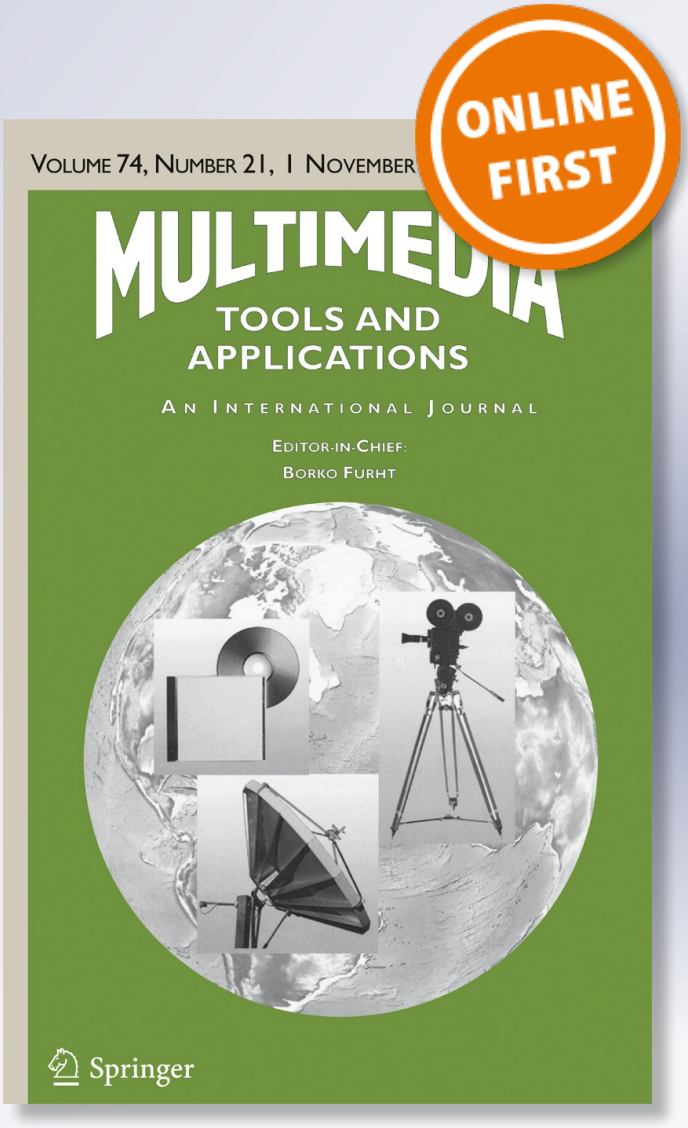
Cognition inspired format for the expression of computer vision metadata

H. Castro, J. Monteiro, A. Pereira, D. Silva, G. Coelho & P. Carvalho

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-015-2974-x



Springer

Multimed Tools Appl
DOI 10.1007/s11042-015-2974-x

Cognition inspired format for the expression of computer vision metadata

H. Castro [1] · J. Monteiro [1] · A. Pereira [1] · D. Silva [1] · G. Coelho [1] · P. Carvalho [1,2]

Received: 7 November 2014 / Revised: 18 September 2015 / Accepted: 29 September 2015

Abstract Over the last decade noticeable progress has occurred in automated computer interpretation of visual information. Computers running artificial intelligence algorithms are growingly capable of extracting perceptual and semantic information from images, and registering it as metadata. There is also a growing body of manually produced image annotation data. All of this data is of great importance for scientific purposes as well as for commercial applications. Optimizing the usefulness of this, manually or automatically produced, information implies its precise and adequate expression at its different logical levels, making it easily accessible, manipulable and shareable. It also implies the development of associated manipulating tools. However, the expression and manipulation of computer vision results has received less attention than the actual extraction of such results. Hence, it has experienced a smaller advance. Existing metadata tools are poorly structured, in logical terms, as they intermix the declaration of visual detections with that of the observed entities, events and comprising context. This poor structuring renders such tools rigid, limited and cumbersome to use. Moreover, they are unprepared to deal with more advanced situations, such as the coherent expression of the information extracted from, or annotated onto, multi-view video

✉ H. Castro
hcastro@inescporto.pt

J. Monteiro
jpsm@inescporto.pt

A. Pereira
ajrp@inescporto.pt

D. Silva
dvsilva@inescporto.pt

G. Coelho
agcoelho@inescporto.pt

P. Carvalho
pedro.carvalho@inescporto.pt

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

[2] Instituto Superior de Engenharia do Porto, Porto, Portugal

Published online: 24 October 2015

resources. The work here presented comprises the specification of an advanced XML based syntax for the expression and processing of Computer Vision relevant metadata. This proposal takes inspiration from the natural cognition process for the adequate expression of the information, with a particular focus on scenarios of varying numbers of sensory devices, notably, multi-view video.

Keywords Metadata · Multi-view video · Multimedia annotation · Computer vision · Cognition

1 Introduction

Over the last three decades the research on Computer Vision (CV) has increased significantly, benefiting from a favourable scientific, technological and commercial context – computers grew in power, individual cameras and camera networks have become ubiquitous and pervasive [17] as well as their applications. In 2013 the British Security Industry Authority (BSIA) estimated there are up to 5.9 million CCTV cameras in the UK [1]. Another example is that of the Chinese city of Chongqing which, as of 2011, had 510,000 surveillance cameras installed [24]. Unlike other types of sensors, video cameras can provide a wide field of view and a good space-time resolution [18], which favours the continuous proliferation of video surveillance systems. Moreover, the access to high quality and inexpensive video cameras and their integration in personal devices, such as mobile phones, has increased their penetration and made them a part of everyday life [5, 7]. There is also a growing desire for detailed content description in the media industry to enrich its current, and archived, products. The research community is also developing a growing interest in the automated interpretation of reality through the processing of video (and other media). This evolution has been accompanied by a proliferation of powerful computers with ever reducing size and by an increase of the speed and ubiquitousness of communication networks.

