jaime S. Cardoso [1,2] and Maria T. Andrade [1,2]

[1] INESC TEC, Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal; jaime.cardoso@inesctec.pt (J.S.C.); maria.t.andrade@inesctec.pt (M.T.A.)
[2] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
* Correspondence: hcastro98@gmail.com

Abstract: The ever-growing capabilities of computers have enabled pursuing Computer Vision through Machine Learning (i.e., MLCV). ML tools require large amounts of information to learn from (ML datasets). These are costly to produce but have received reduced attention regarding standardization. This prevents the cooperative production and exploitation of these resources, impedes countless synergies, and hinders ML research. No global view exists of the MLCV dataset tissue. Acquiring it is fundamental to enable standardization. We provide an extensive survey of the evolution and current state of MLCV datasets (1994 to 2019) for a set of specific CV areas as well as a quantitative and qualitative analysis of the results. Data were gathered from online scientific databases (e.g., Google Scholar, CiteSeerX). We reveal the heterogeneous plethora that comprises the MLCV dataset tissue; their continuous growth in volume and complexity; the specificities of the evolution of their media and metadata components regarding a range of aspects; and that MLCV progress requires the construction of a global standardized (structuring, manipulating, and sharing) MLCV "library". Accordingly, we formulate a novel interpretation of this dataset collective as a global tissue of synthetic cognitive visual memories and define the immediately necessary steps to advance its standardization and integration.

Keywords: dataset; metadata; media; computer vision; machine learning; integration

Citation: Castro, H.F.; Cardoso, J.S.; Andrade, M.T. A Systematic Survey of ML Datasets for Prime CV Research Areas—Media and Metadata. *Data* **2021**, *6*, 12. https://doi.org/10.3390/data6020012

Academic Editor: Joaquín Torres-Sospedra
Received: 22 December 2020
Accepted: 10 January 2021
Published: 22 January 2021

1. Introduction

1.1. Rationale

Our continuous desire for more capable technology, in all fields of action, has led to the ongoing tentative development of Artificial Intelligence (AI). The original pursued strategy to develop AI was aimed at the construction of fully formed "synthetic minds" that would understand, and operate upon, the world based on a pre-established (by humans) set of logical rules. One of the reasons for this was the limited amount of processing power of the initial information processing technology.

The computational power at our disposal has grown incessantly and made a new strategy for the development of AI possible. This strategy, currently being pursued, more closely resembles the way natural cognitions are "built", i.e., through learning, and it is designated as Machine Learning (ML).

Employing this strategy, the goal is the development of tools that can build their "model of the world" on their own by learning from outside informational resources. These resources consist of vast amounts of targeted information (targeted for interpretation or data extraction) and its associated "true" interpretation or "ground-truth" (the data to be extracted). These resources are typically grouped into informational packages called ML datasets, which therefore play the role in ML that "books" play in the process of human learning.

ML-based technological developments have been remarkable, particularly those attained in Computer Vision (ML-based Computer Vision or MLCV), perhaps also because

of the importance that the equivalent sense has for our daily lives. The results of this research have been applied to such areas as pedestrian detection and counting [1]; visual surveillance [2]; human–computer interaction [3]; or autonomous vehicle control [4], etc.

MLCV datasets for the development of visual interpretative tools comprise the base sensory media (images or video) as the target information for interpretation, and metadata describing the aspects of visual reality to be detected, as the "ground-truth". A growing plethora of such MLCV datasets has formed as the research in this area advances. This survey focuses on these datasets.

However, despite the relevance of MLCV datasets for the research in scope, the main MLCV research focus has been on the development of the interpretative tools, and not on the construction or on the details of the datasets. As a result, these have been given very little attention in terms of standardization, uniformization, and of building an integrated tissue for its collective production, storage, and sharing.

Most such datasets were developed by independent and isolated research initiatives, focusing on their own specific and immediate purposes without any kind of strategic planning pertaining to the long-term exploitation and sharing of those resources.

This way, regardless of the labor intensiveness of the production of these resources and their centrality to MLCV research, the overall MLCV dataset tissue is a very disconnected and heterogenous one, in its media component (e.g., static images, video, 2D-depth video, multiview video, etc.), in the described meta-information (e.g., low-level image features, image or video classification, image segments defined by bounding boxes, image segment classifications, etc.), as in the formats employed to describe said meta-information (e.g., plaintext, *.csv*, ViPER [5], VoC [6]).

The means through which (in their repositories) MLCV dataset contents are accessed, retrieved, or shared are also varied. Such contents may be made available in bulk or in varying levels of granularity. They may also be openly shared or accessed at a cost.

MLCV research requires adequate datasets. The higher their quality (larger, more meta-information, rich, and precise) the more labor intensive they are to produce. Naturally, sharing these resources is very relevant for the research/industrial community in scope, as it reduces such costs as well as enables a more uniform training, testing, and comparison of MLCV algorithms and tools.

The above-described current situation means that there is a broad set of synergies for the collaborative production and exploitation of MLCV datasets that is not being exploited. This results in MLCV research initiatives repeatedly incurring in dataset production costs that could easily be avoided, and in obtaining, as a result, datasets of inferior quality than what could easily be achieved in a collaborative environment where these resources would be progressively built up to an ever greater comprehensiveness and quality.

All this means that the progress in MLCV research is not advancing as rapidly, cost-effectively, and efficiently as it could.

Therefore, it is necessary to construct a global tissue of MLCV datasets sharing a common access and manipulation protocol, a common structuring of its internal contents, and a common language (or interconnected set of languages) for the expression of their metadata. This will enable the ML research community to collectively share the production and expansion costs as well as exploit the global pool of such resources, i.e., it will enable the formation and exploitation of a global and ever-growing "library" for machine teaching.

Reaching this objective requires first acquiring a comprehensive knowledge of the contents and formats of MLCV datasets, so that the definition of the necessary unification protocols and tools may then take place. To the best of our knowledge, there is presently no global study of the overall state of the art in MLCV datasets specifically regarding the registered ground-truth information, the formats in which it is expressed, the acquired media information, and the interconnection between media and ground-truth metadata.

Some efforts have already been undertaken along this line, such as the work presented in [2], which focuses on datasets for human action and activity recognition from single or multiview 2D video; Ref. [7] focuses on datasets for action recognition on RGB-D video;

Ref. [8] approaches the most popular databases for object recognition in images; Ref. [9] examines the widely used datasets for salient object detection; and [10] surveys datasets for image description and image captioning. Nonetheless, they are very infrequent and typically focused on datasets pertaining to a single specific type of sensory information (e.g., static images, video, sound) and a specific application domain (e.g., activity detection).

Thus, there is a large set of "blind spots" in this literature and a lack of a global vision. For this, the realization of the global study in scope is something of great importance.

1.2. Objectives

This paper presents a comprehensive survey on the overall panorama pertaining to the media and metadata contents and formats of MLCV datasets. It investigates current and historically relevant datasets pertaining to the interpretation of the visual "sensory information" (image and video) and for a broad range of interpretative purposes, focusing on those that presently deserve the greater scientific attention. Figure 1 presents all the approached datasets inscribed into an organizing taxonomy, which categorizes them according to their employment purpose (expressed through arrow connections) and to the type of media content they comprise (expressed through the placement of inner colored squares).

This paper specifically investigates datasets built from 1994 to 2019 to enable the conduction of research on facial recognition; image segmentation, object and scenario detection and recognition; object tracking; and activity and behavior recognition. Information on them was acquired from MLCV scientific papers, competitions, and dataset repositories found by searching through online databases (e.g., Google Scholar, Google Search, CiteSeerX, Web of Science).

It was this search that enabled identifying those specific application areas as the most relevant MLCV application areas given their predominance in the academic publications pertaining to the overall scope of MLCV.

For each dataset, whenever the information is available, we look into when, how, and by whom it was put together; the precise area of application (e.g., facial reconnaissance, specific activity detection; etc.); any involved licensing aspects; the structure of the dataset (how the different types of information and different types of files are separated and interrelated); the flexibility of access to and manipulation of its contents; the base media type (e.g., single view 2D video, multiview videos, etc.); and the specifics of the ground-truth metadata (contents and format).

As most other aspects, regarding MLCV datasets, the employed metadata formats are also very heterogeneous and ad hoc defined. Given how important the sharing and reutilization of ML datasets is, so too is the interoperability and intelligibility of the associated ground-truth metadata. Thus, the predominant metadata formats employed in existing datasets constitute a key aspect that merits its own analysis. For this, the current survey comprises also a component that focuses on the predominant such formats and their characteristics.

Building on the survey results, we contribute with an analysis of the overall dataset scenario of each of the different MLCV application areas approached. We provide a condensed, quantitative, and qualitative view of these resources and their specificities (type and quantity of media content, type and quantity of annotated aspects, metadata production means dataset contents accessing rights), and of their evolution, so as to discern ongoing and future trends and to assess the possibilities and obstacles to their integration into a homogenous and shared tissue.

The earlier analysis leads us to a set of realizations pertaining to the commonalities and overall nature of the MLCV dataset tissue that results in the formulation of a novel interpretation of this global tissue as one that weaves a global pool of synthetic cognitive memories. Considering all of the above, we then lay forth a set of steps, concerning the data formats, structuring, and accessing of datasets, to facilitate the desired integration. Thus, this paper contributes to the defined novel analogy and the necessary steps forward.

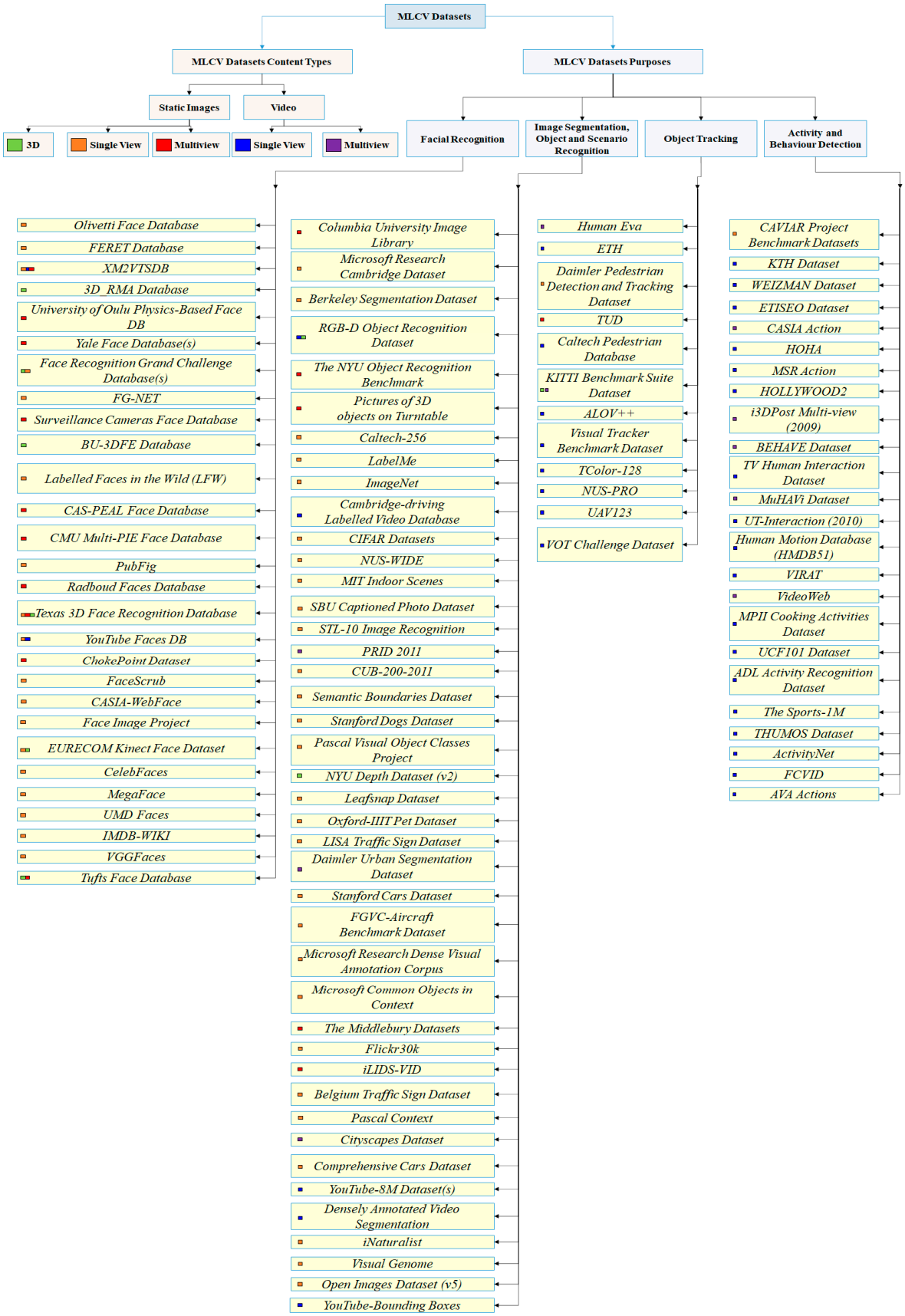

Figure 1. Taxonomy of surveyed datasets.

The rest of the paper is organized as follows. Section 2 presents the research questions that our study seeks to answer and explains the methodology employed in the conduction of this study. Section 3 provides an extensive survey of the most relevant MLCV datasets (for the above-stated application areas). Sections 3.1–3.6 provide a survey of a broad group of MLCV datasets related to the automated interpretation of image for different interpretative objectives. Section 3.7 provides a focused look into the predominant formats employed, in those datasets, for the expression of ground-truth metadata. Section 4 summarizes the qualitative and quantitative findings of this survey and presents our analysis of the overall panorama for each of the approached MLCV application areas for the formats employed for metadata expression in the involved datasets and for MLCV as a whole. Section 5 presents the limitation of our study and conclusions. In Section 6, we put forward our novel analogy to guide the improvement and integration of the overall MLCV dataset tissue, present the necessary steps to attain such interoperability, and present our concluding remarks.

2. Methods

The planification, execution, and reporting of the survey described this paper was conducted in accordance with the action flow described in Figure 2.

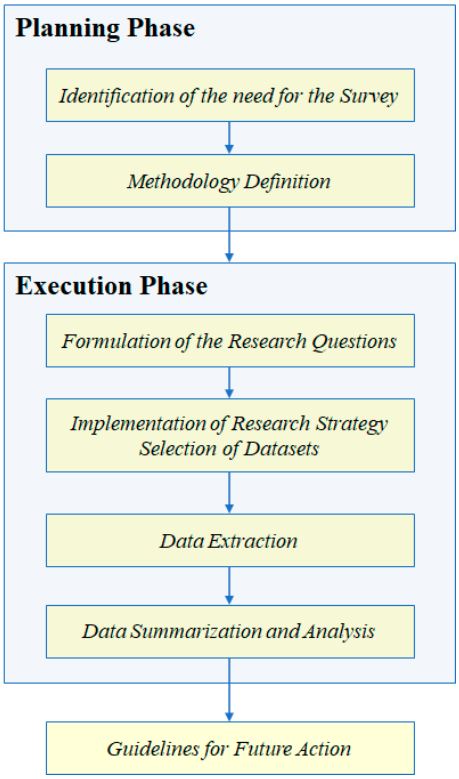

Figure 2. Survey production process.

The planning stage comprised the identification and characterization of the need for this survey, as well as the definition of the survey procedures or methodology. The methodology definition comprised the specification of a first step for research question identification; a strategy for study/MLCV dataset searching and selection, including the identification of search terms and selection of sources; and of the data extraction, synthesis, and analysis process.

The execution stage comprised the implementation of the survey in accordance with the earlier defined methodology. Thus, it consisted of the formulation of the research questions (or objectives) to be addressed by the review; the implementation of the research strategy through the execution of a systematic review pertaining to MLCV datasets and a filtering of the identified results; the extraction of relevant information from the identified

literature (or other sources); and the summarization and analysis of this information to answer the earlier specified research questions.

Based on the attained conclusions, we also derive the necessary future steps, in the definition and structuring of MLCV datasets to enable their coalescence into a coherent standardized whole.

2.1. Research Questions

Table 1 presents the research questions addressed in this survey. From the dataset-related literature, dataset websites, or MLCV competition websites, we acquired the base information on all relevant MLCV datasets. Therefore, we collected information on the acquisition means, types, characteristics (resolution, spectrum, dimensionality, etc.), and amount of their media content. We collected information also on the production means, registered features, perceptive and conceptual aspects, structuring formats, and amount of their metadata (ground-truth) content. Then, this base information was summarized and analyzed to obtain the desired answers. This process was done for Facial Recognition datasets (answering questions RQ1 to RQ5); Object and/or Scenario Detection and Recognition datasets (answering questions RQ6 to RQ9); Object Tracking datasets (answering questions RQ10 to RQ15); and Activity and Behavior Detection datasets (answering questions RQ16 to RQ21). The overall mentioned process enabled gathering the information to answer also questions RQ22 and RQ23.

Table 1. Research question.

RQ#	Research Questions
	Pertaining to Facial Recognition (FR) Datasets
RQ1	Amount of media content in FR datasets—what is the current situation and how has it evolved throughout time?
RQ2	Number of identified individuals in FR datasets—what is the current situation and how has it evolved throughout time?
RQ3	Metadata in FR datasets—what are the main aspects registered in the metadata, the employed formats, and how have they evolved?
RQ4	Image acquisition modes (constrained vs. free) in FR datasets—what is the current situation and how has it evolved throughout time?
RQ5	Modes of access licensing to FR datasets—how have dataset access licensing modes evolved?
	Pertaining to Object and/or Scenario Detection and Recognition (OSDR) Datasets
RQ6	Amount of media content in OSDR datasets—what is the current situation and how has it evolved throughout time?
RQ7	Number of identified objects/scenarios in OSDR datasets—what is the current situation and how has it evolved throughout time?
RQ8	Metadata in OSDR datasets—what are the main aspects registered in the metadata, employed formats, and how have they evolved?
RQ9	Modes of access licensing to OSDR datasets—how have dataset access licensing modes evolved?
	Object Tracking (OT) Datasets
RQ10	Amount of footage in OT datasets—what is the current situation and how has it evolved throughout time?
RQ11	Number of tracked objects classes in OT datasets—what is the current situation and how has it evolved throughout time?
RQ12	Number of individual objects detections in OT datasets—what is the current situation and how has it evolved throughout time?
RQ13	Number of individual objects tracked in OT datasets—what is the current situation and how has it evolved throughout time?

Table 1. *Cont.*

RQ#	Research Questions
RQ14	Metadata in OT datasets—what are the main aspects registered in the metadata, employed formats, and how have they evolved?
RQ15	Modes of access licensing to OT datasets—how have dataset access licensing modes evolved?
	Activity and Behavior Detection (ABD) Datasets
RQ16	Amount of footage in ABD datasets—what is the current situation and how has it evolved throughout time?
RQ17	Amount of images in ABD datasets—what is the current situation and how has it evolved throughout time?
RQ18	Number of activity detections in ABD datasets—what is the current situation and how has it evolved throughout time?
RQ19	Number of activity classes targeted in ABD datasets—what is the current situation and how has it evolved throughout time?
RQ20	Metadata in ABD datasets—what are the main aspects registered in the metadata, employed formats, and how have they evolved?
RQ21	Modes of access licensing to ABD datasets—how have dataset access licensing modes evolved?
	Metadata Formats
RQ22	What are the most widely used, if any, formats for the expression of MLCV datasets metadata?
	Crowdsourcing
RQ23	What are the means employed for crowdsourcing (metadata production)?

2.2. Search Strategy and Study Selection

The predominant information sources for this study were the following databases: Google Scholar; ScienceDirect; Google Search; CiteSeerX; and SpringerLink.

The search was made on the above electronic databases employing composed search queries. These incorporated the different main search terms using the Boolean expression *AND*. For each of the main terms, various different alternatives were included connected with the Boolean expression *OR*.

The following general search query was used for the identification of primary studies: (machine learning OR computer vision OR regression OR classification OR Bayesian network OR decision tree OR support vector machine OR neural network) AND (X) AND (dataset OR ground-truth OR metadata). The X term represents the specific MLCV application area under study. Thus, it consists of (facial recognition OR image segmentation OR object recognition OR object detection OR scenario recognition OR object tracking OR activity detection OR behavior detection).

The search was restricted to the period from 1994 to 2019 as the development of machine learning technology begun in the 1990s, and the years prior to 1994 yielded to little results to be included in the study.

An initial search enabled identifying the candidate primary dataset documental sources (MLCV describing datasets and MLCV dataset websites). The associated full text papers, or website contents, were retrieved and analyzed so as to determine the relevant ones by following the inclusion and exclusion criteria described below. The state-of-the-art and reference sections of the mentioned studies also yielded valuable information on further relevant MLCV datasets, which were also included in the survey.

The candidate documental sources (and respective MLCV datasets) were selected after following the following inclusion and exclusion criteria:

- Inclusion criteria:
- Papers pertaining to the employment of ML techniques for computer vision (regarding the different specific application areas covered by this survey), which describe the employed datasets;
- Papers or websites describing specific datasets for MLCV;

- Papers comprising dataset surveys for specific MLCV application areas;
- Papers describing metadata formats for the expression of MLCV ground-truth information.
- Exclusion criteria:
- Papers/datasets with a small number of citations/mentions (typically bellow 100) in the literature. Exceptions were made for more recently (last 2 years) published papers (less than 60 citations);
- Similar documental sources i.e., studies with similar content done by the by same authors. However, if the results were different in both studies, they were retained.

Finally, the quality assessment criteria were employed in order to identify the final MLCV datasets to be approached. These criteria were:

- The dataset should have scientific relevance, revealed by its uptake by the concerned MLCV research community which translates into papers and citations;
- The information pertaining to the dataset, obtained from the various documenting sources that describe them, should not be overly incomplete or incoherent.

2.3. Data Extraction, Synthesis, and Analysis

We processed all the documental sources for each of the identified MLCV datasets. For each such a dataset, we sought to gather the information regarding the dataset's creation time; its creating originating entity or research team; its specific purpose of application within CV; the source or production mode of the dataset's media content and its quantity; the characteristics of said media content regarding its encoding, resolution, dimensionality, or comprised content; the production mode of the dataset's metadata and its quantity; said metadata's characteristics regarding the features and aspects it describes and the format in which it is expressed; and the media and metadata access licensing details. All the acquired information was exposed in Section 3.

The acquired information was summarized and then subjected to a quantitative and qualitative analysis. The data synthesis process comprised (for each of the specific CV application areas approached):

- The coalescence of all the information, in the respective sub-section of Section 3 into a spreadsheet table with an entry for each dataset. The resulting tables (Tables 2–6) are present in Section 4;
- The synthesis/addition of the similar aspects (columns of the same table), from all datasets on a year-by-year basis, to formulate responses to the research questions. These responses are exposed in Section 4, employing graphs (which chart the above-mentioned calculations) and text. In the next paragraphs, we explain how we proceeded to attain such a synthesis for the different aspects surveyed across the different datasets types.

To do a year-by-year synthesis of the evolution of the amount of media content in datasets, we added, for each year, the number of images (Table 2: column 4; Table 3: column 4, Table 5: column 5) of all the datasets originated in that year (this was done separately for each specific CV application domain). We added only different images (not different versions of the same image), but we added them regardless of their resolution or color characteristics.

To do a year-by-year synthesis of the evolution of the total footage length in datasets, we added, for each year, the total footage length (column 4 of Tables 4 and 5) of all the datasets originated in that year (this was done separately for each specific CV application domain).

To do a year-by-year synthesis of the evolution of the total number of observed identities/persons (Table 2: column 5); detected objects or scenarios (Table 3: column 6); tracked object classes and tracked object class instances (Table 4, columns 5 and 6); and number of activities (Table 5: column 7)—we added, for each year, the total values of such

characteristics for of all the datasets originated in that year (this was done separately for each specific CV application domain).

To do a year-by-year synthesis of the evolution of the total number of object detections (or image segments) described in the metadata, we added, for each year, the number of detection/segments (Table 3: column 5, Table 4: column 7, Table 5: column 6) contained in all the datasets originated in that year (this was done separately for each specific CV application domain).

To do a year-by-year synthesis of the cumulative evolution of the relevance of the different aspects registered in the metadata, we selected the overall most widespread aspects described in the metadata across the datasets of each specific CV application domain (GT Metadata column of Tables 2–6) and then added, for each year, the number of datasets that registered each such aspect in their metadata; then, we calculated, for each year, the cumulative number or registerings of each aspect.

To do a year-by-year synthesis of the cumulative evolution of the modes of dataset access licensing, we identified the predominant licensing modes through which the datasets made their contents available in each specific CV application domain (Licensing column of Tables 2–6); then, we added, for each year, the number of datasets that made their content available according to each licensing mode; finally, we calculated, for each year, the cumulative number that each licensing mode was employed.

The amount of media content comprised by each dataset (be it images or video footage) was sometimes not directly or explicitly provided. In such cases, we had to calculate it or deduce it from the information that was available. For instance, datasets comprising video footage sometimes provide it as a set of individual frames, whilst also indicating the characteristics of their acquisition (namely the frame rate). Based on this information, we calculate the original footage length. On many instances, in the tables of Section 4, we actually present the calculation of the values in scope (instead of the final calculated value). We believe that these numbers are fundamentally right, as the information provided about the datasets is generally sufficient to estimate them correctly, and any small discrepancies that may exist are not significant given the large number (of frames or footage seconds) that we are dealing with.

The overall information richness of images/video is somewhat similar across contemporaneous datasets and has typically grown in time. For this (for instance), a small resolution image in a 1994 dataset was as relevant then as a high-resolution image in a 2019 dataset is now. As the research questions we are addressing with this calculation (RQ1, RQ6, RQ10, RQ16, and RQ17) pertain to the evolution of the amount of media content (i.e., image count or video length) in datasets, we do not find these differences problematic.

The number of object detections (or image segment definitions) present in the datasets' metadata is sometimes not clearly stated. In such cases, these values had to be calculated or deduced. This was done in a variety of ways, employing the available data about each dataset. For instance, some datasets indicate the total number of frames (n), annotation frequency (or fraction of annotated frames f), and the average number of object detections per annotated frame (d). Based on this information, we estimate the value in scope as $n*f*d$. As we always strive to cross different information sources about each dataset (e.g., its describing paper(s) and website), the acquired information is generally sufficient to assure a high degree of accuracy to the above estimates.

The above-mentioned detections are typically described (in the metadata) as some form of an image segment (e.g., bounding box, binary mask, contour shapes, etc.). For the purpose of counting such detections, adding them for each year, and then plotting their evolution across time, we did not distinguish between the different manners though which the detections were described. We understand that they provided information with different precision and density levels, but they all comprise image segmentation data and for the purpose of addressing research questions RQ7, RQ12, and RQ18, they are identical.

The methods we employed to coalesce, estimate, process, and synthetize data were quite straightforward. Given the also straightforward nature of the research questions that we addressed, the relevance of the surveyed datasets, and their overall volume, we

believe that the acquired and synthetized data are representative of the analyzed reality (ML datasets for CV) and provides adequate responses to the research questions defined in Section 2.1.

To express all the acquired and synthesized information, we used visualization techniques such as tables and line graphs, which are all presented in Section 4.

3. MLCV Dataset Assessment Results

3.1. Introduction

In the following sections, we approach a broad set of ML datasets, for CV, looking into a range of different aspects, predominantly focusing on their media and metadata contents, on the employed metadata formats, and on their licensing aspects.

In Sections 3.2–3.6, we look into some of the most relevant datasets built for the training/development of image interpreting provisions. There is a very vast set of specific application areas within the broad scope of image analysis. We analyze some of the most relevant ones. Namely, we look at facial recognition; image segmentation and object and scenario recognition; object tracking; and activity and behavior detection.

We focus on datasets comprising 2D visible spectrum images even if, at some points, datasets with other types of content may be approached.

We also look at some more multipurpose datasets (mixing video and audio interpretation) in Section 3.6.

Our survey focuses on the mentioned sub-areas of CV, because we deem these to be the most relevant ones in terms of density of publications, current academic interest, and overall industrial/commercial employment.

3.2. Facial Recognition Datasets

3.2.1. Olivetti Face Database

The Olivetti Face Database [11,12] was produced between April 1992 and April 1994 at the AT&T Laboratories Cambridge for the development of face identification tools (specifically focused on continuous density Hidden Markov Models).

It comprises 10 different images of each of 40 distinct subjects (four female and 36 male subjects). Their ages range from 18 to 81, the majority being aged between 20 and 35. All images were acquired in a controlled environment against a dark homogeneous background with the subjects in an upright, frontal position, and limited head tilt. These images present varying lighting conditions, facial expressions (open/closed eyes, smiling/not smiling), and facial details (glasses/no glasses). They were manually cropped and rescaled to a resolution of 92 × 112 pixels, with 8-bit gray levels.

This dataset is freely available.

3.2.2. FERET Database

The construction of the FERET [13,14] dataset was supported by the US Department of Defense's Counterdrug Technology Development Program Office (in the context of the FERET program) for the development of tools for facial recognition to be employed in the fields of security, intelligence, and law enforcement.

This dataset's media content was acquired independently from the algorithm developers and in a controlled environment (15 sessions between August 1993 and July 1996). As the image acquisition setup had to be reassembled for each session, this resulted in some minor variations in images collected on different dates.

The dataset comprises 1564 sets of images (365 duplicate sets) for a total of 14,126 images (at a 384 × 256 pixels resolution) of 1199 individuals. For each individual, two frontal views were acquired, and a different facial expression was requested for the second frontal image. Images were also collected from the following perspectives: right and left profiles, right and left quarter profile, and right and left half profile. Additionally, five extra perspectives, irregularly spaced among the basic images, were collected when possible. It was attempted to keep the interocular distance between 40 and 60 pixels.

Some of the individuals were photographed two or more times, and, in some cases, over two years elapsed between their first and last sessions.

Each image was attributed a unique file name that encodes its respective ground-truth. This includes the subject's identity, the nominal pose of the image, the date the image was taken, and the special variations.

The original version of this dataset comprised only eight-bit grayscale images. A newer version of the dataset (the Color FERET Database [15]) was developed comprising colored images.

The FERET database is made available to researchers in face recognition and on a case-by-case basis only.

3.2.3. XM2VTSDB

The XM2VTSDB dataset [16,17] is a large multi-modal face database. It was built in the context of the M2VTS project for the development of tools for personal identification (through analysis of speech and frontal or profile facial images), assuming some subject cooperation.

This dataset comprises four video recordings of 295 subjects (of both genders, varying ages, with and without glasses) performed, in a controlled environment, over a period of four months. This enabled capturing the natural variability of people's appearances. Each such recording contains a speaking head video (with audio) and a rotating head video. Three-dimensional (3D) models of all subjects' heads were also acquired using a high-precision stereo-based 3D camera. Thus, XM2VTSDB comprises also 293 VRML models and texture images. The dataset's static image content was acquired from the high-quality digital video, as this facilitates image processing tasks such as head segmentation, eye detection, lips dynamics assessment, 3D surface modeling, verification of speech/lip shape correlation, and signal synchronization (relevant features for GT metadata production and algorithm development).

The above contents are made available under different sub-groupings (e.g., frontal facial images, lateral facial images, etc.), and as video or as separated sets of frames.

The XM2VTSDB dataset is made available at production cost price.

3.2.4. 3D_RMA Database

The 3D_RMA dataset [18,19] was built by the Signal and Image Center of the Royal Military Academy of Belgium. It is meant to be employed in the development of provisions for facial authentication through the use of 3D facial captures, which enable a facial surface analysis that is less sensitive to viewpoint and lighting conditions than a simple frontal facial image analysis.

The 3D facial captures were achieved through the employment of a 3D acquisition system based on structured light, and the shots were taken with the subjects holding different head orientations: straight forward, left, right, upwards, and downwards. It comprises the 3D captures of the faces of 120 people.

The annotations comprise only the identification of the observed individual.

This dataset is publicly available, and it may freely be used for research purposes.

3.2.5. University of Oulu Physics-Based Face Database

The University of Oulu Physics-Based Face Database [20] (UOPB) was collected at the Machine Vision and Media Processing Unit of the University of Oulu for the development of facial recognition tools but also for color-related studies of faces.

Its image content was acquired in a controlled environment. It comprises photos of the faces of 125 individuals (of different ethnic origins, genders, and with ages ranging from 15 to 65), each of which was acquired under 16 different camera calibration and illumination conditions (an additional 16 photos are acquired if the person is wearing glasses). Each image is 428 $\times$ 569 pixels.

The UOPB dataset also includes three skin spectral reflectance measurements per person, which were measured from both cheeks and forehead as well as illuminant spectral power distribution and camera spectral response.

This dataset's metadata, which is inscribed into the image file names and containing directory structure, includes the identification of the observed individual, the camera calibration and illumination conditions, and whether the person is wearing glasses or not.

The UOPB Face Database (which is physically delivered as a set of CDs) is available for research and verification purposes upon request and for a fee (to cover delivery costs).

