International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 10 (2018) pp. 289-313
© MIR Labs, www.mirlabs.net/ijcisim/index.html

ML Datasets as Synthetic Cognitive Experience Records

H. Castro[1,2] and M. T. Andrade[1,2]

[1] INESC TEC, Campus da FEUP,
Rua Dr. Roberto Frias, 4200 - 465 Porto, Portugal

[2] Faculdade de Engenharia da Universidade do Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto Portugal

***Abstract* – Machine Learning (ML), presently the major research area within Artificial Intelligence, aims at developing tools that can learn, approximately on their own, from data. ML tools learn, through a training phase, to perform some association between some input data and some output evaluation of it. When the input data is audio or visual media (i.e. akin to sensory information) and the output corresponds to some interpretation of it, the process may be described as Synthetic Cognition (SC). Presently ML (or SC) research is heterogeneous, comprising a broad set of disconnected initiatives which develop no systematic efforts for cooperation or integration of their achievements, and no standards exist to facilitate that. The training datasets (base sensory data and targeted interpretation), which are very labour intensive to produce, are also built employing ad-hoc structures and (metadata) formats, have very narrow expressive objectives and thus enable no true interoperability or standardisation. Our work contributes to overcome this fragility by putting forward: a specification for a standard ML dataset repository, describing how it internally stores the different components of datasets, and how it interfaces with external services; and a tool for the comprehensive structuring of ML datasets, defining them as Synthetic Cognitive Experience (SCE) records, which interweave the base audio-visual sensory data with multilevel interpretative information. A standardised structure to express the different components of the datasets and their interrelations will promote re-usability, resulting on the availability of a very large pool of datasets for a myriad of application domains. Our work thus contributes to: the universal interpretability and reusability of ML datasets; greatly easing the acquisition and sharing of training and testing datasets within the ML research community; facilitating the comparison of results from different ML tools; accelerating the overall research process.**

***Keywords*: Machine-Learning, Datasets, Cyber-physical, Synthetic Cognition, Metadata**

I. Introduction

Work on the development of artificial intelligence has begun some decades ago. It has experienced a variable rate of progress, and has gone through different phases, since its inception. The predominant approach today is based on the employment of reasoning mechanisms which are expected to go through a learning/training phase, through which they build their own "model of the world". This approach may be summarily described as Machine Learning (ML).

During the learning/training phase of ML tool development, these are supplied with the information to be evaluated and a Ground-Truth (GT) meta-information describing the interpretations/evaluations to which the ML tool must arrive. Through the learning algorithms these tools learn to link input data to its adequate evaluation. The preparation of the training (and testing) datasets is one of the most labour consuming tasks, in ML research, particularly the GT metadata.

The ML research field is currently characterised by the existence of a broad set of independent initiatives and platforms which pursue different goals, and which are not concerned with a systematic cooperation or integration of their achievements or capabilities. No technical structures or standards exist to facilitate such a cooperation. These initiatives typically adopt their own defined structure and methodology for dataset development, which results in narrow-purposed had-hoc solutions, instead of more comprehensive and versatile ones, hindering their re-utilisations or extension by other parties, and thus, progress.

A number of ML competitions exist which make freely-available, to competitors, relevant datasets. These, albeit having the potential to foster homogenisation of structures and formats across the research community, in reality provide marginal contribution given that they are simplistic and focussed on very specific topics (e.g., object detection [1][2], human activity detection [3], or face recognition [4]).

ML research, in spite of some remarkable results having been obtained, is still in a somewhat immature phase. The focus is still on the development of interpretative algorithms for specific application domains, and not on the integration between platforms. Besides being expensive to produce, datasets are, therefore, also necessarily tailed for specific purposes. These reasons have, thus far, prevented the ML research community from investing much on cooperation at the dataset production stage. The structures of the employed dataset repositories and the formats of GT metadata have thus remained very simple and suffered almost no evolution for a long period of time.

We believe that the ML research (and industrial) community would greatly benefit from the widespread employment of standard means for the storage, structuring, interrelating and sharing of datasets (base data and ground-truth metadata). In an environment of open data sharing (as it is typical of the research environment), this would lead to the coalescence of a very large pool of usable datasets, pertaining to a myriad of application domains. Consequently ML research initiatives, would experience a radical reduction of the costs of training and testing dataset production. The comparison of results between ML tools would also be facilitated. Added manpower would, thus, become available for research in the actual learning algorithms.

The work here presented aims at defining such a standard set of means. Our work focuses on audio-visual datasets. In such datasets: the media information is akin to the sensory information acquired by the sensory organs of natural organisms (e.g. human beings); the GT metadata is akin to the interpretation of the sensory input; and the processing by ML tools is the equivalent to the cognitive interpretation process. This way, the overall ML process may be envisioned as a Synthetic Cognition (SC) process, and the datasets (holding either training data or the actual output data of the automated media processing) can be equated to "memories" or records of Synthetic Cognitive Experiences (SCEs). Such records are thus information objects that, in the latter case, result from synthetic cognitive activity, and in the earlier case are provided as "lessons" to synthetic cognitive activity. Employing this vision in the definition of the structure and formats of datasets (media and GT metadata), is thus a helpful guiding analogy, as it enables to take inspiration from its natural cognitions counterparts.

In light of the above, we have defined a set of solutions to enable true interoperability of ML tools and research work. We present here: a clear vision of how and why ML datasets may, and should, be regarded as SCE records (presented in section II); a survey on the current state-of-the-art regarding the storage, sharing, and structuring of datasets (those related to audio and visual base content) and associated ground-truth metadata (presented in section III); the definition of a versatile structure for a ML dataset repository (in section IV) including its overall capabilities, how it internally stores the different types of data that compose such datasets, and its outward service interface; and (as the core of our work, and guided by the earlier mentioned vision), a tool for the comprehensive structuring of training and testing datasets and for the expression of the multi-level ground-truth metadata describing the interpretations to be learned from the base sensory content. This format (named SynCog) is presented in in section IV. Section V provides a comparative analysis between the proposed SynCog tool and existing tools. In section VI we present our concluding remarks.

This work thus contributes to: the universal interpretability and reusability of ML datasets and output data (in the field of sensory interpretation); greatly easing the acquisition and sharing of training and testing datasets within the ML research community; and greatly facilitating the comparison of result between different ML tools.

II. Envisioning ML Datasets as SCE Records

A. Natural Cognition

Natural cognition is the process through which living animals sense and understand their environment, as well as decide and perform action upon it. At its earlier stages, it consisted only of basic sensing capabilities for the detection of relevant environmental characteristics (such as luminosity or sound), allied to a very limited processing capacity of that sensed information, to build its interpretation of reality, and to also limited means to act back upon the world. Throughout time sensory capabilities progressively expanded, processing and interpretative skills grew in capacity, enabling the formulation of more comprehensive and abstract interpretations of reality, and the means to act back upon the world also grew in capacity.

Even if natural cognition's capabilities may have increased through the ages, its basic blueprint has remained practically the same since its initial stages. It thus consists of an apprehension of, and action upon, external reality which combines: the capture of sensory stimuli by sensory organs; the communication of the acquired information, through the tissue of the nervous system (typically nerves), to a (generally centralized) component of that system; the processing, logical fusion and interpretation of the sensory acquired information by the mentioned component of nervous system; and the acting back upon the world through actuatory means (e.g. arms, hands, legs, claws, tentacles, mouths, etc.).

The, vastly important, processing component/organ builds the higher aspects of cognition: perception, which comprises the identification of reality's spatiotemporal dimensions and characteristics by performing the cognitive segmentation of the sensory input into logically different, but related, parts; and conception which comprises understanding the relationships between real world entities and events and assigning meaning to them [5]. It also handles decision making (as a function of the acquired sensations) and commands the reaction, to the outer world, by controlling the action of the actuating components of the body (e.g., arms, legs, mouth, etc.). The continuous sequence of cognitive events, experienced through life, are stored, in a complex mode, within the core nervous tissue, as its memories.

B. ML and the Coalescing Synthetic Cognition

A number of important, albeit disconnected, technological developments that have been occurring within the last six decades are now converging into the creation of a global cyber-physical structure which is akin to a synthetic cognition. Such developments are:

a) the construction of cheap and light artificial means for the capture and registration of media information. These devices now constitute a pervasive sensory echelon for a global cyber-physical structure;

b) the continuous increase in the global availability of computational processing power (now situated between the upper exascale and the lower zettascale [6]). Collectively, all existing computing devices make up the basis for an information-processing tissue (i.e. cerebral tissue) of an emerging global cyber-physical structure;

c) the communicative interconnection of information processing machines through the relentless growth, of the Internet (presently interconnectin something like 1G hosts [7]). This sprawling tissue enables the communication between the different parts of the emerging cyber-physical structure, and effectively brings it into existence;

d) the development of provisions for the automated interpretation of sensory information (e.g. audio and video interpretation), and automated decision taking. In this regard ML techniques have achieved relevant results in various fields. The growing power of ML solutions represents the growing cognitive/interpretative capability of the processing tissue of the emerging global cyber-physical structure;

e) the development of means for the expression and storage of interpretative information. A myriad of solutions exist for the electronical-digital expression and storage of information. The coherent and interconnected expression, storage and sharing of sensory data and its associated interpretative information, in logically correct and versatile fashion is still taking its first steps, though. Most related work has focused on the expression of ground-truth information or media characterizing metadata. This area consists of the means for the expression and storage of the sensory-interpretative (i.e. cognitive) experiences of the emerging global cyber-physical tissue;

f) the combination of actuating with automated decision taking capabilities in all sorts of machines. In this regard the development of the Internet-of-Things comes to endow the coalescing global cyber-physical structure with ubiquitous sensing and actuator capability.

The combined effect of these developments is the coalescence, and operational integration, of a global cyber-physical structure comprising: a peripheral sensory echelon; a central processing tissue, capable of interpreting the acquired sensory information, reasoning, and storing or sharing its experiences; a peripheral actuatory echelon capable of acting upon the world; and a pervasive tissue that communicationaly interconnects the entire structure. This structure (illustrated in Figure 1) is thus the synthetic counterpart to a natural cognition [8] [9], i.e. a synthetic cognition (SC). For this, it is logically adequate, and practically useful, to analyse the emerging structure and its internal geometry, and plan for its development, employing natural cognition as an inspiration and guiding light.

As depicted in Figure 1, the emerging synthetic cognitive structure may be divided into two areas: the sensory and actuatory periphery and the interpreting and reasoning core. The pivotal area, which presents the greater challenges and rewards, is the interpretative core. It logically comprises all activities related to: spatiotemporal, and semantic, interpretation of acquired sensory information; reasoning; decision making and action control; and the registration of the developed cognitive experiences as coherent memories.

In natural cognition the mentioned tasks are typically handled by the central nervous system (e.g. brain). Within the emerging SC tissue they are supported by information processing hardware and software, organized in the form of isolated and standalone, or distributed and on-line, solutions. The latter ones are those most optimally inscribed within our unifying narrative of an emerging SC, as they are, by definition, interconnected into a global tissue. Within this core tissue, the most relevant technology for the effective development of interpretative and reasoning activities, are ML based tools. This is the key technology for the construction of a true global SC, and it is its development that will enable the maturation of the core area in scope and the effective coalescence of the entire structure.

C. *ML Datasets and SCE Records*

The processing activities developed by SC tools (or ML tools) are equivalent to synthetic cognitive experiences. In natural cognitions (like the one supported by the human brain), records are kept of the endured cognitive experiences, i.e. memories are stored. Similarly, the records kept by synthetic cognitive provisions of their interpretative results (or of other activities), may also be seen as synthetic cognitive experiences records (SCERs).

ML datasets and, specifically, each of their items, do not actually result from sensory information acquisition and interpretation by ML tools. They are, however, the equivalent to “crystalized” lessons that are provided to a SC tool in order to teach it how to interpret some sensory input within some specific context. Unlike human cognition (or any natural one), synthetic cognition may be directly “injected” with precise and unambiguous explanations of reality, which constitute some form of pre-packaged cognitive experience. These objects (ML dataset items) comprise sensory data (audio or visual), metadata describing features calculated from the sensory data, and metadata containing the targeted interpretations of the scenes observed in the sensory content, that the ML tool is to learn. ML dataset items may thus also be seen as the records of synthetic cognitive experiences, but which were not necessarily produced synthetically.

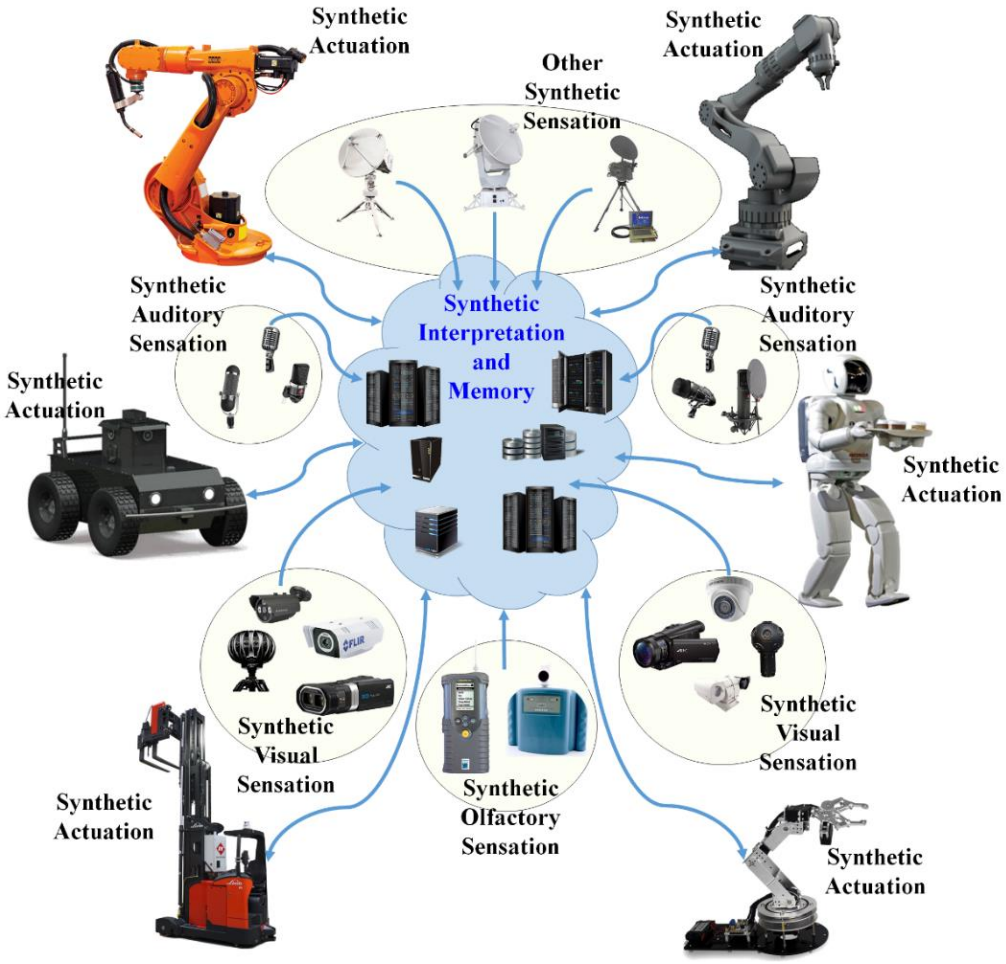

Figure 1. Comparison of Cognitive/Actuatory Tissues

Continuing with our guiding analogy, the repositories which hold ML datasets (sensory content and its associated interpretative metadata), may be collectively equated to a memory cortex of the Global Synthetic Cognition (GSC) structure.