Despite their continuous advances, the practical surveillance systems, media annotation tools and CV provisions currently deployed, or the widespread video camera technology in use, are still unable to provide the desired autonomous analysis. This limitation implies that the billions of hours of video captured worldwide are typically stored and, only in some cases, analysed a posteriori, e.g. for "forensic" purposes [1, 24]. User produced video content is either stored (and frequently ignored), or shared on-line together with, typically poor, user produced describing metadata.

The practical and efficient exploitation of the existing massive quantity of visual information requires an automatic processing and interpretation to extract higher level information. This increasing demand for automated video processing and analysis has led to the development of a growing body of work in automated object detection and tracking. It has also rendered evident the need to associate the extracted meta-information with the video content so as to enable a richer and wider range of uses.

Such an association is also of critical importance for research, especially in the field of CV. The assessment of validity for CV techniques is commonly performed comparing the outcomes obtained through such techniques, with a reference (commonly known as ground truth (GT)). However, it is difficult to generate such information, as it is a cumbersome job, especially when detailed pixel-based references are needed.

Given the aforementioned considerations, there is a thriving demand for the development of tools for automated processing and analysis of video, which are able to perform the detection (segmentation) of relevant visual patterns and the interpretation of the observed scenes. Given

the growing volume of available video content and the automaticity and productivity of the desired mechanisms, the information that will result, from said automated analysis, will also be complex and, inevitably, voluminous. Hence, the associated metadata needs to be structured and registered in a manner which is adequate for a flexible manipulation and exploitation. Such mechanisms enhance the richness of the visual contents and their reuse.

Several solutions have been developed for the expression and structuring of the information produced by automated processing and interpretation of video media. However, the main focus of CV research has been placed on the actual processing of video content and not on the expression, storage and manipulation of the extracted information. As a result, the existing tools for the expression of multimedia metadata are still in a phase of considerable immaturity. They are, commonly, expressively rigid and implement an inadequate separation of the information pertaining to different logical/cognitive levels. For instance, the definition of a region of interest (ROI) in an image (frame) is at a different logical level than that of the identification of the entity or event that is observed in that region. The two types of information should be represented in an interrelated, but separate manner; this is frequently not the case. This results in the formulation of cumbersome information constructs that are not practical to use or exploit.

In the current technological and research context it is clear that the existing tools, for the expression and manipulation of metadata produced by CV algorithms, are far from being agile, flexible and scalable solutions. There is also a scarcity of tools for the expression of meta-information, extracted from visual content, which can support more complex situations in terms of the dimensionality of the contents, specifically multi-view scenarios. In this paper we present a metadata format, named Cognitive Digital Object Format (COG), defined to address the mentioned limitations. Moreover, we describe a tool developed to create and manipulate that metadata, as proof-of-content. The defined metadata format takes a loose inspiration in the natural cognition process to enable a comprehensive but flexible and logically correct expression of all the information extractable from 1-to-n dimensional video object (multiview video comprising 1-to-n videos captured from different viewpoints), at the level of detection (i.e. segmentation), perception (logical association of ROIs across frames and streams) and interpretation (identification of entities and events).

Section 2 describes related work in the field of multimedia characterizing metadata languages and of their manipulation. It points out current insufficiencies and identifies how they should be addressed. Section 3 describes the COG format and section 4 presents its validation through a software tool that was prepared to enable the swift manipulation of the metadata and thus its experimental testing. Section 5 highlights the contributions of this work and section 6 presents our concluding remarks.

2 Expressing computer vision results

2.1 Introduction

CV research focuses on the development of algorithms and mechanisms for the automated processing and interpretation of visual information, captured from the real world, with the aim of producing symbolic (meaningful) representations of the observed realities. Said visual information may be registered as still or moving images, and have varying dimensionalities (varying number of viewpoints). CV encompasses many different areas including: scene reconstruction [13]; event detection or action recognition [26]; video tracking; object recognition [6]; object (e.g. head) pose

estimation [25]; motion estimation; and image restoration [9, 22, 23]. For all these areas, it is important that the resulting information can be expressed and registered for storage, sharing or processing at a later time; another important aspect is the representation of reference information for assessment of the algorithms [2, 3]. However, this expression has received less attention from the scientific community than the processing of the visual information. Nonetheless, some contributions can be found in the literature. In the remaining sub-sections the most relevant of such developments are presented and their shortcomings explained.

2.2 State of the art

MPEG-7 [11], particularly its Audiovisual Description Profile (AVDP) [8], is an obvious candidate language for the expression of CV results. It was defined for describing the results of automatic multimedia analysis and enables the description of a comprehensive set of characteristics, features and contents of visual media, including the ability to tag specific information in each or multiple frames. However, it is a complex tool to use. Its conception was especially focused on the needs and characteristics of the media broadcasting industry, and not on those of the CV community, nor does it try to recreate the logical structure of the natural (i.e. human) cognition process [14]. As a result its constructs are not optimally suited for the expression of CV results.

The work presented in [10], coming out of project CAVIAR [15], proposes an XML based Computer Vision Markup Language (CVML) for the expression of the information extracted by computer vision tools. CVML is a versatile and light tool that enables the expression of a broad range of visually extracted information. However, it intermixes the description of aspects of the observed realities that are at different logical levels; for instance, it groups together the description of image (frame) segments (i.e. the description of detections), with the description of higher level aspects, such as the identity of the observed entity, its role in some specific context and the description of the surrounding scenario. Its capabilities for the description of relationships between entities and events, based on the definition of groups, are also limited and cumbersome. Unlike COG it does not establish any clear separation between the detection, identification and interpretation levels of the understanding of reality. Hence, CVML lacks in global clarity, flexibility and logical correctness.