3.2.6. Yale Face Database(s)

The Yale Face Databases [21,22] was built for the development of tools for facial recognition under varying light conditions. They comprise the Yale Face Database, the Yale Face Database B, and the Extended Yale Face Database B EYFDb. Each of these builds on the latter. As such, the broader and most recent is EYFDb. This is the version we will approach here.

The EYFDb comprises 16,128, close-up, facial images of 28 human subjects under nine poses and 64 illumination conditions. Each of the 64 different images (for the 64 different illumination conditions) of a subject in a specific pose were acquired at camera frame rate (30 frames/second) in about 2 s, and thus, it is expected that there is only a small change in head pose and facial expression between photos. The acquired images (640 × 480 pixel resolution) are 8-bit (grayscale) and were captured with a Sony XC-75 camera.

A subset of the images (45 of the 64 lighting variations for the nine poses and for 10 individuals) was also subjected to a manual alignment, cropping, and re-sizing to a 168 × 192 pixel resolution. Both the original and altered images are made available.

This dataset's metadata consist of the subject's identifier, the pose, and the illumination angles (in the file name of every photo). It consists also (in separate files) of the coordinates of the subject's face and (for photos with frontal poses) the coordinates of the mouth and eyes.

The EYFDb dataset is freely available for research purposes.

3.2.7. Face Recognition Grand Challenge Database(s)

The Face Recognition Grand Challenge Dataset (FRGCD) [23] was developed at Notre Dame University within the context of the Face Recognition Grand Challenge (FRGC) to promote and advance face recognition technology. This dataset consists of 50k recordings divided into training and validation partitions.

The training partition comprises two training sets:

- The still training set is designed for training still face recognition. It comprises 12,776 images from 222 subjects (6388 controlled still images and 6388 uncontrolled ones). It contains from nine to 16 subject sessions per subject;
- The 3D training set contains 3D scans and, controlled and uncontrolled, still images from 943 subject sessions.

The validation partition gathers the images from 4003 subject sessions.

A subject session is the set of all images of a person that is collected at each data collection time (consists of four controlled still images, two uncontrolled still images, and one three-dimensional image). The controlled images are full frontal facial images taken under two lighting conditions. They were acquired with a 4 Megapixel Canon Power Shot G2 and are either 1704 × 2272 pixels or 1200 × 1600 pixels. The 3D images were acquired (by a Minolta Vivid 900/910 series sensor) under controlled illumination conditions and comprise both a range and a texture image.

The metadata of this dataset consist of the identification of the person whose information is acquired at each subject session.

Access to the dataset requires prior approval by the FRGC Program Manager and that is typically granted for research purposes only.

3.2.8. FG-NET

The FG-NET Aging Database [24] was developed to support facial recognition research activities that take facial aging into consideration, such as, specifically, facial age estimation.

It contains 1002 images from 82 different subjects with ages varying between newborns to 69-year-olds. Ages between zero and 40 are the most well represented in the dataset. Most of the images were collected by scanning photographs of subjects found in personal collections, and as such, they present considerable variability in terms of resolution, quality, illumination, viewpoint, and expression.

Each image in the dataset is annotated with the identity and age of the observed person, as well as 68 facial landmark points and a further semantic description (covering such aspects as expression, pose, image quality, and appearance of occlusions (i.e., moustaches, beards, hats, or spectacles)).

The FG-NET dataset is freely available.

3.2.9. Surveillance Cameras Face Database

The Surveillance Cameras Face Database (SCface) [25] was produced at the Faculty of Electrical Engineering and Computing of the University of Zagreb as a means of testing face recognition algorithms in real-world conditions. It may also be used for face modeling or 3D face recognition.

Most images in this dataset were acquired in an uncontrolled indoor environment using five video surveillance cameras of various qualities. The illumination source was the outdoor light, images were taken from various distances, and the observed head poses are the ones typically found in footage acquired by a regular commercial surveillance system (i.e., the camera is placed slightly above the subject's head). Thus, these images mimic real-world conditions and enable the development and testing of robust face recognition algorithms for law enforcement and surveillance use case scenarios.

The dataset also includes a set of high-resolution images acquired with a high-quality digital photography camera at close range in controlled conditions (standard indoor lighting and adequate use of flash to avoid shades). These images were cropped down to 1600×1200 pixels so that the face occupies approximately 80% of the image. This set of images provides nine views of each individual's face, ranging from left to right profile in steps of 22.5 degrees.

Thus, this dataset comprises 4160 static images (in visible and infrared spectrum) of the faces of 130 subjects. Of these, 115 were males and 15 were females. All were Caucasians between the ages of 20 and 75.

This dataset's metadata include the subject's ID, camera number, distance label, and angle label (inscribed into the image file name). It includes also a textual file with coordinates of eyes, tip of the nose, and center of the mouth, as well as date and year of birth, gender, beard presence, moustache presence, and glasses presence, of the subjects of each photo.

The SCFace may be obtained free of charge for academic or scientific research only.

3.2.10. BU-3DFE Database

The Binghamton University 3D Facial Expression (BU-3DFE) Database [26] was created for the development of facial recognition and facial expression recognition tools.

It comprises 3D facial captures for 100 subjects (56% female, 44% male) with ages ranging from 18 to 70 years old and with a variety of ethnic/racial ancestries (White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino).

Each subject's face was 3D captured while performing seven different expressions (happiness, disgust, fear, angry, surprise, sadness, and neutral) with four different intensity levels (for all expressions except neutral). Thus, 25 instant 3D expression models were acquired for each subject, resulting in a total of 2500 3D facial expression models in this database. For each of the earlier models, two corresponding 2D facial texture images were also captured at two views (about $+45^\circ$ and -45°).

The BU-3DFE's metadata includes, for each 3D capture, the subject's identifier, his gender and race, his expression, a set of (83) facial feature points, and a facial pose model.

This dataset is freely available (upon request) for research and non-profit uses. Its employment for commercial purposes may be negotiated.

3.2.11. Labeled Faces in the Wild (LFW)

The Labeled Faces in the Wild (LFW) dataset is supported by the Computer Science Department of the University of Massachusetts [27]. It was built to enable the development of CV solutions for unconstrained face recognition, specifically focusing on face verification and facial pair matching.

The images in this dataset were acquired from the web and comprise a "natural" variability in terms of pose, lighting, expression, background, race, ethnicity, age, gender, clothing, hairstyles, camera quality, color saturation, etc., thus including images with poor lighting, extreme pose, strong occlusions, etc. It is a broad set of images, still, many human sub-groups are underrepresented (e.g., babies and children). Thus, the LFW comprises over 13,000 human face images (base images). Over 1600 of the people identified have two or more distinct photos in the dataset. This dataset contains also other types of images, associated with the earlier one, which include funneled images and deep funneled images. LFW also makes available, for ease of experimentation, parallel versions of the database containing aligned images and superpixel computation.

Associated to each base image, there is a label (meta-information) with the name of the person whose face it depicts (LFW also provides manually verified gender information).

This dataset is publicly available and may be retrieved as a download in bulk.

3.2.12. CAS-PEAL Face Database

The CAS-PEAL Face Database [28] has been developed through the sponsorship of the Chinese National Hi-Tech Program and ISVISION by the Face Recognition Group of JDL for the development of facial recognition tools.

This dataset includes a large set of Chinese (Mongolian) face images with variations pertaining to pose, expression, accessories, and lighting. Thus, it comprises 99,594 images of 1040 individuals (595 males and 445 females). Each subject was simultaneously photographed (in a controlled environment) by nine cameras (Web-Eye PC631 with 640×480 pixels charge-coupled device (CCD)) in three different head poses (frontal, looking up and down). Five kinds of expressions, 6 kinds of accessories (3 glasses and 3 caps), and 15 lighting directions were also taken into consideration.

The metadata pertaining to every image in this dataset are encoded into its filename. Therefore, these include individual identifier; gender and age; lighting variation information; pose information (looking up, looking down, looking forward); expression information (neutral, laughing, frowning, surprising, eyes closed, mouth open); accessory information (none, hat1, hat2, hat3, glasses1, glasses2, glasses3); distance to camera information; time information (first session, second session, third session); background information (blue, red, dark, yellow, white); resolution characteristics (640×480, 320×240).

In addition, the ground-truth of eye locations of all the images are provided in a separate text file.

This database is partly available (30,900 images, converted to grayscale and cropped to size 360×480, of 1040 subjects) for research purposes only on a case-by-case basis.

3.2.13. CMU Multi-PIE Face Database

The CMU Multi-PIE face database [29] is maintained by Carnegie Mellon University (CMU) for the development of facial recognition tools for varying pose and illumination.

It holds over 750k images taken from 337 different individuals. Sixty percent of subjects were European-Americans, 35% were Asian, 3% were African-American, and 2% were other ethnicities. Of these, 264 of the subjects were recorded at least twice and 129 of them were recorded four times. At each session, each individual's face was simultaneously

photographed from 15 different viewpoints and under 19 illumination conditions while the subject was displaying a range of facial expressions. Thirteen cameras were located at head height, in 15-degree intervals. Two other cameras were located above the subject, simulating a typical surveillance point of view. All frontal images were acquired with a Canon EOS 10D (6.3-megapixel CMOS camera). The resulting images are 3072 $\times$ 2048 pixels with the inter-pupil distance of the subjects, typically, in excess of 400 pixels.

Dataset metadata (image labels) are available in text files (subject info) or in the directory structure/file name (expression, illumination, camera view). Such labels are only provided for the lower resolution images, which were taken inside the collection room. Acquired feature points are made available in *.mat* files.

This dataset is available under a license from CMU for internal research purposes. However, as it is only physically distributed (shipped on a dedicated hard drive), its licensing costs include also fees charged to provide the hard drive.

3.2.14. PubFig

The Public Figures Face Database (PubFig) [30] is inspired in the LFW dataset and intends to be somewhat complementary to it in the development of facial recognition tools.

PubFig comprises 58,797 unconstrained images of 200 people collected from the internet. Face and fiducial point detection tools were run on the downloaded images to obtain cropped face images. Then, these were rectified using an affine transform. Nonetheless, the dataset's images present a large variation in pose, lighting, expression, scene, camera, imaging conditions, parameters, etc.

This dataset is divided into two parts. The first is the development subset (includes images of 60 individuals), which is meant to be employed in algorithm development. It presents no overlap with the evaluation subset, nor with the LFW dataset. The second is the evaluation subset (includes images of the remaining 140 individuals), which is meant to be employed in the evaluation of the developed algorithms.

PubFig's metadata are contained in simple *.txt* files, carrying the information in a (*ad hoc*) table-like format. The development subset's metadata consist of the observed person's identity and a set of 73 attributes. These indicate if the person has any of the following characteristics and to which degree: male, Asian, white, black, baby, child, youth, middle-aged, senior, black hair, blond hair, brown hair, bald, no eyewear, eyeglasses, sunglasses, moustache, smiling, frowning, chubby, blurry, harsh lighting, flash, soft lighting, outdoor, curly hair, wavy hair, straight hair, receding hairline, bangs, sideburns, fully visible forehead, partially visible forehead, obstructed forehead, bushy eyebrows, arched eyebrows, narrow eyes, eyes open, big nose, pointy nose, big lips, mouth closed, mouth slightly open, mouth wide open, teeth not visible, no beard, goatee, round jaw, double chin, wearing hat, oval face, square face, round face, color photo, posed photo, attractive man, attractive woman, Indian, gray hair, bags under eyes, heavy makeup, rosy cheeks, shiny skin, pale skin, 5 o'clock shadow, strong nose–mouth lines, wearing lipstick, flushed face, high cheekbones, brown eyes, wearing earrings, wearing necktie, wearing necklace.

The evaluation subset's metadata consist of the earlier one as well as the person's pose (frontal or non-frontal) and expression (neutral or non-neutral), and the lighting characteristics (frontal or non-frontal).

This dataset is available only for non-commercial use. Due to copyright issues, the image files must be retrieved individually from the Internet. PubFig provides the URLs to all images.

3.2.15. Radboud Faces Database

The Radboud Faces Database (RaFD) [31] was built by the Behavioral Science Institute of the Radboud University Nijmegen for the development of provisions focusing on the extraction of information from human faces.

It comprises photos from 67 different individuals (including Caucasian males and females, adults and children, and Moroccan Dutch males), displaying eight facial expressions

(anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral). Each emotional expression was acquired with three different gaze directions (eyes directed straight ahead, averted to the left, and averted to the right.). Each such expression was simultaneously captured by five different cameras (Nikon models D200, D2X, and D300, with resolutions between 10 and 12Mpx) placed at five different angles (placed in steps of 45 degrees) against a uniform white background. Photographed individuals wore black t-shirts and presented no facial hair nor other objects or adornments.

In the post-processing phase, all images were spatially aligned according to facial landmarks (using Matlab tools), cropped, and resized to 1024 × 681 pixels.

Thus, RaFD comprises 1608 sets of five simultaneous photos for a total of 8040 photos.

This dataset is freely available for non-commercial scientific research by researchers associated with an officially accredited university.

3.2.16. Texas 3D Face Recognition Database

The Texas 3D Face Recognition Database (Texas 3DFRD) [32] was assembled with the assistance of research students and faculty from the Laboratory for Image and Video Engineering at the University of Texas at Austin for research in 3D face recognition.

It comprises 1149 pairs of high-resolution, pose normalized, pre-processed, and perfectly aligned color and range images of 118 adult human subjects. The number of images per subject varies between 1 and 89. The subjects' ages range from 22 to 75 years. Subjects include both males and females from the major ethnic groups of Caucasians, Africans, Asians, East Indians, and Hispanics. The subjects' faces present both neutral and expressive modes. Neutral faces are emotionless. The facial expressions present are smiling or talking faces with open/closed mouths and/or closed eyes. All subjects were requested to remove hats and eyeglasses prior to image acquisition.

The dataset's images were acquired with a stereo imaging system, with a 751 × 501 pixel resolution, and (for the range images) at a very high spatial resolution of 0.32 mm along the x, y, and z dimensions. The color and range images were acquired simultaneously and thus are perfectly synchronized. Images were also post-processed to remove extraneous non-facial data and normalized to a frontal pose and a standardized position.

The dataset's metadata include, for each facial image, the observed subjects' gender, ethnicity, facial expression, and the locations of 25 different anthropometric facial fiducial points. The latter were manually located on the facial color images.

The database is freely available for educational and research purposes only.

3.2.17. YouTube Faces DB

The YouTube Faces Database [33] contains a set of videos capturing people's faces for employment in the development of unconstrained face recognition in videos.

The videos were obtained from YouTube employing the same 5749 names of subjects from the LFW dataset to search YouTube for videos. The dataset contains 3425 videos of 1595 different people. On average, 2.15 videos are available for each individual person, and each video clip is 181.3 frames long.

The dataset's metadata are encoded into *.mat* files. It comprises, for each frame, the identification of the person observed, a bounding box enclosing the person's face, and the three rotation angles of the head.

An extension to this dataset is presented in [34], where the original videos have been cropped around the faces, and only consecutive frames of up to 240 frames have been preserved for each original video. Furthermore, facial keypoints have been automatically extracted for every frame of each video.

This dataset is publicly available and may be retrieved as a download in bulk.

3.2.18. ChokePoint Dataset

The ChokePoint Dataset [35] was meant for the development of provisions for person identification (but also for 3D face reconstruction, pedestrian/face tracking, or background estimation and subtraction) under real-world surveillance conditions using existing technologies.

The dataset comprises 48 video sequences and 64,204 face images (extracted from those sequences). Only one subject is visible at one time in every sequence. The first 100 frames in each video are for background modeling, and thus, no foreground objects are present in them. The sequences capture the images of 25 subjects (19 male and six female) going through surveillance portal 1, and 29 subjects (23 male and six female) going through surveillance portal 2. The acquisition of the sequences from portal 1 and portal 2 are one month apart. The acquired video's frame rate is 30 fps, and the image resolution is 800×600 pixels.

A set of three cameras was placed above each portal to acquire footage of the subjects going through in a natural way. When a subject goes through, a sequence of face images (i.e., a face set) can be captured. Such face acquisitions will necessarily present several variations regarding illumination conditions, pose, sharpness, as well as misalignment. Given the positioning of the cameras, one of them is always likely to capture a face set where a subset of the faces is near-frontal.

The dataset's metadata consist of the subject's identification, the portal identification, the video sequence number, the camera identifier, and the weather the subject is entering or leaving. All this information is inscribed into the video file name. The metadata in scope includes also, for every frame of every sequence, the identification of the visible subject (in that frame) and of the position of his/her left and right eyes. For each video sequence, this information is contained in an xml file with an ad hoc format.

This dataset is freely available to the scientific community for non-commercial research purposes.

3.2.19. FaceScrub

The FaceScrub dataset [36] was developed for facial recognition research.

It holds over 100k face images of 530 male and female celebrities, with about 200 images per person. The images in this dataset were compiled by searching the Internet for public figures followed by a cleaning of the results. This implied the automatic detection of face locations, their alignment, and cropping to 96×96.

The dataset's meta-information consists of the identification of the person observed in each image.

FaceScrub is released under a creative commons license (however, they only provide the URLs to the images (plus annotations), as they do not own the content).

3.2.20. CASIA-WebFace

The CASIA-WebFace [37] dataset was built by the Institute of Automation of the Chinese Academy of Sciences for unconstrained face recognition, and in a manner so that it is complementary to LFW.

It was created using automated face detection, cropping and identification mechanisms, and a final manual verification. It is expected to contain some minor mislabeling. This dataset comprises 494,414 images from 10,575 individuals. The metadata consist of the names of the observed identities, which is inscribed into the folders carrying the images.

The dataset is available for research, educational, or non-commercial use, free of charge. Its contents are provided upon request.

3.2.21. Face Image Project

The Face Image Project dataset [38] is meant to facilitate the study of automated age and gender recognition.

This dataset's images were obtained from Flickr albums under a Creative Commons (CC) license. Flickr albums were processed by first running the Viola and Jones face, then

detecting facial feature points, and then manually labeling all images for age, gender, and identity (using both the images themselves and any available contextual information from Flickr). These images are available both in a cropped version and in a cropped and aligned one. Thus, Face Image Project comprises a total of 26,580 images of 2284 subjects.

The dataset is divided into five folds to allow for cross-validation. Each fold contains different subjects to avoid overfitting.

The metadata of each fold are contained in a *.csv* file and consist of a table-like structure containing, for each image, the subject identifier; the name of the original image file; the Flickr face identifier; the subject's age; the subject's gender; the x, y, dx, and dy values describing the face enclosing bounding box in the original Flickr image; facial tilting angle; facial pose; and the score of the fiducial landmark detector.

Access to the dataset is freely available.

3.2.22. EURECOM Kinect Face Dataset

The EURECOM Kinect Face Dataset (EURECOM KFD) [39] was built by the EURECOM Institute for Facial Recognition (but also facial demographic analysis and 3D face modeling) research on 3D facial images acquired employing Microsoft Kinect.

This dataset's content was acquired with Kinect at two sessions in an indoor environment. In each such session, facial images were acquired, of each person, with nine different facial expressions, different lighting and occlusion conditions (neutral, smile, open mouth, left profile, right profile, occlusion eyes, occlusion mouth, occlusion paper, and light). All acquired images comprised three different modalities: an RGB color image (normalized by cropping at the size of 256 $\times$ 256 centered at the face); a depth map; and a 3D point cloud. For each session, an RGB-D video sequence was also acquired.

Therefore, it comprises multi-modal facial images of 52 people (14 females, 38 males). The participants were born between 1974 and 1987, and they present different ethnic backgrounds (Caucasian (21), Middle East/Maghreb (11), East Asian (10), Indian (4), African-American (3), and Hispanic (3)).

The dataset's metadata comprise (in the name of the image files) the person identifier, session identifier, and face status. It includes also (in a separate *.txt* file for each identity) the subject's gender, birth year, ethnicity, image acquisition session time, and whether he/she is wearing glasses. Another file (for each image) comprises six manually assigned facial landmarks (left eye, right eye, the tip of nose, left side of mouth, right side of mouth, and the chin).

The EURECOM KFD dataset is available on a case-by-case basis.

3.2.23. CelebFaces

The CelebFaces dataset [40] was built by the Multimedia Lab of the Chinese University of Hong Kong (MMLAB) for research on face verification in the wild; face attribute recognition; face detection; facial landmark localization; and face editing and synthesis.

The images in this dataset were acquired from the web. It comprises 202,599 celebrity face images (with a 178 $\times$ 218 resolution) for a total of 10,177 identities (which do not overlap those of LFW).

Annotations comprise, for each image, the celebrity identity; the face surrounding the bounding box; five landmark locations (two points for each of the eyes, nose, and mouth); and 40 binary attributes (5 o'clock shadow, arched eyebrows, attractive, bags under eyes, bald, bangs, big lips, big nose, black hair, blond hair, blurry, brown hair, bushy eyebrows, chubby, double chin, eyeglasses, goatee, gray hair, heavy makeup, high cheekbones, male, mouth slightly open, moustache, narrow eyes, no beard, oval face, pale skin, pointy nose, receding hairline, rosy cheeks, sideburns, smiling, straight hair, wavy hair, wearing earrings, wearing hat, wearing lipstick, wearing necklace, wearing necktie, young).

All this information is expressed in an ad hoc format and carried in *.txt* files.

This dataset is available for non-commercial research purposes only and access is granted to its contents upon request. All images in CelebFaces are obtained from the Internet and thus are not the property of MMLAB.

3.2.24. MegaFace

The MegaFace dataset [41] was developed for employment in facial recognition research.

This dataset's images are unconstrained and obtained from Flickr's photo database. The goal of this dataset is to be broad rather than deep (contain many different people rather than many photos of a few people). Therefore, it comprises 4.7 million face images of 672 thousand different identities. This constitutes an average of seven photos per person, with a minimum of three and a maximum of 2469. The images were post-processed: the faces were detected using the Head-Hunter algorithm, and the images were cropped so that the face spans 50% of the photo height.

The dataset's metadata comprise bounding boxes for the face regions of the images. A further 49 fiducial points, as well as yaw and pitch angles, were calculated as computed by the IntraFace landmark model.

For the above, a JSON metadata file exists for each image file containing the coordinates of a box loosely comprising an expanded face detection region, with respect to the full Flickr image; the coordinates of a box tightly comprising the face region (completely contained within the loose box coordinates), with respect to the full Flickr image; and 68 facial landmarks points as x/y coordinates with respect to the tight face bounding box.

The MegaFace dataset is freely available for non-commercial research and educational purposes under a Creative Commons license.

3.2.25. UMD Faces

The UMDFaces dataset [42] was developed for research on facial recognition, head pose estimation, and facial keypoint localization.

The dataset's images were acquired from the web. It comprises both still images and video frames (over 3.7 million video frames).

The dataset's metadata comprise 367,888 face annotations for 8277 individual subjects. These annotations include human-curated (by way of the Amazon Mechanical Turk's crowd-sourced services) bounding boxes for faces, as well as automatically estimated pose information (yaw, pitch, and roll), keypoint location (for 21 keypoints), and gender information. The dataset's metadata, regarding the video frames, comprises the same contents, except for the face delimiting bounding boxes, and it includes annotations for over 3.7 million frames pertaining to 3100 subjects.

This dataset is freely available.

3.2.26. IMDB-WIKI

The IMDB-WIKI dataset [43] was developed for age prediction based on facial images.

This dataset's images (and metadata) were obtained by crawling IMDb and Wikipedia for facial photos and information pertaining to the 100K most popular actors as listed on IMDb. All images without timestamps were discarded. The dataset may present some inaccuracies resulting from imprecision of the crawled information.

Thus, it comprises 260,282 images pertaining to 20,284 celebrities.

The dataset's metadata are included in a *.mat* file. It comprises the following for each photo: date of birth of the celebrity; year when the photo was taken; gender; name; location of the face; detector score for the face location; detector score of the face with the second highest score; IMDB celebrity ID.

The IMDB-WIKI dataset is made freely available for academic research purposes only. Copyright pertaining to the images belongs to their original owners.

3.2.27. VGGFace2

VGGFace2 [44] is a large-scale dataset for facial recognition.

It comprises over 3.3 million images downloaded from Google Image Search, from 9131 celebrities (spanning a wide range of ages, ethnicities, and professions), which have large variations in pose and illumination.

This dataset is approximately gender-balanced, with 59.3% males. The number of images per individual varies between 80 and 843, with an average of 363. It is divided into two splits: one for training with 8631 classes, and one for evaluation with 500 classes.

VGGFace2's metadata include the identity of the observed individual, bounding boxes around faces, and five fiducial facial keypoints. It also include information about the pose (yaw, pitch, and roll) and apparent age. This information is stored in simple *.txt* or *.csv* files employing an ad hoc table-like structure.

The dataset's metadata were produced both automatically and manually. Age and pose information are acquired employing pre-trained pose and age classifiers. The identity information and bounding boxes were first produced through automated means and then subjected to manual verification to attain a 96% degree of accurateness.

This dataset is available for commercial and research purposes under a Creative Commons Attribution-ShareAlike 4.0 International License.

3.2.28. Tufts Face Database

The Tufts Face Database [45] is a multi-modal image database developed to be employed in cross-modality face recognition.

The subjects' images were acquired against a blue background and close to the camera, maintaining a strict control over camera distance to the participant and lighting (diffused lights were employed).

This dataset comprises over 10,000 multi-modal images from 113 different individuals (74 females and 38 males, from more than 15 countries with an age range between 4 and 70 years old). The involved images are of seven image modalities: visible, near-infrared, thermal, computerized sketch, LYTRO, recorded video, and 3D images.

This dataset is freely available for non-commercial research and educational purposes.

3.3. Image Segmentation, Object and Scenario Recognition Datasets

3.3.1. Columbia University Image Library

The Columbia University Image Library [46] (COIL-100) was developed for image recognition.

The dataset comprises images of 100 different objects from 72 different poses. The photographed objects were placed on a turntable and rotated 360 degrees at 5-degree intervals. The camera was tilted down at about 25 degrees to point toward the turntable. The object was clipped from the black background using a rectangular bounding box and resized to 128 $\times$ 128. Images were size normalized and also histogram stretched, i.e., the intensity of the brightest pixel was made 65,535, and the intensities of the other pixels were scaled accordingly. The dataset comprises a total 7200 images.

This dataset's metadata comprise the observed object's identifier and pose (angle) inscribed in the file's name.

The Columbia University Image Library is freely available.

3.3.2. Microsoft Research Cambridge Dataset

Microsoft Research at Cambridge built and released a dataset (MSRCD) [47] to be employed in the development of machine vision algorithms (with supervised and unsupervised training) for the automatic recognition and segmentation of various different object types.

This dataset comprises high-resolution images and associated labeling annotations. Specifically:

- Pixel-wise labeled images from the v1 database (240 images, nine object classes);
- Pixel-wise labeled images from the v2 database (591 images, 23 object classes).

MSRCD's metadata consist (for each image file in the dataset) of the identification of the objects (in the image file names) and the shape mask (as another image file).

This dataset may be freely employed for non-commercial purposes.

3.3.3. Berkeley Segmentation Dataset

The Berkeley Segmentation Dataset (BSD) [48] was produced by the U.C. Berkeley Computer Vision group for the development of tools for contour detection and image segmentation and recognition.

It comprises 800 images of natural scenes (taken from the Corel image database) and 3000 segmentations of such images by 25 different people. On average, five different segmentations (by five different people) are provided for each image.

BSD's GT metadata consist of segmentation maps, which are stored in *.mat* files.

This dataset is freely available.

3.3.4. RGB-D Object Recognition Dataset

This RGB-D Object Recognition Dataset [49] was built by the Computer Science department of the University of Washington for research on visual object category and instance detection from RGB-D data.

The media component of this dataset was produced using a Kinect style 3D camera, which captures synchronized and aligned 640 $\times$ 480 RGB and depth images at 30 Hz. Each filmed object was placed on a turntable, and video sequences were captured for one complete rotation. For each object instance, three video sequences were acquired, each having the camera capture the object from a different angle.

The dataset specifically comprises the visual captures of 300 common household objects (300 instances), pertaining to 51 different object categories, which results in a total of 250K RGB-D images.

The RGB-D Object Dataset also comprises 22 annotated video sequences of natural scenes containing the objects from the dataset.

The dataset's metadata comprise the object identification (inscribed in the archive structure and image file names of the dataset) and the ground-truth pose information for all 300 objects in all images (all 250,000 frames) in the form of segmentation masks (*.png* files). It comprises also the bounding boxes for the objects recognized in the earlier mentioned 22 annotated videos, which are contained within a *.mat* file for each such video.

The RGB-D Object Dataset is freely available for non-commercial research/educational use only.

3.3.5. The NYU Object Recognition Benchmark

The NYU Object Recognition Benchmark (NORB) dataset [50] was built for the development of research on object recognition, from 3D images, independently of the pose, illumination, and background clutter.

This dataset comprises thousands of images of 50 toys. Each toy in the dataset was painted with a uniform bright green to eliminate irrelevant color and texture information; then, it was imaged by two cameras (stereo image pair) under 6 lighting conditions from 9 different elevations and 18 different azimuths (for a total of 194,400 images at a 640 $\times$ 480 resolution). Each captured image was post-processed so that the object is centered in the image and scaled so that its enclosing bounding box is roughly 80 $\times$ 80 pixels and placed on a uniform background, including the cast shadow. Then, three sources of variations were added to the dataset images: the objects were perturbed; the objects were superposed onto a complex background; and distractor objects were added to the background.

The dataset's metadata comprise the labels of the imaged toys and their corresponding bounding boxes. The depicted toys are labeled as belonging to five generic categories: four-legged animals, human figures, airplanes, trucks, and cars.

The dataset's contents are split into various different files of the *.mat* type. All such contents are freely available for research purposes.

3.3.6. Pictures of 3D objects on Turntable

The Pictures of 3D objects on Turntable (P3DTT) dataset [51] was built for research on object feature mapping across viewpoints and lighting conditions and multiperspective object recognition.

This dataset comprises stereo images of 100 different objects acquired from 144 viewpoints under three different lighting conditions. The acquisition system consisted of two cameras taking pictures of objects on a motorized turntable. The change in viewpoint is performed rotating the turntable.

The datasets metadata consist of the object labels and respective identifications, which are inscribed into the dataset's directory structure.

The dataset is freely available for research.

3.3.7. Caltech-256

The Caltech-256 dataset [52] was built on the earlier Caltech-101 dataset to enable research on object recognition.

The dataset's images were acquired from online databases and manually verified. Then, it comprises 30,607 images, pertaining to 256 categories of objects, with a minimum of 80 images per category. The target object in each image is prominent in it, and thus, said images present a small or medium degree of background clutter.

The dataset's metadata consist of the label for each image.

The Caltech-256 dataset is freely available for research.

3.3.8. LabelMe

The LabelMe dataset [53] (maintained by MIT Computer Science and Artificial Intelligence Laboratory or MIT CSAIL) was built for employment in research on object class recognition including when object instances are embedded in a scene.

The maintainers of this dataset provide a web-based tool that enables remote users to provide images and manually annotated them.

This dataset comprises all images from the MIT CSAIL database, in addition to a large number of user uploaded images. This way, it includes 11,845 static pictures and 18,524 sequence frames (all of which have at least one object labeled).

The dataset's metadata consist of bounding boxes, polygons, or segmentation masks (as depicted in Figure 3), which define segments of images, and "free text" tags associated to such segments. This metadata component comprises a total of 111,490 polygons (44,059 of such polygons were annotated using the online tool, and 67,431 of them were annotated offline), which are associated to a total of 2888 different labels (such as car, person, building, road, sidewalk, sky, tree, etc.). In total, 11,571 pictures have less than 10% of their pixels labeled, and around 2690 pictures have more than 90% of labeled pixels. The Pascal VOC format is employed to express the metadata.

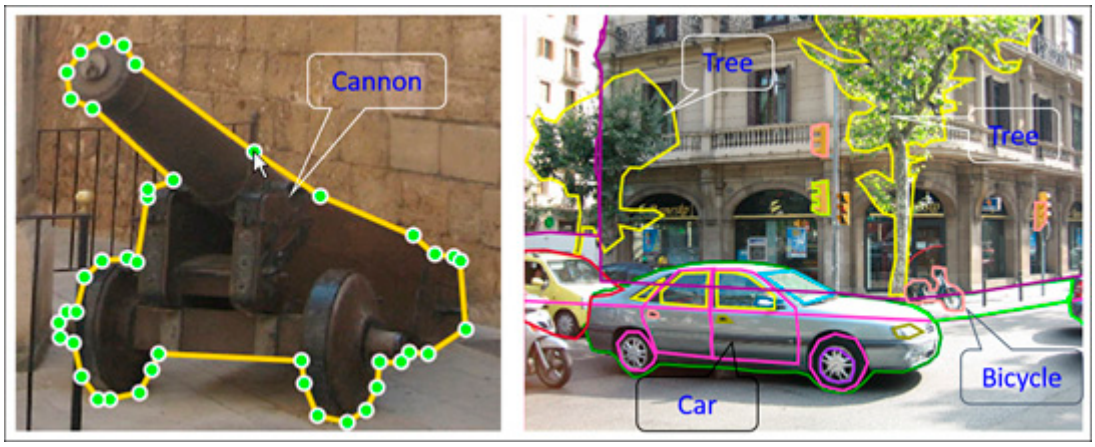

Figure 3. Example of annotated images (images obtained from LabelMe).