D. The Way Forward

Just like natural cognition needs to adequately represent, structure and store its cognitive experiences (i.e. to form memories), so as to cumulatively build their world image, so does SC need to effectively represent and store the records of its experiences. All of the different data that is acquired, its interpretations and the results of all reasoning activities, should be consistently stored, so that they may be re-use and re-exploited.

ML datasets should be viewed as SCE records, as that is the role they play, in the context of the emerging SC. Employing that guiding vision will enable their conception to profit from the variedness of natural cognitive architectures and characteristics, and from natural cognition's inevitable practicality. This will ensure the universality, versatility and logical correctness of the structure and expressive solutions employed in ML datasets and their items.

The way forward, regarding the structuring of ML datasets, and the expression of their interpretative metadata, inevitably implies the definition of a universal protocol to govern them. This protocol should establish how SCE (particularly the interpretative component) should be expressed (i.e. how the metadata of ML dataset items should be expressed), how their different components (sensory an interpretative), should be structured and interrelated (i.e. how the sensory and metadata components of ML dataset items should be structured interlinked), and how the later should be stored (i.e. how ML datasets should overall be structured and stored).

The advancement of ML (consequently, of SC), requires also the definition (and consequent employment) of an adequate type of dataset repository (i.e. a proper synthetic memory holding cortex). Currently, the typical ML dataset repository is little more than a dump of media and metadata files. A new repository type should be defined which: ensures the safe storage of all relevant information; enables a standard communication and cooperation with similar modules, facilitating load distribution; optimizes memory access and manipulation performance; enforces the logical correctness of the synthetic cognitive memories that it holds, and provides a sophisticated and detailed access to them.

In this work we set out to undertake the, above identified, necessary steps for the progress of ML. The devised solutions are presented in section IV.

III. State-of-the-Art of ML Repositories/Datasets and GT Expression

A. Introduction

In this section we review the current state-of-the-art regarding: the structure, operational skills and offered service interface, of ML dataset repositories; the structure of datasets and the manner how their different components are stored; and the formats employed in the expression of the interpretative metadata (i.e. ground-truth metadata). We focus on datasets used by ML tools for sensory content interpretation (e.g. speech recognition, visual object recognition, video tracking of moving objects, etc.), which typically comprise base sensory information (i.e. media content) and the associated interpretation describing metadata. Finally we present our analysis of the current state-of-the-art and its shortcomings.

B. Survey

Various on-line repositories exist for synthetic cognition related metadata and sensory content (i.e. ML datasets for sensory interpretation). The type of information objects made available from these repositories can't actually be described as proper SCE memories, as they do not constitute an adequately integrated set of media and metadata comprising the full cognitive stack (typically, sensation, perception, conception). Some of the most relevant such repositories are presented next.

The MIT Computer Science and Artificial Intelligence Laboratory supports several different repositories and corresponding datasets. Some of the most relevant are:

- Labelme [10] – dataset for computer vison (CV) research, focused on object identification in static images. It is divided into, purpose specific, collections. Each comprises a set of image files (the whole dataset holds tens of thousands of images), for each of which there is an associated metadata file. The dataset's repository enables downloading entire collections or individual image and metadata files. The interpretative metadata format is XML based, and enables the definition or arbitrarily shaped polygons, over an image (depicted in Figure 2), their association to any object category, and the association between shapes. This dataset, therefore, comprises sensory (image) and interpretative (annotation) information for synthetic vision;
- SUN database project [11] – dataset for employment in scenario and object identification in static images. It comprises two collections: one containing static images and metadata for scenario identification; another, images and annotation files for object identification. For each image file there is an associated metadata file. The dataset's repository enables retrieving each collection in bulk or their items (one image and its annotation file) individually. The metadata format is the same as in Labelme. This dataset, therefore, comprises sensory and interpretative information of synthetic vision;
- The Activity Recognition Dataset [12] – is meant for employment in daily life activity detection in video. It consists of a set of video files (over 1M frames) and an archive with all the metadata files. The dataset's repository enables the individual retrieval of the media files and the in-bulk retrieval of the metadata. The metadata describes observed aspects like simple and composed objects, object tracks, hand positions, and activities and interaction events, which may have a long-scale temporal structure. It is expressed in a plaintext form. This dataset comprises sensory and interpretative information pertaining to synthetic vision [13];
- The Flickr Material Database [14] – this dataset is meant to be employed in CV research for material recognition in static images. It comprises a set of static colour images of surfaces, and related meta-information (images with the masks for the area with the target material, label with the category of the observed material). The dataset's repository enables only its retrieval in one block. This dataset, therefore, comprises visual sensory information and associated interpretative data;
- MIT Indoor Scenes [15] – dataset meant for indoor scene recognition from static images. Holds thousands of images and GT metadata. Its repository enables only its retrieval as a single block. The annotation metadata (expressed in a format similar to Labelme's) segments the images, identifies and classifies observed objects. This dataset, therefore, comprises sensory and interpretative information pertaining to synthetic vision [16].

The Computer Science Department of the University of Massachusetts maintains a repository with the Labelled Faces in the Wild dataset [17]. This dataset is meant for employment in CV research for unconstrained face recognition from static images. It comprises thousands of human face images, collected from the web, and associated metadata (super-pixel segmentation of the facial images, and the observed person's name). The dataset's repository enables only its retrieval as a single archive. This dataset, thus, comprises sensory and interpretative information pertaining to synthetic vision.

The ImageNet dataset [18] is meant for employment ML research for scene and object recognition in static images. It comprises a very large set of images and associated meta-information. It's repository enables the individual retrieval of images and associated metadata file. The metadata describes different aspects such as: the identification the visible objects and the location of the objects in the image (bounding boxes, expressed in the PASCAL Visual Object Classes format in XML); attributes of the identified objects (colour, pattern, shape and texture). ImageNet thus comprises sensory and interpretative information pertaining to synthetic vision [19].

The Centre for Research in Computer Vision [20] at the University of Central Florida maintains a repository with a large collection of datasets covering several domains.

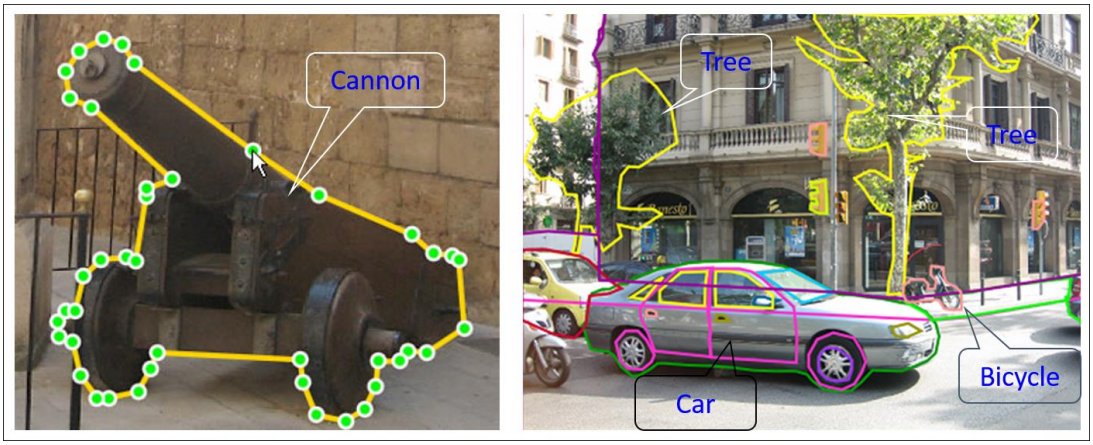

Figure 2. Example of Annotated Images at LabelMe Repository

The most relevant ones are: for object detection and tracking; for people detection in dense crowds (images and 36374 annotations, circumscribing individual persons); for multiview tracking of people in crowded scenes (comprises one set of several videos taken from different points of view from the same event). The repository enables only an in-bulk retrieval of each dataset. This is thus a dataset comprising sensory and interpretative (people and object tracking) information for synthetic vision.

Project CAVIAR [21] maintains a repository with a dataset meant for CV research on people detection and tracking and event detection (multi-agent activity recognition) in video. The dataset has two sections: video clips, acquired with a wide angle camera lens, at INRIA Labs; video clips, acquired with the same camera, along and across the hallway in a shopping centre. Each section comprises an XML file with the respective annotations for its video clips (bounding boxes, and an activity label) for all identified individuals (in each frame). It comprises also, for each frame, a situation and a scenario label. The annotations are expressed in the CVML [21] format which was purposely defined by CAVIAR [22]. It enables the expression of CV extracted information (interpretative), pertaining to people, objects or events, in all frames of a specific video sequence (depicted in Figure 3) and its binding to the sensory data. It, however, inter-mixes the description of logically different levels of cognition, such as: the definition of image (frame) segments, with the identification of the observed entity, or its contextual role. Its capabilities for describing relationships between entities and events are also limited. CVML thus lacks in clarity, flexibility and logical correctness. Each sensory file and its associated interpretative file may be independently retrieved from the repository. This repository thus holds sensory and interpretative information for synthetic vision.

Microsoft maintains a repository with the COCO dataset [23]. It is meant for CV research on object recognition and location, and scene understanding from static images. It comprises over 300k images of complex everyday scenes, and the associated interpretative information (2.5 million labelled instances). The dataset is divided into sections, for each of which the sensory information and associated annotation data is only retrievable in bulk. The annotation information includes segmentation masks and textual data (expressed in JSON): object instance descriptions (object category, location in the image (bounding box), etc.); object key point description; and entire image scenario description. The COCO dataset, therefore, comprises sensory and interpretative information for synthetic vision.

Figure 3. Rendering of CVML Annotated Frames [21]

The CV Department at Caltech maintains a repository with a number of datasets [24], for CV research on various subjects. It specifically comprises:

- CUB-200 – for research on automated bird identification from still images. Contains over 11k images of birds, with respective annotations. The repository makes it retrievable only in two blocks (the media and the annotations blocks). The annotation comprises, for each image, one object circumscribing bounding box, a rough bird segmentation, and set of attribute labels. It is expressed through a set of interrelated plain text files;
- Caltech Pedestrian Database – for CV research on pedestrian detection in video. It comprises approximately 10 hours of video taken from a vehicle driving through regular urban traffic, and associated interpretative metadata. The video information is divided into six, individually retrievable, training sets, each with 6-13 one-minute long video files (seq files – concatenated image frames with a fixed size header). The annotation comprises bounding boxes for pedestrian detections, temporal correspondence between them, and occlusion labels, and is expressed in the vbb format. It is available in a single block (seq and vbb are Matlab related formats);
- Caltech-101 – for CV research object identification in static images. It comprises thousands of images and, for each, an associated metadata file (.mat file with outline and category of the object). Its repository makes both the images and annotations available for retrieval in-bulk;
- Caltech 10,000 Web Faces – for CV research on human facial recognition from static images. It comprises over 10K images and a single file carrying all annotation data. The repository enables only an in-bulk retrieval of the sensory and annotating information. The latter comprises for each image, the coordinates of the eyes, the nose and of the centre of the mouth.

These datasets, therefore, comprise sensory and interpretative information for synthetic vision.

The Cityscapes Dataset [25] is meant to be employed in the development of CV algorithms for semantic urban scene understanding from static images. It comprises a large set of images (frames of videos with street scenes from 50 cities) and associated annotations, and is divided into training and testing subsections. The annotation information comprises the instance-wise delimitation of urban scene components, semantic categorizing data, and dense pixel annotations. This dataset therefore comprises approximate synthetic vision memories, with a sensory and an interpretative (detection and understanding of objects and scenarios) component.

The THUMOS dataset [26], is meant for CV research on activity recognition (101 types) in video content. It holds thousands of videos (from YouTube) and the corresponding annotation information. The dataset is divided into four sections: training; background; validation; and the test section. All sections comprise a file with the annotation metadata, describing the actions observed in each video and their temporal extent (in plain text format or as .mat file – Matlab). The dataset's repository makes the video content (per section) retrievable in bulk or by individual file. Metadata files are retrievable individually. This dataset thus comprises approximate synthetic vision memories.

Google maintains the YouTube-8M Dataset [27], for CV research on action understanding in video content. It comprises 8 million videos (from YouTube) and associated interpretative metadata. It is divided into two sections: one for frame level features and the other for video level features. Both are further divided into training, validation and test partitions. Its repository enables each partition to be retrieved as independent shards of the overall dataset. The annotation information describes frame or video level characteristics (main topics of each video, imagetic and audio features), in accordance with the Inception-V3 image annotation model. This dataset thus comprises data that may be regarded as synthetic vision memories.

A joint project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago maintains a repository which holds the KITTI Benchmark Suite [28]. This dataset is meant for CV R&D on stereo optical flow assessment, visual odometry, 3D object detection and 3D tracking. It comprises sensory information (real-world city driving scenes), acquired with the autonomous automobile platform Annieway [29], and accurate ground-truth information (annotations), acquired through a Velodyne laser scanner and a GPS location system. The repository enables the sensory and annotation data to be retrieved in fragments of the full dataset. This dataset, therefore, comprises sensory (video) and interpretative (3D idealization and object identification) information pertaining to synthetic vision.