The work presented in [21] proposes a metadata model (the Surveillance Application Metadata, or SAM, model), capable of describing the (online and offline) analysis results of sensory input of various types. It focusses on expressing such results as a set of time lines containing events and in a manner that enables the concise definition of dense spatio-temporal information such as object trajectories. However, even if it can employ any controlled vocabulary for the description of the observed reality, this model's basic logical structure is geared towards the description of surveillance relevant aspects. As a result it is event centred, placing a considerable focus on the description of trajectories. Consequently, it is not an optimal tool for the description of every type of observed reality. Furthermore, like most tools in this filed, it intermixes the declaration of logically separate realities (such as events and the sensory devices that captured such events).

The ViPER project [16] has also developed an XML based language for the expression of GT information for visual data to be employed for the evaluation of CV algorithms' performance. This language was intended for the annotation of information pertaining to practically any type of real world entity or event (i.e. reality), over video media. However, it aggregates detection information per observed entity or event. Even if linking it to specific frame spans, it does not aggregate that information on a temporally sequential basis bound to sequential

frames. As a result (like most solutions in this field), it intermixes detection metadata – ROI defining data – with metadata performing the identification and further characterization of the observed realities – identification and characterization data. ViPER's language provides only for a rigid declaration of relationships between detections from different frame spans, pertaining to the same observed reality, as it does so by placing them within a common parent metadata construct that identifies the observed reality. Hence the ViPER language is not adequately structured to access and display the information that it carries (the bounding boxes that define image regions, i.e. detections), in real-time to enable, for instance, the display of bounding boxes over playing video.

None of the existing metadata tools for multimedia annotation or for the expression of computer vision data extraction results, with the possible exception of MPEG-7 AVDP [20], is prepared to deal with multi-view scenarios. Hence, they are under equipped (or not at all) to enable the expression of interrelations between detections from (simultaneous) frames of different video streams, and of the interrelations of such detections and the realities observed in them.

2.3 The way forward

From the analysis of the current state of the art, presented in section 2.2, it becomes clear that there is still ample room and necessity for progress. The development of the existing tools has typically lacked a clear, and logically correct, guiding strategy, and there is also a nearly complete lack of support of multi-view scenarios. The needed tools should essentially be able to describe, or function as registries of, a synthetic cognitive experience. They, hence, stand only to gain from taking inspiration from the human cognitive process, and using it as a guiding analogy. This will enable an adequate arrangement of the expression of the different components of cognition, from the lower level aspects (i.e. the detection of visual shapes and patterns), to the higher level ones (i.e. the semantic interpretation), in a manner that does not just randomly mixes them all up.

Such tools should enable a layered and independent expression of the different levels of cognitive experiences. These should be connected through explicit and dynamic means that enable an agile attachment and detachment of such connections. Furthermore, the metadata tools should be prepared to coherently deal with the characterization/annotation of media from a multi-view scenario. Also, the extracted, or produced, information should be structured in a manner that facilitates its simultaneous reproduction/presentation together with the source media content.

3 Metadata format

3.1 Rationale

The automatic sensing and interpretation of reality has been occurring since immemorial times and has evolved together with living organisms, enabling them to understand and operate in the world [19]. The present quest for the development of cyber-physical systems capable of automatically sensing and interpreting reality, seeks to recreate, and possibly expand, the already native capabilities of biological organisms. Systems capable of sensing (e.g. image capture by cameras), processing and interpreting the outer world (e.g. image processing), may thus be seen as an artificial equivalent, albeit simpler, of the naturally evolved nervous systems.

Under the present analogy, and taking inspiration on the views of [12] on the stages of vision, the information objects resulting from the overall process of image capture and processing can be equated to registries of cognitive events comprising the registry of [4]: 1) sensory stimulae received from the world - the video content captured by cameras; 2) perceptions realized upon the basic sensory content corresponding to the division of shapes, forms and motions in the video content - image regions delineation and interconnection across frames and views; and 3) interpretations built upon the earlier perceptions that associate them to concepts of specific entities and events.

These parallels indicate that an approach based on the cerebral-cognitive operation for the description of the overall information objects in scope is advantageous, as it is enables profiting from the already existing examples of natural cognition and is well prepared to deal with the predictable evolution of these cyber-physical systems, paving the way for future developments. It will allow the development of complex information objects that structure the different logical components of image capture, analysis and interpretation into separate, but intricately related, parts and attain a high degree of flexibility, scalability and ease of manipulation. Based on this insight, we defined a metadata format for the expression of the perceptive and interpretative inferences realized upon the base visual information, possibly composed of multiple views.

3.2 Definition

3.2.1 Overview

The natural apprehension of reality starts from the sensing of stimulae. It then proceeds with the processing of that information for the detection of patterns and shapes, and onto the interpretation and valuation of the latter through their association to concepts of entities and events. The defined metadata format expresses the information resulting from the processing of video divided into three main levels. The most basic level (detection) comprises only the delineation of detections of specific visual patterns in particular points of sensory content (media files), i.e. the definition of image ROIs (frame segments), in which some specific event or entity is visually observable. The top level (interpretation) comprises the expression of semantic concepts which represent entities and events. The middle level (identification) performs the linking between the two previous levels. It associates different, possibly simultaneous, image segments (segments of frames captured from different points of view in a multi-view scenario spanning the same temporal extent), and also associates these to the concepts (of entities and events), defined at the level of interpretation.

The COG format thus enables the weaving of a complex, multi-layered mesh connecting detections realized over different sensations (video streams) captured from different, and synchronized, sensory devices (different cameras with different viewpoints), to the concepts that describe the entities or events that are observed within the mentioned detections. This mesh can be easily manipulated to establish, or remove, connections between specific concepts and spacio-temporal segments, where the entity or event represented by the concept is observable.