Each image is annotated, on average, with 3.3 objects (6876 of the dataset's images have more than five objects annotations).

LabelMe is open to user contributions. Researchers can both download and employ the data available at Labelme as well as provide annotated images to this repository.

The contents of this dataset are freely available for research

3.3.9. ImageNet

The ImageNet dataset [54] holds a very large collection of images and associated annotations, which are meant for employment in scene and object recognition research, specifically, non-parametric object recognition, tree-based image classification, and automatic object localization.

Its data are organized according to the WordNet hierarchy. In this hierarchy, each meaningful concept (which may be comprised of multiple words or word phrases) is called a "*synonym set*" or "*synset*". WordNet comprises over 100k *synsets* (most of which are nouns).

The goal of ImageNet is to gather an average of 1000 images to illustrate each such *synset*, where each image is subjected to quality control and human-annotated. Ultimately, ImageNet should offer tens of millions of adequately classified images for most of the concepts in the WordNet hierarchy. At its current state, ImageNet includes 14,197,122 images for 21,841 *synsets*.

Regarding this dataset's metadata component, it includes both image-level annotations (indicating the presence or absence of specific object classes in an image) and object-level annotations (bounding box enclosing the indicated object). The annotations (object classifications, bounding boxes, and feature-descriptor markers) are expressed in the PASCAL Visual Object Classes format. Presently, 1,034,908 of the dataset's images comprise bounding box annotations. One thousand *synsets* have SIFT features for their characterization and so do 1.2 million images.

The annotation process is based on crowdsourcing.

The contents of the ImageNet dataset are freely available for non-commercial research and/or educational uses. However, ImageNet does not own the copyright of the images. Thus, it typically provides only thumbnails and URLs for them. Nonetheless, for researchers and educators, ImageNet may also provide direct image access.

3.3.10. Cambridge-Driving Labeled Video Database

The Cambridge-driving Labeled Video Database (CamVid) [55] was created to enable research on, and quantitative evaluation of algorithms for, moving object detection and video understanding.

This dataset comprises a collection of videos (960×720 pixels) captured from the perspective of a driving automobile, which increases the number and heterogeneity of the observed object classes. Specifically, CamVid provides four video sequences, totaling more than 22 min of high-quality, 30 Hz, footage. More than 10 of those 22 min are annotated with object class semantic labels.

CamVid dataset's metadata consist of the corresponding (to the video frames) per-pixel semantically segmented images at 1 Hz and, in part, 15 Hz. Thus, it comprises over 700 images with per-pixel semantic segmentation of over 700 video frames, associating each pixel with one of 32 semantic classes (belonging to one of the following groups: moving objects, road, ceiling, fixed objects). These metadata were manually produced. The dataset's metadata comprise also the intrinsic calibration of the cameras and the description of the camera pose trajectories.

CamVid is freely available, and its developers also offer custom-made labeling software.

3.3.11. CIFAR Datasets

The CIFAR-10 and CIFAR-100 datasets [56] are labeled subsets of the "80 million tiny images dataset" [57], and they were developed for research on object recognition.

CIFAR-10 comprises 60k 32 × 32 color images of 10 different classes of things, with 6000 images per class. It is divided into 50k training images and 10k test images. The classes are completely mutually exclusive (i.e., no image belongs to more than one class). This dataset's metadata consist of the labels of each image.

The CIFAR-100 dataset is very similar to CIFAR-10. It has 600 images of each of 100 different classes of things, totaling 60k images. For each class, it comprises 500 training images and 100 testing images. Differently form CIFAR-10, the 100 classes in CIFAR-100 are grouped into 20 superclasses; for this, the metadata associated to each image comprise a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs).

This dataset is freely available as a set of Python ready files, MATLAB ready files, or *c* ready files.

3.3.12. NUS-WIDE

The NUS-WIDE dataset [58] was built for employment in object detection CV research.

The images and most of the metadata in this dataset were acquired from Flickr. Thus, it comprises over 269k images and their associated tags. The metadata in scope include over 425,059 Flickr tags for a total of 5018 concepts (all of which are found in WordNet). It includes as well six types of low-level features calculated from the mentioned images (64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT descriptions). A further component of this dataset's metadata are the manually produced object labels for 81 concepts (for most of the acquired images). These were produced by manual annotators taking advantage of the labels obtained from Flickr and of a previous automatic classification step.

This dataset is freely available for research and/or educational purposes.

3.3.13. MIT Indoor Scenes

The MIT Indoor Scenes dataset [59] (maintained by MIT) is meant for employment in indoor scene recognition CV research.

This dataset's image content was obtained from online image search tools and online photo sharing sites such as Flickr and the LabelMe dataset. Thus, it comprises 15,620 images pertaining to 67 different scene categories (with at least 100 images per category).

The dataset's metadata comprise the definition of ROI (Region of Interest) and the classification of images according to the above-mentioned 67 categories. These metadata are expressed in the same format as that of LabelMe.

The data are freely available.

3.3.14. SBU Captioned Photo Dataset

The SBU Captioned Photo dataset (SBU CPD) [60,61] was developed to be employed in CV research on object recognition in static images with a specific focus on automatic image description generation.

This dataset's images were obtained from Flickr. They were selected through querying and were then both automatically (e.g., employing object detection tools) and manually filtered to produce a collection of over 1 million well-captioned pictures (user-generated captions, obtained also from Flickr).

This dataset's images are hosted at Flickr and must be obtained from that source. SBU CPD's metadata comprise the URLs of all images and their textual descriptions.

This dataset is freely available.

3.3.15. STL-10 Image Recognition

The STL-10 dataset [62] was built for the conduction of research on image recognition comprising unsupervised feature learning, deep learning, and self-taught learning algorithms.

This dataset is inspired by the CIFAR-10 dataset. However, each class has fewer labeled training examples, and a very large set of unlabeled examples is provided to enable learning image models prior to supervised training. It comprises 100k unlabeled images (for unsupervised learning), 500 training images (10 pre-defined folds), and 800 test images per class. Images are 96 × 96 pixels and in color and were acquired from labeled examples on ImageNet.

STL-10's metadata consist of the object class label for each of the labeled images. Ten different object classes are possible: airplane, bird, car, cat, deer, dog, horse, monkey, ship, and truck. The unlabeled images comprise a similar but broader set of objects (than the labeled ones).

This dataset is freely available.

3.3.16. PRID 2011

The PRID 2011 dataset [63] was built to be employed in CV research on person (re)identification.

This dataset consists of a set of images extracted from two video sequences acquired from two different, static surveillance cameras. It comprises images of multiple person trajectories. A total of 475 person trajectories were acquired from one view and 856 from the other one, with 245 persons appearing in both views. Each trajectory comprises between 100 and 150 images, depending on the walking speed of an individual.

The dataset's metadata consist of the individual enclosing bounding boxes.

The PRID 2011 dataset is freely available.

3.3.17. CUB-200-2011

The CUB-200-2011 dataset [64] (which builds on the CUB-200 dataset) was built for the development of CV tools for bird detection and identification.

This dataset comprises 11,788 images of 200 bird species (mostly, North American birds). The images were obtained using Flickr image search and then filtered with the assistance of multiple users of Mechanical Turk

This dataset's metadata consist of the definition (for each image) of bounding boxes, part locations (15 parts, annotated by pixel location and visibility), and attribute labels (28 attribute groupings and 312 binary attributes).

The dataset's contents are freely available for research purposes.

3.3.18. Semantic Boundaries Dataset

The Semantic Boundaries Dataset (SBD) [65] was built for employment in research on semantic contours prediction (of objects in images), as opposed to semantic segmentations (prediction of areas).

This dataset comprises 11,355 images taken from the PASCAL Visual Object Classes (VOC) 2011 dataset.

The dataset's metadata comprise the definition of the boundaries of over 20k object instances belonging to 20 object categories (selected from those of the VOC 2011 challenge). To produce these annotations binary figure-ground segmentations were collected for all the objects and categories in the images of the trainval set of the above stated dataset. To obtain precise boundaries, the cropped bounding box of each instance was rescaled to a size of 500 × 500 pixels and presented to human observers (through the Amazon Mechanical Turk), who then defined the object boundary by marking vertices of a polygon. An average of five annotations were made by different subjects per object instance.

This dataset is freely available for research.

3.3.19. Stanford Dogs Dataset

The Stanford Dogs dataset [66] was originally built for the development of research on, fine-grained, dog image categorization.

This dataset was built employing images and annotations obtained from ImageNet. It comprises 20,580 images of dogs belonging to 120 different breeds from around the world.

The dataset's metadata comprise class labels (with the dog breeds) and bounding boxes (enclosing the animals) for all images. This information is expressed in a custom defined xml format (one metadata file per image).

The contents of this dataset, following from ImageNet, are freely available for non-commercial research and/or educational uses.

3.3.20. Pascal Visual Object Classes Project

The PASCAL VOC Project (Pascal VOC) ran a research challenge from 2005 to 2012, in the area of object detection, classification, and segmentation. For the purposes of those challenges, it maintained a dataset [6] whose contents were progressively expanded with time.

In its most mature version, this dataset comprises a set of 11,530 images of images and their associated annotations (27,450 ROI annotated objects and 6929 segmentations). For each image, its corresponding annotation file describes a bounding box and an object class label for each object, belonging to one of twenty classes, that happens to be present in the image.

The object classes available for labeling are person; animal (bird, cat, cow, dog, horse, sheep); vehicle (airplane, bicycle, boat, bus, car, motorbike, train); and indoor (bottle, chair, dining table, potted plant, sofa, tv/monitor)

A subset of the images is also annotated with pixel-wise segmentation for each observable object. Another set of the images (to be employed for action classification), are partially annotated with people localizations (bounding boxes), reference points, and their actions. A further set of the images (for the person layout taster) has additional annotation describing parts of the people (head/hands/feet).

Contributions to the dataset were effected via the Pascal conference. The VOC2012 dataset is freely available but it includes images obtained from the Flickr, which are available under the Flickr terms of use.

3.3.21. NYU Depth Dataset (v2)

NYU Depth Dataset Version 2 (which builds on version 1) [67] is made available by Nathan Silberman at NYU for research on object detection and image segmentation through the detection of support relationships between visible objects and surfaces in RGB images.

This dataset's media component consists of around 500K RGB-D images of indoor scenes (obtained with Kinect), about 1500 of which are densely annotated. A total of 464 different indoor scenes were captured, belonging to 26 scene types with the identification of 35,064 objects belonging to 894 different classes.

The dataset's metadata comprise a pixel-wise object labeling of the images (i.e., each pixel in the image is attributed to an object).

The annotated component of the dataset is made available as a single *.mat* file. The NYU Depth Dataset is freely available.

3.3.22. Leafsnap Dataset

The Leafsnap Dataset [68] was built to enable the development of CV applications for automated plant species identification by way of leaf image recognition.

Leafsnap's images were acquired from two different sources: 23,147 of them were obtained from the Smithsonian collection. Thus, these are high-quality images of pressed leaves that were taken in controlled backlit and front-lit version (several samples per species); 7719 of them are "field images", which are typically acquired with mobile devices in outdoor environments. They contain varying amounts of blur, noise, illumination patterns, shadows, etc.

This dataset's metadata consist of the species identification (Leafsnap currently covers 185 tree species from the northeastern USA) and the segmented version of the image for each of the original images. The segmentations were automatically generated.

This dataset is freely available.

3.3.23. Oxford-IIIT Pet Dataset

The Oxford-IIIT Pet dataset [69] was built by the Visual Geometry Group at Oxford for research on object (pet) visual detection and location.

This dataset comprises 7349 images of 37 breeds of cats and dogs (12 cat breeds and 25 dog breeds), with roughly 200 images for each such breed. Those images have large variations in scale, pose, and lighting.

The dataset's metadata comprise (for each image) the animal's breed label, a tight bounding box around the head, and a pixel level segmentation marking the body. The segmentation divides images into three regions: foreground (the pet's body), background, and ambiguous (the pet's body boundary and any accessory such as collars).

The dataset's images were downloaded from Catster [70] and Dogster [71] (two social web sites dedicated to pets) as well as from Flickr groups, and Google images. About 2000 to 2500 images were originally downloaded for each of the 37 breeds, which were then were reviewed and filtered by humans.

This dataset is freely available.

3.3.24. LISA Traffic Sign Dataset

The LISA Traffic Sign Dataset [72] was built for the development of CV provisions for US traffic sign detection.

This dataset comprises 6610 images (640×480 to 1024×522) acquired in US locations, containing captures of 49 different types of US traffic signs. These images are frames extracted from videos. The dataset also includes those videos, and the images can be traced back to the latter.

The dataset's metadata (available for all the 6610 images) consist of 7855 traffic sign delimiting bounding boxes and associated traffic sign identifying labels.

The LISA Traffic Sign Dataset is freely available for research.

3.3.25. Daimler Urban Segmentation Dataset

The Daimler Urban Segmentation Dataset [73] (DUSD) is meant to be employed on image segmentation through a combination of appearance (grayscale images) and depth cues (dense stereo vision).

This dataset comprises 5000 rectified stereo image pairs (frames extracted from video sequences recorded in urban traffic) with a resolution of 1024×440 pixels.

The dataset's metadata comprise, for one-tenth of those frames (i.e., 500 frames), pixel-level semantic class annotations (manually labeled pixel-accurate) into 5 possible classes: ground, building, vehicle, pedestrian, sky. A rough estimate of the number of individual segments defined in the dataset (based on the similarities it has with CityScapes [74]), places this value at around 25,000. This metadata includes also dense disparity maps. However, the latter are not manually annotated but computed using semi-global matching.

This dataset is freely available for research.

3.3.26. Stanford Cars Dataset

The Stanford Cars Dataset [75] was built to enable the development of CV provisions for 2D object representation interpretation into the construction of their 3D dimensions.

It contains 16,185 images of 196 classes of cars.

This dataset's metadata consist of the labels (with the car class) and bounding boxes (surrounding the cars) for all the dataset's images. Classes typically consist of make, model and year. This information is inscribed into *.mat* files.

The dataset's construction begun with the acquisition of candidate images, for each of the 196 classes, from Flickr, Google, and Bing. The images were then put though Amazon Mechanical Turk for an initial class verification. The remaining annotation process (bounding-box definition and label attribution) was done through crowdsourcing.

The data are split into 8144 training images and 8041 testing images.

This dataset is freely available.

3.3.27. FGVC-Aircraft Benchmark Dataset

The Fine-Grained Visual Classification of Aircraft (FGVC-Aircraft) Dataset [76] is a benchmark dataset for the development of tools for a fine-grained visual categorization of aircraft.

This dataset contains 10,200 images of aircraft, for 102 different aircraft model variants (100 images for each variant). Most such aircraft variants are airplanes.

FGVC-Aircraft's metadata comprise (for each image) a tight bounding box, around the (main) aircraft, and a hierarchical airplane model label. Aircraft models are organized in a four-level hierarchy consisting of model, variant (102 possibilities); family (70 possibilities); and manufacturer (41 possibilities). This information was expressed in an *ad hoc* format in *.txt* files.

About 70,000 images were downloaded from Airliners.net. The 102 most frequent aircraft variants were retained, resulting in the above mentioned 100 images per variant. The average image resolution is between 1 and 2 mega pixels. The bounding-box annotations were crowdsourced using Amazon Mechanical Turk.

The above data are divided into three equal-sized training, validation and test subsets.

This dataset is freely available for non-commercial research purposes only.

3.3.28. Microsoft Research Dense Visual Annotation Corpus

Microsoft Research's Dense Visual Annotation Corpus [77] (MS DVAC) is a dataset meant for employment in CV research on object recognition in static images.

It comprises 500 images (from Flickr 8K).

This dataset's metadata consist of a vast amount of bounding boxes and facets, provided, for each object in each image, and 100,000 textual labels pertaining to 4000 objects (produced through crowdsourcing with the employment of Amazon's Mechanical Turk).

Both the media and metadata components of this dataset are retrievable separately but in bulk.

MS DVAC is freely available for research.

3.3.29. Microsoft Common Objects in Context

The Microsoft Common Objects in Context dataset (MS COCO) [78] was built for employment in research on object detection and segmentation, and scene understanding.

It comprises over 328,000 images of complex everyday scenes (obtained from Flickr), containing common objects in their natural context.

Its metadata comprise 2.5 million labeled instances (produced through crowdsourcing with the employment of Amazon's Mechanical Turk) pertaining to the detection of 91 objects types. Segments are individually defined in a pixel-wise manner (as explained in Figure 4).

MS COCO is available for research purposes.

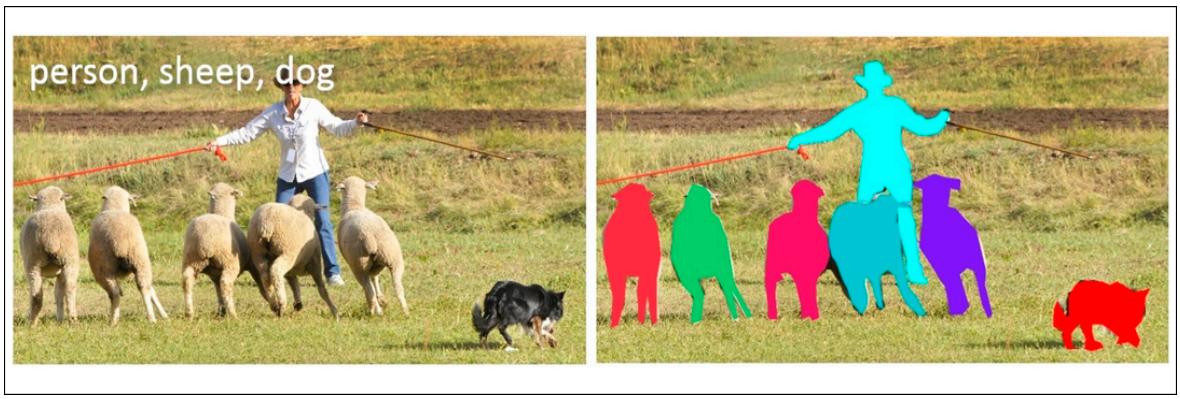

Figure 4. COCO Image Segmentation Example (obtained from [79]).

3.3.30. The Middlebury Datasets

The Middlebury datasets [80] (MDs) include a set of datasets that have been progressively accumulated since 2001, focusing on image understanding based on stereo imagery. Here, we shall focus on the latest of such datasets, as it is the broadest and most up to date.

The 2014 Middlebury Stereo dataset includes 33 different blocks of data. Each such block comprises images acquired from two different views, of a specific scene, under four different illuminations and eight different exposure conditions. Each block comprises also the corresponding, subpixel-accurate, GT metadata (disparity maps). These bocks are divided as 10 test blocks (with no GT); 10 training blocks with GT; and 13 additional blocks with GT.

This dataset is freely available.

3.3.31. Flickr30k

The Flickr30k dataset [81] (which comprehends and extends the Flickr8k [82] dataset) was built to enable CV research on image understanding (and description generation).

This dataset comprises 31,783 photographs of everyday activities, events, and scenes (harvested from Flickr). Its metadata comprise 158,915 captions (obtained via the crowd-sourcing services provided by Amazon Mechanical Turk).

The dataset is freely available for non-commercial research and/or educational purposes.

3.3.32. iLIDS-VID

The iLIDS-VID dataset [83] was built for research on person (re)identification based on image sequences.

It comprises 600 image sequences depicting 300 randomly sampled people, with one pair of image sequences from two camera views for each person. These were created based on two non-overlapping camera views from the iLIDS Multiple-Camera Tracking Scenario, which was captured at an airport arrival hall by a multi-camera CCTV network. These image sequences present variable length (from 23 to 192 frames) with an average of 73 frames.

The dataset's metadata comprise the ground-truth information, which consists of the bounding boxes circumscribing the visualized person and a range of other aspects. This metadata are expressed as a ViPER compliant XML construct.

Benchmarked training/testing splits are also provided.

The dataset is freely available for research purposes only

3.3.33. Belgium Traffic Sign Dataset

The Belgium Traffic Sign (BelgiumTS) Dataset [84] was built to be employed in traffic sign recognition research.

This dataset comprises over 145k (1628×1236) images, acquired on Belgian roads (with a car-mounted, eight-camera apparatus). The dataset's metadata include over 13,444 traffic sign annotations for more than 9006 images, corresponding to 4565 physically distinct traffic signs (with an average of three views/annotations for each physical traffic sign) visible at less than 50 m from the camera. The annotations consist of traffic sign delimiting bounding boxes; camera IDs: and camera poses.

This dataset is freely available for research.

3.3.34. Pascal Context

The Pascal Context dataset [85] is meant to be employed in CV research on image segmentation and object detection.

This dataset builds upon the trainval dataset provided for the PASCAL VOC 2010 detection challenge. Thus, it comprises the 10,103 images of that dataset (as its own training and validation dataset) and a further 9637 images for the testing dataset. Its metadata consist of the pixel-wise segmentation of the training/validation images and the labeling of each such segment according to one of 540 possible categories (each image contains, on

average, 12 segments). These categories fall under three main classes: objects, stuff (e.g., sky); hybrids. Annotation was performed by six in-house annotators.

This dataset goes beyond the original PASCAL dataset as it provides semantic segmentation and labeling for the whole scene in every image.

This dataset is freely available for research.

3.3.35. Cityscapes Dataset

The Cityscapes Dataset [74] was built to enable the development of vision algorithms for semantic urban scene understanding including scene labeling, instance-level scene labeling, and object detection.

It comprises a large and diverse set of stereo video sequences (containing street scenes from 50 different cities) and an extracted subset of their frames. Such frames are the annotated ones and their preceding and trailing ones.

This dataset's metadata comprise high-quality dense pixel annotations for 5000 frames and coarser polygonal annotations for another 20,000 images as shown in Figure 5. Half of the annotated frames are extracted from long video sequences, while the rest are the 20th of 30-frame video snippets (1.8 s long videos). These annotations segment the images into the different observable objects/spaces from a set of object/space classes: flat (road, sidewalk, parking, rail track); human (person, rider); vehicle (car, truck, bus, on rails, motorcycle, bicycle, caravan, trailer); construction (building, wall, fence, guard rail, bridge, tunnel); object (pole, pole group, traffic sign, traffic light); nature (vegetation, terrain); sky (sky); void (ground, dynamic, static).

The Cityscapes' metadata include also (for each video) precomputed depth maps, the GPS coordinates of the video acquisition processes, ego-motion data from vehicle odometry, and measurements of the environment temperature at the video acquisition location.

The Cityscapes Dataset is freely available to academic and non-academic entities for non-commercial purposes.

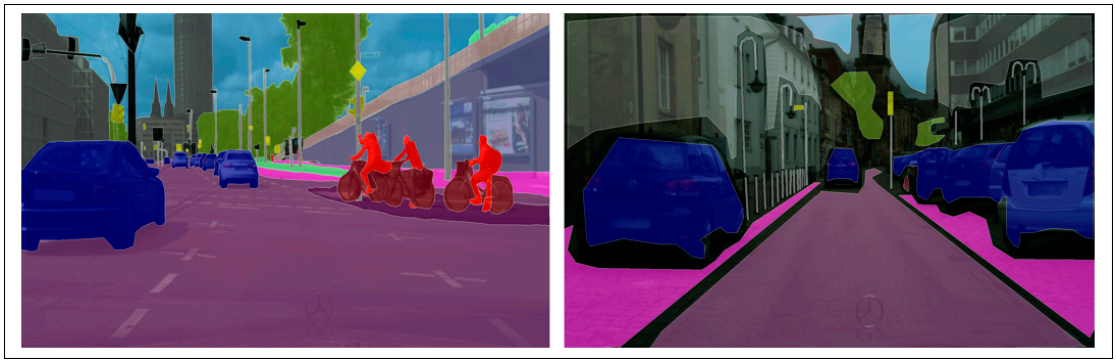

Figure 5. Finely (**left**) and coarsely (**right**) annotated images (partially obtained from [74]).

3.3.36. Comprehensive Cars Dataset

The Comprehensive Cars (CompCars) dataset [86] was created by the MMLAB of the Chinese University of Hong Kong to enable the development of CV tools for car model verification, fine-grained car classification, and car attribute prediction.

This dataset comprises both car images acquired from the web and those obtained from a surveillance context. Images acquired from the web total 164,344. Of these, 136,726 capture the entire car, while 27,618 of them capture car parts (headlight, taillight, fog light, air intake, console, steering wheel, dashboard, and gear lever). These later images are roughly aligned for the convenience of further analysis. Images acquired from a surveillance context total 50,000, and all capture frontal car views.

This dataset's metadata comprise for most of the web context images the viewpoint (front, rear, side, front side, and rear side), car make and model, and five car attributes (maximum speed, displacement, number of doors, number of seats, and type of car).

The images acquired from the web contain 163 car makes and 1716 car models; for the surveillance images—bounding box, car make and model, and car color. These metadata are expressed in an ad hoc format in *.txt* files.

The CompCars dataset' is available for non-commercial research purposes only. Its images were obtained from the Internet and are thus are not the property of MMLAB.

3.3.37. YouTube-8M Dataset(s)

The YouTube-8M is a large-scale labeled video dataset maintained by Google [87] for the development of provisions for automated video understanding.

This dataset holds 8 million (6.1 million, after the 2018 clean-up) YouTube video IDs (representing a total of 350,000 h of video).

Its metadata comprise precomputed audio-visual features from billions of the frames and audio segments of those videos. The visual features were extracted using the Inception-V3 image annotation model trained on ImageNet. Audio features were extracted using a VGG-inspired acoustic model.

These metadata comprise also video-level labels, which are the main themes of each video (assessed by a YouTube video annotation system using content, metadata, contextual, and user signals). The label vocabulary consists of 3862 Knowledge Graph entities. Each such entity is observable in at least 200 videos (with 3552 training videos per entity on average). The average number of video-level labels per video is 3.01.

The YouTube-8M Segments dataset is an extension of the YouTube-8M dataset with segment annotations, with the aim being to temporally localize the entities in the videos. Thus, it comprises human-verified labels on about 237,000 segments pertaining to 1000 classes (a subset of the classes employed for the YouTube-8M dataset vocabulary).

The dataset is split into three partitions: training (70%); validation (20%); and testing (10%). Features are published for all splits. However, labels are published only for the training and validation partitions.

Both datasets are made available by Google under a Creative Commons Attribution 4.0 International license.

3.3.38. Densely Annotated Video Segmentation

The Densely Annotated Video Segmentation (DAVIS) [88] dataset was constructed to facilitate research on object segmentation in video footage.

This dataset comprises 50 video sequences, captured at 24 fps (frames per second) and full HD with 1080p spatial resolution. Each sequence has a short temporal extent (about 2–4 s) but spans multiple occurrences of common video object segmentation challenges such as occlusions, motion blur, and appearance changes. Each such sequence contains at least one target foreground object, which is separable from the background regions, or two spatially connected objects. The total amount of the dataset's images is about 3600.

The dataset's metadata (manually created and provided for 3455 frames) consist of densely annotated, pixel-accurate and per-frame ground truth segmentation information in the form of binary masks, pertaining to four evenly distributed classes (humans, animals, vehicles, objects) and several actions. It is provided for each sequence.

This dataset was released under the BSD License.

3.3.39. iNaturalist

The iNaturalist dataset [89] was produced for the visual identification of living species.

This dataset's media content comprises a total of 859,000 images from over 5000 different species of plants and animals. This consists of 579,184 images for the training set and 95,986 images for the validation set. iNaturalist's metadata consist of the label (observed species, date, and location) for all species and the bounding box surrounding the identified specimen for a subset of the images. A total of 561,767 bounding boxes were created for 2854 different classes of plants and animals. The annotation format employed (for both the

image labels and bounding boxes) closely follows that of the COCO dataset (annotations are stored in the JSON format).

The contents of this dataset were produced in the context of the iNaturalist citizen science effort. This enables volunteer naturalists to map and share their biodiversity observations from across the globe through a custom-made web portal and mobile apps.

The iNaturalist dataset is available for free for non-commercial research and educational purposes.

3.3.40. Visual Genome

The Visual Genome dataset [90] was built to enable the development of CV provisions capable of attaining a more complex visual scenario understanding.

This dataset comprises 108,249 images obtained from the intersection of MS-COCO's 328k images and Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M)'s [91] 100 million images. All such images were obtained from Flickr and range in width from 72 to 1280 pixels, with an average width of 500 pixels.

The Visual Genome dataset's metadata consist of seven main components: region descriptions, objects, attributes, relationships, region graphs, scene graphs, and question–answer pairs. Each image in the dataset is annotated, on average, with 42 region descriptions (and a directed graph representation for each of the 42 regions). Such a region is defined by a bounding box and a descriptive phrase. Those descriptive phrases range from one to 16 words in length (the average length is five words). On average, each image (in the context of the mentioned region descriptions) has annotations for 21 objects and 18 attributes.

Close to 45% of the defined objects in the dataset's metadata are further characterized with at least one attribute. Overall, the dataset's metadata comprise 16 million total attributes with 13,041 unique ones.

The metadata in scope comprise also the definition of 13,894 unique relationships (between objects), with over 1.8 million total relationships described. They also include 1,773,258 question answering (QA) pairs regarding the content of the dataset's images. Each such pair includes a question and its correct answer. On average, every image has 17 QA pairs associated to it.

All objects, attributes, relationships, and noun phrases in region descriptions and questions–answer pairs were canonicalized to WordNet synsets.

Visual Genome metadata were collected and verified entirely by crowd workers from Amazon Mechanical Turk.

This dataset is licensed under a Creative Commons Attribution 4.0 International License.

3.3.41. Open Images Dataset (v5)

The Open Images Dataset [92,93] is the product of a collaboration between Google, CMU, and Cornell universities. It was built to enable research on image classification, object detection, and visual relationship detection at scale.

This dataset comprises over nine million images (more precisely, URLs to images) that have been annotated with labels. These categories cover more real-life entities than those of ImageNet.

Its metadata include 36.5 million image-level labels for 19,900 concepts; 15.4 million bounding boxes for 600 object classes in 1.9 million images, and 375,000 visual relationship annotations involving 57 classes. On average, each image has about 8 labels assigned and 8.4 boxed objects. It comprises also segmentation masks for 2.8 million object instances in 350 classes. Such masks describe the outline of objects, thus characterizing their spatial extent to a much higher level of detail.

The Open Images Dataset is split into a training set (9,011,219 images), a validation set (41,620 images), and a test set (125,436 images). The annotations in scope have been automatically produced (with a model similar to Google Cloud Vision API); however, the image-level labels for the validation and test sets, as well as part of the training set,

have been human-verified. Most such verifications were done by in-house annotators at Google, and a few were done by crowd-sourcing; 90% of the bounding boxes for the training set were manually drawn by professional annotators at Google, while for the validation and test sets, all boxes were manually drawn; segmentation masks, for the validation and test splits, were manually produced, while for the training split, they were produced by a state-of-the-art interactive segmentation process, where professional human annotators iteratively correct the output of a segmentation neural network; visual relationships between objects, for all splits of the dataset, were added manually.

The Open Images Dataset Extended [94], as its name implies, extends the Open Images Dataset. Thus, it complements the core Open Images Dataset with additional images and image-level annotations. It specifically adds 478K images annotated across 6000 categories. These images and annotations were contributed by global users of the Google Crowdsource Android app. Most such images focus on India, the Middle East, Africa, and Latin America, and on some key categories such as household objects, plants and animals, food, and people in various professions. Many of the donated images and annotations were also verified by human annotators at Google.

All above-mentioned images (which have been collected from Flickr) have a Creative Commons Attribution license that allows their sharing and adapting. The annotations are licensed by Google LLC under CC BY 4.0 license.

3.3.42. YouTube-Bounding Boxes

The YouTube-Bounding Boxes (YouTube-BB) dataset [95] was built at Google to enable the advancement of the state of the in CV for video understanding.

This is a vast video dataset with densely sampled, high-quality, single-object bounding-box annotations. It comprises around 380,000, 15 to 20 seconds long, video segments (extracted from 240,000 different publicly available YouTube videos, of a quality like that of cell phone obtained video). These segments were selected to feature objects in natural settings without editing or post-processing.

The dataset's metadata consist of object classification information and object locating bounding boxes. This information was added to all videos at a rate of 1 frame per second. Thus, YouTube-BB comprises 10.5 million classification annotations (pertaining to 23 different classes of objects, which are a subset of the COCO classes) and 5.6 million tight bounding boxes around tracked objects, on video frames. YouTube-BB's metadata (which is stored in simple *.csv* files) is human-curated, and it was produced by crowdsourcing through Amazon Mechanical Turk. The attained accuracy for every object classification and bounding boxes placement is above 95%.