The Computer Science department of the University of Washington maintains a repository with the RGB-D Object Recognition Dataset [30]. This dataset is meant for CV research on object recognition from static images (2D+depth). It comprises a large collection of 2D+depth images (pertaining to 300 household objects), and GT information for all 300 objects. For each instance of each object type there is: a set of images (from various angles); a set of metadata files describing the object's location, the object's segmentation mask and the object's pose. The repository enables the dataset's retrieval in bulk, or by individual object instance. This dataset, therefore, comprises sensory and interpretative (tree-dimensional idealization) information pertaining to synthetic vision.

Carnegie Mellon University maintains a repository with the CMU Multi-PIE face database [31], which is meant for employment in CV research on the interpretation of facial expressions from static images. It comprises over 750k images of 337 people with different facial expressions and associated metadata. For each person (for four different facial expressions) images where acquired from 15 different viewpoints and under 19 illumination conditions. The annotation describes the portrayed emotion in each set of 15x19 images. The overall dataset can only be retrieved, physically, through the acquisition of a hard drive with a copy of it. This dataset, therefore, comprises sensory (images) and

interpretative (emotions) information pertaining to synthetic vision, for emotion recognition.

The University of California, Irvine (UCI) Machine Learning Repository [32] holds a collection of databases available for use by the ML community, for a variety of purposes. It stores a vast repertoire of images, sounds and other types of data, and a broad set of associated GT metadata, typically expressed in purpose specific plaintext format. The retrieval of sensory or interpretative information is, typically, only possible in bulk. This repository, therefore, holds what may be regarded as sensorial and interpretative information pertaining to senses such as vision and audition, besides other types of metadata.

The University of Pennsylvania maintains a repository with the TIMIT dataset [33], for ML research on automatic speech recognition. It comprises recordings (in individual files), of spoken American English, by 630 individuals of different genders and dialects. It comprises also annotations consisting of time-aligned phonemical and lexical transcriptions (stored in separate files) of the recorded speech. This dataset therefore holds synthetic audition memories, comprising a sensory and an interpretative (semantic) component.

The NY University maintains a repository holding the UrbanSound Dataset [34], meant for ML research on sound recognition. It comprises over 1k sound recordings (obtained from [35]) and corresponding annotations. Each recording may contain multiple sound events of a variety of classes (drawn from the urban sound taxonomy [35]). The annotation information includes: the start and end times of sound events, their salience, and their class, expressed as a CSV file; and the metadata provided by the Freesound API (description, tags, id, format, etc.), expressed as a JSON file. This dataset therefore comprises data describable as synthetic audition memories, comprising a sensory and a (logically) low level interpretative component.

VoxForge [36] is a free online repository for speech corpus and acoustic models, holding a dataset for ML research on speech recognition. Said dataset is user provided and comprises audio content (people reading some text excerpts out loud) and its textual description (the text being read). It is divided into individually retrievable blocks, each containing a number of audio recording files, and a set of metadada files with the time circumscribed definition of the spoken sentences, and other information (in plain text). This repository thus holds data that may be regarded as synthetic audition memories, comprising a sensory (speech records) and an interpretative (human produced text) component.

The WEIZMANN dataset [37] (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 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 type of recorded action). Each set carries 9 videos (each with a different person performing the same activity). The annotating, for this sub-dataset, is: the definition of the foreground silhouettes of each moving person (a single file in MATLAB format); the background sequences used for background subtraction (retrievable in bulk); and the identification of the performed activity. The dataset in scope, therefore, comprises sensory and interpretative information pertaining to synthetic vision.

The University of Edinburgh's School of Informatics maintains a repository with the BEHAVE dataset [38], which is meant for CV research on behaviour identification and analysis of interacting groups of people in video. It comprises various video sequences of groups of people having different interactions (ten specific types), and GT information with the tracking of the observed individuals (bounding boxes enclosing each interactive person and the corresponding labels, in the ViPER XML format). Both the video and metadata files are individually retrievable. This dataset, thus, comprises sensory and interpretative information pertaining to synthetic vision.

The KTH Royal Institute of Technology maintains a repository with the KTH dataset [39], for CV research on human activity recognition from video. Its video sequences where acquired with a static camera, over a homogeneous background, and capture six types of human actions, performed several times by 25 people in four different scenarios (overall set of 600 videos). The annotations are provided as an ASCII file. The dataset is divided into 6 (individually retrievable) sections (one for each human action type), each comprising the respective videos and metadata file. This dataset, therefore, comprises sensory and interpretative information pertaining to synthetic vision.

The INRIA institute maintains a repository with the ETSIO dataset [40], for ML research on human activity recognition on video. It is divided into five distinct sections, one for each of the contexts of video acquisition. The annotation (in Viper-GT xml format [41]) for each of the videos is stored in its separate file. Both the sensory and interpretative contents can only be retrieved in their entirety. This is thus a dataset comprising sensory and interpretative information pertaining to synthetic vision.

The Imagelab Laboratory, of the University of Modena, set up a repository with the ViSOR dataset [42]. It is to be employed in ML research on: human behaviour analysis; human tracking; event analysis; people counting; pedestrian crossing; human identification; smoke detection; and human action recognition. It comprises hundreds of (surveillance related) videos and their individually associated metadata files (bounding boxes and semantic characterization), expressed in the ViPER GT format [41]. All dataset files are individually retrievable. This dataset, thus comprises sensory and interpretative information pertaining to synthetic vision.

The Faculty of Science, Engineering and Computing of Kingston University maintains a repository with the

MuHAVi dataset [43], for 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 7 parts (corresponding to 7 actors) each of which contains 8 sub-parts (corresponding to 8 cameras), which carry the videos. The metadata of this dataset block is a set of manually produced annotations describing the silhouettes of the actors in frames. Each available combination of person/camera/action annotations may be retrieved individually. The latest part of this dataset consists of 8 sections, each comprising a continuous, individually retrievable, video file, from one of the cameras, which captures all of the actions from that point of view. The metadata (contained in a single spreadsheet file) consists of: the description of start and end times of each sub-action in each video by each actor; and the actor delineating silhouettes for each of the videos. This dataset, therefore, comprises sensory and interpretative information pertaining to synthetic vision.

The Visual Geometry Group of the Oxford University maintains a repository with the TV Human Interaction Dataset [44], for CV research on interaction recognition between two people in video. This dataset comprises 300 video clips, depicting 4 interaction types, as well as clips with no footage of any of such interaction types. It contains also (for every video) a metadata file describing (for every frame, in plain text format): the upper body of observable people (with a bounding box); their head orientation; and their interaction label. The sensory and interpretative components of this dataset can only be retrieved in bulk. This dataset, therefore, comprises sensory and interpretative information pertaining to synthetic vision.

The Serre lab at Brown University maintains the HMDB51 dataset [45], for CV research on human action recognition in video content. It comprises over 6k video clips divided into 51 action categories. Annotations (for each clip) comprise the video's action category, the visible body parts/occlusions, the camera motion state and view point relative to the actors, and the number of people involved in the action. Both the sensory and the interpretative information can only be retrieved in bulk. This is thus a dataset comprising sensory and interpretative information pertaining to synthetic vision.

The MILtrack dataset [46] [47] is meant for CV research on visual tracking under changes of object appearance, illumination conditions, object occlusions and cluttered background conditions, over video. It comprises twelve video sequences as well as GT information consisting of the targeted objects' location (bounding boxes) in a purpose defined text format. The dataset's holding repository enables each video sequence and associated metadata to be retrieved as one block. This dataset, therefore, comprises sensory and interpretative information pertaining to synthetic vision.

Microsoft maintains a repository with the Microsoft Research Dense Visual Annotation Corpus [48], which is meant for CV research on object recognition in static images. It comprises 500 images, annotated with bounding boxes and facets for each object in each image. The sensory and interpretative components of the dataset are retrievable separately but in bulk. This dataset is thus composed of sensory and interpretative information pertaining to synthetic vision.

The SBU Captioned Photo dataset [49] is meant for CV research on object recognition in static images. It comprises 1M images and associated annotations. The dataset's repository enables each image to be individually retrieved, but the interpretative information can only be retrieved in bulk (a file with one million short descriptions). This dataset is thus composed of sensory and interpretative information pertaining to synthetic vision.

The PASCAL VOC project [50] ran a research challenge, from 2005 to 2013, in the area of object recognition. For the purpose of that challenge it maintained a dataset, to be employed in CV research on object recognition, classification, detection and segmentation, in static images. The (training part of the) dataset comprises as set of images and the associated annotations of each, describing a bounding box and an object class label for each observable object, belonging to one of twenty classes. Another set of the images (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). The entire dataset (for each of the editions of the challenge) is retrievable only in bulk. This dataset may thus be considered to comprise sensory and interpretative information pertaining to synthetic vision.

The Computer Science department of the Boston University maintains a repository holding a dataset [51], for CV research. It comprises image and video content as well as annotation information. Specifically it includes: over 70 (individually retrievable) video sequences and associated interpretative information (describing head position and orientation) to be employed in research on 3D head tracking; a set of (individually retrievable) video sequences, and associated labelling information, to be employed in CV research on skin colour segmentation; a set of (over 100k) computer graphics rendered hand images, and the associated ground truth for each image describing the articulated pose, to be employed in research on hand pose estimation. This section of the dataset is only retrievable as a single object. This dataset thus comprises sensory and interpretative information for synthetic vision.

The Advanced Digital Sciences Centre of the University of Illinois maintains a repository with the FaceScrub dataset [52], which is meant for research on face recognition. It comprises a set of images of human faces (over 106k images of 530 male and female celebrities retrieved on-line), and the necessary metadata to identify those persons. The dataset can only be retrieved in bulk. This dataset thus holds data that may be regarded as synthetic vision memories, comprising a sensory and an interpretative component.

The Arabic Speech Corpus dataset [53] is to be employed in computer hearing (CH) research on: speech synthesis; alignment of speech corpora with their phonetic transcript;

and speech recognition. It comprises over 3.7 hours of Modern Standard Arabic speech and the corresponding phonetic and orthographic transcriptions, aligned with the recorded speech on the phoneme level. It specifically comprises: over 1.8k files containing spoken utterances; over 1.8k files with text utterances; over 1.8k files containing the phoneme labels with time stamps of the boundaries where these occur in the spoken utterance files; a file describing the phoneme sequence for every spoken utterance file; a file containing the orthographic transcript for every spoken utterance file. The dataset's repository enables only its in-bulk retrieval. This dataset thus comprises data that may be regarded as synthetic audition memories, comprising a sensory and an interpretative (semantic) component.

The Million Song Dataset (MSD) [54], is meant for ML research on automated music information retrieval. It comprises a, freely-available, collection of audio features and associated metadata for a million contemporary music tracks. It does not include any audio itself, only its derived features. The dataset's repository enables only its retrieval as a whole. The MSD thus comprises data that may be regarded as synthetic audition memories. It does not however store the audio sensory component; it stores only the low-level interpretative one.

The Cardiff School of Computer Science maintains a repository with the 4D Cardiff Conversation Database [55], for CV and affective computing research in areas like deception detection or behaviour analysis. It contains 3D video footage of 17 natural conversations between pairs of people. It comprises also (for each conversation), annotations pertaining to: speaker and listener activity; conversational facial expressions; head motion; and verbal/non-verbal utterances. The annotation format allows for multiple annotation tracks and hierarchical tracks, and time-accurate text annotation of speech sections. This is thus a dataset comprising sensory an interpretative metadata pertaining to synthetic vision and hearing.

Up to this point we have focused on complete solutions for the expression and storage of SCE, and in the context of ML Repositories/Datasets. Some other tools exist, however, which even if their employment context is typically not related to ML datasets, are still relevant for the expression of some SCE components (typically the interpretative one). The development of these tools was not guided by a vision of them as SCE records, and thus, they are not effectively devoted to express such experiences.

In the next paragraphs we approach some of the most relevant such tools, typically developed for multimedia enrichment with metadata, with a particular focus on their abilities for the registration of the interpretative components (perception and conception), of synthetic cognition.

The Resource Description Framework (RDF) [56] is a set of specifications by W3C, and a cornerstone of its Semantic Web initiative. It is meant to allow the conceptual description (for automated tools) of real or web resources or knowledge management applications. The Web Ontology Language (OWL) [57], is a knowledge representation markup language for the definition of ontologies. It has formal semantics and it is built as a vocabulary extension of RDF. The RDF/OWL complex may thus be used for the expression of some parts of the interpretative component of SCE – the conceptual/semantic component and some perceptive aspects. It is, however, ill-suited for the description of the later as well as for the interconnection of the different levels of interpretative information and for their detailed binding to the sensory content from which it derives.

MPEG-7 [58] is an MPEG Group standard. Its purpose is to enable the description of multimedia content to provide support for the interpretation of its meaning. MPEG-7's main components are the [59][60]: Multimedia Description Schemes (MDS); Visual Description Tools for describing low-level visual features; Audio Description Tools for describing audio content; and Data Definition Language for extending MPEG-7's overall ontology. MPEG-7 therefore enables the expression of interpretative information (perceptive and conceptive), and it's binding to the sensory component (media files). However it does not provide an adequate separation of the perceptive and conceptive components and does not provide for their versatile and logically correct interconnection. It thus has no notion (and provides no way to make it explicit) of the different abstraction levels that may be identified within both the perceptive and, especially, conceptive information. Furthermore, Mpeg-7's employment has proven cumbersome, and its expansion complex [61].

Video Event Representation Language (VERL) and Video Event Markup Language (VEML) were developed within the Advanced Research and Development Activity 2004 Event Taxonomy Challenge Project. VERL is a formal language (built on OWL) for the definition of complex event ontologies involving composed events and inter-event temporal relationships. VEML defines the specific markup for the description of event instances (for events defined in the ontology), and for their biding to the points in the video that they pertain to [62]. Together, VEML and VERL are a relevant tool for describing the semantic interpretation (by synthetic cognition), of video content and for its binding to its originating media. This toolset is, however, not adequately prepared for the definition of other levels of the interpretative component of cognition, and for their interrelating.