The proposed metadata format is also intended to support the perceptual and semantic annotation of multi-view video objects. These are composed by multiple video streams that result from the simultaneous capture, by different cameras, of visual information pertaining to the same scene or physical space. Each captured stream is composed by a sequence of frames and the total set of frames from all the streams may be transversely divided (in regards to the

time axis), into (approximately) simultaneous frame groups. Each such group is composed by a frame of each stream.

Each individual frame corresponds to a base visual registry, in which a specific set of detections may be realized. Each detection may be graphically defined by a polygon located within the space of the frame, and is associated to higher level information that performs the expression of the identification/interpretation of the realities (entities or events) observable in the defined regions.

As a result, the metadata's structure is divided into four levels: one level technically characterizing the media files and three levels carrying the interpretative metadata. Figure 1 provides a graphic representation of these levels, and their relationships, which are:

- **Capture** – declares and describes the most relevant technical aspects of the sensory (visual) information. It declares the captured streams and describes their characteristics, (e.g. frame rate, resolution, etc.);
- **Detection** – registers the data coming out of the visual detection of objects of interest, which correspond to effective realities captured on frames. It registers the polygons that circumscribe the identified visual objects in the different frames of a simultaneous frame group, along with the corresponding frames;
- **Identification** – registers the associations detected between visual information contained within regions (circumscribed by polygons) of different simultaneous frames (the frames in a frame group). It also registers the connection between these groups of associated polygons to the symbols (located at the ***interpretation*** section), that identify the real events or entities that are observable within them. This way, the ***identification*** metadata performs the connection between the more abstract concepts, which symbolize real world things, to the segments of visual information where such things are observable;
- **Interpretation** – performs the definition of the set of events and entities which are effectively observed within the captured visual information.

The following sub-section provide a description of the proposed COG metadata format, its sections and elements.[1]

3.2.2 COG root

At the root of the defined metadata format is the ***Data*** element (presented in Fig. 2). It carries sub-elements ***Capture***, ***Detection***, ***Identification*** and ***Interpretation*** which comprise the previously referred metadata sections.

All elements of the metadata model possess an attribute that uniquely identifies them within the model instance and enables inter-element referencing.

3.2.3 Capture

Each multi-view media object will be made up of video streams captured by n cameras (identifiable as $\{C_0 \ldots C_n\}$).

[1] A more detailed description of the metadata model can be found in http://mat.inescporto.pt/wp-content/uploads/2014/01/COG_metadata_model_schema.pdf

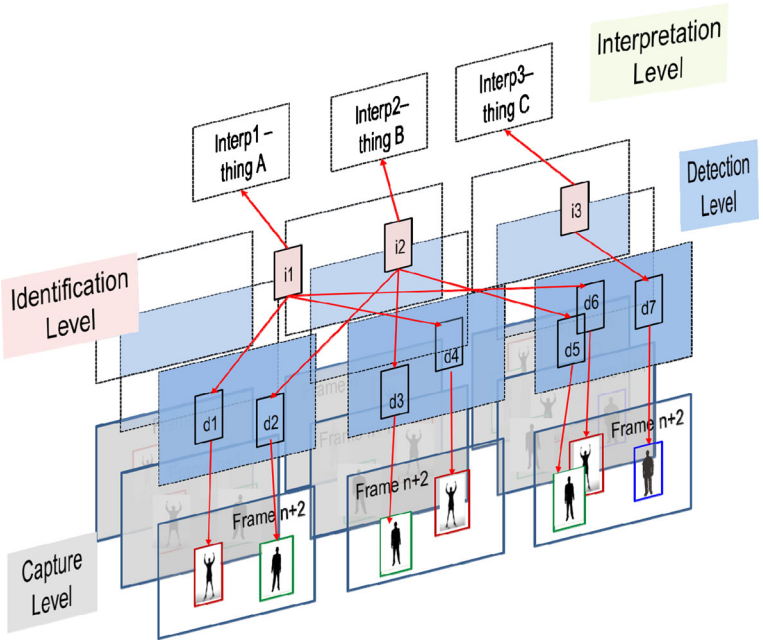


Fig. 1 Relationships between COG Levels

Each camera (C_i) produces a video stream (possibly stored in a specific file), composed by multiple (m) consecutive frames (identified as $\{f_1^{C_i} \ldots f_m^{C_i}\}$). Information describing these aspects of the multi-view video object is stored in the ***Capture*** part of the COG format.

The ***Capture*** element (Fig. 3) includes attributes to express the duration of the videos and their frame rate, ***duration*** and ***setspersecod*** respectively. Each ***FrameStream*** child elements represents one of the captured video streams.

3.2.4 Detection

Detection corresponds to the most basic level of reality interpretation. It essentially comprises the delineation of detections of specific visual patterns and shapes in the frames.

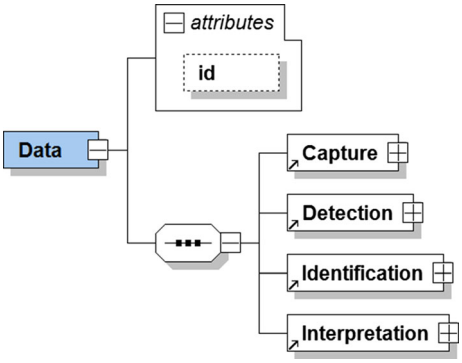


Fig. 2 COG Metadata root schema

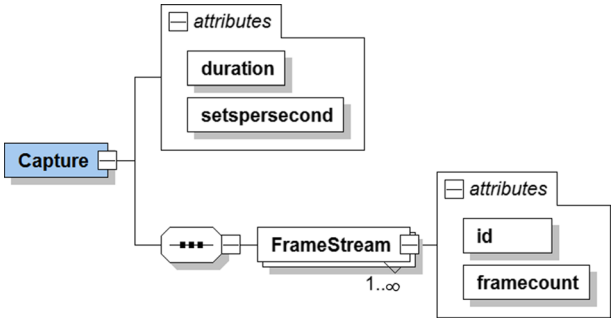


Fig. 3 Capture metadata schema

The proposed model provides support for a multi-view detection. In such multi-camera scenarios it is important (and expected), that they have a common time (or clock signal). Only with this requirement is it possible to identify true relationships in the captured frames, at any given time (i.e. identify that the captured realities refer to the same thing). Hence, cameras should be configured so that frames from different cameras pertaining to the same moment, have marginal time differences. As a result, the acquired frames may be grouped into simultaneous sets, $\{f_i^{C_1} ... f_i^{C_n}\}$, where each frame comes from a different stream/camera. During the image analysis process, the captured frames of the same set, will be analysed for the detection of specific visual objects. The data resulting from this process is expressed in the ***Detection*** part of the COG metadata structure.