This dataset is licensed by Google Inc under a Creative Commons Attribution 4.0 International License.

3.4. Object Tracking Datasets

3.4.1. Human Eva

The Human Eva dataset [96] was built to aid in the development of visual human pose estimation and tracking provisions.

This dataset's sensory (video) and GT data (motion data) were captured simultaneously using multiple high-speed video capture systems and a calibrated marker-based motion capture system, respectively. Specifically, the dataset comprises seven calibrated video sequences (four grayscale and three color) synchronized with 3D body pose information (GT metadata). The video sequences capture four subjects performing a set of six predefined actions (e.g., walking, jogging, gesturing, etc.) three times. In total, the Human Eva dataset includes 50,000 frames (640 $\times$ 480 resolution) of synchronized video information along with their respective motion capture ground-truth data, which was collected at 60 Hz.

This dataset is partitioned into training, validation, and testing subsets. It comprises also support software for manipulating the dataset's contents and evaluating results.

The contents of this dataset are available free of charge for research purposes only.

3.4.2. ETH

The ETH dataset [97] was developed for employment in CV research on pedestrian tracking—specifically, simultaneous pedestrian detection and ground-plane estimation from video data.

The dataset comprises a total of 2293 video frames (with a 640 × 480 resolution) from four video sequences and 10,958 pedestrian annotations (bounding boxes). Said annotations are defined in a plaintext format in a single file per sequence.

This dataset is divided into training and testing subsets. The training subset comprises a video sequence (490 frames) acquired (at 15 fps) with a stereo pair of cameras mounted on a children's stroller. Its metadata consist of 1578 annotations with pedestrian detections. The testing subset comprises three video sequences and associated metadata. The first test sequence includes 999 frames and 5193 annotations. The second test sequence (which shows a stroll over a busy square) comprises 450 frames and 2359 annotations. The third test sequence (taken on a sunny day on a sidewalk) is composed of 354 frames and has 1828 annotations.

The ETH dataset is freely available for the research community.

3.4.3. Daimler Pedestrian Detection and Tracking Dataset

The Daimler Pedestrian Detection Benchmark dataset [98] was built for CV research on pedestrian detection.

It comprises training and testing subsets. The earlier includes 6744 full images containing no pedestrians and 15,660 cut-outs of pedestrians produced by manually extracting 3915 rectangular position labels (bounding boxes) from video images (i.e., the datasets's metadata). From each label, four pedestrian samples were created by mirroring and randomly shifting the bounding boxes by a few pixels in horizontal and vertical directions. The latter subset includes 21,790 full images (640 × 480 pixels) with 56,492 manual labels, including 259 trajectories of fully visible pedestrians. These were acquired as the frames of a 27-min-long video taken from a vehicle driving through urban traffic.

The dataset is made freely available to academic and non-academic entities for research purposes.

3.4.4. TUD

The TUD dataset [99] was developed for research on pedestrian detection and tracking.

It comprises testing and training subsets. The earlier (TUD-Brussels) was acquired from a driving car in the inner city of Brussels. It includes 508 image pairs (one pair per second and its successor of the original video) at a resolution of 640 × 480, annotated with a total of 1326 pedestrian detections (each described by a bounding box).

The latter (TUD-MotionPairs) includes a positive and a negative training subset. The earlier (acquired with a handheld camera at a resolution of 720 × 576 pixels), comprises 1092 image pairs annotated with 1776 pedestrian detections (resulting in 3552 positive samples with mirroring). The latter includes 192 image pairs. Eighty-five such pairs were captured using a handheld camera at a resolution of 720 × 576 pixels, while another 107 such pairs were recorded from a moving car.

The TUD dataset is freely available.

3.4.5. Caltech Pedestrian Database

The Caltech Pedestrian Database [100] was built to enable CV research on pedestrian detection.

This dataset comprises approximately 10 h of (30 Hz) video footage (about 1M frames) taken from a vehicle driving through regular traffic in an urban environment.

The dataset's metadata comprise annotations for about 250,000 frames (in 137, approximately, minute long segments), in 132,000 of which pedestrians are visible. These annotations consist of 350,000 bounding boxes (describing pedestrian detections in frames).

Each bounding box (BB) tightly delimits a pedestrian. When occlusion occurs, the BB delimits the estimated full extent of the pedestrian's body, and an additional BB is provided for the visible part of the pedestrian. The labels assigned to BBs may be of three different types: Person, for individual pedestrians; People, for large groups of pedestrians; and Person?, when a clear identification of a pedestrian is ambiguous or easily mistaken.

This dataset's metadata also include information describing the temporal correspondence between bounding boxes and detailed occlusion labels.

Over 2000 individual pedestrians are detected in this dataset.

This dataset is freely available for research.

3.4.6. KITTI Benchmark Suite Dataset

The KITTI Benchmark Suite dataset [4,101] is part of a project maintained by Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago. It is meant for employment on CV research on object detection and tracking in the context of autonomous driving research.

The dataset's base media content was generated through the employment of the autonomous driving platform Annieway [102], which was driven through Karlsruhe, in rural areas and on highways. This enabled the generation of high-resolution color and grayscale video with real-world city driving scenes.

This dataset comprises 6 hours of footage of traffic scenarios (divided into 151 video sequences), taken at 10–100 Hz, using a variety of sensor modalities, including high-resolution color and grayscale stereo cameras; a Velodyne 3D laser scanner; and a high-precision GPS/IMU inertial navigation system. The acquired content is calibrated, synchronized, and timestamped. The raw dataset is divided into the following categories: City, Residential, Road, Campus, and Person (28, 21, 12, 10, and 80 video sequences respectively). In addition to the raw image sequences, the dataset also comprises post-processed sequences ("synced data"), i.e., rectified, and synchronized video streams.

The KITTI Benchmark Suite dataset comprises also metadata. These consist of spatial–temporal object labels in the form of 3D tracklets. This way, for each sequence (for each frame of each sequence), and for each dynamic object within the reference camera's field of view, a 3D bounding box is described in Velodyne coordinates. These bounding boxes register the object's class, its 3D size (height, width, length), its translation, and its rotation in 3D. The available classes for the detected objects are Car, Van, Truck, Pedestrian, Person (sitting), Cyclist, Tram, and Misc (e.g., trailers, segways). The accurateness of this ground truth information is assured by the employment of the Velodyne laser scanner and the GPS localization system. This information is stored in an XML file (for each video sequence) with an ad hoc format defined for the purpose at hand.

This dataset registers also, for each frame, 30 different GPS/IMU values. These describe the geographic coordinates including altitude, global orientation, velocities, accelerations, angular rates, accuracies, and satellite information.

The dataset also includes information pertaining to optical flow and visual odometry.

The KITTI Benchmark Suite dataset is published under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License. Therefore, it may not be used for commercial purposes, and any derivative work must be distributed under the same license.

3.4.7. ALOV++

The ALOV++ (Amsterdam Library of Ordinary Videos for tracking) dataset [103] was built for CV research on object trackings with the aim of covering as diverse a set of circumstances as possible pertaining to illuminations, transparency, specularity, confusion with similar objects, clutter, occlusion, zoom, severe shape changes, different motion patterns, low contrast, etc.

It comprises a total of 315 video sequences (with a total of 89,364 frames). Eleven of these are standard video sequences frequently used in tracking research, and 65 other sequences have been reported earlier in the PETS workshop. The remaining 250 are new.

They are mostly real-life videos from YouTube, their content pertains to 64 different types of targets (human face, a person, a ball, an octopus, microscopic cells, a plastic bag, or a can), and their length may vary between 9.2 s and 2 min.

The dataset's ground-truth metadata consist of rectangular bounding boxes (and their labels). These were typically added to every fifth frame, and the ones for the remaining frames were acquired by linear interpolation.

The ALOV++ dataset is freely available.

3.4.8. Visual Tracker Benchmark Dataset

The Visual Tracker Benchmark (VTB) [104] dataset (which builds on [105]) provides the means for research on visual tracking.

This dataset comprises 100 video sequences (obtained from research work from the same context), which may present nine different types of attributes (illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutters, low resolution). All sequences are split into their composing frames.

This dataset's metadata consist of bounding boxes locating the target object of every sequence, in every one of its frames. This information (for each sequence) is expressed employing a very simple ad hoc format and contained within a *.txt* file.

The VTB is freely available.

3.4.9. TColor-128

The TColor-128 dataset [106] was built for CV research object tracking with a focus on taking advantage of color information.

It comprises 128 color video sequences along with ground-truth metadata. Fifty of the video sequences were obtained from previous studies. The remaining 78 were newly collected from the Internet. The dataset's video sequences pertain to a variety of contexts (highway, airport terminal, railway station, concert) and present many challenging factors (full target occlusion, large illumination change, significant target deformation, and low resolution)

TColor-128's metadata consist of bounding boxes that circumscribe the identified object in each frame and also a description of the challenge factors in each sequence. These metadata are expressed in a simple ad hoc format stored in *.txt* files.

This dataset is freely available for research purposes.

3.4.10. NUS-PRO

The NUS-PRO database [107] was built for the development of CV research on object tracking.

This dataset comprises 365 image sequences collected from YouTube. As these were acquired by handheld cameras, they contain sudden object movements caused by hand movement. Their average length is 300 frames (1280×720), but it may vary between 146 and 5040 frames. Such sequences may belong to five different categories (face, pedestrian, sports, rigid object, and long sequences).

NUS-PRO's metadata consist of bounding boxes locating the target object in a specific frame. In this regard, and depending on the type of targeted object, 220 of the sequences (pedestrian, helicopter (rigid object), basketball, gymnastics, racing, soccer, tennis (sports)) are annotated by torso-based bounding boxes, while the remaining are labeled by boundary-based bounding boxes. The bounding boxes for each target comprise its full extent, including both the visible and invisible (inferred) parts. The dataset also provides the occlusion level of the target object in each bounding box. This may be classified into three categories: no occlusion, partial occlusion, and full occlusion. The dataset comprises also foreground masks.

The bounding boxes and occlusion level annotations are provided for the first frame of each sequence. The foreground masks of non-face objects, and the fiducial points of face

images, in the first frames of each sequence are also provided. For 73 of the sequences, the complete bounding boxes, foreground masks, and occlusion level annotations are provided.

The NUS-PRO database is freely available.

3.4.11. UAV123

The UAV123 dataset [108] was produced to be employed in CV research on target tracking from low altitude UAV footage.

It comprises 123 new, and fully annotated, HD video sequences (with a total of over 110,000 frames and an average duration of 30 s) captured from a low-altitude aerial perspective. These sequences have a minimum of 109 frames and a maximum of 3085 frames. On average, the acquired videos comprise 915 frames.

The UAV123 dataset comprises three different subsets: set 1 includes 103 sequences captured with an off-the-shelf professional-grade UAV (DJI S1000), which follow different objects at altitudes varying from 5 to 25 m. All 103 sequences were acquired with 720p and at 30 fps. They are annotated with upright bounding boxes also at 30 fps. Annotation was done manually at 10 fps and then linearly interpolated to 30 fps; set 2 comprises 12 sequences, acquired with a boardcam (with no image stabilization), on-board a small low-cost UAV following other UAVs. These 12 sequences have lower quality and resolution and present considerable noise. Annotation for set 2 was built in the same way as for set 1; set 3 contains eight synthetically produced sequences. Annotation was automatically produced at 30 fps, and it also includes a full object mask/segmentation.

Some of the tracked objects types are cars, trucks, boats, persons, groups, and aerial vehicles.

The UAV123 dataset is available under request.

3.4.12. VOT Challenge Dataset

Throughout its various editions, the VOT challenge [109] has built a broad dataset and a precisely defined and consistent benchmark for the field of visual tracking.

In its latest version (2019), this dataset comprises the following sub-datasets: VOT-ST (built for research on short-term tracking in RGB images); VOT-RT (built for research on "real-time" short-term tracking in RGB images; VOT-LT (built for research on long-term tracking, specifically dealing with target disappearance and reappearance); VOT-RGBT (built for research on short-term tracking in RGB and thermal imagery); and VOT-RGBD (built for research on long-term tracking in RGB and depth imagery).

The VOT-ST subset contains 60 public sequences (12 of those obtained from [110]), with a total of 101,956 frames. It comprises also (manually created) segmentation masks for tracking targets in all frames of the sequences, and rotated bounding boxes were fitted to these segmentation masks. Each frame is also (semi-automatically) annotated with visual attributes such as occlusion; illumination change; motion change; size change; and camera motion.

The VOT-RT subset is presented for logical purposes (given the different challenges within the competition). It is the same as the VOT-ST.

The VOT-LT subset is the same as in [111]. It comprises 50 challenging sequences of diverse objects with the total length of 215,294 frames. On average, each sequence contains 10 long-term target disappearances, each lasting on average 52 frames. All such sequences are annotated (per sequence, not per frame) with the following visual attributes: full occlusion; out-of-view; partial occlusion; camera motion; fast motion; scale change; aspect ratio change, viewpoint change, similar objects.

The VOT-RGBT subset comprises 60 sequences (an average length of 335 frames) obtained from [112]. All frames of all sequences have been annotated with the visual attributes: occlusion; motion change; size change; camera motion. They have also been annotated with (semi-automatically generated) segmentation masks and rotated bounding boxes.

The VOT-RGBD subset is the Color and Depth Visual Object Tracking Dataset and Benchmark (CDTB) dataset from [113]. It comprises 80 sequences acquired with three

different setups: a Kinect v2 RGBD sensor; a pair of Time-of-Flight (Basler tof640) and an RGB camera (Basler acA1920); and a stereo pair (Basler acA1920). Kinect was used for 12 indoor sequences, the Time-of-Flight pair was used in 58 indoor sequences, and the stereo pair in 10 outdoor sequences. This subset comprises also (for all sequences) aligned RGB frames and dense depth frames. In terms of metadata, this dataset contains tracking boxes for various household and office objects. The total number of its frames is 101,956 in various resolutions.

All bounding-box metadata are expressed as simple sets of eight value lines in a *.txt* file for each sequence.

The VOT Challenge also provides a toolkit for the manipulation of, and experimentation with, its datasets. All of VOT's contents are freely accessible through the VOT toolkit, under various different licenses.

3.5. Activity and Behavior Detection Datasets

3.5.1. CAVIAR Project Benchmark Datasets

Project CAVIAR maintains a dataset [114] meant for employment in research on people detection and tracking and event detection.

CAVIAR comprises 52 videos (of scripted and real-life activities) divided into two subsets (which were acquired) in two different locations: the entrance lobby of the INRIA Labs at Grenoble (six different scenarios acted out by the CAVIAR team members); a shopping center in Lisbon (real and scripted events taking place in a corridor and at a shop entrance).

The ground-truth metadata (expressed in XML, employing CVML [115]) consist of bounding boxes describing the positions of people and objects in the frames, and of the description of observed events, such as walking, browsing, meeting, fighting, window shopping, entering/exiting stores, etc. In some cases, information is also provided regarding the bodily positions of observed people, such as head position, gaze direction, or hand, feet, and shoulder positions.

The CAVIAR dataset is freely available.

3.5.2. KTH Dataset

The KTH dataset [116,117] (maintained by the KTH Royal Institute of Technology) is meant to be employed in CV research on human activity recognition from video.

It comprises 2391 video sequences acquired with a static camera (at 25 fps with 160×120) over a homogeneous background, which captures six types of human actions performed several times by 25 people in four different scenarios (overall set of 600 videos). These sequences have an average duration of 4 seconds.

The dataset's metadata consist of the action labels and corresponding frame-spans and are provided as an ASCII file. The dataset is divided into six (individually retrievable) sections (one for each human action type), each comprising the respective videos and metadata file.

The KTH dataset is freely available.

3.5.3. WEIZMAN Dataset

The WEIZMANN dataset [118,119] (one of the first created) is meant to be employed in ML research on different, but related, aspects of video interpretation.

It comprises two sub-datasets: the Weizmman Event-Based Analysis dataset; and the Weizmann Actions as Space-Time Shapes dataset.

The earlier (for research on clustering and temporal segmentation of videos) comprises a single, long, sequence of approximately 6000 frames, displaying different people, wearing different clothes, and performing four types of activities. The annotation information is simply the description of the action observed at each frame.

The latter (for research on human action recognition from video) comprises about 90 videos (static viewpoint), grouped into 10, individually downloadable, sets (one for

each of the ten types of recorded action). Each such set includes about nine videos (each with a different person performing the same activity). The annotating information for this sub-dataset consists of the definition of the foreground silhouettes of each moving person (expressed in a single file in MATLAB format) and the background sequences used for background subtraction (retrievable only in bulk) as well as the identification of the performed activity.

This dataset is freely available.

3.5.4. ETISEO Dataset

The ETISEO dataset [120,121] (maintained by INRIA) is meant to be employed in CV research on human activity recognition on video.

This dataset comprises about 40 video sequences. These are divided into five distinct sections, one for each of the contexts of video acquisition (building, corridor, building entrance, metro, and road), and they may depict actions of 15 different types (walking, running, sitting, lying, crouching, holding, pushing, jumping, pick up, puts down, fighting, queueing, tailgating, meeting, and exchanging an object).

The dataset's metadata comprise (for each video sequence) three different types of data: ground truth (e.g., object bounding box, object class, event, etc.) produced by human operators using the ViPER format [122]; general data on the video sequences concerning video processing difficulties and recording conditions; and camera calibration data and contextual information describing the topology of the scene. The two latter types are expressed in the PETS format. The metadata for each video are stored in its separate file.

The repository holding the ETSIO dataset enables its video and metadata contents to be retrieved only in their entirety.

This dataset is available to the research community and for non-commercial use (on a case-by-case approval).

3.5.5. CASIA Action

The CASIA Action dataset [123] was created by the Center for Biometrics and Security Research (of the Chinese Academy of Sciences) for research on algorithms for the visual identification of human activities.

It comprises 1446 video sequences which were captured simultaneously with three static non-calibrated cameras from different viewing angles (horizontal, angle, and top down views), at a frame rate of 25 fps and a resolution of 320 $\times$ 240. The sequences last from 5 to 30 s. The above video sequences capture human activities (in an outdoor context) pertaining to eight types of actions of a single person (walk, run, bend, jump, crouch, faint, wander, and punching a car) each performed by 24 subjects; and seven types of two-person interactions (rob, fight, follow, follow and gather, meet and part, meet and gather, overtake) performed by every two subjects.

The dataset's metadata consist of the identification, for each sequence, of the action taking place and of the involved subject(s).

The dataset is freely available for scientific research and possibly for commercial use upon request.

3.5.6. HOHA

The Hollywood Human Actions (HOHA) dataset [124] was built for the conduction of research on human action recognition in realistic and unconstrained videos such as in feature films, sitcoms, or news segments.

The dataset comprises two video training subsets and as well as a video testing subset. One of the training subsets was manually annotated, and the other was annotated through the automated generation of annotations from movie scripts. The manually and automatically annotated training sets contain action video sequences from 12 movies (219 action samples and 231 labels for the earlier and 233 action samples with 143 labels

for the latter), and the test subset actions pertain to sequences from 20 different movies (211 action samples with 217 labels).

The annotations in scope describe the targeted human actions (belonging to eight possible types: AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, StandUp), which were observed in each video sequence and the frame range to which each such action pertains.

This dataset is freely available.

3.5.7. MSR Action

The MSR Action dataset [125], which builds on the KTH dataset, was created for research on human action recognition in video, with a focus on pattern matching-based action detection.

It comprises 16 video sequences (lasting from 32 to 76 s) depicting three types of actions: hand clapping (14 samples), hand waving (24 samples), and boxing (25 samples), which were performed by 10 people. Each such sequence contains multiple examples of such actions, and some contain actions performed by different people. There are both indoor and outdoor scenes.

The dataset's metadata consist of (manually produced) spatio-temporal bounding boxes defined for each action in each sequence.

This dataset is freely available for research.

3.5.8. HOLLYWOOD2

The HOLLYWOOD 2 dataset [126] (which builds on the HOHA dataset described in Section 3.5.6) aims at providing a comprehensive benchmark for human action recognition in realistic and challenging scenarios.

It comprises over 3669 video sequences, with approximately 18 h of video, depicting 12 classes of human actions and 10 classes of scenes. Each sequence may contain instances of several actions.

The dataset's metadata consist of the indication, for each sequence, if it contains each specific type of target activity. There is a file for each target activity type, which indicates, for every video sequence, if it depicts a specific activity or not. The employed annotation format is similar to that of PASCAL VOC for the image classification task.

The collection of the video sequences and their labeling was done by means of an automatic process of script-to-video alignment in combination with text-based script classification. The publicly available scripts of 69 movies were processed; then, specific actions and scenes were identified in the scripts and clipped from the video at the corresponding time interval.

Based on the earlier subset, another segment of the dataset was constructed, employing a manual verification of the labels.

This dataset is freely available.

3.5.9. i3DPost Multi-View (2009)

The i3DPost dataset [127] is a database of multiview video and 3D descriptions of human action/interaction. It is meant for employment in research on human action recognition from multiview videos or 3D posture model sequences.

The dataset's video content was acquired in a prepared indoor setting with a convergent eight-camera setup, acquiring synchronized videos at a 1920 $\times$ 1080 resolution and at 25 Hz rate. It comprises a total of 104 multiview videos, which corresponds to 832 (8 $\times$ 104) single-view videos. Each such video depicts one of eight people performing one of 13 different human actions (walking, running, jumping, bending, handwaving, jumping in place, sitting–stand up, running–falling, walking–sitting, running–jumping–walking, handshaking, pulling, and facial expressions). The facial expression multiview video depicts a person performing, sequentially, the six basic facial expressions separated by the neutral expression.

The acquired scenes involve one person performing a specific action; a person executing different actions in a succession; and two individuals interacting with each other.

The dataset's metadata consist of, for each video sequence, the identification of the person involved and of the activity taking place; a binary mask sequence, omitting all background, which is stored as video; and a 3D mesh, at each frame, describing the respective 3D human body surface.

The database is freely available for research purposes only.

3.5.10. BEHAVE Dataset

The BEHAVE dataset [128,129] (built by the University of Edinburgh's School of Informatics) is meant to be employed in CV research on behavior identification and analysis of interacting groups of people in video.

The dataset comprises various video sequences (over 90,000 frames) taken from two views of various groups of people having different interactions (ten specific types of interactions). The video content was acquired at 25 fps with a resolution of 640 × 480 and is available either as AVI encoded files or (for only one of the views) as a numbered set of JPEG single image files.

For all the sequences but one (for only one of the views), ground-truth information with the tracking of the observed individuals is available. This information is composed of bounding boxes (expressed in the ViPER XML format) enclosing each of the interacting pedestrians. The dataset's metadata include also the activity labels and their (frame delimited) ranges for all sequences. Both the video and metadata files are individually retrievable.

The BEHAVE dataset is freely available for research.

3.5.11. TV Human Interaction Dataset

The TV Human Interaction Dataset [130,131], maintained by the Visual Geometry Group of the Oxford University, is meant to be employed in CV research on the recognition of interactions between two people in videos.

This dataset comprises 300 video clips (ranging from 30 to 600 frames) collected from over 20 different TV shows, depicting four interaction types (50 clips for each), as well as 100 clips with no footage of any of such interaction types. It contains also (for every video) a metadata file describing (for every frame) the upper body of observable people (with a bounding box); their head orientation; and their interaction label. This information is expressed in a purpose designed plain text format.

This dataset is freely available for research purposes only.

3.5.12. MuHAVi Dataset

The MuHAVi dataset [132,133] (maintained by the Faculty of Science, Engineering, and Computing of Kingston University) is meant to be employed in CV research on silhouette-based human action recognition from multiview video.

This dataset comprises two blocks (added at different times). The earliest is divided into 17 individually retrievable sections (one for each covered human action type). Each section contains seven parts (corresponding to seven actors) each of which contains eight sub-parts (corresponding to eight cameras). These comprise the videos, which are split into individual frames, corresponding to their specific combination of action/actor/camera. The metadata of this dataset block are a set of manually produced annotations describing the silhouettes of the actors in frames (for only two actors and two camera views of five of the 17 action types). Each available combination of person/camera/action annotations may be retrieved individually.

The latest part of this dataset comprises uncut video (split into individual frames). It consists of eight sections, each comprising a continuous, individually retrievable, video file, from one of the cameras, which captures all of the actions (and also the gaps and breaks in between) from that point of view. The metadata (contained in a single spreadsheet file) consist of the description of start and end times (frame numbers) of each sub-action in each

video by each actor and the actor-delineating silhouettes for each of the videos. Each set of silhouettes is individually retrievable.

This dataset is freely available.

3.5.13. UT-Interaction (2010)

The UT-Interaction dataset [134] was built for the High-Level Human Interaction Recognition Challenge and has evolved with time. Its purpose is to foster the development of algorithms for the recognition of complex human activities from continuous videos taken in realistic settings.

This dataset comprises a total of 20 main video sequences whose average length is 1 minute. The videos were acquired with a resolution of 720 × 480 and a frame rate of 30 fps (the average height of a person in these video sequences is about 200 pixels).

Each main video sequence contains footage of several human–human interactions (occurring sequentially and/or concurrently), belonging to a set of six classes of such actions (shake hands, point, hug, push, kick, and punch). Each video contains at least one instance of each interaction type, providing an average of eight executions of human activities per video. Each such execution is composed of several atomic actions (each of which may be of 10 different types). Overall, the dataset comprises footage of 60 interactions and more than 180 atomic actions.

This dataset's metadata (condensed into a single XLS file) consist of the description of all observed activities. This includes (for each interaction or atomic action of each video) the activity label, its time interval, and the bounding boxes that circumscribe the activity area (defined once for the entire frame range of the activity) or relevant segments of it.

The UT-Interaction dataset is divided into 10 subsets. Each such subset contains videos of a pair of different persons performing all six interaction types. In subsets 1 to 4, only two interacting individuals are present. In subsets 5 to 8, both interacting persons and pedestrians are observable. In subsets 9 and 10, several pairs of interacting persons perform the activities simultaneously. Each subset has a different background, scale, and illumination.

This dataset is freely available.

3.5.14. Human Motion Database (HMDB51)

The HMDB51 dataset [135] is meant to be employed in research on human activity recognition from video data.

The dataset in scope comprises nearly 7000 video clips (obtained from various sources such as movies, public databases, and YouTube) divided into 51 activity categories (101 videos per category on average). Each video clip depicts a single action, the height of the main actor in it is at least 60 pixels long, and their minimum duration is 1 s.

The meta-information (manually produced annotations) for each video describes the action category to which it belongs. This may be general facial actions (smile, laugh, chew, talk); facial actions with object manipulation (smoke, eat, drink); general body movements (cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave); body movements with object interaction (brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw); and body movements for human interaction (fencing, hug, kick someone, kiss, punch, shake hands, sword fight).

Furthermore, the annotations also describe, for each video, the body parts (of the intervening humans) that are visible (the options being: head, upper body, full body, lower body); the number of people involved in the action (the available options are single, two, three); the camera motion status (mobile, static), and camera viewpoint (front, back, left, right).

This dataset is freely available under a CC Attribution 4.0 License.

3.5.15. VIRAT

The VIRAT dataset [136] is a large-scale, outdoor, surveillance video dataset. It was built for employment in research on visual event recognition.

This dataset comprises 29 h of video, which was collected from both stationary ground cameras and moving aerial vehicles, depicting 23 different event/activity types. These activities are grouped into the following super-types: single person event (8); person and vehicle events (7); and person and facility events (2).

The first part of the video content of the dataset consists of 25 h of stationary ground camera footage. This was acquired across 16 different scenes, amounting to an approximate average of 1.6 h of video per scene. Twenty-one of said video hours capture natural events, the remaining 4 hours capture staged activities (subset of four scenes).

The second part of the video content in scope includes 4 h of aerial video footage. The captured events are staged by hired actors and vehicles at a designated site. Aerial footage was acquired at a 640 × 480 resolution and 30 Hz frame rate.

VIRAT's metadata comprise the specification of two different types of ground truth: tracks consisting of logically related bounding boxes across frames for moving objects; and localized spatiotemporal events. Regarding the earlier type, only the visible part of moving objects is tagged with a bounding box, and that box is not extrapolated beyond occlusion by guessing. These metadata were mostly produced employing Amazon Mechanical Turk. Annotators tagged objects periodically and used automatic interpolation to recover the annotations in between key frames. Regarding the latter type of ground-truth metadata (activity labeling with precise start and end moments), it was produced by experts.

The dataset is freely available for research or commercial purposes.

3.5.16. VideoWeb

The VideoWeb multiview dataset [137] was developed for research on recognizing non-verbal communication among multiple persons.

This dataset comprises 2.5 h of video, which is divided into 368 video clips (with an average length of 4 min), capturing 51 scenes. Each scene was simultaneously captured by a minimum of four and a maximum of eight cameras (from a subset of 37 outdoor wireless cameras) at a resolution of 640 × 480 and at an approximate frame rate of 30 fps. The videos from the different cameras are approximately synchronized.

The scenes involve up to 10 actors interacting in various ways (with each other, with vehicles, or with facilities), acting out nine types of everyday activities (people meeting, people following, vehicles turning, people dispersing, shaking hands, gesturing, waving, hugging, and pointing).

All the videos for the 51 scenes are hand-annotated (frame numbers and camera ID for each activity label). This information is stored in XLS format.

The dataset is available under request.

3.5.17. MPII Cooking Activities Dataset

The MPII Cooking Activities Dataset [138] was built by the Max Planck Institute for Informatics, for employment on research on fine-grained activity recognition, focusing on cooking activities.

This dataset comprises 44 videos, with a total length of more than 8 h or 881,755 frames, acquired at a resolution of 1624 × 1224 pixels and at a rate of 29.4 fps. These pertain to 65 different cooking activities and vary in length from 3 to 41 minutes.

It comprises also, for each video, annotation information describing the observed activity(s), the frame at which that activity begins, and the frame at which it ends (5609 annotations of 65 activity categories). A subset of the frames was also annotated with human pose information describing aspects such as the positions of the shoulder, elbow, wrist, and hand joints as well as head and torso (1071 frames annotated pertaining to 10 subjects).

This dataset is freely available for research purposes only.

3.5.18. UCF101 Dataset

The University of Central Florida has compiled a string of progressively more comprehensive datasets for action recognition. These are UCF Sports, UCF11, UCF50, and UCF101 [139]. Thus, the latter is a human action recognition dataset comprising realistic action videos (containing, for instance, camera motion and cluttered background), collected from YouTube.

This dataset includes 13,320 videos (27 h of video data), each depicting one of 101 action categories. These videos have a frame rate of 25 FPS and a resolution of 320 × 240. The average video duration is 7.21 s.

The dataset's metadata consist of the action label for each video. The action categories are divided into five types: Human–Object Interaction, Body-Motion Only, Human–Human Interaction, Playing Musical Instruments, and Sports.

The dataset is freely available.

3.5.19. ADL Activity Recognition Dataset

The Activity Recognition Dataset [140] (maintained by MIT CSAIL) was built for the development of CV research on daily activity recognition on footage acquired from an "egocentric" perspective.

This dataset comprises over one million frames (over 10 h of video) with (20 different) people performing (18 different types of) unscripted, everyday activities. The original footage (from which the frames were extracted) is high-definition quality video (1280 × 960) with 30 FPS and with 170 degrees of viewing.

The dataset's metadata comprise activity labels (18 activity types); object bounding boxes (42 object types); tracks of objects in view; hand positions; and person–object interaction events. Some of the annotated features require a temporally long structure (preparing breakfast can take a few minutes) and complex object interactions (a kitchen cabinet looks different when its door is open). For this, the employed annotation format enables the representation of temporal pyramids (which generalize the spatial pyramid) and composite object models that take into account the different appearance of objects when being interacted with.

The ADL Dataset is freely available for research.

3.5.20. The Sports-1M

The Sports-1M dataset [141] was built for the development of research on sports activity detection and classification in video.

It comprises (links to) 1,133,158 YouTube videos annotated with 487 sports labels. These annotations were automatically produced using the text metadata associated to the videos. The dataset comprises 1000 to 3000 videos per class, and approximately 5% of the videos are annotated with more than one class.

The Sports-1M dataset is licensed under Creative Commons 3.0.

3.5.21. THUMOS Dataset

The THUMOS dataset [142] was developed (building on the UCF101 dataset [139]) for employment on research on activity recognition and (temporal) location in video content.

This dataset includes thousands of videos comprising over 430 h of video and 45 million frames, depicting the 101 activity types present in the UCF101 dataset. It is divided into testing, background, and validation subsets. The first comprises 2104 untrimmed videos, with an average of 20 videos for each of the 101 activity classes. The second set includes 2980 videos that do not to contain any instances of said 101 actions. The third comprises 5613 videos depicting 20 of the activities in scope.

The dataset comprises also annotation information expressed in the ViPER format. This describes (for the testing and validation sets) the activities that may be observed in each video and the temporal span of each such activity.