The ViPER project [41][63][64] has developed ViPER GT, an XML-based language for the expression of visual data ground-truth information. It enables the annotation of various types of information, over video media. However it aggregates detection information per observed entity or event, and not on a temporally sequential basis bound to sequential frames. It also mixes detection information with the identification and further characterization of the observed realities. ViPER's language also enables only a rigid declaration of relationships between detections from different frame-spans, pertaining to the same observed reality. The ViPER language thus enables the construction and interconnection of only some components of SCE records and in a very sub-optimal manner.

C. Analysis

The approached datasets are all meant to be employed in the development of provisions for some type of automated interpretation of some base sensory (audio or video) content.

As we argued earlier, the full set of existing, sensory based, datasets (and associated repositories) may be seen as the memorial tissue of the emerging global synthetic cognitive structure. However, it is still far from being an integrated tissue. It consists of a myriad of different components (datasets and their repository provisions) which are emerging, typically, in isolation from one another, through the work of separated initiatives and focusing on very specific individual objectives. There is no global cooperation protocol, or coordinating structure, enabling the operational integration of the repositories/datasets in scope, and they are not regarded (by their holders) as part of a global body.

The structuring of datasets is simplistic, and solely focused on the immediate objectives established for them. The structure actually employed is only meant to provide some basic order to the stored information. Typical examples of this are the separation between sensory and interpretative data (in virtually every dataset), and the separation between training, validation and testing sections of the dataset (in Cityscapes and THUMOS datasets). Said structuring has no vision of this information as SCE records.

Most existing datasets typically comprise: one or two types of sensory information (predominantly video and audio); and interpretative metadata describing very specific aspects of reality (observed in the sensory content) of relevance to the dataset's domain of application. The interpretative component is typically structured employing the simplest possible means, for expressing very specific ground-truth information for training, validating and testing ML applications (e.g. bounding boxes and identification labels employed in ImageNet for object recognition, or temporal delimitation, salience and class information employed by UrbanSound for sound recognition). It is thus simplistic, not prepared to be expanded to truly large volumes, expressively limited and very narrow focused (e.g. PASCAL VOC, CVML or ViPER, are some of the most advanced formats employed by the datasets, and still are very simplified). It also incorrectly mixes the expression of logically different interpretative aspects (e.g. PASCAL VOC, CVML, as well as MPEG7), and is invariably incomplete. Furthermore, the interpretative components of datasets are also poorly interconnected with the sensory content from which they derive (e.g. RDF/OWL), and also lack any provisions to enable weaving a global interpretative (perceptive/conceptive) tissue across different modules of the memorial tissue (i.e. across different repositories/datasets) in a robust and universally understandable fashion.

This way, the solutions/formats employed, by existing datasets, for the expression of what we argue/envision to be SCEs, are in no way inspired by such a vision, and the resulting information objects are no representations of SCEs.

The dataset repositories make their contents available in very simplistic and rigid ways. Their capabilities are limited and unsophisticated (e.g. CMU Multi-PIE Face Database, Labelled Faces in the Wild, ImageNet).

For all the above the memorial GSC tissue is clearly far from being an integrated structure. Its immediate advancement requires, minimally, that: a standard internal data structure, and a sophisticated interface are defined for ML dataset repositories to enable an agile and universal interaction with these "memorial lobes"; and also that a universal protocol is defined for the expression of the metadata component of datasets and for the overall the interrelation/structuring of all their informational components, i.e. for the full expression of SCE records. In this later regard what most current ML research initiatives predominantly need is and adequate separation between the different logical levels existing within the metadata component of their datasets and an effective and extensive interconnection of the logically interconnected sections of such levels, and this is what our protocol delivers.

IV. SCE Repository and Expression Format

A. Repository

A repository for ML datasets (comprising sensory information and associated ground-truth data) should be viewed and designed as a memory storing "lobe" (or a memorial module) of the emerging GSC. Its contents are the synthetic equivalent of memories, i.e. SCE records. They thus hold sensory and interpretative information.

A memorial module must therefore: ensure the safe storage of all relevant components of SCE records (sensory and interpretative records); maintain a standard communication and cooperation with similar modules (and other modules of the GSC), thus facilitating synergies and load distribution and the overall operational integration of the memorial tissue (and, consequently, of the GSC core tissue); optimize memory access and manipulation performance; enforce the structural and logical correctness of the synthetic cognitive memories that it holds (in accordance with the SynCog protocol); and provide a detailed and sophisticated access to the stored memories.

A memorial module must comprise two main parts (as presented in Figure 4): the sensory information holding part; and the interpretative information holding part. The sensory component of SCE records is the largest and least structured (just sequences of data blocks, encoding image or audio). Also, sensory records (i.e. media files), shall have a unique universal identifier, and will contain image (video or static) or audio information. These records should be mono sensory (i.e. storing only one type of media), as this enables a decoupling of that information making it easier to manipulate it and to define sensorially complex SCEs.

The sensory information holding part of the memorial module should: be a type of media file repository with big storage capacity and rapid access bandwidth; and be equipped with the capability to extract segments of the sensory records it holds. This component of the memorial module may be collocated with the part holding the interpretative data or placed remotely. It should thus offer an interface for local and remote access enabling earlier mentioned operations.

The interpretative component, of SCE records, comprises less information, but is endowed with greater meaning, i.e. it comprises the logical idealizations that arise from the sensory component and is thus endowed of logic interconnectedness. It will not require very high storage capacity. Access to it, however, will be more frequent (revisited for searches, delivery to the periphery, and for further interpretation), and non-continuous (this information is profusely interconnected and its use implies accessing distant points in it).

The interpretative information holding part should: prize access speed and agility over of storage capacity; store the interpretative metadata in file form and (for the relevant parts) in database format; enforce the logical correctness of the interpretative component of SCE records, as well as of its connection to the sensory component; enable the detailed searching, reading and writing of interpretative information in varying degrees of granularity. This part of the module should also have the ability to, locally or remotely, resort to the services of the sensory information holding part. To provide its services to the client side the interpretative component should thus provide a servicing interface.

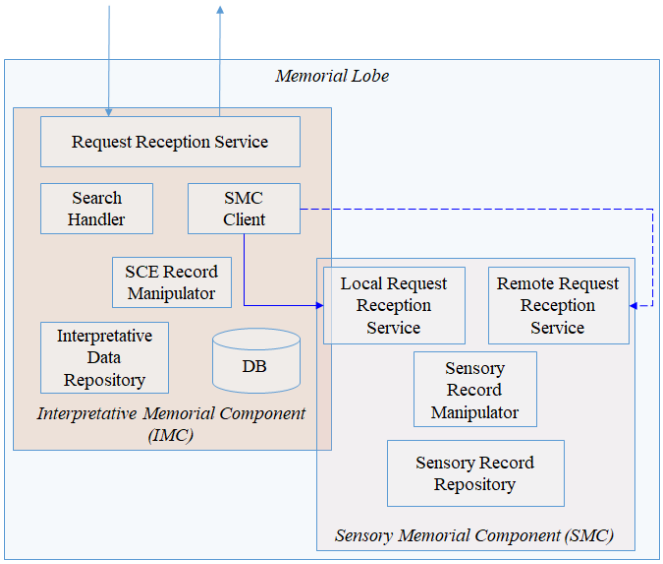

Figure 4. Structure of a Memorial Lobe

The two components, of the memorial module, combined will enable the client entities (e.g. ML research and development initiatives), to seamlessly and dynamically, collaborate in the production, share, and use broad and versatile sensory/interpretative dataset.

B. SynCog

1) Introduction

Here we present a novel metadata model to overcome the present limitations pertaining to the expression, registration and transmission of synthetic sensory-interpretative experiences (i.e. the expression of ML datasets). The SynCog, builds on previous work [65], extending it to allow the detailed description of experiences involving various simultaneous sensory captures of different types of senses (video and audio), and not only visual sensations as the previous work did. Moreover, it introduces new means of expressing higher levels of complexity in the interpretation of those interrelated sensations. Moreover, it introduces new means of expressing higher levels of complexity in the interpretation of those interrelated sensations.Moreover, it introduces new means of expressing higher levels of complexity in the interpretation of those interrelated sensations.

The apprehension of reality, by natural cognitions, begins from the sensing of stimuli. That information is processed, enabling the detection of patterns, shapes and spatiotemporal relationships. This permits the development of an idealization of the spatiotemporal characteristics of the outer observed reality, i.e. the mental reconstitution of the outer environment. On top of this interpretative layer natural cognition then builds its semantic interpretation and valuation of reality, through the merger of sections of the idealized construct into concepts of entities and events.

The synthetic counterpart to natural cognition will include similar components. SynCog must thus enable the: expression of captured (imagetic and audio) sensory information; expression of the idealized spatiotemporal reality; logical linking of the different components of the idealized construct to the different segments of the sensory information from which it derives; expression of the conceptual field that arises from the idealized construct; and the logical linking of the conceptual components to the sections of the idealized construct from which they emerge.

SynCog was therefore conceived as a multi-layered structure, as exemplified in Figure 5 that accommodates cognition's different logical levels. Focusing on Figure 5b (which may be seen as a simplified version of SynCog object), the bottom layer comprises the acquired sensory information. The two top levels comprise the interpretative

information. The lower, of these two, contains the idealizing data. The top one carries the conceptual level information.

ML datasets frequently comprise also information which results from the deterministic calculation of some features characterizing the sensory content. Frequently it is over these features that the actual synthetic interpretative process is executed. However, there is a tremendous variety of features that may be calculated and employed. We thus believe that it is more relevant to store the base sensory content, and leave the representation and storage of features for a later version of SynCog. ML research initiatives using SynCog as their dataset expression and structuring tool will always have the sensory component at their disposal and may calculate the associated features at will.

A future version of SynCog should enable the expression of the features in scope as yet another interpretative layer located bellow the idealization layer or as extra information to be attributed to the spatiotemporal idealizations defined within the idealization layer.

Regarding their storage SynCog objects may be represented and stored in file form (a single file archive, or the separated set of sensory and interpretative files), or stored into an adequate database (for the most part of the interpretative metadata).

2) SynCog Object Sections

The SynCog format enables the expression of SCE records whose sensory base is either image or audio. Sensory content may have a variable length temporal dimension (e.g. a static image has a zero length temporal dimension while a video as a non-zero length temporal dimension). The temporal extension of a SynCog object may thus extend from zero to infinity (theoretically).

To deal with this, SynCog objects are divided into sections of a specific temporal length (T seconds). This way the sensory and interpretative components will be cut into sections of T seconds of duration. For each T seconds long interval a SynCog object section exists which comprises a set of sensory files (one for each of the simultaneous sensory captures that make up the sensory level of the SynCog object), and two interpretative files, one carrying the perceptive data and the other the conceptive data. If the sensory content in scope has no temporal extension, then the SynCog object will consist of a single section.

All sections of a SynCog object are universally identifiable through the id of their parent SynCog object and their chronologic sequence number.Figure 5 presents an exemplifying structure of a SynCog object. Part a) of the figure presents the full SynCog object as a chain of sequential sections. Part b) presents a closer view of one of the sections of the object (it may also be seen, alternatively, as a simplified view of a complete SynCog object, which fuses the section-wise and global conceptive levels into one).

Some of the idealized constructs may stretch across SynCog sections (i.e. some of the sensory observed realities may be present in several consecutive sections of the sensory records, resulting in the need, at the spatiotemporal interpretation level, to reference points, in the sensory records, of several different sections of a SynCog object).

The same is true for the conceptual level (i.e. some of the perceived spatiotemporal constructs may be present across several consecutive sections of the perceptive metadata, resulting in the need, at the conceptive interpretation level, to reference points, in the spatiotemporal metadata, of several different sections of a SynCog object). As all sensory records and their segments, and all instances of perceptive and conceptive constructs are universally identified (within SynCog), the references in scope are easily established. This means, however, that the sections of SynCog objects are not entirely independent of one another. Furthermore, the concept instances that arise from the observed reality (instances of events or entities), and the relationships involving them, may emerge at many different instants (i.e. sections), of an overall SynCog object, and will need to be referenced from every such segment. This way, all the instances of all the event, relationship and entity concepts observed within a SynCog object's sensory track, are defined in a unique conceptive metadata document (represented by the Global Conception Level at the top of Figure 5). Throughout the conceptive files of the sections of a SynCog object, the relevant concept instances are referenced whenever necessary.

3) Sensory Component

The sensory component (i.e. media content) of SynCog objects (and of each of its sections), is expressed at the sensory registration level (bottom layer of each section of Figure 5b). This may comprise static or moving visual information, audio information, structured light image captures, olfactory sensory acquisitions, etc. We shall for now limit it to video and audio. This component may be of varying dimensions (e.g. multiview video) and integrate different senses (e.g. audio and video), that maintain between each other (between their capture points and instants) different spatial-temporal relationships.

The sensory part of a SynCog information object (of each section) shall therefore be composed by different media files (an N sized collection). Each must contain a mono-sensory acquisition (S_i), from a specific capture point, throughout time (e.g. a video feed acquired by a specific camera, or an audio recording acquired from a specific point). These may have different overall time durations and informational characteristics (resolution, frame rate, sample rate, etc.). The recordings of different senses, or taken from different capture points, are therefore stored separately.

In the case of natural cognitions (such as the human one), the spatiotemporal relationships between the different sensation captures are somehow hardcoded into the fabric of the brain, which allows for the correlation of the information collected through the different sensory organs and the construction of the mental image of the world. In the case of synthetic cognition, it is also necessary to know these relationships, for the interpretation of reality, and to register them, along with the rest of SCE, to enable the posterior

"remembering" of the SCE or the posterior reinterpretation of the sensory data. The sensory component of SCEs shall, thus, contain a metadata file carrying the definition, within a specific referential, of the relative positions of each sensory capture point, from which a certain context is being observed, and the time interval (within a common time referential), within which each sensory capture takes place. This way all sensory records are inscribed into a spatiotemporal grid that interrelates them.