The ***Detection*** element (see Fig. 4a.) contains a sequence of ***DetectionSet*** elements, with each element representing the set of visual detections: $\left\{ \left(d_1^{f_1^{C_1}} ... d_n^{f_1^{C_1}} \right) \| ... \right\}$ for frame 1 (i.e. $\{i=1\}$), performed over a set of simultaneous frames $\left(\{f_i^{C_1} ... f_i^{C_n}\}\right)$, coming from different streams. For example, these could be detections of people in different frames.

The ***id*** attribute of the ***DetectionSet*** uniquely identifies the element and is the time point, (counting from capture starting time and expressed in milliseconds), at which the capture of the frames of the ***DetectionSet*** takes place. Each ***Frame*** element within the ***DetectionSet*** represents a single frame (of a specific video stream), from a set of simultaneous frames.

Each such element carries the description of the graphical polygons that circumscribe the detected visual objects (i.e. $\left\{ d_1^{f_1^{C_1}} ... d_n^{f_1^{C_1}} \right\}$), within that frame ($f_x^{C_1}$ respectively). Hence, a ***Frame*** element contains a ***ts*** attribute that indicates the frame's actual capture time (in milliseconds), typically an offset with regard to the origin of the stream), and ***Box*** elements, each representing a specific visual detection, delimited by a polygonal box. For the ***Box*** element we have adopted a literature well established form, and included the attributes ***x***, ***y***, ***w***, ***h*** – coordinates of the centre, width and height. Presently only quadrilateral polygons are supported, but that restriction may be overcome in the future.

3.2.5 Identification

Identification comprises the recognition of associations between different detections, corresponding to the same reality, and the establishment of connections to the concepts that pertain

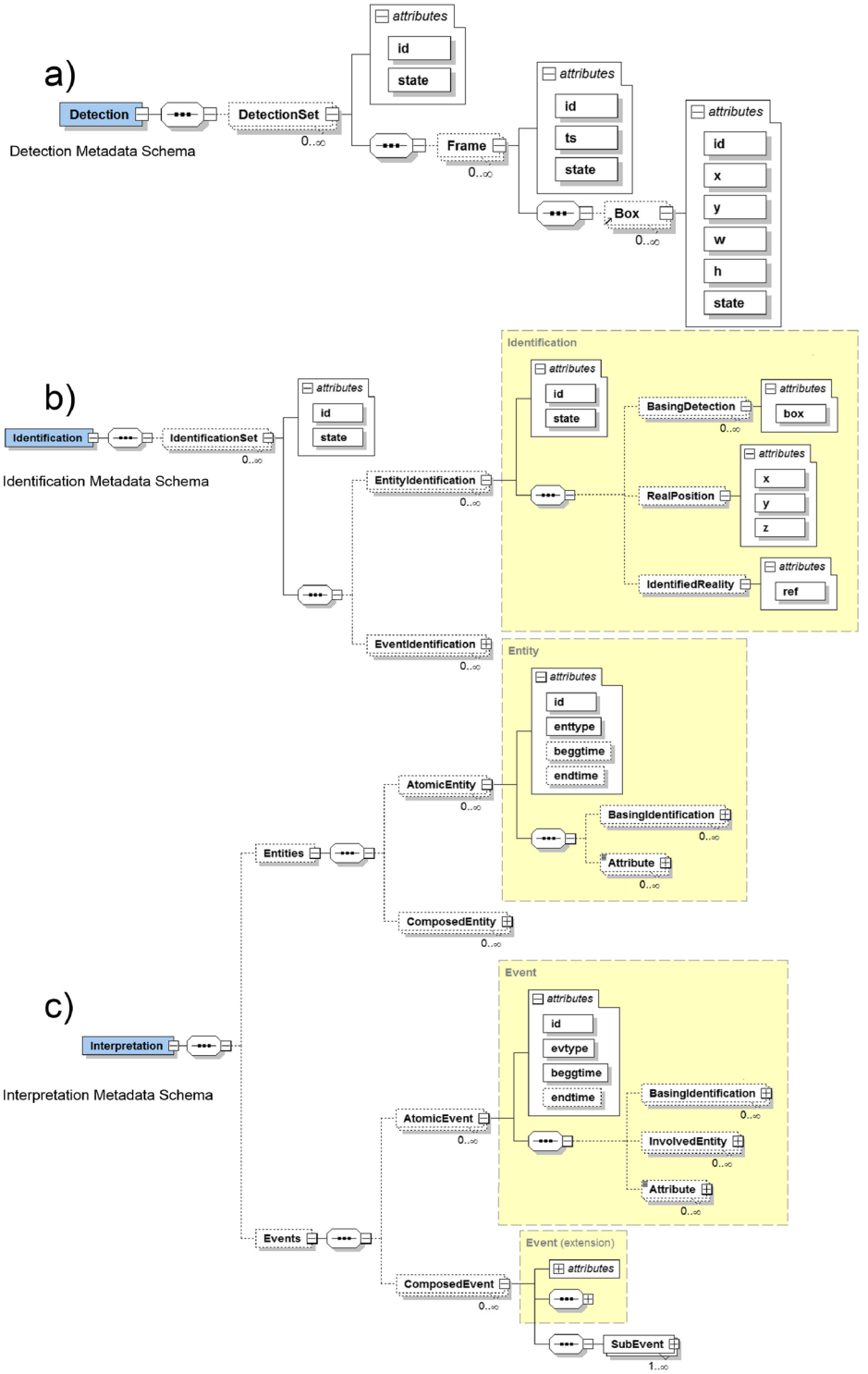


Fig. 4 Detection, identification and interpretation metadata

to them. It comprises, for instance, the association of different detections (from different simultaneous frames), of the same person, across different views (i.e. frames), and their linking to the metadata that represents the concept of that person.