This dataset is available for research.

3.5.22. ActivityNet

The ActivityNet dataset [143] is the dataset provided for the ActivityNet challenge for human activity understanding.

Various editions of it have been prepared throughout the years. In its most recent versions, it provides video sequences pertaining to 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video for a total of 849 video hours. It comprises 10,024 training videos annotated with 15,410 activity instances and 4926 validation videos with 7654 activity instances. Half (50%) of the videos are in 1280×720 resolution, and the majority have a frame rate of 30 FPS.

The dataset's metadata consist of the description, for each sequence, of activities observed in them, and their starting and ending times. These metadata were manually produced specifically thought the services of Amazon Mechanical Turk.

The ActivityNet dataset is available for research purposes.

3.5.23. FCVID

The FCVID dataset [144,145] was built for CV research on activity recognition on unconstrained video.

It comprises 91,223 videos (obtained from YouTube), portraying 239 categories of things (183 are events and 56 are objects, scenes), organized under a hierarchy of 11 high-level super-categories. Globally, FCVID contains 4232 h with an average video duration of 167 s. Each video sequence is annotated (manually) with one or more activity labels (with no temporal delimitation).

The dataset is freely available for research purposes only and upon request.

3.5.24. AVA Actions

The AVA Actions dataset [146] was created for the development of spatio-temporal action recognition CV provisions.

It comprises 437, 15-min long, movie clips (obtained from YouTube) densely annotated with 80 atomic visual action types, which are localized in space and time. These movies clips consist of the 15th to 30th minute time intervals of 437 movies.

The dataset's annotation, even if focusing on the labeling of actions, is person-centric. The annotations are provided at a 1 Hz frame sampling frequency (which yields 900 keyframes for each movie clip). At each keyframe, every person is localized using a bounding box and labeled with all relevant actions from the AVA action vocabulary. These label actions pertaining to the actor's pose (standing, sitting, walking, swimming etc.), interactions with objects, or interactions with other persons.

This results in 1.62 million action labels. Multiple labels per human occur frequently. All bounding boxes have one pose label, 28% of bounding boxes have at least one person–object interaction label, and 67% of them have at least one person–person interaction label.

The dataset's metadata were generated by crowd-sourced annotators. Each video clip was reviewed by three independent annotators. An initial set of bounding boxes was generated using the Faster-RCNN person detector. Then, human annotators add the remaining bounding boxes missed by the detector. They also interconnect the bounding boxes over short periods of time, pertaining to the same action, to obtain ground-truth person tracklets. The dataset comprises 81,000 tracklets, whose duration ranges from a few seconds to a few minutes.

AVA's video content is available under YouTube access conditions. Its metadata are freely available.

3.6. Multipurpose Datasets

3.6.1. YFCC-100M

The Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) [91] was built to enable research on a wide spectrum of CV tasks.

It comprises 100 million media objects (99.2 million are photos and 0.8 million are videos) obtained from Flickr (uploaded between 2004 and 2014).

The dataset's metadata comprise, for each media object, its Flickr identifier, the user that created it, the camera that took it, the time at which it was taken, the location where it was taken (if available), the CC license it was published under, its title, user tags, machine tags, and description, the closed captions (extracted from the videos), motion features such as dense trajectories and motion boundaries, as well as direct links to its page and its content on Flickr.

In total, 68,552,616 photos and 418,507 videos in the dataset have been annotated with user tags (or keywords). Of these, 3,343,487 of the photos and 7281 of the videos carry machine tags (data automatically added by a camera, computer, application, or some other automated system).

The dataset's visual content is available under different, Creative Commons, licenses. Approximately 31.8% of the dataset is licensed as appropriate for commercial use, and 17.3% has been assigned the most liberal license that only requires the photographer that took the photo to be attributed. The dataset's metadata are freely available.

3.6.2. SUN Database

The Scene UNderstanding (SUN) database [147] (maintained by MIT CSAIL) was constructed for research in scene and object recognition. It is opened to researcher contributions through the LabelMe toolbox.

This dataset comprises two parts: the Scene Recognition Benchmark and the Object Detection Benchmark. The earlier comprises 130,519 images of environmental scenes (399 categories). The latter includes 16,873 images of objects (thousands of categories). The dataset's metadata consist of labels and image segment defining polygons (which is available, also, in the PascalVOC format).

This dataset is freely available for research.

3.6.3. MIT Flickr Material Database

The Flickr Material Database (FMD) [148] (maintained by MIT CSAIL) is to be employed in CV research on material recognition.

This dataset comprises a broad set of color photographs of surfaces. Each image in the dataset contains surfaces belonging to one specific type of material.

The metadata, associated to each image, assign the material depicted in each image to one of ten categories: fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and wood. One hundred images are available for each category: 50 close-ups and 50 regular views.

The contents of this dataset are available under different Creative Commons licenses.

3.6.4. VidTIMIT

The VidTIMIT [149] dataset was built for the conduction of research on a number of CV and Computer Audition (CA) areas, such as automatic lip reading; multiview face recognition; multi-modal speech recognition; or person identification.

This dataset comprises associated audio-visual acquisitions of 43 different volunteers (19 female and 24 male), each reciting 10 different short sentences. The average sentence duration is 4.25 s or approximately 106 video frames (using 25 fps). They were acquired in a noisy office environment, and the audio content was stored as a mono 16bit, 32 kHz WAV file. Two of the 10 sentences are common to all speakers, and the remaining are different. All sentences were selected from the test section of the NTIMIT corpus [150].

In addition to the audio-visual acquisitions pertaining to the sentence recitations, each person performed an extended head rotation sequence for video acquisition, allowing for the extraction of profile and 3D information.

The dataset's metadata consist of the text with the sentences and the identification of the head poses for each audio-visual acquisition.

The dataset is freely available for research.

3.6.5. Berkeley DeepDrive

The Berkeley DeepDrive (BDD100k) [151] dataset was built for research on object and area detection and object tracking in the context of automobile driving.

This dataset comprises 100,000 videos of 40 s each (acquired through crowdsourcing).

The dataset's metadata are provided for each frame at 10 s intervals. It comprises image-level description of the weather conditions (six types); time of day (three possibilities); bounding-box annotations of 10 object categories, including the attributes "occluded" and "truncated"; lane markings (eight main categories—road curb, crosswalk, double white, double yellow, double other color, single white, single yellow, single other color), with the attributes of continuity (full or dashed) and direction (parallel or perpendicular); and drivable areas, divided into two different categories: directly drivable areas and alternatively drivable areas.

This metadata comprise also fine-grained, pixel-level annotations for images, from each of 10,000 video clips randomly sampled from the whole dataset. Each such pixel is attributed to a specific object label (40 object classes).

This dataset comprises also, for 2000 of the videos (for about 400,000 frames), object tracking metadata. Each of these videos is approximately 40 s, and it is annotated at 5 fps, resulting in approximately 200 frames per video. This section of the metadata includes 130,600 track identities and 3.3 million bounding boxes.

This dataset is freely available for research upon request.

3.6.6. Oxford Robotcar Dataset

The Oxford Robotcar [152] dataset was developed for employment in research on ML-based tools for long-term road vehicle autonomy.

This dataset was produced using the RobotCar platform. This is a regular vehicle equipped with a sensor suite, consisting of the following sensory devices:

- Cameras—1 Point Gray Bumblebee XB3 trinocular stereo camera, acquiring images with 1280 × 960 resolution at 16Hz; 3 Point Gray Grasshopper2 monocular cameras, acquiring images with 1024 × 1024 resolution;
- LIDAR: 2 SICK LMS-151 2D LIDAR, operating at 50 Hz with 50 m range; 1 SICK LD-MRS 3D LIDAR operating at 12.5 Hz with a 50 m range;
- GPS/INS—1 NovAtel SPAN-CPT ALIGN with inertial and GPS/GLONASS navigation system.

Sensory acquisition was performed by traversing the same route, through central Oxford, for over 100 times during the period of a year.

This dataset comprises video acquisitions of over 1000 km of driving activity, summing up to almost 20 million images collected from the six cameras of the platform.

The dataset's metadata consist of the LIDAR, GPS, and inertial data, and further data performing logical and chronological the association between this information and the acquired images.

This dataset is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3.7. Overview of Metadata Formats Employed in CV Datasets

The broad set of MLCV competitions and challenges has led to the emergence of a few languages/formats for the expression of the metadata in the associated datasets. As MLCV researchers leverage these datasets to perfect their tools, the formats of their annotations become somewhat commonly used but are not yet close to formal protocols.

In the sections below, we present some of the most relevant such formats.

3.7.1. Pascal VoC

The PASCAL VoC [6] emerged from the PASCAL annually released object detection datasets and reported benchmarks.

The PASCAL VoC format is an XML-based format for image annotation. Each PASCAL VoC file comprises the annotations for a single image in a dataset. Those annotations comprise the name of the folder that contains the images; the name of the target image file; and the size of the target image file, in terms of width, height, and depth. The depth for a black and white image is 1, and for a color image, it is 3; there are a series of object detection declarations. Each one describes the detection of an object in the target image. This description comprises the object's name/label, its pose, the indication of whether it is truncated, the indication of whether the object is of difficult recognition, and an object-enclosing bounding box (axis-aligned rectangle specifying the extent of the object visible in the image).

Thus, this format enables only a simplistic annotation of image content, without any separation of the different logical levels of visual information interpretation (detection, spatial idealization, and semantic interpretation), or even the means to express some of that information. It also does not perform an integrated expression of visual meta-information, as the annotations (interpretations) of each image are store in a separate file.

3.7.2. COCO JSON

The COCO dataset was originally developed by Microsoft in 2014. COCO JSON is the format employed, within COCO, for the structuring of its metadata.

A single COCO JSON file may carry all the metadata of a dataset. The COCO JSON format enables the description of aspects about images and their content, which are of use to the following CV problems: keypoint detection, object detection, segmentation, and caption creation. Thus, it comprises the following annotations: object detection; keypoint detection; stuff segmentation; panoptic segmentation; image captioning.

A COCO JSON file comprises the following sections:

- Info—high-level information about the dataset
- Licenses—list of image licenses that apply to images in the dataset.
- Categories—list of categories. Categories can belong to a supercategory
- Images—all the image information in the dataset without bounding-box or segmentation information
- Annotations—list of every individual object annotation from every image in the dataset

The COCO JSON format enables a fine-grained segmentation of images. Similar to other tools, it does not perform a separation between the different logical layers of the interpretation of reality.

3.7.3. HDF-5

The Hierarchical Data Format (HDF5) [153] was developed for the storage and organization of large amounts of heterogeneous data, specifically datasets. HDF5 comprises a data model, a library, and a file format for storing and managing data.

An HDF file may contain a variety of file types, and it supports an unlimited number of datatypes. It also enables a flexible and efficient access and manipulation of the data it stores. The HDF5 format does not prescribe any specific metadata format. Instead, it enables any metadata scheme to be defined and built by the user and processed according to its own logic.

The main components of an HDF-5 file are (as presented in Figure 6):

- Group—similar to a folder, within an HDF5 file, that may contain other groups or datasets within it;
- Dataset—the actual data contained within the HDF5 file. Datasets are often (but do not have to be) stored within groups in the file.

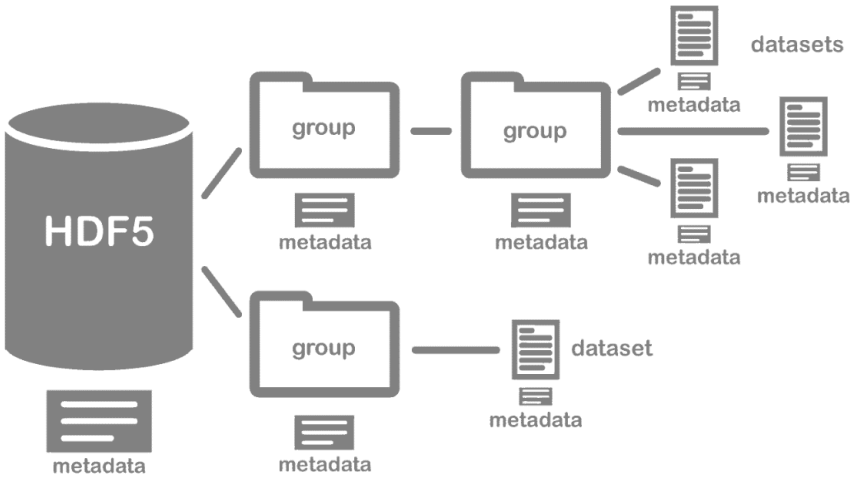

Figure 6. An example Hierarchical Data Format (HDF5) file structure (from [154]).

HDF5 is a self-describing file format. Each such file, each of its inner groups and datasets, may have associated metadata that describes exactly what its data are. This way, information may be added describing, for instance, how the data were collected, the employed sensor, etc.

3.7.4. CAVIAR's CVML

The XML-based Computer Vision Markup Language (CVML) [115] was developed within the context of the CAVIAR project [114] along with its manipulating software. CVML is a tool for the expression of computer vision results, and it is meant to enable the cooperation between separate research groups as well as make their research results more easily accessible to other areas of science and industry.

CVML enables the expression of CV-extracted information, pertaining to people, objects, or events, in all frames of a specific video sequence (i.e., enables the binding between interpretative and sensory data), such as their identification; appearance; position (circumscribing the object in a bounding box as presented in Figure 7); and activity type. It also enables the aggregation of information in different forms such as temporally defined feature vectors or sequences of frames, or entity identifications.

CV-extracted information expressed in CVML is of easy processing for the collection of statistical information and the semi-automatic analysis of a video stream, facilitating human activity recognition and the detection of unusual events, which may be exploited, for instance, for the early warning of human security staff.

CVML inter-mixes the description of aspects of the observed realities from different logical levels, such as the definition of image (frame) segments, with the description of the identity of the observed entity or its contextual role. Its capabilities for the description of relationships between entities and events, based on the definition of groups, are also limited. Thus, CVML presents some shortcomings in what regards clarity, flexibility, and logical correctness.

3.7.5. ViPER

ViPER GT is an XML-based language for the expression of visual data ground-truth information. It was developed by the ViPER project [5,122,155], together with software tools for producing such data and rendering it over the annotated video, as presented in Figure 8.

Figure 7. Rendering of Computer Vision Markup Language (CVML) annotated frames [114].

The ViPER GT format enables the annotation of information, pertaining to a multitude of realities, over video media, at the scene and object level.

It aggregates detection information per observed entity or event and not on a temporally sequential basis bound to sequential frames. It mixes detection information with the identification and further characterization of the observed realities. It presents some capability, even if limited, to declare relationships between detections from different frame-spans, pertaining to the same observed reality. It does so by placing such detections within a common parent metadata construct that identifies the observed reality.

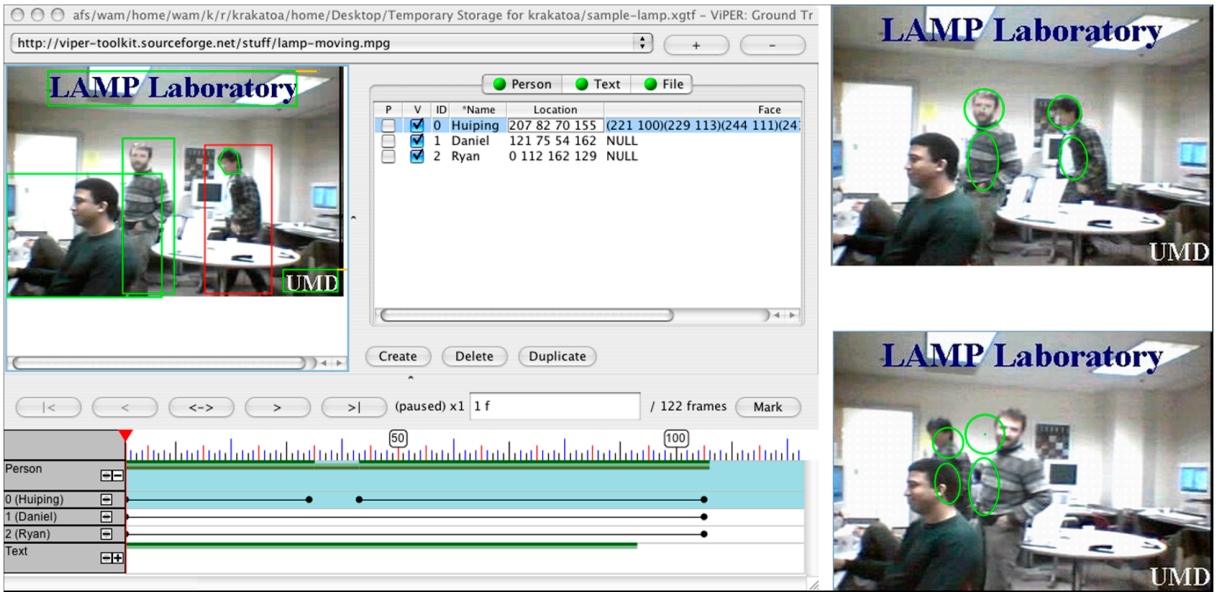

Figure 8. ViPER Editing Tool and Further Annotated Images (from [122]).

This way, this language is structured in a manner that intermixes the description of different logical levels of reality, for instance, signaling the visual detection of an object, within a frame, intermixed with the identification of that object and with further data about

it. Its structure makes ViPER also not suited, for instance, for the real-time display of the information it carries (i.e., the bounding boxes defining image segments).

4. Results Summary, Analysis, and Discussion

In this section, we shall now look at all the MLCV datasets for each of the specific application areas approached, in a collective manner, in order to assess their typical contents and characteristics; how they have evolved with time and their ongoing evolution trends; their predominant strengths and shortcomings; and to assess the possibilities and obstacles to the integration of such datasets into a homogenous and interoperating tissue.

It should be noted that for each of the specific areas approached, we base our analysis on the datasets that we surveyed. Even if such subsets constitute only a sample of the full scenario in each such area, they nonetheless include the most relevant datasets (in terms of academic/research references and uptake) and thus constitute a representative sample.

4.1. Analysis of Datasets for Facial Recognition

In this section, we present the summary and analysis of the 28 surveyed datasets for facial recognition. These results provide the answers for RQ1 to RQ5.

Table 2 presents a summary of the most relevant aspects of all surveyed datasets developed for facial recognition. It describes, for each such dataset, its creation/publication date; the relevant pertaining bibliographic reference(s); the number of images it comprises (typically, facial images); the number of different individuals whose faces are captured in the dataset's image pool; the predominant features of GT metadata and how these metadata are produced; the type of licensing involved in accessing the dataset's contents; and some further notes on the specificities of the dataset, typically regarding the way in which the dataset's images were produced or the manner through which the dataset's contents may be retrieved.

Table 2. Datasets for facial recognition.

Name	Creation	Refs.	Nr of Images	Identities	GT Metadata	Licensing	Notes
Olivetti Face Database	1994	[11,12]	400	40	EntityID, facial expression, lighting, eye glasses	Freely available	Controlled images Download in bulk
The FERET DB	1996	[13,14]	>14 K	1199	EntityID, pose	Free for research Case-by-case	Controlled images Download in bulk
XM2VTSDB	1999	[16,17]	-	295	-	Available at a payment	Video content Physical distribution (CD ROM)
3D RMA	2000	[18,19]	720	120	Person name	Free for research	3D captures from structured light
UOPB	2000	[20]	>2000	125	Entity ID, camera calibration, illumination	Available for a fees for delivery costs	Controlled images Physical distribution (CD ROM)
Extended Yale Faces DB	2001	[21,22]	16,128	28	Entity ID, pose, illumination in file/dir name	Free for research	Controlled images Download in bulk
FRGCD	2002	[23]	50,000	222	Person Name	Case-by-case	Controlled images
FG-NET	2004	[24]	1002	82	EntityID, age, gender, facial landmark points, etc.	Freely available	Unconstrained images Download in bulk
SCFace	2006	[25]	4160	130	Entity ID, birth date, gender, facial occlusions; camera number, distance, and angle; coordinates of eyes, nose and mouth	Free for research	Unconstrained images
BU-3DFE	2006	[26]	2500	100	Subject ID, gender race, facial expression, feature point set and pose model.	Free for research. Negotiable for commercial use	Controlled images 3D facial captures

Table 2. *Cont.*

Name	Creation	Refs.	Nr of Images	Identities	GT Metadata	Licensing	Notes
LFW	2007	[27]	13,233	5749	Person Name	Publicly available	Unconstrained images Download in bulk
CAS-PEAL Face DB	2008	[28]	>30 K	1040	Entity ID, gender and age, lighting, pose, expression, accessories, distance, time, resolution, eye locations.	Free for research on case-by-case basis	Constrained images
CMU Multi-PIE	2009	[29]	>750 K	337	Subject, expression, illumination, camera view	Available under paid license	Constrained images Physical distribution (disk drive)
PubFig	2009	[30]	58,797	200	Entity ID, age, gender, facial landmarks, accessories, pose, facial expression, lighting	Free for non-commercial use	Download of the metadata. Individual retrieval of images from Internet
Radboud Faces Database	2010	[31]	8040	67	Entity ID, emotion, and gaze in file name	Free for research and non-commercial	Controlled images Download in bulk
Texas 3DFRD	2010	[32]	2298	118	Entity ID, gender, ethnicity, expression, 25 fiducial points	Free for research	Constrained images, multi-modal images
YouTube Faces DB	2011	[33]	>600 K	1595	Identity, bounding box, head pose in *.mat* files	Publicly available	Unconstrained images Download in bulk
ChokePoint	2011	[35]	64,204	54	Subject ID and eye position	Free for research	Unconstrained images
FaceScrub	2014	[36]	>100 K	530	Person name	CC License for metadata	Unconstrained images Download of file URLs
CASIA-WebFaces	2014	[37]	>494 K	10,575	Person name	Free for research and non-commercial	Unconstrained images Download in bulk
Face Image Project	2014	[38]	>26 K	2284	Subject ID, age, gender, facial BB, pose and tilt	Publicly available	Unconstrained images Download in bulk
EURECOM KFC	2014	[39]	>2.8 K	52	Subject ID, face status, gender, age, occlusions, facial landmarks	Available on case-by-case basis	Constrained images
CelebFaces	2015	[40]	>202 K	10,177	Celebrity identity face b. box landmark locations binary attributes (ad hoc format)	Free for research Case-by-case	Piecemeal download from Google Drive
MegaFace	2016	[41]	4.7 M	672K	Entity ID, facial BB and landmarks	Free for research	Unconstrained images Download in bulk/chunks
UMDFaces	2016	[42]	4 M	8277	Entity ID, facial BB, gender, pose, 21 keypoints (A. Mechanical Turk)	Freely Available	Unconstrained images
IMDB-WIKI	2016	[43]	>520 K	20,284	Entity ID, birth, name, gender, year image was acquired, facial location and location scores	Free for academic purposes	Unconstrained images Download in chunks
VGGFace2	2018	[44]	>3.3 M	>9000	Entity ID, facial BBs and keypoints, pose, age in *.txt* and *.csv* files.	CC ASA Licence	Unconstrained images Download in bulk
Tufts Face Database	2018	[45]	100 K	112	Entity ID, gender, age	Free for non-commercial research and education	Seven image modalities Download in chunks (per image modality)

Facial recognition datasets typically comprise a vast set of more or less close-up facial images of a certain number of individuals and the necessary metadata to identify (and characterize) the person visible in each photo and, frequently, to localize or describe that person's face or some characteristics of it (e.g., keypoints, presence of facial props, etc.).

The media component of these datasets is typically comprised of static 2D images. Nonetheless, various other modalities of visual media are also employed, such as 3D image captures (e.g., 3D_RMA, BU-3DFE); stereo 3D images (e.g., XM2VTSDB); 3D+D images (e.g., Texas 3DFRD); infrared images (e.g., SCface), etc.

Even if not linear, there is a continuous trend (visible in Figures 9 and 10) for the increase in the amount of images (or media content) in such datasets. Images have typically evolved from being low-resolution acquisitions of peoples' faces to high-resolution acquisitions of entire scenarios where the individual faces occupy only a portion of the image. The number of pictured individuals has also grown; i.e., these datasets have steadily grown in quantity (of images and people) and quality.

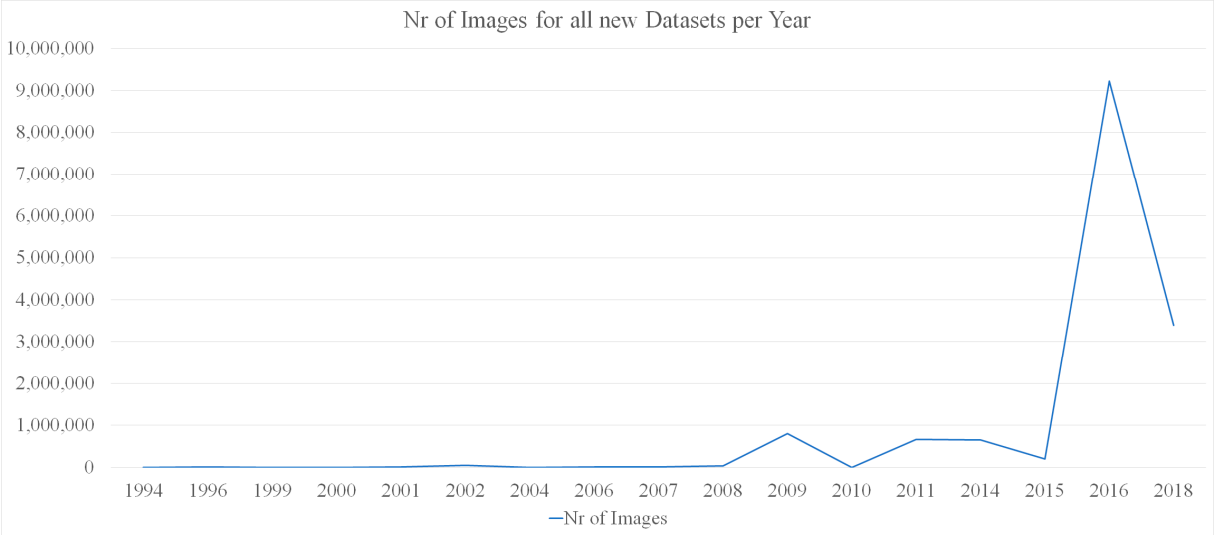

Figure 9. Evolution of the number of images in surveyed facial recognition datasets.

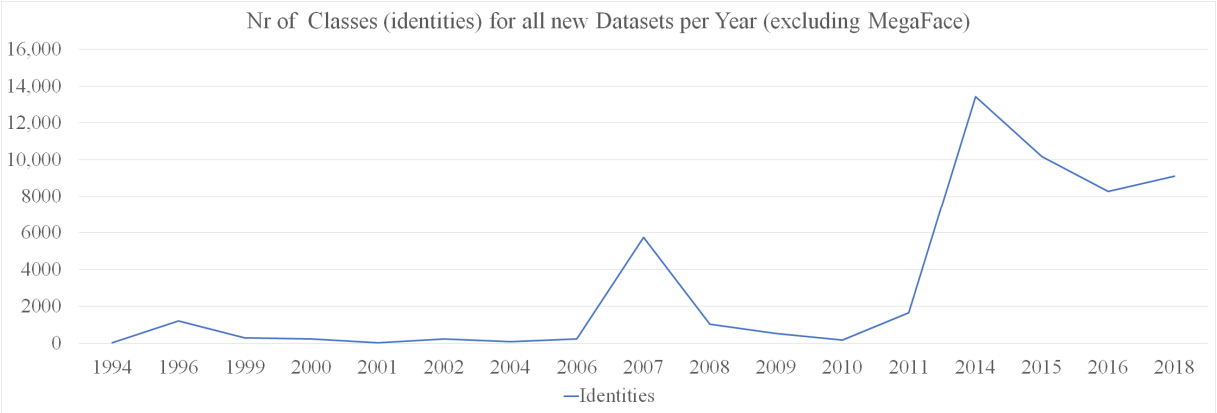

Figure 10. Evolution of the number of identified persons in surveyed facial recognition datasets.

Image acquisition mode has also changed from predominantly staged/controlled scenarios to uncontrolled ones (as show in Figure 11).

The metadata comprised by the dataset's in scope invariably include the identification of the person whose face is depicted in each image. This has always been the most frequent component of said metadata. However, over time, the metadata component of datasets for facial recognition has expanded to include various other aspects (as shown in Figure 12). Most such aspects are related to the location of the face/head in the image and its characteristics (bounding boxes, facial landmarks, head position, eye gaze, etc.).

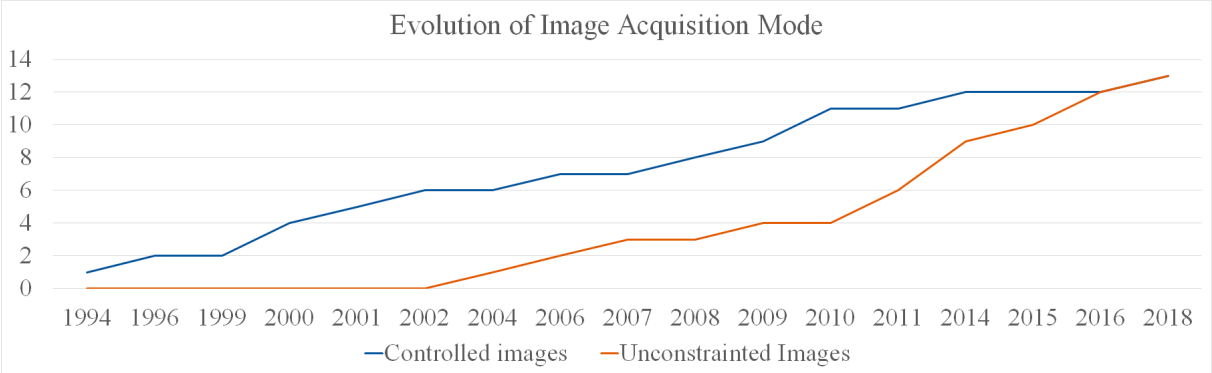

Figure 11. Evolution of image acquisition modes in surveyed facial recognition datasets.

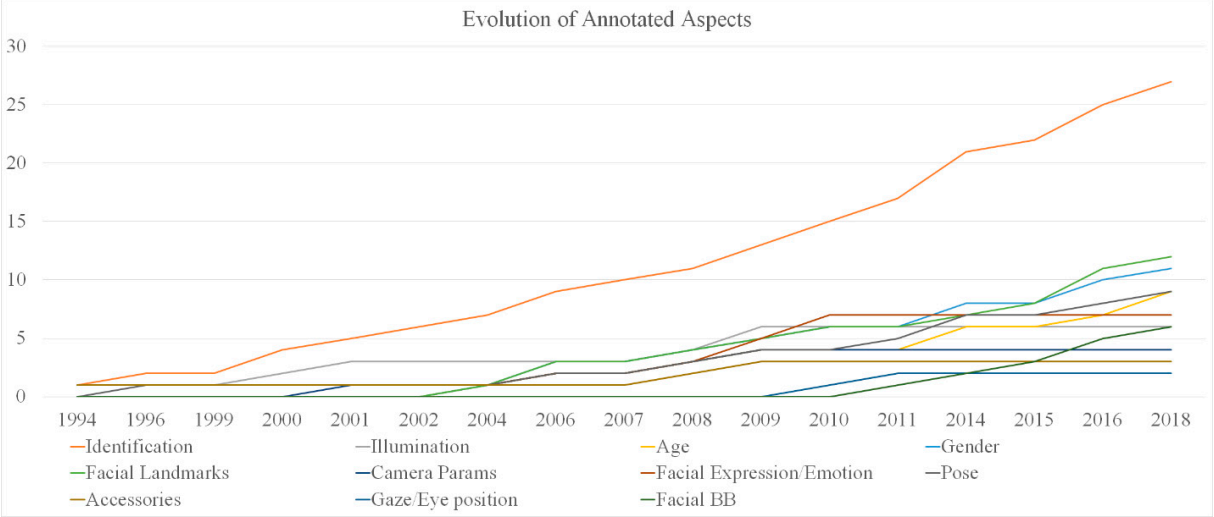

Figure 12. Cumulative evolution of annotated aspects in surveyed facial recognition datasets.

This evolution of the image and metadata contents has followed the progress in automated facial recognition and associated capabilities. These tools target increasingly complex images with more background noise; with a smaller portion of the image occupied by the face; acquired in the wild, and not staged; with facial accessories (e.g., glasses); with partial occlusions. The associated metadata evolve to describe the location of the face, other facial landmarks, and facial context-related aspects (accessories, pose, etc.).

Facial recognition tools have also expanded their capabilities to tasks such as the identification of gender, eye gaze, expressed emotion, or age. The metadata of the involved datasets have evolved to comprise the necessary GT for such objectives.

The access licensing types, to the datasets in scope, initially varied broadly, from paid to completely free. This situation has evolved, and presently, these datasets are predominantly free for research purposes or for any purposes (as seen in Figure 13).