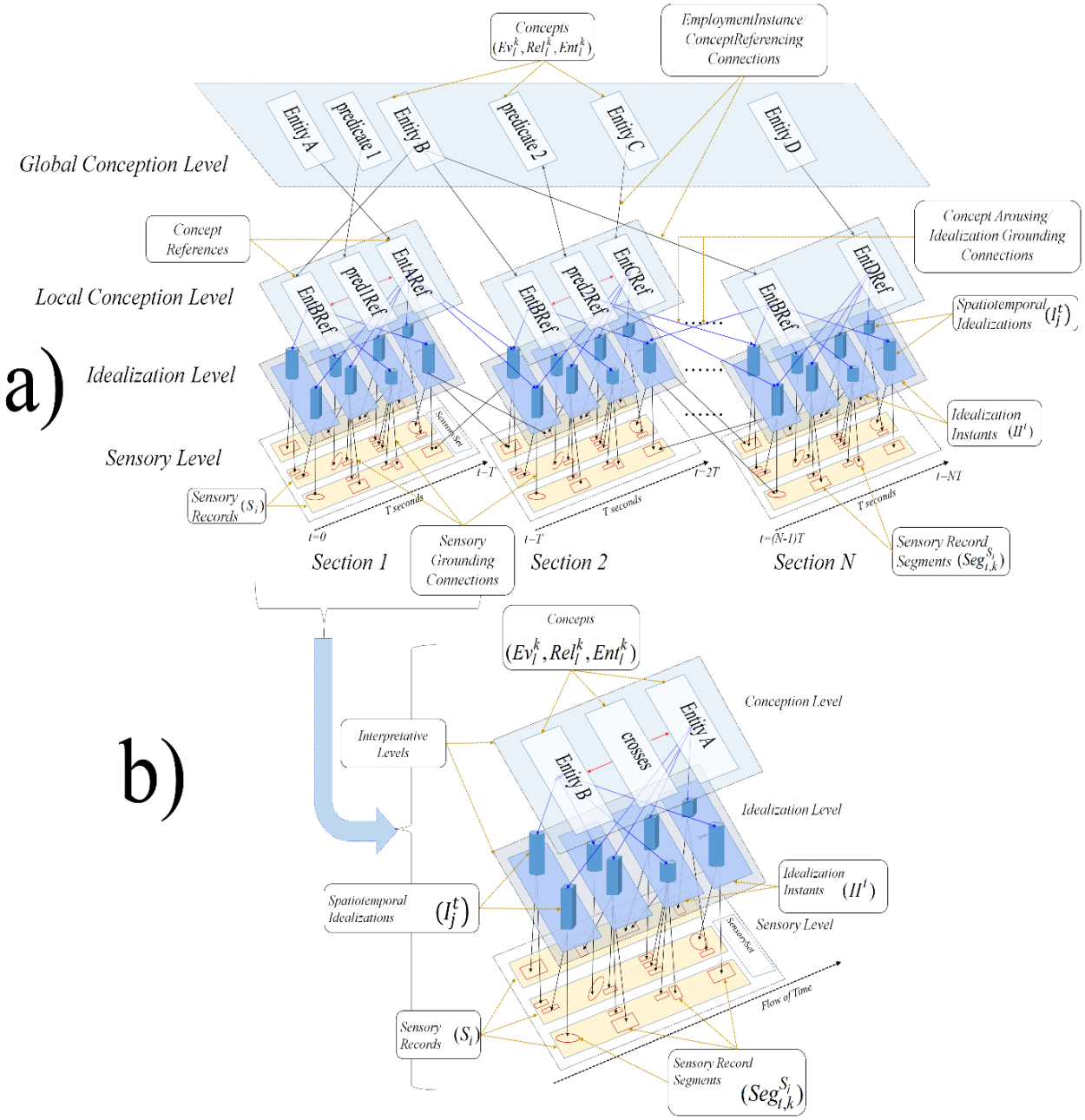

Figure 5. SynCog Structure

The file carrying this information (SensorySet.xml) employs an XML syntax, and is contained within the first segment of a SynCog object. The root element is *SensorySet*. For each sensory capture point (each individual sensory device) it carries a *Sensor* element. The internal structure of the latter is the following:

- a *Type* element – indicates the type of sensory acquisition in scope (video, audio, etc.);
- a series of *Characteristic* elements– each indicates (through the *name* attribute) the informational characteristic to which it pertains (e.g. resolution), and through its content, the value of that characteristic;
- a *Position* element – it indicates, though the *x*, *y* and *z* attributes, the location of the sensor device within a stipulated coordinates space. A single value is employed

for a sensor's position as for now we consider only fixed sensors;

- an *Interval* element – indicates the beginning and ending moments (attributes *start* and *end*) of the sensory record, within a common time referential for all the sensory records para of a specific synthetic cognitive experience record.

4) *Interpretative Component*

The interpretative component consists of two levels: the perception and idealization level (middle layer of each of the section blocks of Figure 5a), and the conception level (global top layer of, and top layer of every section block of, Figure 5a).

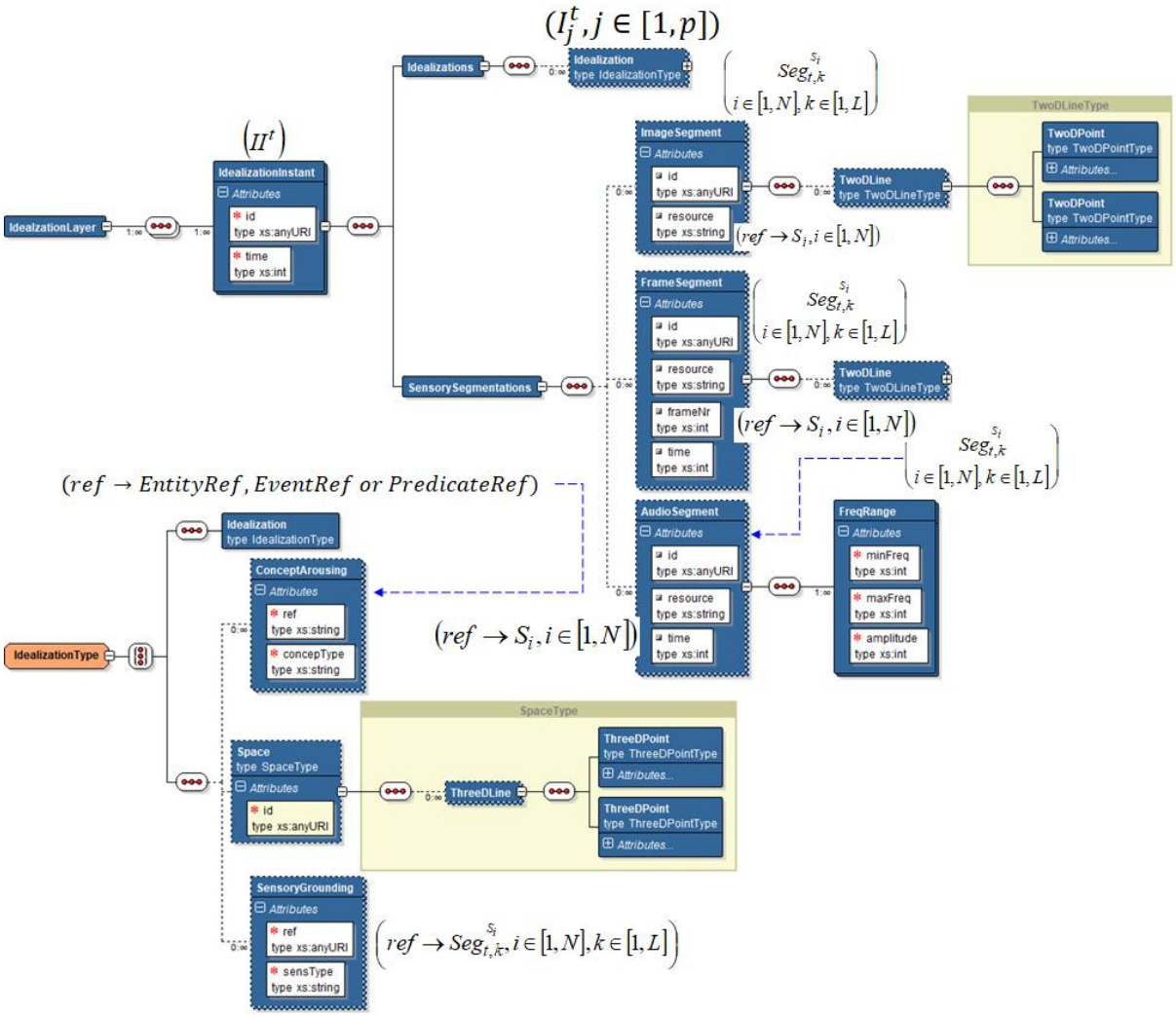

Figure 6. Idealization Metadata Schema

The earlier one comprises information describing the spatiotemporal interpretation and segmentation of the scenario observed by the sensors. The latter sub-level describes the conceptual panorama that emerges from the idealized reality, i.e. the concepts (of entities or events), spatiotemporally idealized at the previous level (of entities or events), and the relationships between them.

The idealization information includes two different parts (its schema is presented in Figure 6):

- the purely idealizing component which performs the tetra-dimensional (spatiotemporal) description of the scenario (one instant at a time) under sensory observation, as well as its segmentation into contiguous individual entity or event instances (represented by the middle layer in Figure 5b). In terms of human cognition this component is the equivalent to the mental image that we build of the spatiotemporal characteristics of the outer world;
- the idealization grounding component (represented by the arrows bellow the middle layer and by the perimeters projected over the sensory records of the lower layer in Figure 5b) which logically links the idealized spatiotemporal realities to the segments of the sensory records from which they are interpreted. It enables, for instance, to specify the connection between the idealization of a car's three-dimensional structure moving down a road for 3 seconds, to the segments of each of the frames (visual sensory record) where that vehicle is visible performing the activity in question.

Idealization information will be expressed employing an xml syntax, and it is storable into a single file (for each section of the overall SynCog object). The root element of the perceptive (spatiotemporal idealizing) metadata of each section of a SynCog object is the *IdealizationLayer* element. Each such element is identified by the combination of the identifier of its parent SynCog object and the sequence number of its comprising section of that object. Every *IdealizationLayer* element carries a series of *IdealizationInstant* elements. Each of these represents the three-dimensional spatial idealization of a scenario observed at a specific time instant (specified in the time attribute), orII^t, comprised within the time interval encompassed by the comprising section of the overall SynCog object. Each *IdealizationInstant* carries an *Idealizations* element and a *SensorySegmentations* element. The earlier comprises the definition of a one or more idealizations (by way of, a P sized series, of inner *Idealization* elements), and the latter comprises the definition of the all the segments of sensory records on which all the idealizations of the current idealization instant are based upon (by way of its inner *Segment* elements).

Each *Idealization* element performs the actual definition of a spatiotemporal idealization (I_j^t), links it to the segments of sensory records from which it derives, as well as to the concepts that arise from it. It may carry either: a series of inner *Idealization* elements, and thus be a composed idealization; or a series of *ConceptArousing* elements, a *Space* element, and a series of *SensoryGrounding* elements. Each *ConceptArousing* element links the comprising idealization to a specific concept (defined in the conceptual metadata layer, presented below), which emerges from it, at the conceptive cognitive level. It performs that association by referencing the identifier of the concept instance representing element in its *ref* attribute. The target elements will be those present in the conceptive metadata of the comprising section of the SynCog Object (i.e. *EntityRef*, *PredicateRef* or *EventRef*, elements). The *concepType* attribute of *ConceptArousing* indicates the type of concept in scope (event, entity or relationship).

The *Space* element represents the specific, contiguous and closed, space which is occupied by one (or a set of them) entity, event or relationship instance (someEv_l^k, Rel_l^k, Ent_l^k, the l[th] instance of the k[th] entity, relationship event concept), cognitively identified in the space and time interval/instant whose overall cognition is represented by *IdealizationInstant*. That space is defined by a set of *3DLine* elements. Each such element defines one of the edges of the border planes of the space in scope. It does so through two *3DPoint* elements which define two points in three-dimensional space (attributes *x*, *y* and *z*).

Each *SensoryGrounding* element performs the connection (i.e. logical grounding) of the idealization in scope to one of the segments of the overall sensory component (defined within the *SensorySegmentations* element), from whose interpretation said idealization emerges (i.e. each *SensoryGrounding* performs the connection between an I_j^t and some $Seg_{t,k}^{S_i}$, the k[th] segment at the t[th] idealization instant of the i[th] sensory record). The target sensory record may be visual (image or video) or auditory (indicated by the *sensType* attribute), and the *SensoryGrounding* element references the definition of such segments at the *ref* attribute.

Therefore, if the target record is:

- a segment of a static image – an *ImageSegment* element will be referenced. This will carry a set of *2DLine* elements which will jointly define a 2D space (or set of them) over the target image where the idealized reality in scope is visible. Each *2DLine* has two *2DPoint* child elements defining (though attributes *x* and *y*) the *2DLine's* endpoints. The *ImageSegment* element also indicates the sensory record (media file of the sensory layer of the SCE record) to which it pertains, through the resource attribute;
- a segment of a video frame – a *FrameSegment* element will be referenced. It will carry the same contents as in the previous case, with the added attributes of *frameNr* (the frame's sequence number) or *time* (temporal location of the frame within the video), which must be used, alternatively, to identify the target frame within the video. The frame should be identified by its location within the entire video, and not just within the portion of the video pertaining to the enclosing SynCog section;
- an audio segment – an *AudioSegment* element will be referenced. This element indicates the sensory record to which it pertains through the *resource* attribute. Its *time* attribute indicates the specific temporal point, within said record, to which it pertains. The time point should defined in terms of the entire duration of the audio record and not in the referential of the SynCog section in scope. A series of *FreqRange* child elements will jointly define the signal's component, at the specified temporal point, to which the *AudioSegment* element pertains. Each *FreqRange* carries the *minFreq* (lower bound frequency), *maxFreq* (upper bound frequency), and *amplitude* (amount of signal within this frequency range).

The *SensorySegmentations* element defines all the segments of sensory records from which the idealized constructs (defined within the earlier presented *Idealization* elements) derive (an *L* sized collection). It carries all the elements of the *ImageSegment*, *FrameSegment*, and *AudioSegment* types that are referenced by the earlier mentioned *SensoryGrounding* elements (i.e. the *SensorySegmentations* element defines all $Seg_{t,k}^{S_i}$ of all S_i that are relevant to all the I_j^t defined in the *Idealizations* element). The referencing is made through the unique identifiers that those elements possess (like all other elements of the metadata for the expression of synthetic cognitions), carried by the *id* attribute.