The detections performed at the previous phase ($\left\{\left(d_1^{f_1^{C_1}} \ldots d_n^{f_1^{C_1}}\right) \| \ldots\right\}$) in each set of frames enable the identification of the real entities or events registered in the detections and their tracking across different, and consecutive, sets of frames. Hence, each set of detections, performed over a simultaneous set of frames $\left\{f_i^{C_1} \ldots f_i^{C_n}\right\}$, enables the identification of a set of realities (events or entities), definable as:

$$\left\{r_1^{\left(f_1^{C_1} \ldots f_1^{C_n}\right)} \ldots . r_n^{\left(f_1^{C_1} \ldots f_1^{C_n}\right)}\right\} = \left\{ent_1^{\left(f_1^{C_1} \ldots f_1^{C_n}\right)} \ldots . ent_n^{\left(f_1^{C_1} \ldots f_1^{C_n}\right)}\right\} + \left\{ev_1^{\left(f_1^{C_1} \ldots f_1^{C_n}\right)} \ldots . ev_n^{\left(f_1^{C_1} \ldots f_1^{C_n}\right)}\right\}$$

This information is conveyed within the ***Identification*** section, which contains a series of ***IdentificationSet*** elements. Each such element registers the identifications of a set of real entities and events, observed within a specific set of simultaneous frames. The structure of COG metadata, at the level of ***Identification***, is graphically depicted in Fig. 4b. Each ***EntityIdentification*** within the ***IdentificationSet*** element performs the declaration of the identification of an entity across a specific set of detections ($ent_x^{\left(f_y^{C_1} \ldots f_y^{C_n}\right)}$), realized over simultaneous frames (from the different streams that compose the multi-view video object). In turn, each ***EventIdentification*** element performs the declaration of the identification of a specific event across a specific set of detections($ev_z^{\left(f_y^{C_1} \ldots f_y^{C_n}\right)}$), realized over simultaneous frames.

The ***EntityIdentification*** and ***EventIdentification*** elements have the same internal structure, which includes ***IdentifiedReality,*** a series of ***BasingDetection*** and ***RealPosition*** elements. The ***IdentifiedReality*** element references the element at the *interpretation* level that describes the identified entity or event. Each ***BasingDetection*** element identifies a detection (a section of a frame circumscribed by a box), which serves as a base to the identification expressed by its parent element. It does so by referring the corresponding Box element's identifier (the value of the Box's id attribute). The ***RealPosition*** element expresses the estimated real position of the identified reality, according to some specific reference coordinate system.

In this way, the detections made over images are bound to the definition of the entities and events present in them by way of a separate level (the identification level). This enables a decoupling between the two levels that makes it very easy to effect changes on either side, without any need to alter the other.

3.2.6 Interpretation

Reality may be described as a set of events participated by entities. Events are temporally limited occurrences, i.e. they have a beginning and an end time. Events may also contain sub-events (for instance, a "*Retail Shop Action*" event may comprise several "*Product Browsing*" events). Entities may be simple "objects" (such as "*Client*" or "*Product*"), regarded as being without internal sub-entities, or complex ones (such as "*Client Group*"), which do comprise internal sub-entities (the individual "*Clients*").

In a succinct manner, one may identify: temporally limited occurrences in which entities are involved (events); and real or abstract objects (entities). Events may be sub-divided into atomic (events devoid of any internal events) and composed ones (events which comprise internal events). Entities may also be sub-divided into atomic and composed.

During the COG metadata production process, as new entities and events are detected and identified, their representation is added to the ***Interpretation*** part of the metadata. As a result, within the course of the processing of a multi-view video object, a set of entities $\{ent_1 .....ent_n\}$ and events $\{ev_1 .....ev_m\}$ (belonging to a limited set of domain specific entities and events), will be identified and declared within the ***Interpretation*** section. In the ***Identification*** part of the metadata, information will be added performing the connection between the ***Detection*** and ***Interpretation*** components, enabling a clear but flexible association between image segments (defined at the ***Detection*** level), to concepts representing entities and events.

The ***Interpretation*** element comprises an ***Entities*** element and an ***Events*** element, which respectively contain the declaration of all relevant entities and events identified throughout the full duration of the video streams. These elements contain a list of ***AtomicEntity*** and ***ComposedEntity***, in the case of the ***Entities*** element, and ***AtomicEvent*** and ***ComposedEvent***, in the case of the ***Events*** element. Each element of the list declares an atomic or composed entity/event.

An ***AtomicEntity or ComposedEntity*** has several attributes including: ***enttype*** (the type of entity in scope, e.g. "*Client*" or "*Product*"); ***beggtime*** (time instant in which it is considered that the entity begins its presence within the context under observation); ***endtime*** (time instant in which it is considered that the entity ends its presence within the context under observation). It also includes a set of ***BasingIdentification*** where each elements indicates, a specific identification, declared at the ***Identification*** level, of the entity in scope realized over a set of simultaneous frames. The full set of all such identifications constitutes the identification of the global participation of the entity in scope in the reality under observation. Finally, a set of ***Atribute*** sub-elements *convey* the value of a characterizing parameter of the entity in scope. That payload may also be semantic metadata (e.g. RDF) for furthering the conceptual characterization of the comprising entity. The name of that parameter is specified in the ***name*** attribute of the ***attribute*** element.