Overall, facial recognition datasets, even if focused on similar problems, are very heterogeneous both in media specificities and in metadata contents and formats. They may contain single-view or multiview images acquired in a controlled fashion or "in the wild", with a 3D component or just 2D, stored in a myriad of formats, etc. The metadata may just describe the visible person or also the position of the face and various other facial characteristics and calculated image features, etc.

The contents of these datasets are typically designed to address the immediate needs of a specific research initiative/problem and are expanded, when and if necessary, in a completely ad hoc manner. There is typically no special concern about the overall structure

of the dataset's data, the format of the metadata, the integration of the metadata with the media content, and with the interoperability with the contents of other datasets.

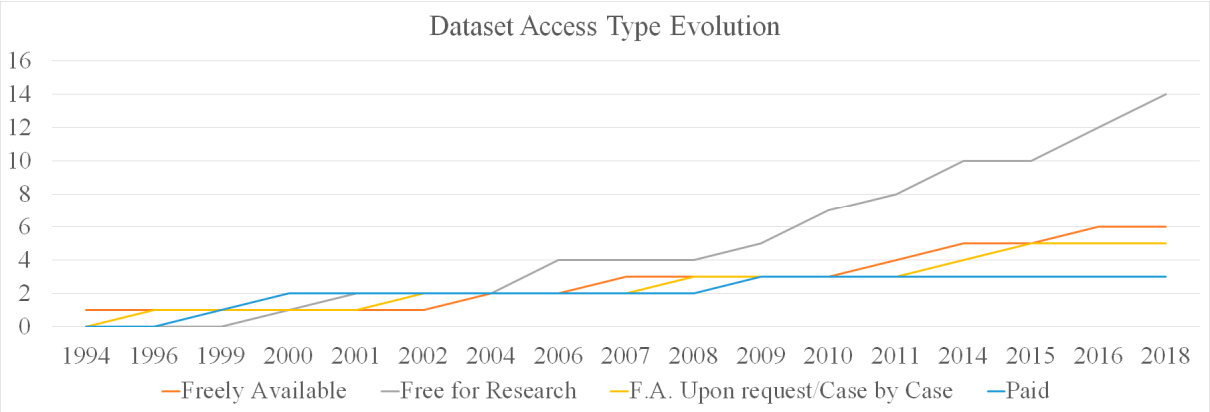

Figure 13. Evolution access modes to surveyed facial recognition datasets.

The repositories that store these datasets are simplistic. They typically provide for an in-bulk (or very high grained) retrieval of media and metadata and only minimal other manipulation capabilities.

Thus, the current panorama in facial recognition datasets is very distant from one of uniformity of data contents, structures, and formats. There is a shortage of interoperability that impedes a seamless reutilization of data and exploitation of synergies.

4.2. Analysis of Datasets for Object/Scenario Detection and Recognition

In this section, we present the summary and analysis of the 43 surveyed datasets for object and/or scenario detection and recognition. These results provide the answers for RQ6 to RQ9.

Table 3 presents a summary of the most relevant aspects of all surveyed datasets developed for image segmentation, object detection, and scenario recognition. It describes, for each such dataset, its creation/publication date; the relevant pertaining bibliographic reference(s); the number of images it comprises (typically, those for which segmenting/detection annotations are provided); the number of different individual object detections, or segment definitions, defined in the overall metadata; the number of individual objects/scenario types captured in the dataset's metadata; the number of annotators or means of annotation; the predominant components of the GT metadata; and some further notes with a particular focus on licensing aspects.

Table 3. Datasets for segmentation, object recognition, and scenario recognition.

Name	Creation	Refs.	Number of Images	Number of Detection/Segments	Number of Objects/Scenarios	Number of Annotators	GT Metadata	Licensing/Notes
COIL-100	1996	[46]	7.2 K	-	100	-	Object label and pose	Freely available
MSRCD	2000	[47]	>800	>800	34	-	Object ID and shape masks	Freely available for non-commercial use Download in bulk
BSD	2001	[48]	800	3000	-	25	Segmentation maps in *.mat* files	Freely available
RGB-D ORD	2001	[49]	250 K	250 K	300	-	Segmentation mask image file BB for video frames in *.mat* file	Freely available for non-commercial use Piecemeal download
NORB	2004	[50]	>194 K	>194 K	5	-	Object labels and BBs	Freely available for research Download in chunks
P3DTT	2005	[51]	2 × 144 × 3 × 100	-	100	-	Object labels and perspectives	Freely available for research

Table 3. *Cont.*

Name	Creation	Refs.	Number of Images	Number of Detection/Segments	Number of Objects/Scenarios	Number of Annotators	GT Metadata	Licensing/Notes
Caltech-256	2007	[52]	30,607	-	256	-	Object labels	Freely available for research
LabelMe	2008	[53]	30,369	111,490	2888	-	Object labels, BBS, polygons, segment. Masks (Pascal VoC)	Freely available for research
ImageNet	2009	[54]	>14 M	>1 M	21,841	Crowdsourced	Object classification, BBs, features in (Pascal VOC)	Freely available for non-commercial use Piecemeal download
CamVid	2009	[55]	>39 K	$\approx 700 \times 32$	32	-	Pixel-level object segmentations	Freely available
CIFAR-10/100	2009	[56]	60 K	-	10/100	-	Object label	Freely available
NUS-WIDE	2009	[58]	269,648	425,059	5018(Flkr) 81(man)	-	Object labels	Freely available
MIT Indoor Scenes	2009	[59]	15,620	-	67	-	Scene label (Pascal VOC)	Freely available
SBU CPD	2011	[60,61]	1 M	1 M	-	Crowdsourced/ automated	Image captions	Freely available
SLT-10	2011	[62]	100 K	500	10	-	Object label	Obtained from ImageNet Download in bulk
PRID 2011	2011	[63]	$\approx (475 + 856) \times 125$	$\approx (475 + 856) \times 125$	245	-	Bounding boxes	Freely available
CUB-200-2011	2011	[64]	11,788	>>11,788	200	-	Label, BBs, parts, and attributes	Freely available for research
SBD	2011	[65]	11355	>20 k	20	Crowdsourced (Amazon Mechanical Turk)	Object boundaries and labels	Freely available for research
Stanford Dogs	2011	[66]	20,580	>20 K	120	-	Object label and BBs	Freely available for non-commercial use
Pascal VOC	2012	[6]	>11 K	>33 K	20	Crowdsourced	Object label, BBs, pixel-wise masks, reference points and actions (Pascal VOC)	Freely available, Flickr terms of use Download in bulk
NYU Depth D.	2012	[67]	(500 K) 1.5 K	>35 K	894	Crowdsourced (Amazon Mechanical Turk)	Pixel-wise object labels and masks	Freely available Download in bulk
Leafsnap	2012	[68]	>30 K	>30 K	185	-	Tree species, segmented images	Freely available Download in bulk
Oxford-IIIT Pet	2012	[69]	>7 K	>7 K	23	-	Animal breed label, head BB, body segmentation	Freely available
LISA TSDB	2012	[72]	6610	7855	49	-	BBs and label	Freely available for research
DUSD	2013	[73]	10 K	25 K	5	-	Pixel-level semantic class annotations	Freely available for research
Stanford Cars	2013	[75]	>16 K	>16 K	196	Crowdsourced (Amazon Mechanical Turk)	Car make, model, year, BB (*.mat* files)	Freely available Download in bulk
FGVC-Aircraft	2013	[76]	>10 K	>10 K	102	Crowdsourced (Amazon Mechanical Turk)	Aircraft model and bb (*.txt* files)	Freely available for research Download in bulk
MS DVAC	2014	[77]	500	100 K	4K	Crowdsourced (Amazon Mechanical Turk)	Object labels and BBs	Freely available for research Download in chunks
MS COCO	2014	[78]	>328 K	2.5 M	91	Crowdsourced (Amazon Mechanical Turk)	Object labels and segmentations (JSON)	Freely available for research Download in chunks
MDs	2014	[80]	>2 K	-	-	-	Disparity maps	Freely available Piecemeal downloads
Flickr30k	2014	[81]	>30 K	>158 K (captions)	-	Crowdsourced (Amazon Mechanical Turk)	Image textual description	-

Table 3. *Cont.*

Name	Creation	Refs.	Number of Images	Number of Detection/Segments	Number of Objects/Scenarios	Number of Annotators	GT Metadata	Licensing/Notes
iLIDS-VID	2014	[83]	≈600 × 73	≈600 × 73 (presumably)	300	-	BBs and various other info in (XML) ViPER compliant format	Freely available for research
BelgiumTS	2014	[84]	145 K	13,444	4565	-	BBs, camera ID, and pose	Freely available for research
Pascal Context	2014	[85]	10,103	10,103 × 12	540	6	Pixel-wise segmentation masks, labels (Pascal VOC)	Freely available for research
Cityscapes	2015	[74]	>200 K total 25 K annot.	≈(5000 × 30) + (25,000 × 20)	30	-	Finer and coarser pixel-wise object annotations/segmentatic	Freely available for research
CompCars	2015	[86]	>214 K	>50 K	1716 + 163	-	Car make, model, part, attribute, view	Freely available Piecemeal downloads
YouTube-8M	2016	[87]	Millions	237 K	3862	Crowdsourced,	Audio-visual features, video level labels	Freely available CC BY 4.0
DAVIS	2016	[88]	3600	3455	>4	-	Binary masks	BSD License
iNaturalist	2017	[89]	>850 K	>560 K	>5K	Crowdsourced (iNaturalist effort)	Species label and BB (same format as COCO)	Freely available for research Download in chunks
YouTube-BB	2017	[98]	10.5 M	>5.6 M	23	Crowdsourced (Amazon Mechanical Turk)	Object label and bounding box	Freely available for research Download in bulk
Visual Genome	2017	[90]	>108 K	>4.5 M	≈13,041 + 13,894	Crowdsourced (Amazon Mechanical Turk)	Region descriptions, objects, attributes, relationships, region graphs, scene graphs, and question–answer pairs	Freely available CC BY 4.0 Download in chunks
Open Images Dataset (v5)	2018	[92,93]	9 M	36.5 M(img-l) 15.4 M(BBs) 375K(rels)	19.9 K(img-l) 600(BBs) 57(rels.)	-	Image-level labels, bounding boxes	Hosted at Github Freely available CC BY 4.0
YouTube-8M Segments	2019	[87]	Millions	237 K	1000	Crowdsourced,	Audio-visual features, video and frame level labels	Freely available CC BY 4.0

Datasets for image segmentation and object and scenario recognition typically comprise a vast set of static images, depicting the targeted objects or scenarios, from one or more points of view. These datasets comprise, as well, the necessary metadata to locate (in the images) and label the objects and scenarios in scope.

The predominant media content employed in these datasets is 2D static images. Other media modalities are also employed, such as 2D+D video (e.g., RGB-D Object Recognition Dataset); 2D video (a growing number datasets such as Cityscapes or YouTube-8M); or 3D images (e.g., NORB).

The evolution of the total number of such images, and of the total number of individual object/scenario detections, shows a clear growth trend with time (Figures 14 and 15). However, that trend is not linear. A larger growth occurred in 2009, but the earlier growth rate was resumed afterwards. However, this is only so because of the appearance of the Imagenet Dataset in said year. This was an exceptionally broad effort that set itself apart from the predominant trend. If Imagenet is left out of the picture, the growth trend becomes more or less continuous with a certain acceleration in time.

This reflects the continuous development of image acquisition and collection platforms as well as of image annotation tools. Its reflects also the development of crowdsourced annotation platforms (such as Amazon's MechanicalTurk). All these developments combined have enabled the production of truly vast datasets.

The metadata, in image segmentation and object recognition datasets, comprise always the identification of the targeted object (or scenario, in the case of scenario recognition datasets) and, typically, some form of locating that object in the image. This localization is done with bounding boxes (or polygons) or some type of segmentation mask.

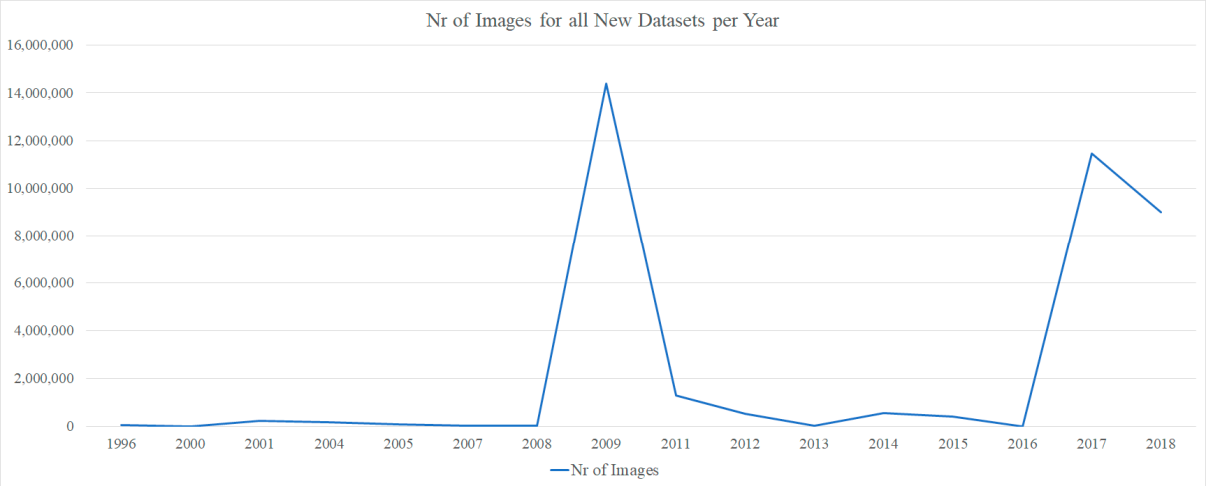

Figure 14. Evolution of the number of images in surveyed segmentation and recognition datasets.

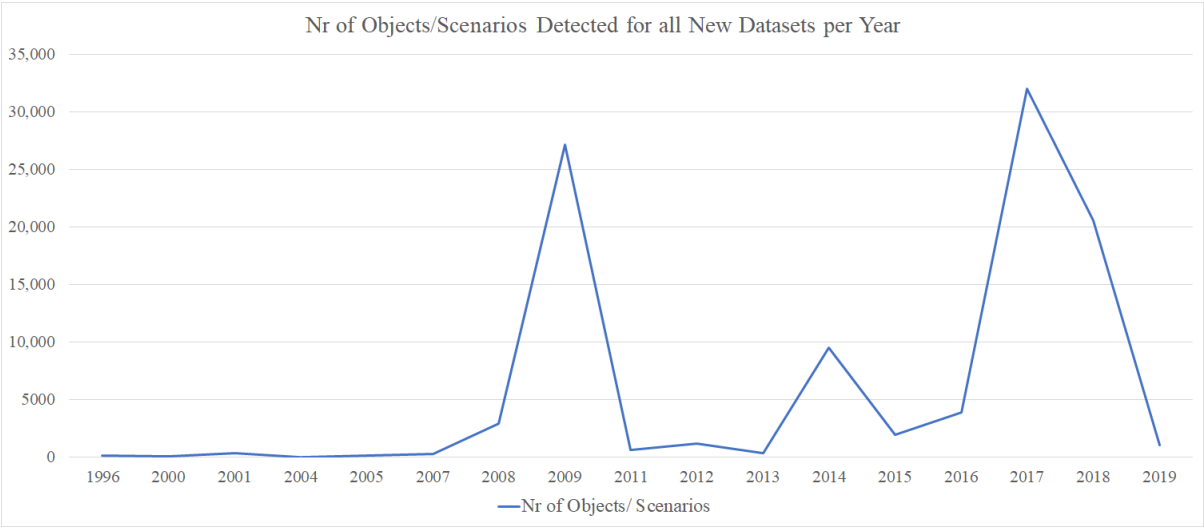

Figure 15. Evolution of the number of individual object/scenario detections in surveyed segmentation and recognition datasets.

With the evolution of these datasets, bounding boxes have come to be the predominant localization metadata, even if masks have also grown in use (as shown in Figure 16). Various other metadata components have also appeared; however, the overall types of metainformation present datasets in scope have remained remarkably constant.

Datasets of this type have typically been made available under very open conditions (as shown in Figure 17). This way, they are either freely available for any use or for academic use only, the latter having become the predominant form.

The datasets in scope present the same heterogeneity seen in those approached in Section 4.1. At the level of metadata, said heterogeneity is even bigger given the greater variety and complexity it presents in segmentation and detection datasets (when compared to facial recognition ones).

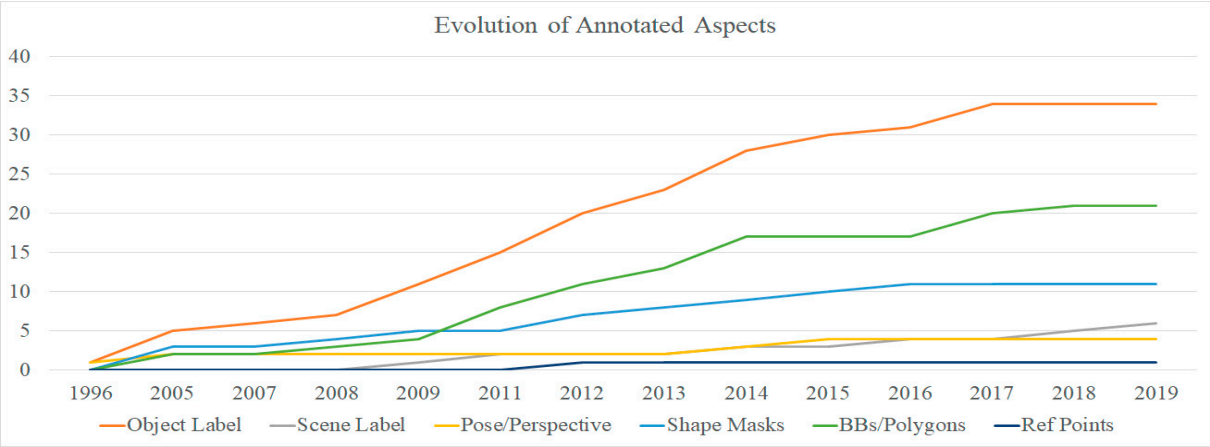

Figure 16. Evolution of annotated aspects in surveyed segmentation and recognition datasets.

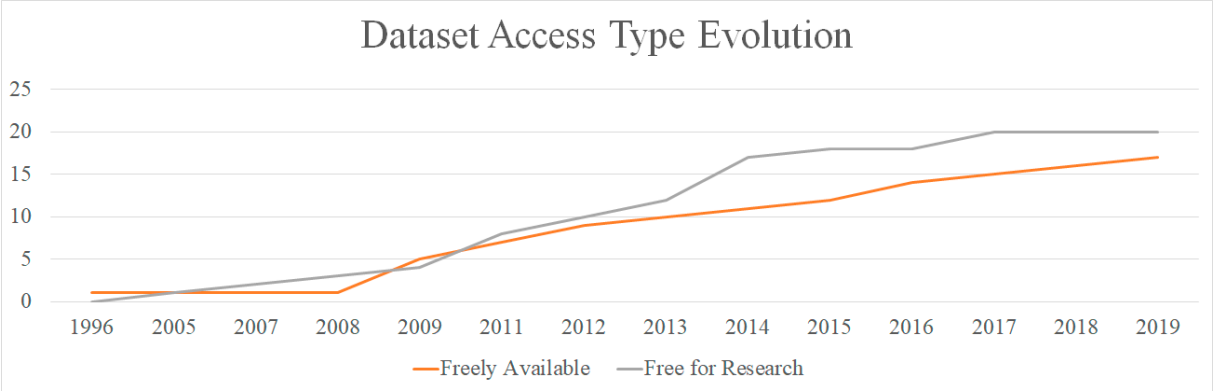

Figure 17. Evolution of access types to surveyed segmentation and recognition datasets.

These datasets and their contents are typically designed to address the specific needs of some particular research objective. No special concern goes into their design pertaining to standardization, ease of reuse by other initiatives, or integration into a greater tissue.

The repositories that store these datasets are also simplistic and provide for only minimal dataset manipulation capabilities.

Thus, the current panorama in pertaining to these datasets is not one of uniformity of data contents, structures, and formats. Interoperability and reutilization are far from optimal.

4.3. Analysis of Datasets for Object Tracking

In this section, we present the summary and analysis of the 12 surveyed datasets for object tracking. These results provide the answers for RQ10 to RQ15.

Table 4 presents a summary of the most relevant aspects of all surveyed datasets developed for object tracking. It describes, for each such dataset, its creation/publication date; the relevant pertaining bibliographic reference(s); the total amount of footage it comprises expressed in seconds (s), minutes (m), or hours (H); the number of different tracked object classes; the number of individual objects tracked; the number of all detections of all tracked objects in all frames; the predominant components of the GT metadata; and some further notes with a particular focus on licensing aspects.

Datasets for object tracking typically comprise video content, depicting the objects (whose tracking is desired), in motion, and seen from one or more points of view. They comprise, as well, the necessary metadata to identify, locate and track across sequential frames the objects in scope.

These datasets must necessarily comprise a temporal dimension as motion occurs in time. For this, the predominant media content they comprise is 2D video. In some cases,

that video is split into its composing frames, but the temporal variation information is still preserved in the image sequences.

Table 4. Datasets for object tracking.

Name	Creation	Refs.	Footage Length	Number of Tracked Classes	Number of Tracked Objects	Number of Detection/Segments	GT Metadata	Licensing/Notes
Human Eva	2006	[96]	833 s	6	4	50k	3D body poses descriptions (motion capture)	Freely available for research
ETH	2007	[97]	153 s	1	Hundreds	10,958	Bounding Boxes	Freely available
Daimler	2008	[98]	428 s + 27 m	1	259	72,152	Bounding Boxes	Freely available for research
TUD	2009	[99]	-	1	311	1326 + 1776	Bounding Boxes	Freely available
Calthech D	2009	[100]	10 H	1	>2k	350k	Bounding Boxes, occlusion labels	Freely available for research
Kitti	2011	[4,101]	6 H	8	>2160	>300k	Object labels and 3D BB across time	Available under CC ShareAlike 3.0
ALOV++	2011	[103]	≈315 × 9.2 s	64	315	>89K	Object type and BB	Freely available
VTB	2015	[104]	-	-	>100	Tens of Thousands	Object label and 3D BB	Freely available
TColor-128	2015	[106]	-	tens	128	Thousands	Bounding Boxes	Freely available for research
NUS-PRO	2016	[107]	-	17	160	≈73 × 300 + 292	Object type, BBs, pixel-wise mask, occlusion type, fiducial points	Freely available
UAV123	2016	[108]	≈ 123 × 30s	6+	≈123	>110K	Object type and BB	Available under request
VOT Challenge	2019	[109]	-	-	>250	>450K	Object labels, BBs. Frame visual attributed (*.txt* file)	Freely available under various licenses Provides toolkit

The general evolution trend is for the steady growth of the amount of footage contained in these datasets as well as of the total amount of its annotations (i.e., the number of classes of objects, number of individually tracked instances of such object classes, and volume of actual annotations), as shown in Figures 18–21. However, this trend is not linear, as the period from 2008 from 2015 saw a sharp increase in such contents that was not continued afterwards. This is due to both our sampling of the overall panorama and to the release of some noticeable datasets in that period.

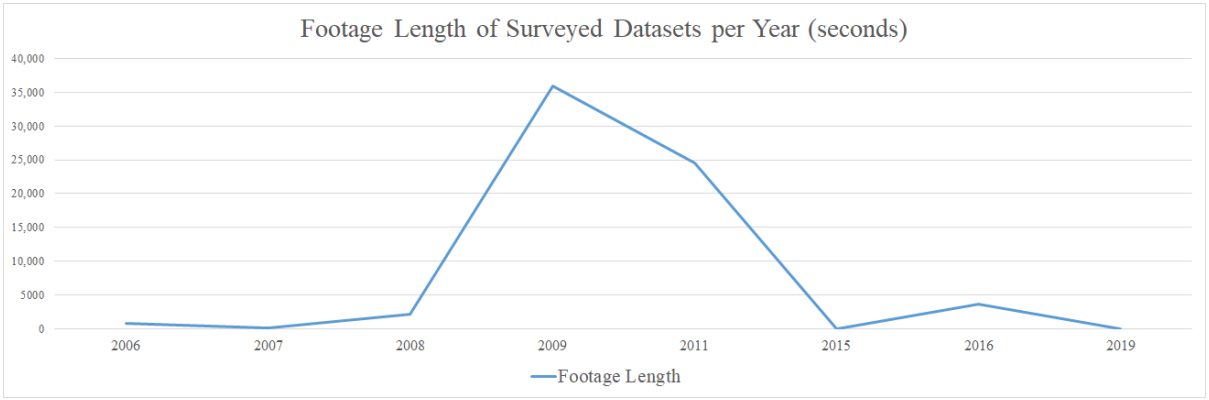

Figure 18. Evolution of the total footage length in surveyed object-tracking datasets.

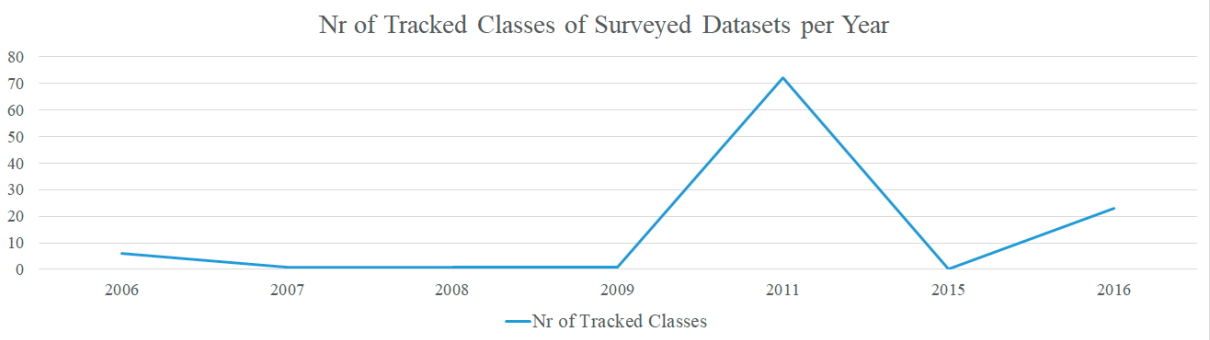

Figure 19. Evolution of the number of tracked object classes in surveyed object-tracking datasets.

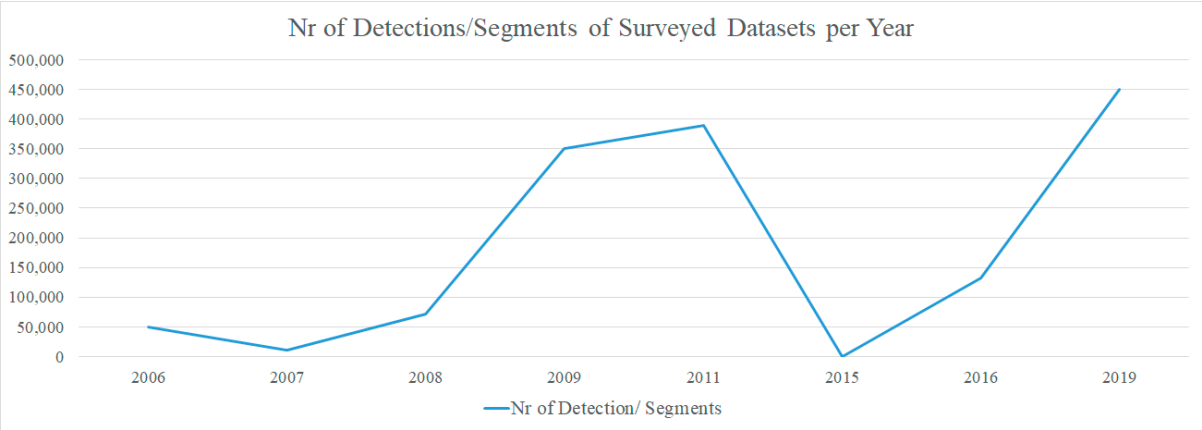

Figure 20. Evolution of the number of individual detections in surveyed object-tracking datasets.

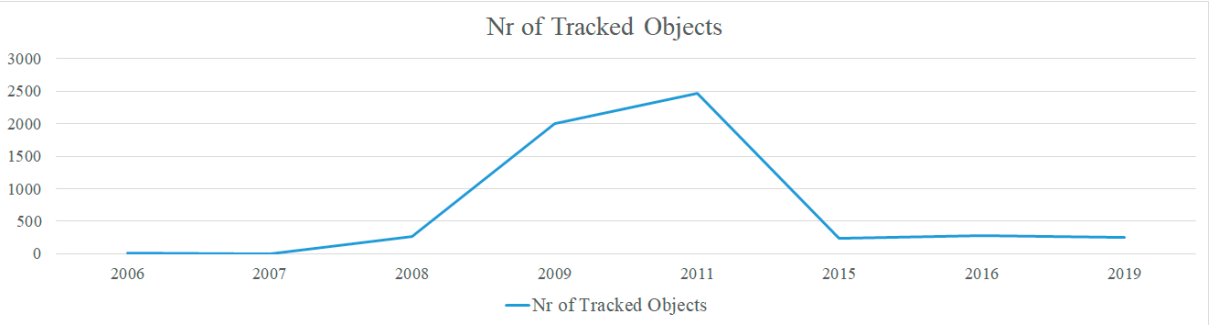

Figure 21. Evolution of the number of tracked individual objects in surveyed object-tracking datasets.

In addition to the object label, the predominant metadata in these datasets consist of the location of the tracked objects in the images, which is typically done with 2D bounding boxes (or polygons). Other metadata components have also appeared, with 3D bounding boxes and occlusion labels gaining relevance (as shown in Figure 22).

Datasets of this type have typically been made available under relatively open conditions (Figure 23). The predominant access mode has typically been "free for research" with "completely free" gaining more relevance in the recent past.

The datasets in scope are relatively more homogeneous, pertaining to their media component and compared to the earlier approached datasets, but they still present the same overall heterogeneity as the one described in Section 4.2.

They were designed to address specific object tracking-related needs without any concern for universality, logical correctness (in terms of the separation of the different levels of cognitive information), standardization, or seamless reuse by other initiatives. The same is true for their repositories.

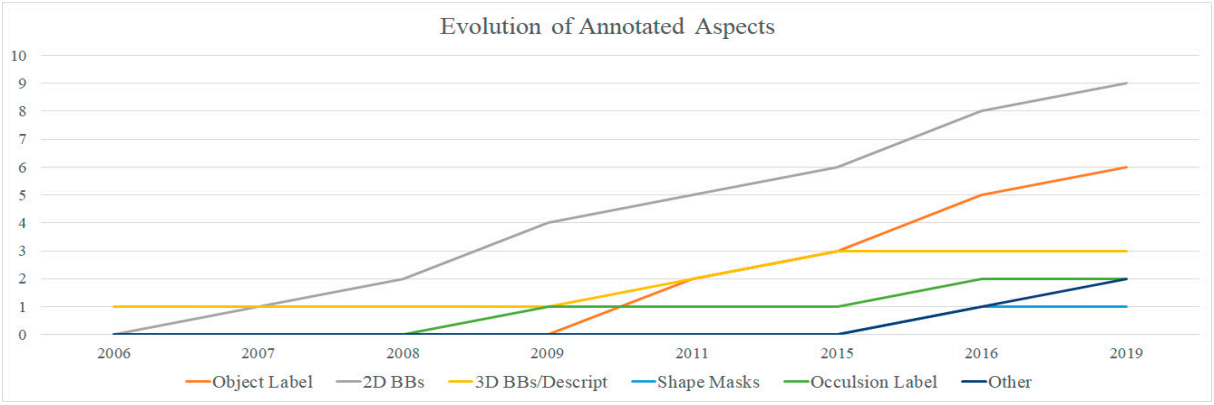

Figure 22. Evolution of annotated aspects in surveyed object-tracking datasets.