The conceptive part of a SynCog object expresses the semantic interpretation of the sensorialy observed and cognitively idealized reality. To do so the SynCog format employs a logical structure based on the concepts of entity, event and predicate (depicted as rectangles in the top global layer of Figure 5a), and on references to such concepts (depicted as rectangles in the top layer of each section).

Entities are real or abstract objects that may, or may not, include internal entities, the predicates are the relationships that exist between entities or events, and events are sets of entities and predicates or sets of sub-events. Their instances are the individual occurrences of such realities (e.g. a specific person, like John Smith, or car like John Smith's car; a specific event like John's car crash). The taxonomies of entities, events and predicates is, for now, outside the scope of our work. SynCog conceptive metadata defines instances of entities, events and predicates and binds them to the remainder of the interpretative information.

Conception information is expressed employing the xml syntax. It is stored in a single file (represented by the Global Conception Level in Figure 5a), which will thus comprise the entire conceptive interpretation of a specific set of acquired sensory records. Each section of a SynCog object, however, will have its own sub-set of the conceptive information (the Local Conception Levels of Figure 5a). This sub-set (stored in its specific file), will carry references to the entity, event and relationship concepts (expressed at the global conceptive information file), which are relevant to the time interval of the section of the SynCog record in scope.

The *ConceptionLayer* element is the root of the conceptive component of the interpretative information. Its internal structure may consist of either a sequence of *EntityRef*, *PredicateRef* and *EventRef* elements or a sequence *Entity*, *Predicate* and *Event* elements (as illustrated in Figure 7). In the earlier case, it specifically comprises:

- a (*M* sized) sequence of *EntityRef* elements – each of these elements references (via the *ref* attribute) an *Entity* element in the global conceptive metadata file (each *EntityRef* references an Ent_l^k);
- a (*R* sized) sequence of *PredicateRef* elements – each of these references (via the *ref* attribute) a *Predicate* element in the global conceptive metadata file (each *Predicate* defines a Rel_l^k);
- a (*S* sized) sequence of *EventRef* elements – each of these elements references (via the *ref* attribute) an *Event* element in the global conceptive metadata file (each *Event* defines an Ev_l^k).

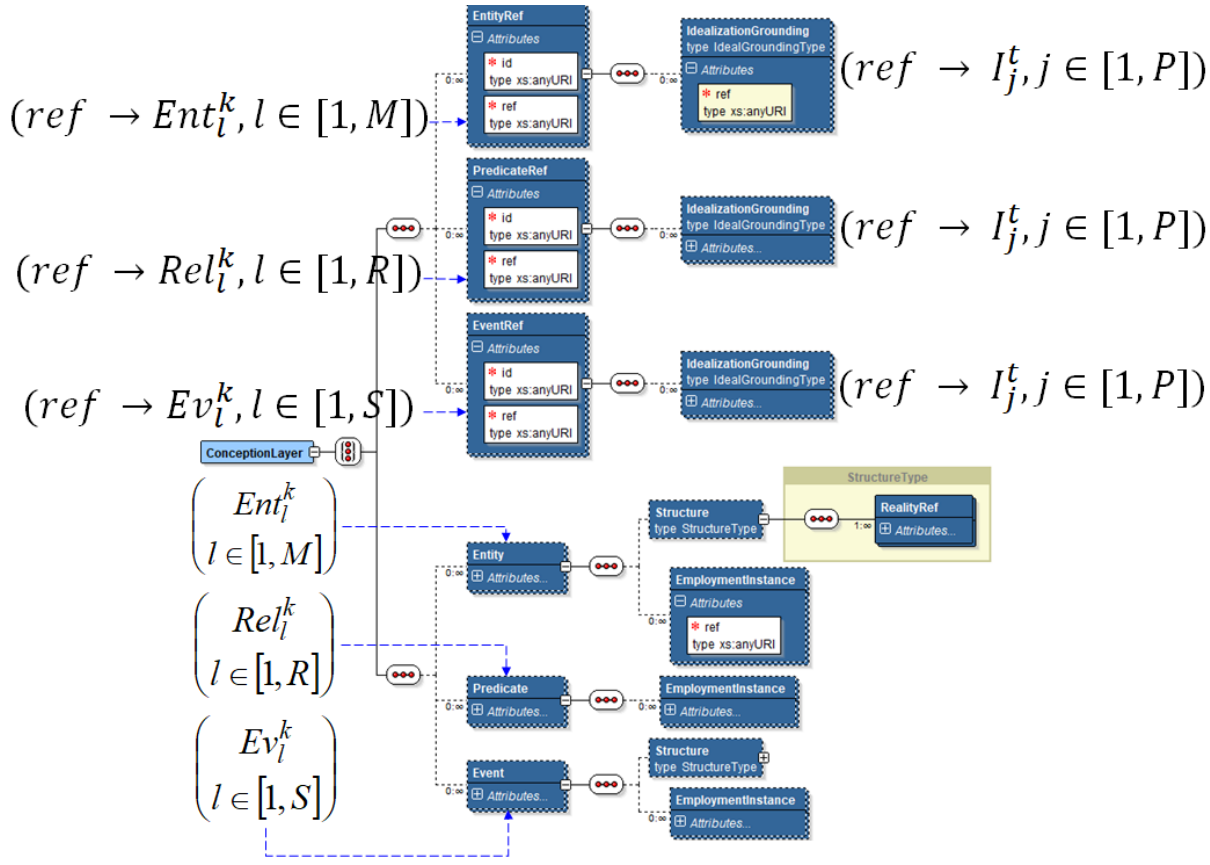

Figure 7. Conception Metadata Schema

Each of the earlier elements comprises series of *IdealizationGrounding* elements. Each of the latter performs the connection of the (reference to the) concept instance in scope (event, predicate, or entity instance) to one of the idealizations (I_j^t) from which it originates, by indicating, though the *ref* attribute, the value of the identifier of an *Idealization* element from the idealizing interpretative section. This way, any entity, relationship or event instance may be associated (via the *Ref* elements) to a set of idealizations performed over three-dimensional space throughout a set of temporal points (a *IdealizationGrounding* element binds one of Ev_l^k, Rel_l^k, Ent_l^k to one of to one of I_j^t from one of II^t).

In the latter case, the *ConceptionLayer* element comprises the following inner structure:

- a (*M* sized) sequence of *Entity* elements – all entity instances sensed within some sensory record (or set of them) of an SCE, spatio-temporaly idealized and conceptually understood are defined by an *Entity* element. The *type* attribute indicates the type of entity (within the context of a taxonomy identified by the *taxonomy* attribute), and the *id* attribute carries the unique identifier of the entity instance in scope (each *Entity* defines an Ent_l^k);
- a (*R* sized) sequence of *Predicate* elements – all instances of relationships sensed within some sensory record (or set of them) of an SCE, spatio-temporaly idealized and conceptually understood are defined by a *Predicate* element. The *type* attribute indicates the type of relationship (within the context of a taxonomy identified by the *taxonomy* attribute), and the *id* attribute carries the unique identifier of the specific Predicate instance. Each of these elements indicates, through the *actor* and *object* attributes, the entity or event instances that perform the agent and object roles within the relationship instance described by the *Predicate* (each *Predicate* defines a Rel_l^k);
- a (*S* sized) sequence of *Event* elements – all event instances sensed within some sensory record (or set of them) of an SCE, spatio-temporaly idealized and conceptually understood are defined by an *Event* element. The *type* attribute indicates the type of event (within the context of a taxonomy identified by the *taxonomy* attribute), and the *id* attribute carries the unique identifier of the event instance (each *Event* defines an Ev_l^k).

The three previous elements comprise the following inner structure:

- zero or one *Structure* element – the *Predicate* element does not possess this internal element, only the *Entity* and *Event* elements do. It carries a series of *RealityRef* elements. Each of these refers (through the *ref* attribute) an inner reality of the specific concept instance in scope. If it is an entity instance, only other entity instances may be referred as internal realities of it. If it is event instance, it may refer as internal realities other event instances, or alternatively, entity and predicate instances;
- a series of *EmploymentInstance* elements – each of these elements performs the connection of the concept instance in scope (event, predicate, or entity instance) to one of the *Ref* elements (*EntityRef*, *PredicateRef* and *EventRef*), from where it is referenced.

C. Employment Example

To ease the understanding of the mode (and advantages) of employment of the ML dataset repository defined in section IV.A and of the SynCog format (defined in section IV.B), we present in this section a practical example, pertaining to the conduction of research in ML, which demonstrates how, in that context, the services of such a repository will be resorted to, and how the mentioned format will be used for structuring the sensory and interpretative contents of the dataset held at the repository.

We thus envision the *Multi Perspective Video and Audio Repository* and its comprised *Dataset* (*MPVARepository* and *MPVADataset* respectively).

The purpose of *MPVARepository* is to function as a central hub for the storage and for the coordination of the cooperative construction and sharing of the information objects necessary for conducting ML research on automated interpretation of multiview video and multisource audio. *MPVARepository* enables client applications to add SynCog objects (to the *MPVADataset*) and to alter the existing ones (typically to expand their interpretative component), with varying levels of detail. The *MPVADataset* is to be employed for the development of ML applications for different specific objectives (within the overall context of automated interpretation of video and audio acquired within a retail space). Different research teams may employ the dataset and collaborate in its continuous expansion and enrichment.

In specific terms, the sensory contents of *MPVADataset* were acquired at *ShopX*. Its commercial space is monitored by: three different cameras, placed at different points, which perform a continuous video acquisition (producing sensory records S_1, S_2, S_3); and two microphones which capture the shop's overall sound panorama at two different points (producing sensory records S_4 and S_5). These five information feeds were acquired for *N* different intervals (each lasting *L* hours), at different days of the week and at different hours of the day. The *MPVADataset* comprises also interpretative contents (the metadata describing the realities observable in the sensory records). However, within the present context, the process through which this interpretative metadata is produced is of no relevance, as this work focuses on the expression of ML datasets as SCE records (and on the exploitation of such datasets), and not on how such experiences are actually produced.

The information resulting from each sensory acquisition episode, and the associated interpretative metadata, is structured as a SynCog object. Consequently the *MPVARepository* contains *N* SynCog objects, one for each mentioned sensory capture episode. Each such object comprises three simultaneous video records of the shop's interior and two simultaneous sound records. Each SynCog object contains also interpretative metadata comprising the spatiotemporal idealization of the shop's interior and of its contents (including the idealization's grounding to the sensory records) throughout the time.

The *MPVADataset*, therefore, originally comprises *N* SynCog objects, each including sensory and spatiotemporal interpretative information. Figure 8 presents the resulting logical structure of such a SynCog object (more precisely, the structure of the first section of such an object). It comprises sensory records S_1 to S_5 from t=0 to t=*T*, and $II^{t=0}$ to $II^{t=T}$.

It therefore comprises, at the base, three sensory information blocks of the visual type (i.e. S_1, S_2, S_3) and two more sensory information blocks of the audio type (i.e. S_4 and S_5). This logical level also comprises the information describing the spatiotemporal interrelations between, and

characteristics of, the sensory acquisition points (but it is not represented in Figure 8).

In the middle layer of the image is the idealized 3D reality throughout T seconds of the SynCog section (which correspond to some X sensory capture, and subsequent idealization, instants). At each such instant, we also represent the connection between one relevant segment of the 3D idealization of reality (a subspace of it, i.e. some I_j^t) and the segments of the sensory information (image and sound) from which its idealization derives (several $Seg_{t,k}^{S_i}$). That segment is the part of the idealization that corresponds to the body of a person who is moving through the shop. It is bound (lower dashed blue arrows) to the segments of the frames where the person is visible (from different points of view) and to the component of the audio data that is interpreted as being produced by the person and also assists in the formulation of the person's spatiotemporal idealization.

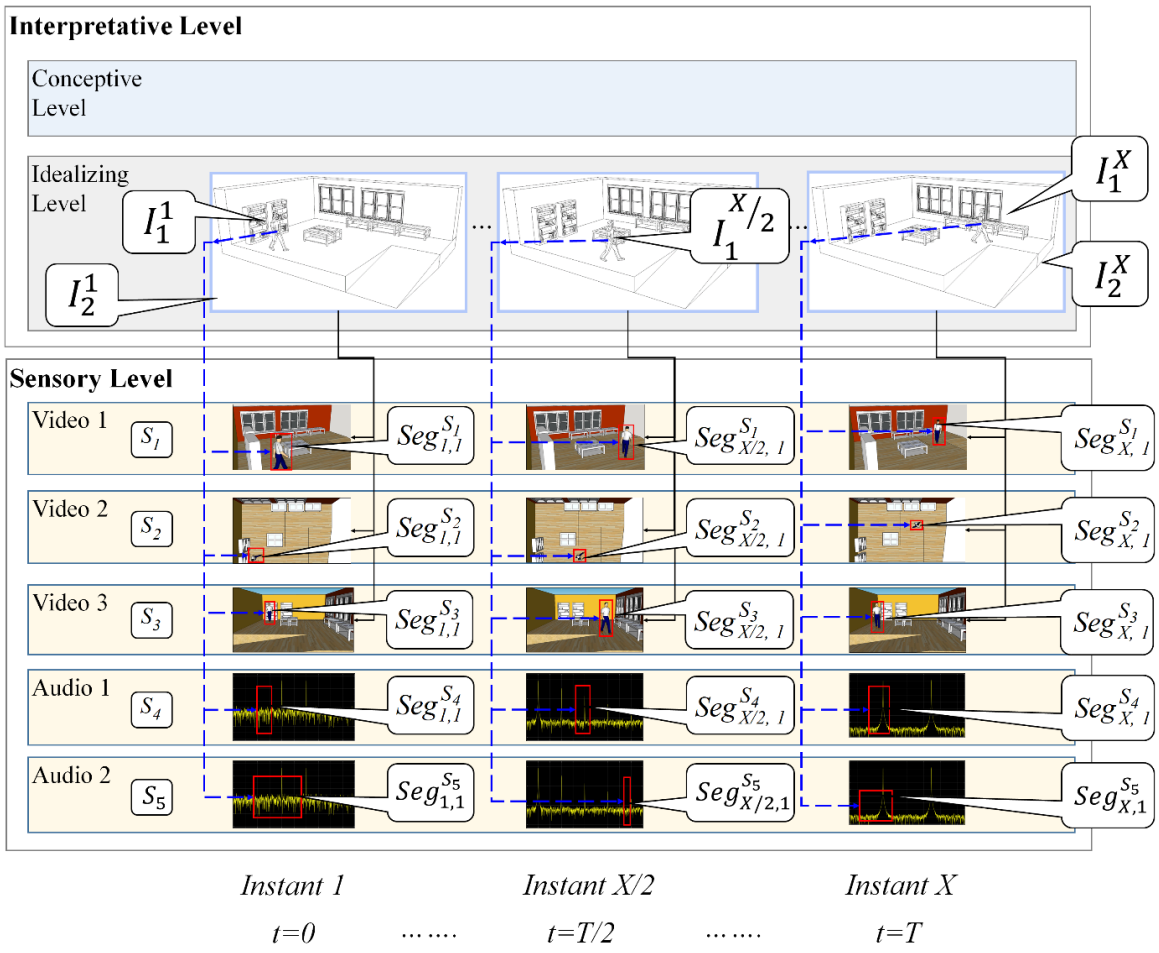

Figure 8. Original SynCog Section

At the top of the image, for informative purposes, we present an empty conceptual information layer. The original SynCog objects of this dataset include only sensory and idealization information, and thus have no conceptive information layer.