A ***ComposedEntity*** element has an additional set of ***SubEntity*** elements, whose attributes contains the identifier of the sub-entity in scope, for instance, the identifier of a "*Client*" (atomic entity) and the identification of the role played by the sub-entity in scope, within the context of its including entity, e.g. "*Buyer*".

An ***AtomicEvent*** *or* ***ComposedEvent*** has several attributes, including: ***evtype*** (the type of event in scope, such as "*Purchase*"); ***beggtime*** and ***endtime*** (same meaning as in ***AtomicEntity*** stated, but applied to the atomic event). It also includes a set of ***BasingIdentification*** sub-elements with the same meaning as in the ***AtomicEntity*** element. The ***AtomicEvent*** *or* ***ComposedEvent*** have a set of ***InvolvedEntity*** sub-elements, with each element referencing an entity, (atomic or composed), declared at the ***Identification***, which is involved in the event in scope. This element contains attributes to identify the involved entity, such as the identifier of a "*Client*" atomic entity, and the corresponding role within the context of the event in scope, e.g. "*Buyer*".

Additionally, a ***ComposedEvent*** element has a set of ***SubEvent*** elements that reference another event (atomic or composed), that is present within the *Interpretation* section and which is part of the composed event in scope.

The earlier described structure of the ***Interpretation*** part of the COG metadata, is graphically depicted in Fig. 4c. Figure 5 illustrates the overall metadata structure that was defined and how it integrates with the captured video content.

3.3 Employment example

Given its versatility the earlier described metadata format may be employed in a variety of different scenarios. It may therefore be used for: annotating audio-visual content with enriching, graphical and semantic, information (e.g. identification of characters and situations, product identification for product placement based advertisement, etc.); definition of segments or regions of interest within individual images, and tagging of semantic information to them describing such things as, the entities (or their parts) visible in them, the position of such parts (e.g. headpose), etc.; declaration of perceptual relationships between different image segments from the same or different images.

In the specific field of object tracking on video, our format may be employed to express (annotate), for each frame the dimensions and locations of the tracked objects, and, across frames, the identities, characteristics and the events (actions) in which such entities are involved.

Within this particular context, the COG format may be employed, for instance, for the expression (and subsequent sharing), of the results of automated video tracking of customers, across time and multiple viewpoints (cameras), in a retail space. In this regard, Fig. 6 graphically describes the logical structure of the COG metadata describing a situation where two clients are observed conversing, across a certain frame interval, from two different cameras. In more technical terms, Fig. 6, describes a situation where two entities (of the "*Client*" type), and an event (of the "*Client Conversation*" type), are visually detected, and identified, in two sequential time instants (t = 880 ms and t = 920 ms), and from two different viewpoints (recorded in two simultaneous frames). The visual detections are expressed as ***Boxes*** within ***Frames*** within ***DetectionSets*** identified with ids 880 and 920, at the ***Detection*** part of the metadata. The ***Identification*** part of the example performs (for both time instants) the connection between related detections (visual detections of the same entity or event in simultaneous frames), and between these and the ***Interpretation*** level concepts that identify them (i.e. entities g1:entity1 and g1:entity2, and event g1:ev1). This example thus describes a, simple, observed reality, in a multi-view video object, at two specific moments in time.