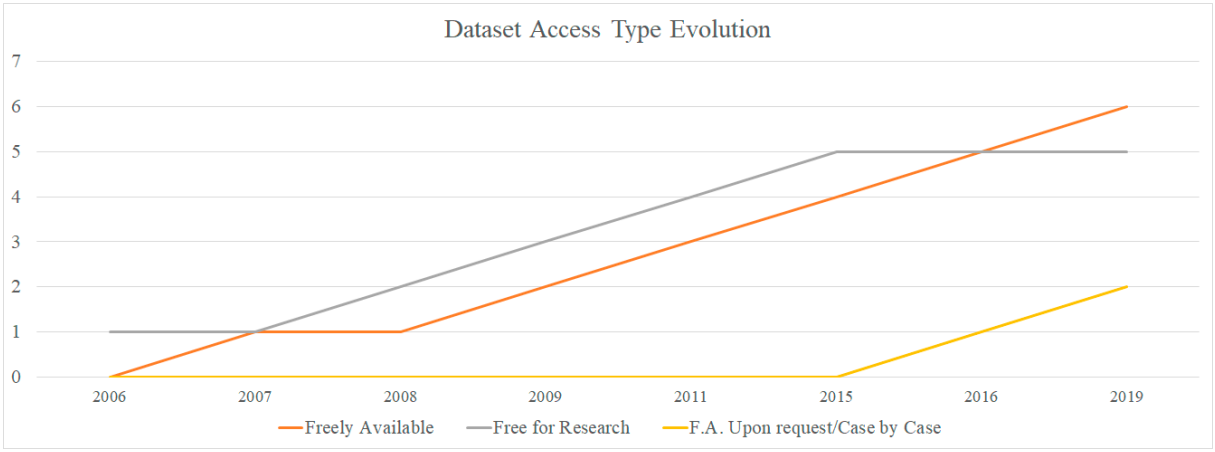

Figure 23. Evolution of access types to surveyed object-tracking datasets.

Thus, the tissue of tracking datasets is far from optimal for integration, interoperability, and reutilization.

4.4. Analysis of Datasets for Activity and Behavior Detection

In this section, we present the summary and analysis of the 24 surveyed datasets for activity and behavior detection. These results provide the answers for RQ16 to RQ21.

Table 5 presents a summary of the most relevant aspects of all surveyed datasets developed for activity and behavior detection. It describes, for each such dataset, its creation/publication date; the relevant pertaining bibliographic reference(s); the total amount of footage it comprises (in either hours (H) or seconds (s)); the full number of images/frames (which is not always associated with the metadata available); the total number of targeted activity types; the predominant components of the GT metadata; and some further notes with a particular focus on licensing aspects.

Datasets for activity and behavior detection typically comprise video content, as activities are something that occurs in time and thus require a temporal dimension (to its sensory acquisition) to be identified. In some cases, the media component consists of multiview video (e.g., i3DPost or MuHAVi), and only for a few exceptional cases, it is made up of static images.

Table 5. Datasets for activity and behavior detection.

Name	Creation	Refs.	Footage Length	Nr of Images	Nr of Detection/Segments	Nr of Activities	GT Metadata	Licensing/Notes
CAVIAR	2004	[114]	≈3853 s	≈96,325	>>96,325	8	Bounding boxes, observed behavior, head position, gaze direction, or hand, feet, and shoulder positions	Freely available
KTH	2004	[116,117]	≈2391 × 4s	≈2391 × 4 × 25	-	6	Action labels and frame spans	Freely available
WEIZMAN	2005	[118,119]	-	Thousands	-	10	Activity label (AL) per video. Background and foreground masks per image.	Freely available
ETISEO	2007	[120,121]	-	-	-	15	AL, BBs	Freely available for research (on a case-by-case basis)
CASIA Action	2007	[123]	≈1446 × 18 s	≈1446 × 18 × 25	>1446	15	AL, subjects, per sequence	Free for research (for commercial upon request)
HOHA	2008	[124]	-	-	≈231 + 143 + 217	8	AL label per sequence	Freely available
MSR Action	2009	[125]	≈325 s	≈325 × 25	63	3	Spatio-temporal BB per action	Freely available for for resarch
HOLLYWOOD2	2009	[126]	18 H	≈1.59 M	-	22	AL per sequence	Freely available
I3DPost	2009	[127]	-	-	104	13	Person ID, AL per seq., binary masks, 3D mesh per frame	Freely available for research
BEHAVE	2010	[128,129]	≈3600 s	>90 K	>90K	10	AL per frame range, BB per frame	Freely available for research
TVH ID	2010	[130,131]	-	≈85,500	>>85,500	4	BB, head orientation, interaction label	Freely available for research
MuHAVi	2010	[132,133]	-	-	-	17	People silhouettes per frame, AL frame ranges	Freely available
UT-Interaction	2010	[134]	20 × 1	20 × 60 × 30	60 + 180	6	AL, frame range, BBs, per activity per sequence	Freely available
HMDB51	2011	[135]	-	-	-	51	AL, visible body parts, number of people, per sequence	Freely available
VIRAT	2011	[136]	29 H	≈29 × 3600 × 30	>>29 × 3600 × 30	23	BB per frame, AL frame range (A. Mechanical Turk)	Freely available
VideoWeb	2011	[137]	2.5 H	≈2.5 × 3600 × 30	>51	9	AL, frame range per sequence XLS format	Available under request
MPII	2012	[138]	>8 H	>881 K	5609/1071	65	AL per frame range, bodily part position per frame	Freely available for research
UCF101	2012	[139]	27 H	≈27 × 3600 × 25	≈13,320	101	AL per sequence	Freely available
ADL	2012	[140]	>10 H	>1 M	≈1 M/30	18	Activity and object labels, BBs, object tracks, interaction events	Freely available for research

Table 5. *Cont.*

Name	Creation	Refs.	Footage Length	Nr of Images	Nr of Detection/Segments	Nr of Activities	GT Metadata	Licensing/Notes
Sports-1M	2014	[141]	-	-	-	487	ALs per sequence	Freely available
THUMOS	2015	[142]	430 H	>45 M	-	101/20	AL and range	Freely available for research
ActivityNet	2015	[143]	849 H	≈849 × 3600 × 30	23,064	203	AL per sequence (A. Mechanical Turk)	Freely available for research
FCVID	2015	[144,145]	4232 H	-	>91,223	239	AL per sequence	Freely available for research upon request
AVA Actions	2018	[146]	437 × 15 MN	437 × 900	1.62 M	80	BB and AL per keyframes, tracklets	YouTube video, metadata freely available

The footage in these datasets typically consists of video acquisitions of the targeted activities taking place.

The informational volume of the datasets in scope has generally been growing (as show from Figures 24–27) in terms of the amount of the footage they contain, the total number of images (particularly those with annotations), and the total number of annotations (activity detections indicated either per video or on a more precise, frame-span-based way).

The same is true for the total number of activities targeted for detection in each dataset. Initial activity detection datasets were mostly focused on a reduced set of specific activities. They have evolved to comprise increasingly broader sets of activities.

This overall growth has become more accentuated in the recent past. However, this trend has not been linear. Instead, growth has occurred in spurts. The biggest such spurt has occurred in 2015, but it was omitted from the graphics (in Figures 28 and 29), as it was statistically aberrant and obscured the identification of the overall trend.

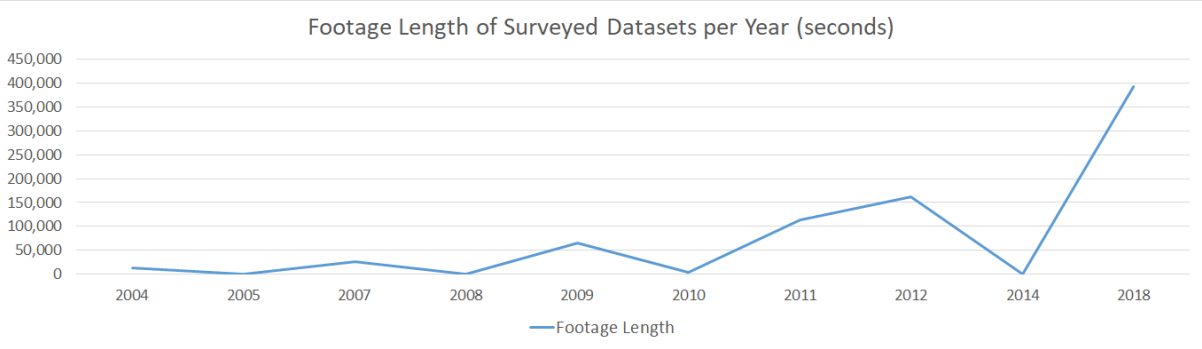

Figure 24. Evolution of the total footage length in surveyed activity recognition datasets.

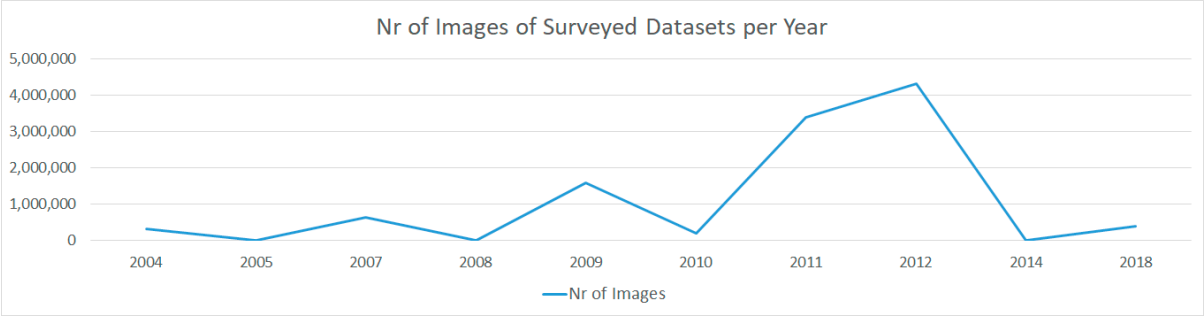

Figure 25. Evolution of the number of images in surveyed activity recognition datasets.

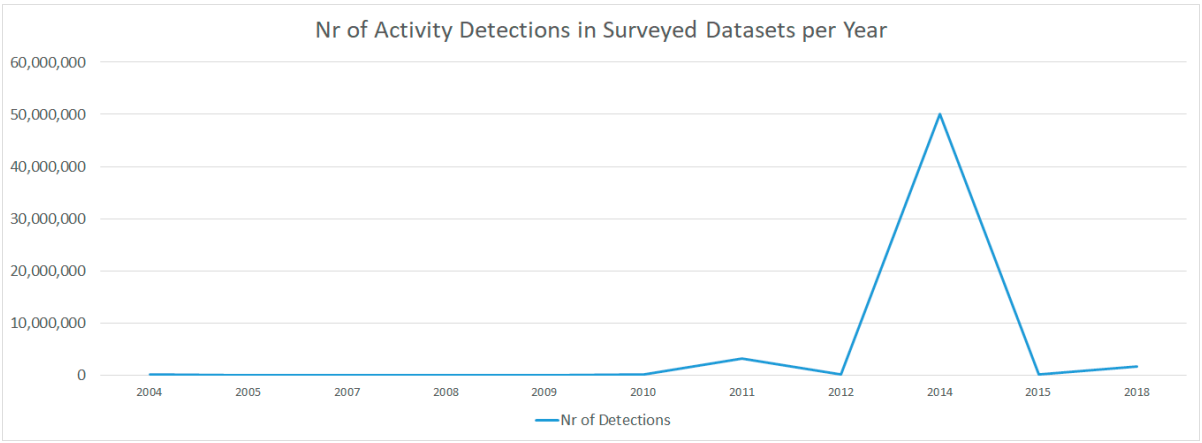

Figure 26. Evolution of the number of activity detections in surveyed activity recognition datasets.

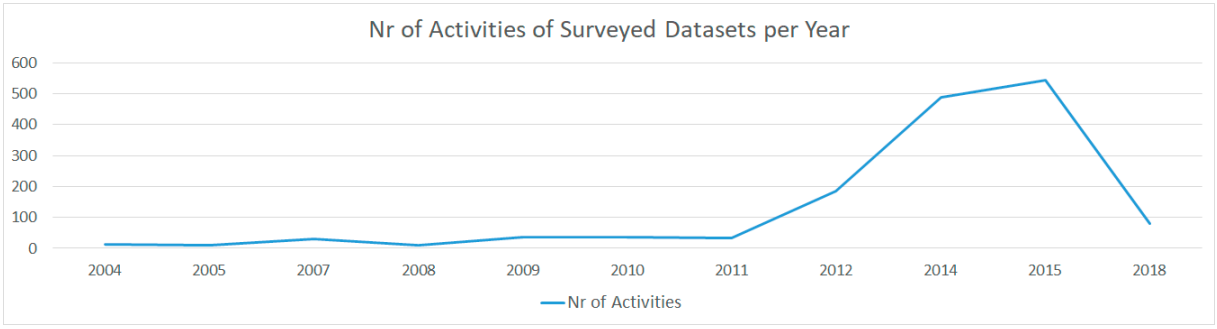

Figure 27. Evolution of the number of activity types targeted in surveyed activity recognition datasets.

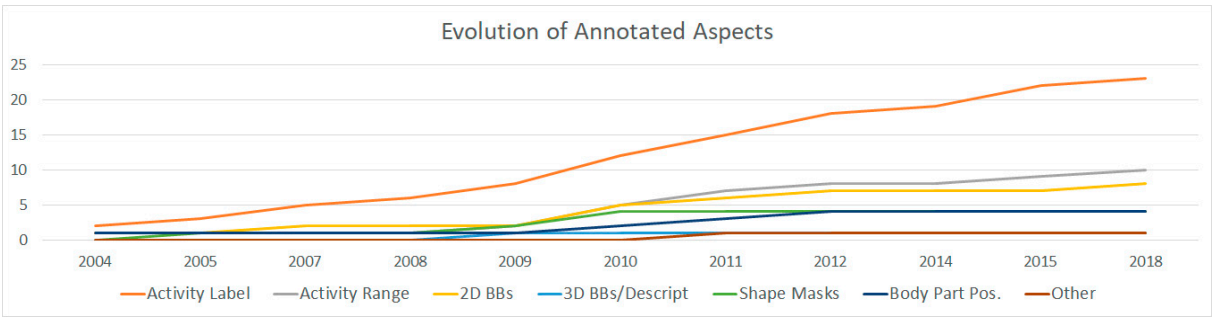

Figure 28. Evolution of annotated aspects in surveyed activity recognition datasets.

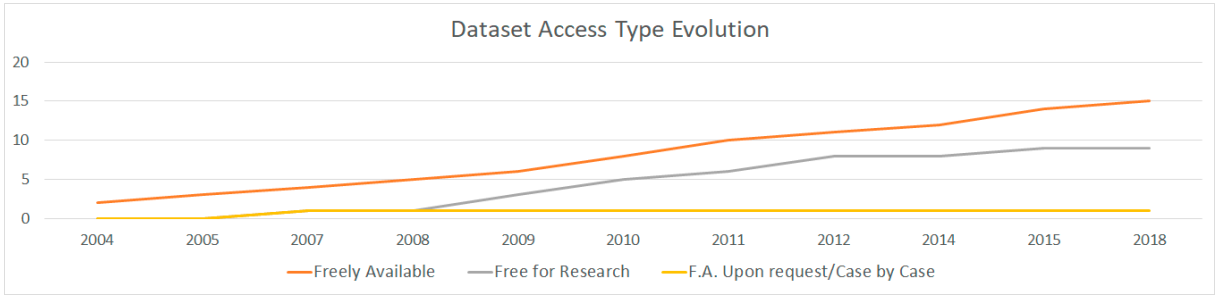

Figure 29. Evolution of access types to surveyed activity recognition datasets.

4.5. Analysis of Multipurpose Datasets

Table 6 presents a summary of the most relevant aspects of all surveyed multipurpose datasets. It describes, for each such dataset, its creation/publication date; the relevant pertaining bibliographic reference(s); the total amount of footage it comprises (in either hours (H) or seconds (s)); the total number of images; the total number of individual detections; the predominant aspects of the GT metadata; and some further notes with a particular focus on licensing aspects.

Table 6. Multipurpose datasets for Machine Learning (ML).

Name	Creation	Refs.	Footage Length	Number of Images	Number of Detection/Segments	Number of Classes	GT Metadata	Licensing/Notes
YFCC-100M	2015	[91]	-	>68 M	>>68 M	-	User tags, machine tags, etc.	Creative Commons
SUN Database	2010	[147]	-	>146 K	-	thousands	Labels and polygons (Pascal VOC)	Freely available for research
FMD	2014	[148]	-	1000	1000	10	Material labels	Creative Commons
VidTIMIT	2003	[149]	≈1935 s	≈45,580	-	-	Spoken content, head pose	Freely available for research
BDD100k	2020	[151]	>111 H	-	>>3.3 M	-	Object label, BBs, weather, lane markings, drivable areas, pixel-level annotations	Freely available for research upon request
Oxford Robotcar	2017	[152]	-	>20 M	-	-	LIDAR, GPS, and INS data	CC BY-NC-SA 4.0

As their designation implies, these datasets present an extreme variety of media contents, metadata contents, and formats.

Given their extreme variability, it is not simple to define/identify clear evolutive trends in these types of datasets. Nonetheless, there is still an apparent trend for a continuous expansion of the amount of media and metadata in these datasets.

Datasets of this type are also available in a typically open manner.

More than any other types of approached dataset, the one in current scope presents extreme heterogeneity and great obstacles for integration into a regular and standard tissue.

4.6. Analysis of Employed Metadata Formats

In this section, we present the summary and analysis of our survey of metadata formats (focusing on the five major ones) for the expression of MLCV ground truth. These results provide the answer for RQ22.

In Table 7, we summarize the main characteristics of the most relevant formats for metadata expression in MLCV datasets.

Table 7. ML dataset metadata formats.

Name	Base Format	Metadata Type	Vocabulary Expressiveness	Granularity	Employment
PASCAL VoC	XLML	low-level media features, segmentation, content semantics	free text	scenario and region	LabelMe, ImageNet, Pascal Visual Object Classes Project, Pascal Context, MIT Indoor Sceenes, SUN Database, Hollywood2
COCO JSON	JSON	segmentation, content semantics, administrative metadata	free text	scenario and region	iNaturalist, MegaFace and UrbanSound (JSON but not COCO JSON)
HDF-5	Binary	-	custom	custom	
CVML	XML	low-level media features, segmentation, content semantics	free text	frames, regions in frames, moving regions, shots	CAVIAR Project
ViPER	XML	segmentation, content semantics	free text	scenario and region	ViPER Project, iLIDS-VID, ETISEO Dataset, BEHAVE Dataset, THUMOS

The characteristics we look at are:

- Base format—the base textual format employed for the specific metadata format/language;
- Metadata type—this represents the dimension and focus of the enabled annotations. The possible such types are low-level media feature metadata; image segmenting metadata; spatial, temporal, and spatio-temporal dimensions describing metadata; content descriptive metadata that address the semantics of the media; and administrative metadata that describe such aspects as creation date, creator, etc.
- Vocabulary expressiveness—structuredness and comprehensiveness of the annotation vocabulary. The employment of an ontology adds to the vocabulary's structuredness. In their turn, ontologies may support only object concepts or also (and thus being more comprehensive) relationships concepts;
- Granularity—whether the permitted annotations may only apply to entire content assets as a whole or if they may focus on specific sub-parts of it. In static images, the tools in scope may enable scene or global-level annotations (for the entire image), or region-based, local, and segment-based annotations (image segments). For video content, annotation may refer to the entire video, temporal segments of it (shots), individual frames, regions within frames, or to moving regions, i.e., a region followed across a sequence of frames;
- Employment—the datasets in which each format is employed.

The current scenario regarding the metadata formats employed in MLCV datasets is a very fragmented one. Most initiatives define very simplistic metadata solutions that fit their specific needs. Only a few effectively devote some effort to the definition of more comprehensive and thought-out provisions for expressing MLCV GT.

The small number of metadata formats that resulted from such efforts (approached in this survey) has a very low employment rate. Only a few datasets actually use them. The typical situation is that they are employed within the datasets produced by the same research initiative that defined them (e.g., CVML) and within a few other derived or associated resources. Frequently, that employment is also done with abundant creative freedom, and so the format is not effectively employed but only a derived version of it.

Furthermore, the formats in scope were typically not defined with a concern for universality, comprehensiveness, versatility, or standardization. They were conceived with a focus on the needs of the research initiative from which they were spawned, while attempting to minimize complexity and maximize the expedience of implementation. No global vision of the ML field was employed to guide their definition.

Thus, the formats under analysis enable only a simplistic annotation of their target media content; they provide no means for annotating multi-sensory media content, enforce no separation of the different logical levels of the GT information according to the different levels of cognitive interpretation of reality that they pertain to (detection, spatial idealization, and semantic interpretation), and frequently miss altogether the means for the expression of some such levels.

When they do comprise the means for the description of conceptive metadata, these are very poor, particularly in terms of the definition of an overall interconnected semantic landscape that adequately weaves the relationships between events and objects.

These formats also provide very poor means for the spatio-temporally precise interconnection/interrelation between media and metadata. Media is simply stored in some files, metadata in others, and the latter references the earlier (or segments of it) through some arbitrary means.

Therefore, the identified metadata formats for MLCV GT are at a very initial phase of what would be a true universal format for the structuring and expression of CV metadata. One that may be used at virtually any dataset, regardless of its interpretative objectives, and the resulting metadata would be seamlessly reused by any other initiative later.

4.7. Overall Analysis and Discussion

In this section, we present the overall analysis of all surveyed datasets and related aspects. These results complete the answers for RQ1 to RQ22 and answer also RQ23.

With the evolution of MLCV science and technology, our interpretative objectives grow more ambitious. The input information (media content) becomes more voluminous and multi-dimensional, and the desired output information (machine interpretations) is necessarily more complex and abstract.

The media and metadata contents of MLCV datasets have changed to accommodate the needs for such an evolution. As our study shows, there is a general trend for the increase in volume and complexity of both the media content and its associated metadata.

Earlier work done on MLCV was predominantly focused on such interpretative objectives as image segmentation or object detection in relatively simple static images or facial recognition from small but close-up facial images. No temporal dimension was comprised by the realities to be detected.

This evolved to the processing of bigger, more information-rich visual content, and onto the processing of 2D video, and then onto the processing of both image and video with greater dimensionalities (2D+D, 3D, multispectral, multiview, etc.). All this is at the service of ever more ambitious interpretative objectives such as the simultaneous detection and location of faces, objects, or persons, the estimation of their 3D shapes, their tracking across time and space, and the identification of activities and complex behaviors.

Consequently, in the associated datasets, the sensory content and information have steadily become more rich, complex, realistic, and noisy. This means bigger images; longer and more information dense video; multi-dimensional visual media; complex visual scenery with multiple visible realities, increased variety of realities, occlusions, shadows, background noise; etc.

The metadata component of MLCV datasets has not been as complexified as the media one, but it still grew in sophistication and became more varied and precise to cater to the needs of MLCV tools with growing discerning capabilities and training requirements.

Thus, it evolved from simple tags describing the relevant visual or auditory content of an entire recording/acquisition to more complex constructs comprising a more precise definition of the detection of the targeted realities by defining the specific fragments of images or video registries where detections occur (e.g., bounding boxes, bit masks, contour masks); a semantically richer identification of the detected realities; and further contextualizing information (e.g., explanations of occlusions, visible facial expressions, head pose).

Some level of content reutilization exists among the observed datasets, as shown in Figure 30 (this figure focuses only on the minority of cases where such reusage occurs), but it is not the predominant case. This situation typically consists of the reutilization of general-purpose visual content (and, sometimes, of the associated metadata or parts of it) hosted at large commercial media sites (e.g., YouTube, Bing, Google, Flickr). In these cases, the reutilized media and metadata content is typically subjected to a relatively extensive processing before employment or of more specialized visual content, produced by large collaborative efforts developed, typically, in the context of, ML-based, CV research competitions (e.g., COCO, LabelMe, Labeled Faces in the Wild, ImageNet, or PascalVoC). In such cases, the media and metadata are closer to the precise needs of the reutilizing entity (dataset) and, as such, they are typically exploited without requiring extensive prior transformation.

The production of the metadata component of MLCV datasets has been moving toward a greater employment of crowdsourcing (particularly through the services of Amazon Mechanical Turk). The datasets whose metadata were crowdsourced are summarized in Table 8. Nonetheless, they are still a minority.

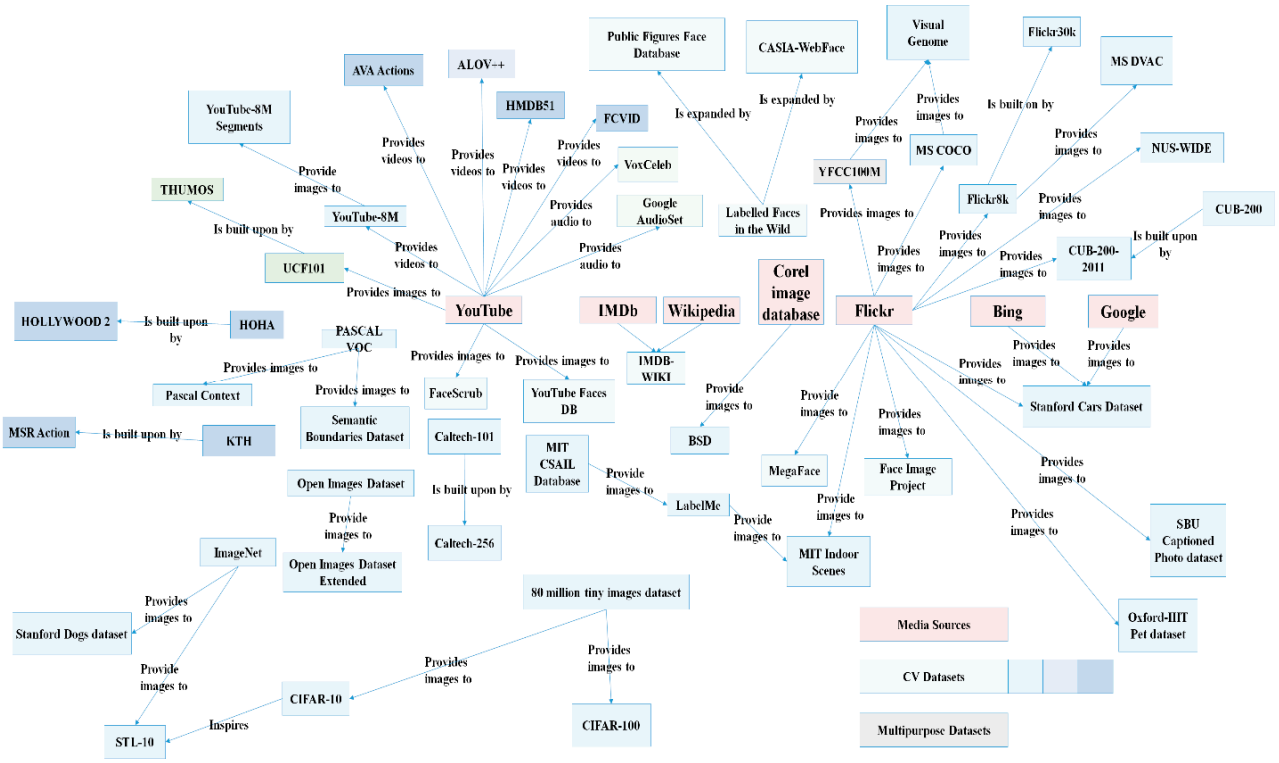

Figure 30. Relationships between datasets pertaining to flow/reutilization of media content.

Table 8. Crowdsourcing means employed for ML dataset metadata production.

Name	Application Domain	Crowdsourcing Means	Notes
UMDFaces	Facial Recognition	Amazon Mechanical Turk	
CUB-200-2011	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
SBD	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
Stanford Cars Dataset	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
FGVC-Aircraft	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
MS DVAC	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
MS COCO	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
Flickr30k	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
YouTube-BB	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
Visual Genome	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
NYU Depth D.	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
iNaturalist	Segmentation, Object and Scenario Recognition	Amazon Mechanical Turk	
ImageNet	Segmentation, Object and Scenario Recognition	Other crowdsourcing means	
Open Images Dataset	Segmentation, Object and Scenario Recognition	Google Crowdsource Android app	Partial crowdsourcing
VIRAT	Activity and Behavior Detection	Amazon Mechanical Turk	
ActivityNet	Activity and Behavior Detection	Amazon Mechanical Turk	
AVA Actions	Activity and Behavior Detection	Other crowdsourcing means	
Mozilla Common Voice	Speech Recognition	Other crowdsourcing means	
Fluent Speech Commands	Speech Recognition	Other crowdsourcing means	

The majority of reutilization scenarios consist of the employment, by specialized datasets, of specific subsets of more general-purpose datasets (e.g., Public Figure Face Database re-employing content from Labeled Faces in the Wild or Stanford Dogs Dataset reusing content from ImageNet) in a predominately opportunistic fashion, without contributing much to the extension of the overall ML dataset tissue.

The current scenario of MLCV dataset media content and metadata is one of specialization. However, this is the result of a simplistic focus on the requirements of immediate research goals, and it also stems from the fact that such tools, albeit their growing power, are still narrowly focused in their capabilities. It is not the specialization that would occur after the development of a mature general-purpose solution is achieved (general purpose machine intelligence and general-purpose dataset) and then different specializations take place. The current specialization is merely a simplistic solution to address small portions of the overall objective of reality interpretation by machines, aiming at a merger (hopefully) at a later time.

Therefore, all the components of MLCV datasets are still very much focused on specific application areas, and the overall dataset tissue is near completely disconnected.

The design of these resources is not conditioned by any concern for universality or interoperability nor guided by any vision conducive to the formation of a global tissue for machine teaching.

There is no established common/standardized structure for the benchmarking of ML algorithms. The only means that exist to enable such a process are the study and replication of published research works, and the participation in disconnected and unsystematic research initiatives/competitions (e.g., VOT challenge [156] or the COCO Image Segmentation Challenge [79]) that periodically pop in and out of existence, enabling researchers, in that period, to compare their work against common shared datasets.

The panorama regarding the actual online repositories for the datasets in scope is similarly simplistic and heterogeneous. Typically, they present limited and unsophisticated interfacing capabilities and are unable to provide a precise, detailed, and dynamic reading and/or writing access to the information that they hold. In many cases, these repositories allow only for a simple download of entire datasets (e.g., Labeled Faces in the Wild, CMU Multi-PIE Face Database, or ImageNet). These repositories were set up for an expeditious sharing of media and metadata resources amongst a small, interested community of researchers and developers. Thus, they are distant from being an interoperating component of a larger integrated tissue.

The overall scenario pertaining to the expression of GT metadata (in MLCV datasets) is one of total absence of standards or even common and shared tools. The overwhelming majority of research initiatives (i.e., their associated datasets) employ ad hoc metadata solutions specifically tailored for their purposes that are as simplified as possible. Little to no concern is had regarding the structuring of the information in a logically layered and integrated manner, nor with its ease of interpretation, universality, or long-term reusability. Employed metadata formats are simple, straightforward, and minimalist in structure and vocabulary. Some evolution, in terms of the relatively growing formality of the metadata, has been observed, as this is dictated by the growing density and preciseness required for MLCV metadata. However, this has been slow.

Only a few such formats were actually envisioned as formal solutions and enjoy some diffusion in their employment. Even so, these formats are still simplistic and not conceived as a universal tool for the expression and sharing of MLCV metadata.

They typically focus on the expression of some very domain-specific ground-truth information (e.g., CVML, ViPER, Pascal VOC) for employment in the training and testing of ML applications. Even if such formats do comprise, sometimes, a sensory and an interpretative component, the following limitations apply: they are always very narrow focused; provide inadequate bindings between sensory and interpretative data, or none at all; comprise and intermix, without any criteria, information of various different types and levels of abstraction, i.e., information pertaining to different levels of cognition; and

are invariably very incomplete. Frequently, the sensory component is completely unreferenced by the metadata, and thus, no effective binding of the interpretative metadata to it is provided.

Furthermore, their use has been mostly sporadic, and there is no present trend that is conducive to an increase in their usage or in their universality and versatility.

Summarizing what we have stated above, the level of media content reutilization between MLCV datasets is low; the logical connection between such datasets is very small; the dataset repositories are heterogeneous and non-interoperable; practically no standards exit to enforce a common expression of the metadata component of the datasets; few attempts have been made to develop such common metadata expression formats; and their uptake is very small. The result of this is an overall scenario of heterogeneity, disconnectedness, ad hoc solutions, and a complete absence of a view that is conducive to the development of a global ML teaching tissue and a universal language for the expression of ML metadata. Existing resources present excessive specificity, low versatility, and a non-homogenous documentation of their structure and inner logic. This greatly reduces the possibility for a seamless cooperative production and easy and efficient sharing of these resources for an exchange and comparison of results. It fails to take advantage of the obvious potential that exists for the formation of a global synergistic ML tool training and testing data structure. Research in MLCV stands only to benefit from the formation of such a standardized and integrated datasets tissue.

Our view is that the progress of ML will require a continuous growth of ML dataset contents in both volume and information richness. They will become increasingly complex, and it will become necessary to employ adequately defined formats and advance to an effective integration and standardization of this tissue. This way, in a manner somewhat like a sinusoidal movement between specialization and integration, a universal general-purpose format for the structuring of ML datasets and the expression of their metadata will need to be attained.

As evidenced by the survey, the metadata component of MLCV datasets may contain a large variety of information, depending on what is intended that machines may "learn" from it. Thus, it may consist of low-level features of images, image segment definition by bounding boxes, image segment classifications, descriptions of the semantic landscape, etc.