Research team A is developing a ML tool for customer 3D tracking across, and gender identification within, ShopX, through the employment of the *MPVADataset*. They first contribute to the expansion of the dataset by adding needed conceptive information to every SynCog object in the dataset. Team A thus employs the services of *MPVARepository* to add the necessary conceptive information (to each SynCog object) to perform the identification/differentiation of all the individuals (visible in any of the views of the sensory component of the SynCog object, and three dimensionally interpreted in the idealizing metadata), across the time length in which they are observable, as well as the conceptive metadata to enable the identification of the gender of such people.

Team A them, logically, divides *MPVADataset* into training and testing parts and proceeds to train and test their application. They thus retrieve the sensory components (video) of every SynCog object as well as the relevant parts of the interpretative component. The latter information includes the idealizing information (ground-truth for the visual detection, and three dimensional tracking of people across the shop), and the conceptive information (as ground-truth for the assessment of customer gender).

Both at the training and testing phases, the output of team A's application is also a SynCog object (for each training or testing SynCog object), with its spatiotemporal interpretation of the multiview footage and its semantic interpretation (built on top of the earlier one) of the customers' genders. To assess

the accuracy of its tool, at each developmental phase, team A has only to compare the latter SynCog (which has only an interpretative component), with the earlier mentioned ground-truths.

After team A's addition of conceptive metadata to *MPVADataset*'s SynCog objects, their structure is expanded (conceptive content is added). Building on Figure 8, we exemplify, in Figure 9, the resulting structure of a (section of a) SynCog object.

The conceptive layer of the observed section of a SynCog object now comprises (references to) one instance of the *Person* concept (i.e. *PersonY*), which is defined in the overall conceptive information block (independent of any section). Specifically, at every time instant the (section's) conceptual level contains a reference to one instance of the *Person* concept (i.e. a ref to Ent_1^{Person} or *PersonY*).

The overall conceptive information block comprises, besides the *PersonY* instance of the *Person* concept, one instance of the *hasProperty* relationship concept (i.e. $Rel_1^{hasProperty}$ or *hasProperty1*).

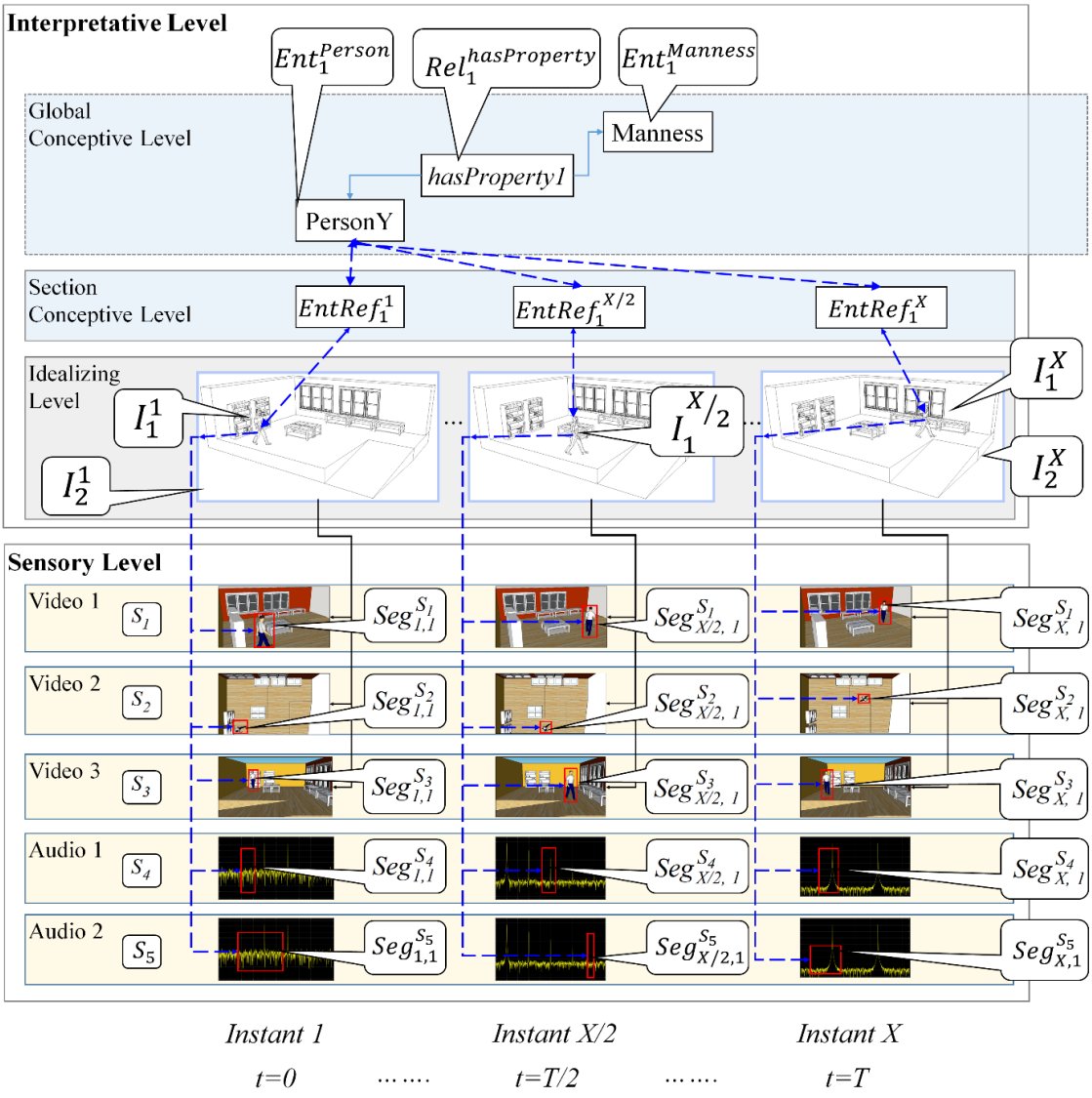

Figure 9. SynCog Section after Team A's Contribution

The descending direction of the connections between the elements (located at the conceptive layer) referencing the *PersonY* concept instance and the instances of that person's spatial idealization throughout the *X* time instants (top, dual-sided, dashed blue arrows), represents the grounding of the targeted concept instance on the segments of the reality idealization (I_j^t) from which it derives. The ascending direction of the dashed blue arrows represent the *ConceptArousing* logical links (the inverse of the grounding ones), establishing the connection between spatiotemporal idealizations and the concepts that emerge from them (*PersonY*), via the referencing elements at the conceptive layer of the SynCog object section.

The *hasProperty1* relationship concept instance connects the *PersonY* entity concept instance to the *Manness* entity concept (abstract entity which is the property of being a man). Both the *hasProperty1* relationship concept and the *Manness* abstract entity concept have no grounding connections to the

idealization information, as they do not derive directly from any of the observed scenario.

Employing the services of *MPVARepository* and the *MPVADataset*, team A is thus able to develop a ML application capable of determining the three-dimensional path of costumers across *ShopX*, as well as their gender. In the process team A contributes to the expansion of the dataset with further interpretative metadata.

Team B is developing a ML tool for customer behaviour identification, through automated analysis of audio and video information, within *ShopX*, by employing the *MPVADataset* (after its expansion by team A). Their desired ML application should process multiview video information and audio information and automatically detect the following types of customer behaviour: shop traversing; conversing between customers; and product browsing.

To be of use *MPVADataset* first needs to have some interpretative information added, so that a complete ground-truth is available. Team B then employs the services of *MPVARepository* to add the necessary idealizing and conceptive information (to each SynCog object). It consists of: the conceptive identification of events of the mentioned types; the grounding of that information to the spatiotemporal idealizations from which it originates; the mentioned spatiotemporal idealizations; the grounding of the earlier idealizations on the segments of the sensory records from whose interpretation they stem.

After team B's addition of idealizing and conceptive metadata to *MPVADataset* its structure is expanded. Figure 10 (expands Figure 9) shows the resulting structure of a SynCog object (section).

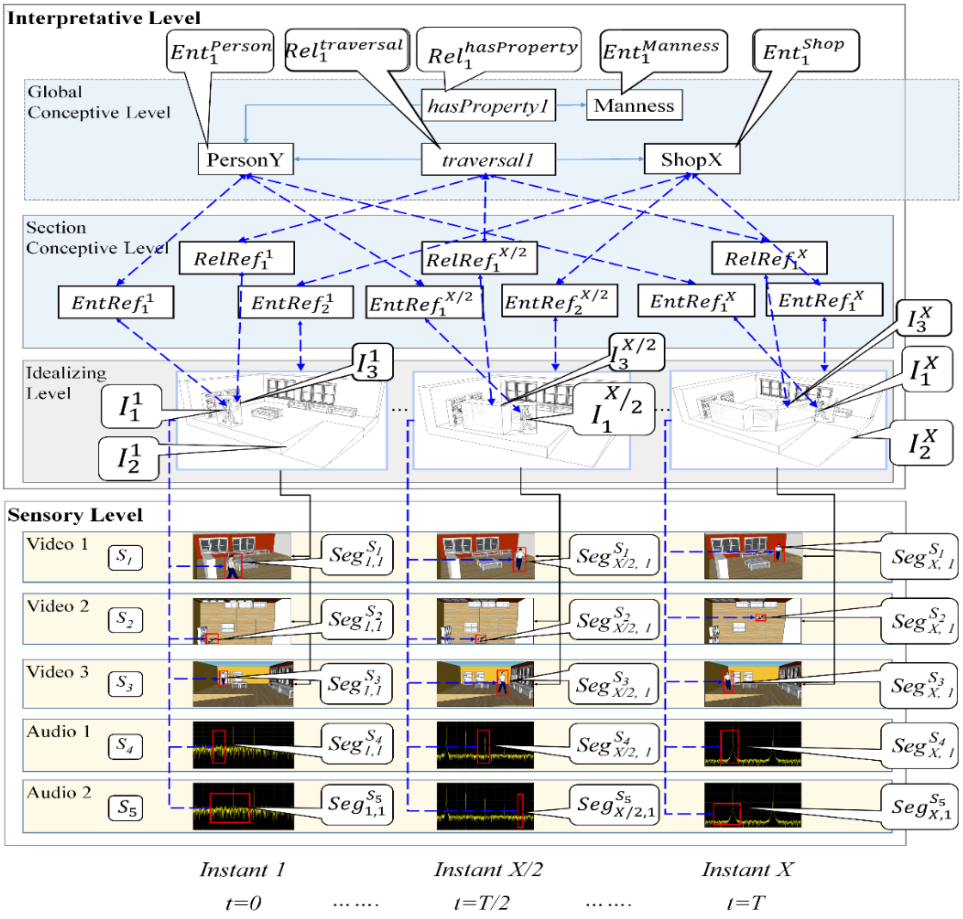

Figure 10. SynCog Section after Team B's Contribution

The conceptive layer of the observed section of a SynCog object now comprises (references to) the predicate concept instance of *traversal1* (an instance of the *traversal* concept), as well as references to the *ShopX* and *PersonY* entity concept instances (*PersonY* had already been added by team A). All concept instances are defined in the Global Conceptive Level (GCL), and the *traversal1* concept references *PersonY* and *ShopX* as its actor and object. Specifically at every time instant the (section's) conceptual level references one instance of the *Person*, *traversal* and *ShopX* concept (i.e. Ent_1^{Person} or *PersonY*, $Rel_1^{traversal}$ or *traversal1*, and Ent_1^{shop} for *ShopX*), which are defined in the GCL.

The connection between the elements referencing the *Person* concept instance (*PersonY*) and all the instances of its spatial idealization throughout the X time intervals (top dashed blue arrows), represents the grounding of that concept

instance on the segments of the reality idealization (several I_j^t) from which it derives. As such the *traversal* concept instance (*traversal1*) is also grounded on various segments of the idealization information throughout time. These segments (presented in Figure 10 as a stretching volume of space comprising the path of *PersonY*) represent the approximate space where it is considered that the *traversing1* is taking place. Just like the *traversing1* cognitive instance includes the *PersonY* cognitive instance so does the *traversing1* space include the *PersonY* space. The *PersonY* cognitive instance emerges from every instant of sensory capture where the person is observed. In its turn, the idealization of the *traversal* activity arises (at each time instant) from segments of the visual sensory record (the segment in each frame of each viewpoint where *PersonY* is visible) and segments of the audio sensory record (e.g. the segments corresponding to the sound of *PersonY*'s footseps). Just like the *Person* concept instance (*PersonY*), so is the *Shop* concept instance (*ShopX*) grounded in the idealization of the whole shop (top solid black arrows), and such idealizations are grounded on the sensations (video frames) that capture the shop environment (lower solid black arrows).

V. Comparison

The defined repository and, particularly, the SynCog tool, represent a clear step forward in the representation of synthetic cognitive experiences, and their storage, relatively to all comparable tools (approached in section III).