Superficially, it may appear that there is no overall structure or logic interconnecting these different types of information, but that logic does exist. The annotation of media information with interpretative metadata consists of the enrichment of the base media content with information describing the interpretative results of higher levels of cognition. This interpretation may be contextually dependent; however, all information resulting from the interpretation of reality (through the interpretation of images or sounds) conforms to a layered structure of progressively higher levels of abstraction.

The cognitive interpretative information stack begins with low-level features characterizing media content (e.g., color characteristics), and it continues onto the basic segmentation of the observed reality, to the construction of a coherent space–time perception of the observed reality, and onto a higher semantic interpretation of reality comprising the realization as object detection and identification, facial recognition, or activity identification.

The universal (desirably) tool that we deem necessary for the expression of ML dataset metadata should be defined in accordance with the described cognitive stack.

5. Limitations

The survey, summary, and analysis work here presented considered a vast number of documental sources pertaining to MLVC datasets with a preference for peer-reviewed papers. Nonetheless, it faces some limitations, or obstacles, resulting from the specific characteristics of the MLCV research field in its present state.

The core aspect of present MLCV research is the actual algorithms and mechanisms for information interpretation and/or inference. The employed datasets and, particularly, their metadata component, are regarded as a necessary but auxiliary part of the research

work without much scientific added value. For this, the datasets do not merit a great deal of attention nor a very extensive or precise description.

It is expected that researchers that need to employ a specific dataset will figure out or deduce whatever details are missing based on the observation of the dataset resources themselves and on their research experience.

This way, the provided descriptive information is typically succinct and infrequently poor. Many times, the information is available in a fragmented way, and some aspects need to be pieced together or deduced. Thus, there is a general lack of precise information on the media and, especially, on the metadata content of MLCV datasets. Consequently, there is also a lack of explicit information about the employed formats for metadata expression.

The information that is available about MLCV datasets is not uniform. It varies greatly between datasets. In addition, it sometimes happens that different sources about the same MLCV provide somewhat conflicting information.

In addition, some of the encountered datasets are statistical outliers in some of the observed aspects (e.g., number of images or registered object detections in the metadata). These are uncommon and typically consist of datasets built by some very large-scale initiatives. Their inclusion in the overall data tends to obscure the predominant evolution patterns.

We seek to remediate this overall situation by:

- Crossing information obtained from various different documental sources about each MLCV dataset so as to acquire information that is complete and coherent;
- Approaching a very broad group of datasets (for each of the surveyed application domains) to average out any partial information insufficiencies and increase the representativeness of our sample;
- Leaving out of the synthesis and analysis, statistically aberrational data, whilst explicitly mentioning that exclusion (Section 4) and justifying it.

Nonetheless, all that we state above constitutes limitations to our survey.

Even though we have done an exhaustive search (employing the information sources mentioned in Section 2.2), there is a non-negligible possibility that relevant MLCV datasets may have been left out.

6. Conclusions

6.1. The Way Forward

Natural cognition is the process through which animals apprehend and understand their environment. This process implies the acquisition of information pertaining to the outer world through the reception of stimuli (visual, auditory, tactile, olfactory, etc.), by the sensory organs; the processing of such stimuli, at different levels, by the different components of the nervous system; and their interpretation for the construction of a mental image of the world (specifically, at the central part of that system) that allows animals to act upon that world.

Cognition results from the integration of all sensory interpretations into a single and multidimensional interpretation of reality.

Our information technology has abundant capability to collect (store and transact) sensory information from the outer world (image, sound, etc.). It is presently acquiring the capability to learn to interpret that sensory information in increasingly flexible and abstract ways. Thus, we are witnessing the emergence of something that may be equated to a Synthetic Cognition (SC).

Considering this, the contents of MLCV datasets may be seen as the representation of the informational outcomes of a cognitive process; i.e., they consist of acquired visual sensory information and of its corresponding (correct) interpretation. They are both equitable to registries of a synthetic cognitive processes and to "books" to teach a synthetic cognition.

This is a powerful and prolific analogy that should be employed as a guiding vision to future developments of ML dataset contents and formats, building on such works as [157–159].

As we show throughout the entire survey, and in Section 4 in particular, the current panorama regarding MLCV datasets is extremely fragmented, heterogeneous, unstandardized, and metadata poor. Good MLCV datasets are costly resources to produce, but the immense potential that exists for their global cooperative production and exploitation is being wasted.

Current and future progress in MLCV research requires that this loss of synergy be addressed. The same argument is made, for instance, in the UK's government-defined industrial strategy for the artificial intelligence industry [160], in the USA's National Science and Technology Council recommendations regarding artificial intelligence [14,161]. The Open Data Institute [162] also puts the resolution of this issue forward as a necessary component of the data trust infrastructure that they believe should be constructed.

MLCV datasets should be integrated into a global dataset tissue to be developed cooperatively using standard formats and tools. The (media and metadata) formats employed within this tissue should seek to be universal to make said tissue easily exploitable, expandable, or customizable into any direction. Anyone should be able to get a desired segment of the dataset tissue (content and metadata), add any further necessary information (and contribute it to the global tissue), and run with it. MLCV provisions should output their results also in standard, universal easily shareable and comparable formats.

The online repositories for these resources should provide a sophisticated and detailed access to their contents, enabling precise and fine-grained writing and reading from them through a standardized interface. They must also enforce the structural and logical correctness of the content that they hold.

The tools for the expression of the metadata component of MLCV datasets should be standardized and tentatively universal. They enable a layered and independent expression of the different logical levels of the interpretation of sensory (audio and visual) information. These should be connected through explicit and dynamic means that enable an agile attachment and detachment of such connections. Furthermore, said metadata tools should be prepared to coherently deal with the characterization/annotation of base sensory media of various different types and different dimensionalities. In addition, the interpretative information should be structured in a manner that facilitates its simultaneous reproduction/presentation together with the source media content, and the entire information objects (merging media and associated metadata) should be constructed in a manner that facilitates their sharing, manipulation, and exploitation.

The process of constructing a global integrated ML dataset tissue should be guided by the earlier presented vision. Employing an analogy associating ML dataset contents with synthetic cognitive experience records enables gaining extremely valuable insight from the human cognitive process (the cognitive process best known to us). This will promote the enforcement of logical and structural correctness in the definition of those information objects and on the protocols for communicating and interacting with their repositories.

This vision will ease the development of logically adequate and normalized/standardized tools for the expression of the different components of ML dataset metadata (i.e., interpretative cognitive information), from the lower-level aspects (i.e., the detection of visual shapes and patterns) to the higher level ones (i.e., the semantic interpretation), in a manner that not only does not randomly mix them all up but also enables their complete expression and storage as sharable information objects.

Employing this vision will also ease the definition of adequate and capable means for sharing the information objects in scope across the ML community, thus enabling a standard and efficient communication and cooperation and the emergence of a true global ML dataset tissue.

6.2. Final Remarks

There is a plethora of datasets developed for MLCV research on image and video interpretation. This survey approaches a representative subset of such resources. This subset stretches in time from the early beginnings of datasets for MLCV research to the

present day, looking into those of greater relevance for both technical and scientific reasons but also for historical ones.

What the surveyed datasets reveal is an abundance of different types and dimensionalities of sensory information and an equal variety of associated annotations, extracted features, and meta-information in general.

There is a general and time-consistent growth trend regarding the volume and complexity of both the media and metadata content, which is a function of the expanding capabilities of MLCV tools and, consequently, of their training and testing requirements.

Datasets may now comprise hundreds of thousands or even millions of images, countless hours of video, and millions of increasingly complex individual annotations. Some of the observed datasets are general purpose, but most are still focused on a specific MLCV objective, or on a set of related ones, and thus so is their media and metadata content.

Together, the datasets in scope form a very heterogenous tissue regarding the comprised media content, its acquisition mode, dimensionality, and format; the comprised metadata (typically, the GT information), its expression format and the means for its production; and the storage and sharing of both media and metadata.

Only ad hoc solutions are employed, particularly regarding the expression of metainformation. No standards exist to govern the overall structuring of datasets, their effective contents, or the formats employed. In fact, there is no guiding vision of a global integrated research effort and informational tissue for the teaching of MLCV provisions.

Production efforts (particularly of the GT metadata) are increasingly sustained though crowdsourcing, but there is still no vision of an integrated system for its cooperative production and reusage.

This way, all the existing potential for the cooperation and reuse of MLCV datasets is not really being taken advantage of. For this to occur, MLCV datasets should be developed in a manner that weaves them into a global integrated tissue, which may be built and exploited cooperatively, in a straightforward manner, using standard formats (for media and metadata) and tools. Dataset repositories should become more sophisticated and maintain a standardized operation.

The standardization and integration process that needs to take place should be guided by a common comprehensive vision of the ML field that facilitates its seamless weaving together. That vision should be one that establishes a parallel between natural (cognition and its components) and ML technology as its emerging synthetic counterpart. This will foster and guide the development of the latter by providing a reliable and time-tested basis from which we may learn and derive extremely valuable insight, particularly for assessing the adequateness, completeness, and logical correctness of the means for the expression and manipulation of ML dataset metadata.

Thus, this survey contributes with a comprehensive view of the panorama pertaining to MLCV datasets: its past, present, and predictable immediate future. It identifies the lack of standardization integration and synergies (particularly in the production and exploitation of datasets) as the main limitation in this context and lays forth the necessary developments to address such limitations and the inspiring vision to guide such developments.

Author Contributions: H.F.C. and M.T.A. and J.S.C. were responsible for the ideation of this work. The literature search and data collection were performed by H.F.C. The data analysis was done by H.F.C. The drafting and revision of the paper was done by all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work was developed with the financial support of the Fundação para a Ciência e Tecnologia (FCT), Portugal, within the scope of the post-Doctoral grant with the reference number SFRH/BPD/108329/2015. This work was also partially financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e Tecnologia within project POCI-01-0145-FEDER-028857. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761. [CrossRef]
2. Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Comput. Vision Image Underst.* **2013**, *117*, 633–659. [CrossRef]
3. Jaimes, A.; Sebe, N. Multimodal human-computer interaction: A survey. *Comput. Vision Image Underst.* **2007**, *108*, 116–134. [CrossRef]
4. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
5. Mariano, V.Y. Performance Evaluation of Object Detection Algorithms. In Proceedings of the 16th International Conference on Pattern Recognition, Quebec City, QC, Canada, 11–15 August 2002; pp. 965–969.
6. Everingham, M.; Eslami, S.A.; van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision* **2015**, *111*, 98–136. [CrossRef]
7. Zhang, J.; Li, W.; Ogunbona, P.O.; Wang, P.; Tang, C. RGB-D-based action recognition datasets: A survey. *Pattern Recognit.* **2016**, *60*, 86–105. [CrossRef]
8. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vision* **2020**, *128*, 261–318. [CrossRef]
9. Borji, A.; Cheng, M.M.; Hou, Q.; Jiang, H.; Li, J. Salient object detection: A survey. *Comput. Visual Media* **2019**, *5*, 117–150. [CrossRef]
10. Bernardi, R.; Cakici, R.; Elliott, D.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N.; Keller, F.; Muscat, A.; Plank, B. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.* **2016**, *55*, 409–442. [CrossRef]
11. Samaria, F.S.; Harter, A.C. Parameterisation of a stochastic model for human face identification. In Proceedings of the 1994 IEEE Workshop on Applications of Computer Vision, IEEE, Sarasota, FL, USA, 5–7 December 1994; pp. 138–142.
12. Olivetti Face Database Website. Available online: http://www.cam-orl.co.uk/facedatabase.html (accessed on 19 January 2021).
13. The FERET Database WebPage. Available online: https://www.nist.gov/programs-projects/face-recognition-technology-feret (accessed on 19 January 2021).
14. National Science and Technology Council, Preparing for the Future of Artificial Intelligence. 2016. Available online: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (accessed on 19 January 2021).
15. Messer, K.; Matas, J.; Kittler, J.; Luettin, J.; Maitre, G. XM2VTSDB: The extended M2VTS database. In Proceedings of the Second International Conference on Audio and Video-Based Biometric Person Authentication, Washington, DC, USA, 22–24 March 1999; Volume 964, pp. 965–966.
16. XM2VTSDB Website. Available online: http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/ (accessed on 19 January 2021).
17. Beumier, C.; Acheroy, M. Automatic 3D face authentication. *Image Vision Comput.* **2000**, *18*, 315–321. [CrossRef]
18. 3D_RMA Database Website. Available online: http://www.sic.rma.ac.be/~{}beumier/DB/3d_rma.html (accessed on 19 January 2021).
19. Marszalec, E.A.; Martinkauppi, J.B.; Soriano, M.N.; Pietikaeinen, M. Physics-based face database for color research. *J. Electron. Imaging* **2000**, *9*, 32–39.
20. Georghiades, A.S.; Belhumeur, P.N.; Kriegman, D.J. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 643–660. [CrossRef]
21. Yale Face Databases Website. Available online: http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html (accessed on 19 January 2021).
22. Phillips, P.J.; Flynn, P.J.; Scruggs, T.; Bowyer, K.W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; Worek, W. Overview of the face recognition grand challenge. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 947–954.
23. Panis, G.; Lanitis, A. An overview of research activities in facial age estimation using the FG-NET aging database. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 737–750.
24. Grgic, M.; Delac, K.; Grgic, S. SCface–surveillance cameras face database. *Multimed. Tools Appl.* **2011**, *51*, 863–879. [CrossRef]
25. Yin, L.; Wei, X.; Sun, Y.; Wang, J.; Rosato, M.J. A 3D facial expression database for facial behavior research. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), IEEE, Southampton, UK, 10–12 April 2006; pp. 211–216.
26. Huang, G.B.; Ramesh, M.; Berg, T.; Learned-Miller, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*; Technical Report 07–49; University of Massachusetts: Amherst, MA, USA, 2007.
27. Gao, W.; Cao, B.; Shan, S.; Chen, X.; Zhou, D.; Zhang, X.; Zhao, D. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2007**, *38*, 149–161.
28. Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S. Multi-pie. *Image Vision Comput.* **2010**, *28*, 807–813. [CrossRef] [PubMed]

29. Kumar, N.; Berg, A.C.; Belhumeur, P.N.; Nayar, S.K. Attribute and simile classifiers for face verification. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, IEEE, Kyoto, Japan, 27 September–4 October 2009; pp. 365–372.
30. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; van Knippenberg, A.D. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
31. Gupta, S.; Castleman, K.R.; Markey, M.K.; Bovik, A.C. Texas 3D face recognition database. In Proceedings of the 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI), IEEE, Austin, TX, USA, 23–25 May 2010; pp. 97–100.
32. Wolf, L.; Hassner, T.; Maoz, I. Face recognition in unconstrained videos with matched background similarity. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), IEEE, Colorado Springs, CO, USA, 21–23 June 2011; pp. 529–534.
33. Wong, Y.; Chen, S.; Mau, S.; Sanderson, C.; Lovell, B.C. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *2011 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR 2011 WORKSHOPS)*; IEEE: Colorado Springs, CO, USA, 2011; pp. 74–81.
34. Ng, H.W.; Winkler, S. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*; IEEE: Paris, France, 2014; pp. 343–347.
35. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv* **2014**, arXiv:1411.7923.
36. Eidinger, E.; Enbar, R.; Hassner, T. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. For. Secur.* **2014**, *9*, 2170–2179. [CrossRef]
37. Min, R.; Kose, N.; Dugelay, J.L. Kinectfacedb: A kinect database for face recognition. *IEEE Trans. Syst. Man Cybern. Syst.* **2014**, *44*, 1534–1548. [CrossRef]
38. Sun, Y.; Wang, X.; Tang, X. Hybrid deep learning for face verification. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1489–1496.
39. Kemelmacher-Shlizerman, I.; Seitz, S.M.; Miller, D.; Brossard, E. The megaface benchmark: 1 million faces for recognition at scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4873–4882.
40. Bansal, A.; Nanduri, A.; Castillo, C.D.; Ranjan, R.; Chellappa, R. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*; IEEE: Denver, CO, USA, 2017; pp. 464–473.
41. Rothe, R.; Timofte, R.; van Gool, L. Dex: Deep expectation of apparent age from a single image. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 11–18 December 2015; pp. 10–15.
42. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*; IEEE: Xi'an, China, 2018; pp. 67–74.
43. Tufts Face Database Webpage at Kaggle. Available online: https://www.kaggle.com/kpvisionlab/tufts-face-database (accessed on 19 January 2021).
44. Nene, S.A.; Nayar, S.K.; Murase, H. Columbia Object Image Library (coil-100)-Technical Report No. CUCS-006-96. 1996. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.360.6420&rep=rep1&type=pdf (accessed on 19 January 2021).
45. Microsoft Research Cambridge Dataset Website. Available online: https://www.microsoft.com/en-us/research/project/image-understanding (accessed on 19 January 2021).
46. Martin, D.; Fowlkes, C.; Tal, D.; Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th IEEE International Conference on Computer Vision. ICCV 2001*; IEEE: Vancouver, BC, Canada, 2001; Volume 2, pp. 416–423.
47. Lai, K.; Bo, L.; Ren, X.; Fox, D. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE International Conference on Robotics and Automation*; IEEE: Shanghai, China, 2011; pp. 1817–1824.
48. LeCun, Y.; Huang, F.J.; Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004*; CVPR 2004; IEEE: Washington, DC, USA, 2004; Volume 2, p. II-104.
49. Moreels, P.; Perona, P. Evaluation of features detectors and descriptors based on 3D objects. In *Tenth IEEE International Conference on Computer Vision (ICCV'05)*; IEEE: Beijing, China, 17–21 October 2005; Volume 1, pp. 800–807.
50. Griffin, G.; Holub, A.; Perona, P. Caltech-256 object category dataset (Self-published). 2007. Available online: https://authors.library.caltech.edu/7694/1/CNS-TR-2007-001.pdf (accessed on 19 January 2021).
51. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vision* **2008**, *77*, 157–173. [CrossRef]
52. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 248–255.
53. Brostow, G.J.; Fauqueur, J.; Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.* **2009**, *30*, 88–97. [CrossRef]
54. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. 2009. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf (accessed on 19 January 2021).
55. Chua, T.S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. NUS-WIDE: A real-world web image database from National University of Singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval, Santorini, Greece, 8–10 July 2009; pp. 1–9.

56. Quattoni, A.; Torralba, A. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 413–420.
57. SBU Captioned Photo Dataset Webpage. Available online: http://vision.cs.stonybrook.edu/~{}vicente/sbucaptions (accessed on 19 January 2021).
58. Ordonez, V.; Kulkarni, G.; Berg, T.L. Im2text: Describing images using 1 million captioned photographs. In Proceedings of the 24th International Conference on Neural Information Processing Systems, Granada, Spain, 12–17 December 2011; pp. 1143–1151.
59. Coates, A.; Ng, A.; Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft., Lauderdale, FL, USA, 11–13 April 2011; pp. 215–223.
60. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 91–102.
61. Caltech-UCSD Birds-200-2011 Dataset Website. Available online: http://www.vision.caltech.edu/visipedia/CUB-200-2011.html (accessed on 19 January 2021).
62. Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; Malik, J. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*; IEEE: Barcelona, Spain, 2011; pp. 991–998.
63. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.F. Novel dataset for fine-grained image categorization: Stanford dogs. In Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Colorado Springs, CO, USA, 25 June 2011; Volume 2. No. 1.
64. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
65. Kumar, N.; Belhumeur, P.N.; Biswas, A.; Jacobs, D.W.; Kress, W.J.; Lopez, I.C.; Soares, J.V. Leafsnap: A computer vision system for automatic plant species identification. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 502–516.
66. Parkhi, O.M.; Vedaldi, A.; Zisserman, A.; Jawahar, C.V. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Providence, RI, USA, 2012; pp. 3498–3505.
67. Mogelmose, A.; Trivedi, M.M.; Moeslund, T.B. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Trans. Intell. Transp. Syst.* **2012**, *13*, 1484–1497. [CrossRef]
68. Scharwächter, T.; Enzweiler, M.; Franke, U.; Roth, S. Efficient multi-cue scene segmentation. In *German Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 435–445.
69. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3d object representations for fine-grained categorization. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 1–8 December 2013; pp. 554–561.
70. Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv* **2013**, arXiv:1306.5151.
71. Microsoft Research Dense Visual Annotation Corpus Download Page. Available online: https://www.microsoft.com/en-us/download/details.aspx?id=52523 (accessed on 19 January 2021).
72. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
73. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; Westling, P. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*; Springer: Cham, Switzerland, 2014; pp. 31–42.
74. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
75. Wang, T.; Gong, S.; Zhu, X.; Wang, S. Person Re-Identification by Video Ranking. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*; Springer: Zurich, Switzerland, 2014.
76. Timofte, R.; Zimmermann, K.; van Gool, L. Multi-view traffic sign detection, recognition, and 3D localisation. *Mach. Vision Appl.* **2014**, *25*, 633–647. [CrossRef]
77. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.G.; Lee, S.W.; Fidler, S.; Urtasun, R.; Yuille, A. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 891–898.
78. Cordts, M.; Omran, M.; Ramos, S.; Scharwächter, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset. In *Proceeding of the 28th IEEE Conference on Computer Vision and Pattern Recognition, Workshop on the Future of Datasets in Vision*; IEEE: Boston, MA, USA, 11 June 2015; Volume 2.
79. Yang, L.; Luo, P.; Loy, C.C.; Tang, X. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3973–3981.
80. YouTube8M Dataset Webpage at Google Research Website. Available online: https://research.google.com/youtube8m (accessed on 19 January 2021).
81. Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; van Gool, L.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 724–732.
82. van Horn, G.; Aodha, O.M.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S. The inaturalist species classification and detection dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8769–8778.

83. Enzweiler, M.; Gavrila, D.M. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 2179–2195. [CrossRef] [PubMed]
84. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision* **2017**, *123*, 32–73. [CrossRef]
85. Open Images Dataset Website. Available online: https://ai.googleblog.com/2016/09/introducing-open-images-dataset.html (accessed on 19 January 2021).
86. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv* **2018**, arXiv:1811.00982. [CrossRef]
87. Sigal, L.; Black, M.J. *Humaneva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion*; Brown Univertsity TR: Providence, RI, USA, 2006; p. 120.
88. Ess, A.; Leibe, B.; van Gool, L. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th International Conference on Computer Vision*; IEEE: Rio de Janeiro, Brazil, 2007; pp. 1–8.
89. Wojek, C.; Walk, S.; Schiele, B. Multi-cue onboard pedestrian detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 794–801.
90. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 304–311.
91. KITTI Benchmark Suite Dataset Website. Available online: http://www.cvlibs.net/datasets/kitti (accessed on 19 January 2021).
92. Smeulders, A.W.; Chu, D.M.; Cucchiara, R.; Calderara, S.; Dehghan, A.; Shah, M. Visual tracking: An experimental survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1442–1468.
93. Visual Tracker Benchmark Dataset Webpage. Available online: http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html (accessed on 19 January 2021).
94. Liang, P.; Blasch, E.; Ling, H. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5630–5644. [CrossRef] [PubMed]
95. Li, A.; Lin, M.; Wu, Y.; Yang, M.H.; Yan, S. Nus-pro: A new visual tracking challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 335–349. [CrossRef]
96. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 445–461.
97. Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Pflugfelder, R.; Kamarainen, J.K.; Zajc, L.C.; Drbohlav, O.; Lukezic, A.; Berg, A.; et al. The seventh visual object tracking vot2019 challenge results. In Proceedings of the 12th IEEE International Conference on Computer Vision Workshops, Kyoto, Japan, 27 September–4 October 2019.
98. CAVIAR Project Website. Available online: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm (accessed on 19 January 2021).
99. KTH Dataset for Recognition of human actions HomePage. Available online: http://www.nada.kth.se/cvap/actions (accessed on 19 January 2021).
100. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004*; ICPR 2004; IEEE: Cambridge, UK, 2004; Volume 3, pp. 32–36.
101. WEIZMANN Dataset HomePage. Available online: http://www.wisdom.weizmann.ac.il/%7Evision/SpaceTimeActions.html (accessed on 19 January 2021).
102. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05)*; IEEE: Beijing, China, 17–21 October 2005; Volume 2, pp. 1395–1402.
103. ETSIO Dataset HomePage. Available online: http://www-sop.inria.fr/orion/ETISEO (accessed on 19 January 2021).
104. Nghiem, A.T.; Bremond, F.; Thonnat, M.; Valentin, V. ETISEO, performance evaluation for video surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*; IEEE: London, UK, 2007; pp. 476–481.
105. CASIA Action Dataset Website. Available online: http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp (accessed on 19 January 2021).
106. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Anchorage, AK, USA, 2008; pp. 1–8.
107. Yuan, J.; Liu, Z.; Wu, Y. Discriminative subvolume search for efficient action detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 2442–2449.
108. Marszalek, M.; Laptev, I.; Schmid, C. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 2929–2936.
109. Gkalelis, N.; Kim, H.; Hilton, A.; Nikolaidis, N.; Pitas, I. The i3dpost multi-view and 3d human action/interaction database. In *2009 Conference for Visual Media Production*; IEEE: London, UK, 2009; pp. 159–168.
110. BEHAVE Dataset HomePage. Available online: http://groups.inf.ed.ac.uk/vision/BEHAVEDATA (accessed on 19 January 2021).
111. Blunsden, S.J.; Fisher, R.B. The BEHAVE video dataset: Ground truthed video for multi-person behavior classification. *Ann. BMVA* **2010**, *4*, 1–12.

112. TV Human Interaction Dataset HomePage. Available online: http://www.robots.ox.ac.uk/~{}alonso/tv_human_interactions.html (accessed on 19 January 2021).
113. Patron-Perez, A.; Marszalek, M.; Zisserman, A.; Reid, I. High Five: Recognising human interactions in TV shows. In Proceedings of the British Machine Vision Conference (BMVC), Aberystwyth, UK, 31 August–3 September 2010.
114. MuHAVi Dataset HomePage. Available online: http://velastin.dynu.com/MuHAVi-MAS (accessed on 19 January 2021).
115. Singh, S.; Velastin, S.A.; Ragheb, H. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*; IEEE: Boston, MA, USA, 2010; pp. 48–55.
116. Ryoo, M.S.; Aggarwal, J.K. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th International Conference on Computer Vision*; IEEE: Kyoto, Japan, 2009; pp. 1593–1600.
117. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A Large Video Database for Human Motion Recognition. In Proceedings of the 13th International Conference on Computer Vision (ICCV 2011), Barcelona, Spain, 6–13 November 2011.
118. Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.C.; Lee, J.T.; Mukherjee, S.; Aggarwal, J.K.; Lee, H.; Davis, L.; et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*; IEEE: Colorado Springs, CO, USA, 2011; pp. 3153–3160.
119. Denina, G.; Bhanu, B.; Nguyen, H.T.; Ding, C.; Kamal, A.; Ravishankar, C.; Roy-Chowdhury, A.; Ivers, A.; Varda, B. Videoweb dataset for multi-camera activities and non-verbal communication. In *Distributed Video Sensor Networks*; Springer: London, UK, 2011; pp. 335–347.
120. Rohrbach, M.; Amin, S.; Andriluka, M.; Schiele, B. A Database for Fine Grained Activity Detection of Cooking Activities. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), IEEE, Providence, RI, USA, 16–21 June 2012.
121. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Action, Classes from Videos in the Wild (Technical Report CRCV-TR-12–01), Centre for Research in Computer Vision from the University of Central Florida. 2012. Available online: https://arxiv.org/pdf/1212.0402.pdf (accessed on 19 January 2021).
122. Pirsiavash, H.; Ramanan, D. Detecting activities of daily living in first-person camera views. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Providence, RI, USA, 16–21 June 2012; pp. 2847–2854.
123. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
124. Idrees, H.; Zamir, A.R.; Jiang, Y.G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The THUMOS challenge on action recognition for videos in the wild. *Comput. Vision Image Underst.* **2017**, *155*, 1–23. [CrossRef]
125. Heilbron, F.C.; Escorcia, V.; Ghanem, B.; Niebles, J.C. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
126. Jiang, Y.G.; Wu, Z.; Wang, J.; Xue, X.; Chang, S.F. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 352–364. [CrossRef] [PubMed]
127. Jiang, Y.G.; Wu, Z.; Wang, J.; Xue, X.; Chang, S.F. FCVID: Fudan-Columbia Video Dataset. Available online: http://www.yugangjiang.info/publication/TPAMI17-supplementary.pdf (accessed on 19 January 2021).
128. Gu, C.; Sun, C.; Ross, D.A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6047–6056.
129. Thomee, B.; Shamma, D.A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L.J. The new data and new challenges in multimedia research. *arXiv* **2015**, arXiv:1503.01817.
130. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: San Francisco, CA, USA, 2010; pp. 3485–3492.
131. Sharan, L.; Rosenholtz, R.; Adelson, E.H. Accuracy and speed of material categorization in real-world images. *J. Vision* **2014**, *14*, 12. [CrossRef]
132. Sanderson, C. Automatic Person Verification Using Speech and Face Information. Ph.D. Thesis, School of Microelectronic Engineering of the Faculty of Engineering and Information Technology Griffith University, Brisbane, Australia, 2003.
133. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 15–16 June 2020; pp. 2636–2645.
134. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 year, 1000 km: The Oxford RobotCar dataset. *Int. J. Robot. Res.* **2017**, *36*, 3–15. [CrossRef]
135. FERET Colour Database Website. Available online: https://www.nist.gov/itl/products-and-services/color-feret-database (accessed on 19 January 2021).
136. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1021–1030.

137. Torralba, A.; Fergus, R.; Freeman, W.T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1958–1970. [CrossRef]
138. Catster Website. Available online: http://www.catster.com/ (accessed on 19 January 2021).
139. Dogster Website. Available online: http://www.dogster.com/ (accessed on 19 January 2021).
140. COCO Image Segmentation Challenge Website. Available online: https://cocodataset.org/#home (accessed on 19 January 2021).
141. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [CrossRef]
142. Open Images Extended–Crowdsourced Dataset Website. Available online: https://research.google/tools/datasets/open-images-extended-crowdsourced/ (accessed on 19 January 2021).
143. Real, E.; Shlens, J.; Mazzocchi, S.; Pan, X.; Vanhoucke, V. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5296–5305.
144. Website for Team AnnieWAY. Available online: http://www.kit.edu/kit/english/pi_2011_6778.php (accessed on 19 January 2021).
145. Wu, Y.; Lim, J.; Yang, M.H. Online object tracking: A benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2411–2418.
146. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: New York, NY, USA, 2019.
147. Lukežič, A.; Zajc, L.Č.; Vojíř, T.; Matas, J.; Kristan, M. Now you see me: Evaluating performance in long-term visual tracking. *arXiv* **2018**, arXiv:1804.07056.
148. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [CrossRef]
149. Lukezic, A.; Kart, U.; Kapyla, J.; Durmush, A.; Kamarainen, J.K.; Matas, J.; Kristan, M. CDTB: A color and depth visual object tracking dataset and benchmark. In Proceedings of the 12th IEEE International Conference on Computer Vision, Kyoto, Japan, 24 September–4 October 2019; pp. 10013–10022.
150. List, T.; Fisher, R.B. CVML–An XML-based Computer Vision Markup Language. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004. ICPR 2004.
151. Project ViPER Website. Available online: http://viper-toolkit.sourceforge.net (accessed on 19 January 2021).
152. Jankowski, C.; Kalyanswamy, A.; Basson, S.; Spitz, J. NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database. In *International Conference on Acoustics*; Speech and Signal Processing: Albuquerque, MA, USA, 1990; Volume 1, pp. 109–112.
153. HDF5 Support Page. Available online: http://portal.hdfgroup.org/display/HDF5/HDF5 (accessed on 19 January 2021).
154. NeonScience Webpage on HDF5. Available online: https://www.neonscience.org/about-hdf5 (accessed on 19 January 2021).
155. Doemann, D.; Mihalcik, D. Tools and techniques for video performances evaluation. In Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, 3–7 September 2000; pp. 167–170.
156. Visual Object Tracking Challenge Website. Available online: https://www.votchallenge.net/ (accessed on 19 January 2021).
157. Castro, H.; Alves, A.P. Cognitive Object Format. In *International Conference on Knowledge Engineering and Ontology Development*; Funchal: Madeira, Portugal, 2009.
158. Castro, H.; Monteiro, J.; Pereira, A.; Silva, D.; Coelho, G.; Carvalho, P. Cognition Inspired Format for the Expression of Computer Vision Metadata. *Multimed. Tools Appl.* **2016**, *75*, 17035–17057. [CrossRef]
159. Castro, H.; Andrade, M.T. ML Datasets as Synthetic Cognitive Experience Records. *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.* **2018**, *10*, 289–313.
160. Hall, W.; Pesenti, J. Growing the artificial intelligence industry in the UK. In *Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy*; OGL: London, UK, 2017.
161. Gal, M.S.; Rubinfeld, D.L. Data standardization. *NYUL Rev.* **2019**, *94*, 737. [CrossRef]
162. Open Data Institute Website. Available online: https://theodi.org/ (accessed on 19 January 2021).