The proposed repository (section IV.A) far surpasses any of the existing ML dataset repositories, in terms of the ease of manipulation of the stored contents it allows. Existing repositories were predominantly conceived as dumps of such information. At most, they enable, the upload or download of individual, but complete, sensory and interpretative records. Very frequently they enable only an in bulk download of entire datasets and have close to no provisions for a collaborative construction of datasets. The repository we lay forth provides a much more detailed access its datasets, enabling a much finer grained submission, retrieval and alteration of stored records (e.g. addition of idealizing and conceptive metadata). What we propose is thus much closer to an effective memorial lobe of the emerging GSC.

The SynCog format, as it defines ML datasets as records (or sets of records) of SCEs (and given the superior adequateness of such a strategy, as explained in section II.D), enables a much more logically correct, and versatile structuring of such information. SynCog adequately separates the representation of the different levels of cognition, and assures the logical grounding of the information in each level on the information of the immediately inferior (interpretation wise) level.

Some of the most elaborate formats, for structuring the contents of datasets in the surveyed repositories, are the: Labelme format (employed in MIT's Labelme and SUN datasets); PASCAL VOC XML format (employed in ImageNet and PASCAL VOC Project datasets); the CVML format (employed in the CAVIAR project); or the ViPER GT xml format (employed in the BEHAVE, ETSIO, and ViSOR datasets). All of these are very simplistic and in no way enable the construction of information objects comparable to records of synthetic cognitive experiences. CVML, ViPER GT and also MPEG7, for instance, mix the expression of the three different components of cognitive experiences (sensation, perception and conception) in an indiscriminate manner, typically performing a direct association of, varying levels of, conceptual interpretations onto the sensory records.

SynCog enables expressing all levels of cognitive experiences (as opposed to RDF or VERL which focus only on conceptual aspects), and links them in a way that eases their logical decoupling and the navigation through their informational tissue, as well as its piecemeal manipulation. SynCog's inspiration in natural cognition makes it very versatile. It is capable of expressing a variety of synthetic sensory and interpretative geometries, and may thus express synthetic cognitions characterized by: multiple simultaneous sensory captures, acquired under different sensor placement geometries, and comprising different senses; the perceptual interpretative fusion of the acquired sensory base by means of complex spatio-temporal idealizations that are grounded onto multiples senses; and the association of multilevel conceptual information to the idealizing information, and therefore, to the sensory base. Existing tools, like MPEG-7 and CVML, permit only the association of meta-information to visual content (imagens e video), and in scenarios involving only one sensory capture and, therefore, without any multi-capture e multi-sensory information crossing at the higher cognitive levels.

RDF, like SynCog, also enables a versatile and expandable representation and interconnection of conceptual information. However, SynCog performs a much better grounding of the conceptive information onto the idealization information (and therefore, onto the sensory one as well), and has the added capability to define events, besides entities and predicates.

The employment of SynCog to structure and express the contents of datasets will result in a higher degree of complexity of such datasets (specifically of the GT metadata) than existing ones, and (mainly because of the earlier factor) will also result is a larger overall data size of their metadata component. The larger data size as negligible effects given the advancements in memory storage capacity of computing machines. The greater complexity is clearly compensated by the preciseness and expressive power and versatility it enables. On the long run it actually leads to greater clarity and less complexity than the simple extension of current solutions to deal with the progressively greater needs of technological cooperation in the field of ML.Also, the conversion of most existing datasets to the SynCog format is not a complex matter as only the metadata needs to be transformed.

In light of what is expressed above, SynCog is clearly more adequate, than any existing tool, for a versatile expression of SCEs in their multiple components, being thus a clear contribution for the development of ML and for the emergence and coalescence of the global synthetic cognition.

Summarily our work contributes to the development of a standard and interoperable type of ML dataset repository along with a universal and versatile format for structuring the information stored in such datasets. As exemplified in section IV.C, these tools enable a seamless cooperation between different research initiatives, the collaborative construction of shared datasets (particularly regarding the expensive work of ground truth production) and their continuous expansion and enrichment, as well as their selective reemployment by different initiatives with different objectives. We are thus contributing to enable the exploitation of synergies, within the field of ML, to the reutilization of work, and thus to the acceleration of ML progress.

VI. Conclusions

Artificial intelligence, or synthetic cognition, has progressed at a variable rate and has gone through different phases, each characterized by the pursuit of a specific technological venue. Presently the most promising technological venue is ML.

ML employs a bottom-up strategy for achieving cognition. It aims at developing tools that are able to learn and thus, to build their own image of the world. These tools are typically subjected to a training phase where they learn to perform some association between some input data and some output evaluation of it. The input data is frequently media information (at least in the context of our work), like audio and video content (i.e. data which is akin to sensory information) and the output evaluation corresponds to some interpretation of the observed audio or visual scene. The preparation of these training/testing datasets is one of the most labour intense aspects of ML research.

The ML research field comprises a myriad of independent initiatives and platforms (maintaining an equally diverse and heterogeneous set of datasets and associated repositories) which develop no systematic cooperation or integration of their achievements. There are no technical structures or standards for the facilitation of cooperation and for the development of synergies. Most relevantly, there are no tools for the systematic cooperation in the production of training/testing datasets, and no formats for a universal structuring of such datasets. This disconnected and uncooperative scenario, where *ad hoc* solutions are the rule, does not allow for an optimal progress in the field of ML.

In this work we argue that there is a pressing need for the development of: far more capable ML dataset repositories; and a universal format for the structuring of ML datasets. Furthermore, we argue also that the development of the mentioned tools should be guided by a vision that equates: the overall cybernetic tissue running ML provisions to a global synthetic cognition; the ML dataset repositories to memorial lobes of a said global cognition; and the datasets to records of synthetic cognitive experiences. In line with our argumentation we specify the basic functionalities that such a ML dataset repository should support. We define also a format (SynCog) for the universal structuring of ML datasets as records (or sets of records) of SCEs. We also provide an example of how the defined repository and format may be exploited for an optimal cooperation between ML research and development initiative.

This work thus contributes with a guiding vision and with the definition of two fundamental tools for fostering the progress in ML research, by enabling a universal cooperative production and sharing of ML datasets, as well as the sharing of the output results of ML provisions.

Future work should comprise the implementation of the defined repository, and of a software kit for the manipulation of SynCog objects (SynCog ML datasets), as well as the development of some specific and practical employment case to validate the developed tools. SynCog should also be adapted to allow the expression of pre-interpretative features calculated from the sensory data that are frequently employed by ML tools to lighten the interpretative work.

Acknowledgments

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.

References

[1] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, “The pascal visual object classes challenge: A retrospective”, International journal of computer vision, 111(1), 98-136, 2015.

[2] Imagenet, Large Scale Visual Recognition Challenge website, http://www.image-net.org/challenges/LSVRC

[3] ActivityNet website, http://activity-net.org

[4] Face Recognition Grand Challenge website, https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc

[5] D. H. Sanes, T. A. Reh, W. A. Harris, “*Development of the Nervous System*”, Elsevier Academic Press, 2006.

[6] AI Inpacts, "*Global computing capacity*", http://aiimpacts.org/global-computing-capacity

[7] ISC, Internet Domain Survey, http://www.isc.org/network/survey/ (retrieved 25/07/2017)

[8] F. Heylighen, "*Conceptions of a Global Brain: an historical review*", Evolution (Uchitel Publishing House) 1 (2011): 274-289.

[9] F. Heylighen and J. Bollen, "*The World-Wide Web as a Super-Brain: from metaphor to model*", Cybernetics and Systems' 96. R. Trappl (Ed.). Austrian Society For Cybernetics. 1996.

[10] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, "*LabelMe: a database and web-based tool for image annotation*", International journal of computer vision, 77(1), 157-173, 2008.

[11] SUN Database Website, http://groups.csail.mit.edu/vision/SUN

[12] The Activity Recognition Dataset Website, http://people.csail.mit.edu/hpirsiav/codes/ADLdataset/adl.html

[13] H. Pirsiavash, D. Ramanan, "*Detecting activities of daily living in first-person camera views*", In Computer Vision

and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 2847-2854). IEEE.
[14] Flickr Material Database Website, https://people.csail.mit.edu/lavanya/fmd.html
[15] MIT Indoor Scenes Website, http://web.mit.edu/torralba/www/indoor.html
[16] A. Quattoni, and A. Torralba, "*Recognizing indoor scenes*", In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 413-420). IEEE.
[17] Labeled Faces in the Wild Home Page, http://vis-www.cs.umass.edu/lfw
[18] ImageNet Website, http://www.image-net.org
[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma & A. C. Berg, "*Imagenet large scale visual recognition challenge*", International Journal of Computer Vision, 115(3), 211-252, 2015.
[20] Centre for Research in Computer Vision at the University of Central Florida Website, http://crcv.ucf.edu
[21] T. List & R. B. Fisher, "*CVML-an XML-based computer vision markup language*", In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 1, pp. 789-792). IEEE.
[22] Project CAVIAR website, http://homepages.inf.ed.ac.uk/rbf/CAVIAR
[23] Microsoft COCO Website, http://mscoco.org
[24] Computational Vision at Caltech HomePage, http://www.vision.caltech.edu/archive.html
[25] CityScapes Dataset Website, https://www.cityscapes-dataset.com
[26] THUMOS Challenge 2015 Website, http://www.thumos.info
[27] YouTube-8M Dataset Website, https://research.google.com/youtube8m
[28] KITTI Vision Benchmark Suite Website, http://www.cvlibs.net/datasets/kitti
[29] Annieway Platform Website, https://www.mrt.kit.edu/annieway
[30] RGB-D Object Dataset HomePage, http://rgbd-dataset.cs.washington.edu/index.html
[31] The CMU Multi-PIE Face Database Website, http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html
[32] UCI Machine Learning Repository Website, http://archive.ics.uci.edu/ml/datasets.html
[33] TIMIT Dataset HomePage, https://catalog.ldc.upenn.edu/LDC93S1
[34] UrbanSound Dataset Website, https://serv.cusp.nyu.edu/projects/urbansounddataset/urbansound.html
[35] Freesound Homepage, https://freesound.org
[36] VoxForge Website, http://www.voxforge.org
[37] WEIZMANN Dataset HomePage, http://www.wisdom.weizmann.ac.il/%7Evision/SpaceTimeActions.html
[38] BEHAVE Dataset HomePage, http://groups.inf.ed.ac.uk/vision/BEHAVEDATA
[39] KTH Dataset for Recognition of human actions HomePage, http://www.nada.kth.se/cvap/actions
[40] ETSIO Dataset HomePage, http://www-sop.inria.fr/orion/ETISEO
[41] Project ViPER website, http://viper-toolkit.sourceforge.net
[42] ViSOR Dataset HomePage, http://www.openvisor.org/index.asp
[43] MuHAVi Dataset HomePage, http://velastin.dynu.com/MuHAVi-MAS
[44] TV Human Interaction Dataset HomePage, http://www.robots.ox.ac.uk/~alonso/tv_human_interactions.html
[45] HMDB Database HomePage, http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database
[46] S. Dubuisson, & C. Gonzales, "*A survey of datasets for visual tracking*", Machine Vision and Applications, 27(1), 23-52, 2016.
[47] MILtrack Dataset HomePage, https://bbabenko.github.io/miltrack.html
[48] Microsoft Research Dense Visual Annotation Corpus Download Page, https://www.microsoft.com/en-us/download/details.aspx?id=52523
[49] SBU Captioned Photo Dataset Webpage, http://vision.cs.stonybrook.edu/~vicente/sbucaptions
[50] The PASCAL Visual Object Classes Homepage, http://host.robots.ox.ac.uk/pascal/VOC
[51] HomePage of the CV DAtaset of the Computer Science department of the Boston University, http://www.cs.bu.edu/groups/ivc/data.php
[52] FaceScrub Dataset HomePage, http://vintage.winklerbros.net/facescrub.html
[53] Arabic Speech Corpus HomePage, http://en.arabicspeechcorpus.com
[54] Million Song Dataset HomePage, https://labrosa.ee.columbia.edu/millionsong
[55] Cardiff Conversation Database Website, http://www.cs.cf.ac.uk/ccdb
[56] RDF definition page at W3C website, "*RDF 1.1 Concepts and Abstract Syntax*", W3C, https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225, 2014.
[57] OWL definition page at W3C website, "*OWL Web Ontology Language Reference*", W3C, https://www.w3.org/TR/owl-ref, 2004.
[58] L. Chiariglione, "*Summary of the MPEG-7 standard*", MPEG.
[59] J. M. Martínez, "*MPEG-7 Overview*", 2004, https://web.archive.org/web/20100214101510/http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm
[60] J. Annesley, J. Orwell, "*On the use of MPEG-7 for visual surveillance*", Sixth IEEE International Workshop on Visual Surveillance, May, Graz, Austria 2006.
[61] F. Pereira and R. Koenen, "*MPEG-7: A Standard for Multimedia Content Description*", International Journal of Image and Graphics Vol 1(3), pp. 527—547, 2001.

[62] Alexandre R. J.F. et al, "*VERL: an ontology framework for representing and annotating video events*", IEEE multimedia 12.4 (2005): 76-86.
[63] V. Y. Mariano et al, "*Performance evaluation of object detection algorithms*", 16th International Conference on Pattern Recognition, pp.965-969, 2002.
[64] D. Doemann and D. Mihalcik, "*Tools and techniques for video performances evaluation*", International Conference on Pattern Recognition, pp. 167-170, 2000.
[65] H. Castro et al, "*Cognition inspired format for the expression of computer vision metadata*", Multimedia Tools and Applications 75.24 (2016): 17035-17057.

Author Biographies

First Author Helder Castro holds a PhD in Electrical and Computers Engineering awarded, in 2013, by the University of Porto. He has, for the last fifteen years, been working on scientific research within the context of various National and EC funded research projects. His main research interests are distributed information systems, metadata production and exploitation particularly in the scope of synthetic cognition assistance.

Second Author Maria Teresa Andrade is an Assistant Professor at FEUP, at DEEC. She obtained a degree in Electrotechnical and Computing Engineering in 1986, the MSc in 1992 and the PhD in 2008, at FEUP. She participates in research activities at INESC Porto, integrated in the Multimedia Systems Area. Main interests include context-awareness, mobile and adaptable multimedia applications in heterogeneous environments; 3D and multiview video streaming; quality of service and of experience in multimedia services; semantic technologies and content recommendation; digital television, digital cinema and new media.