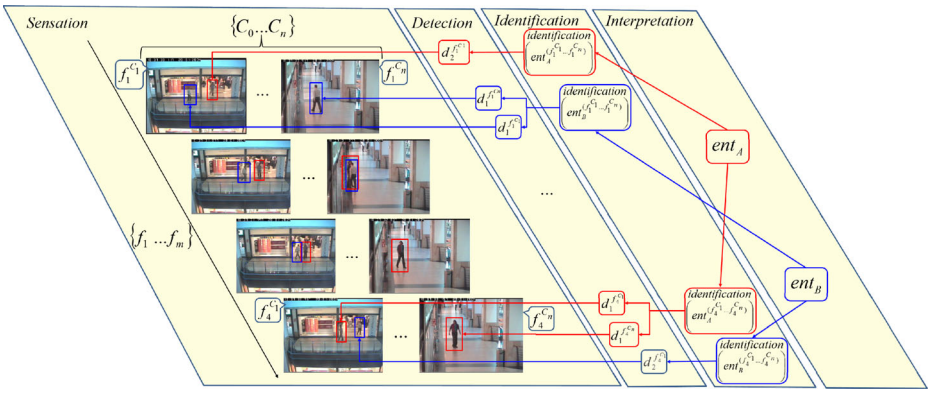


Fig. 5 Illustration of COG metadata structure

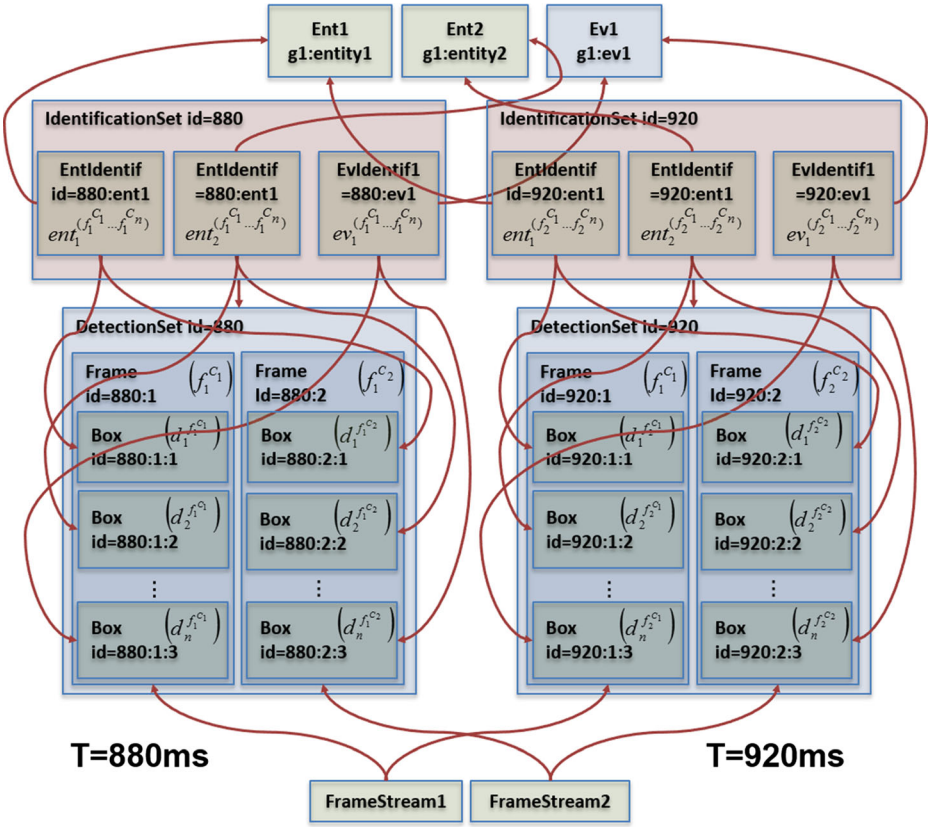


Fig. 6 Graphical depiction of a cog metadata structure

4 Format validation

4.1 Introduction

The validation of the COG format and demonstration of its contribution to the CV field, implies testing its practical employment in an illustrative usage domain, so as to show its adequacy, usefulness and ease of use.

This way, during the process of definition and implementation of the COG format, a software tool was developed to enable experimenting with it, in the specific domain of multiview video surveillance of retail spaces. Specifically, the developed tool, the Metadata Manager (MM), which comprises a COG metadata manipulation engine and a GUI (for user interaction), enables the employment of the COG format for the expression of the information produced (by automatic provisions outside the scope of this paper) through the analysis of the earlier multiview video data.

This tool has already enabled the employment of our model in, past and present, research projects, for the expression of tracking information obtained from the automated analysis of multiview video objects.

The MM prototype was equipped with the necessary capabilities to enable the production of detection, identification and interpretation metadata. It was also equipped with the required

provisions to enable the visualization of the metadata in scope, such as the graphical display of enriching meta-information over video content (for instance, the graphical delimitation of image segments, in motion, and in multiple different views, and the presentation of the identity of the observed entity or event).

The MM is thus capable of building and changing COG metadata documents, through input manually provided by users (which operate over the videos of the different views), or through input that is automatically produced. It is also capable of accessing and retrieving, the produced information, and displaying it over a view under reproduction, in real time, enabling, for example, the manual correction of automatically produced data.

4.2 Format employment

At the beginning of the annotation process, the MM automatically creates the metadata's skeleton in accordance with the number of views involved. This is composed of the base capture information and by a set of empty ***DetectionSet***s and ***IdentificationSet***s. An empty ***DetectionSet*** and ***IdentificationSet*** will be added for each set of simultaneous frames $\left(\left\{f_i^{C_1} \ldots f_i^{C_n}\right\}\right)$ of each time instant where a capture of frames occurs for the duration of the videos.

In the next step the operator adds all the specific concepts (describing events, or entities), which may be provided upstart (e.g. the concepts pertaining to each of the individual "*Staff Members*" involved in the operation of a store, as well as to the available "*Products*", and to the "*Purchase*" events). This information is inserted at the ***Interpretation*** level and becomes available to be associated to specific regions of specific frames, throughout the duration of the videos.

At the next stage, the human operator, or an automatic mechanism, proceeds to annotate the frames of each of the videos (illustrated in Fig. 7). This operation consists of defining quadrangular segments over each individual frame, throughout each video, and of associating them to the relevant entities or events, that were previously added to the interpretation section. If the target entity or event, of an annotation of a specific frame segment, is not yet present in the interpretation section, it is simply added.

Additionally, the quadrangular segment definition process, by a human operator, may be made more expeditious by means of automatic completion mechanisms that enable the operator to annotate (and associate to an entity or event) only some spaced-out (time-wise) frame segments throughout the video.

This overall process results in the addition of ***Box*** elements (one for each quadrangular segment), to the ***Frame*** elements (that correspond to the adequate view), of the ***DetectionSet***s (that correspond to the adequate time instant within the videos).

Once the annotation of all views is done, the process of annotating the multi-view video object, with information describing the visual detection of entities and events and their identification, is completed. At this time it is possible to playback any of the views, showing the annotated polygonal segments (superimposed over the video image). Figure 8 presents a still of such an annotation enriched playback.

4.3 Results and performance

The developed software permitted the testing of the COG data model through its employment within the context of different research projects (mentioned in the acknowledgments section)

Fig. 7 Image Segment Definition

enabling its validation in the fields of ambient assisted living and retail space surveillance (having the latter been chosen as the illustrative example of section 4).

Exploiting the capabilities of our data model, the developed software achieves an agile visual annotation of the multiple views, as well as an easy and efficient association and dissociation of

Fig. 8 Metadata Enriched Playback

the visual annotations to the entities and events that are observable in them. This is attained because of the structural dissociation between the different logical levels of interpretative metadata that it comprises. As the expression of detection data is structurally independent from that of interpretation data (being connected to it by identification data), changing the identity of the entity (e.g. identity of a specific customer) or event (specific observed customer action) associated to a specific frame segment (ROI), or set of frame segments, is as simple as changing a specific portion of identification data, or a set of them. This dissociation also enables a large reduction in data redundancy as interpretation level data (entity and event declarations) need only be present once, and all instances of their (specific entities or events) appearance in the views, will be connected back to the same interpretation data portion.

The annotation process enabled by the application leads to the production of a metadata file whose size depends on the duration and number of views, that comprise the multi-view video object, as well as on the reality that is captured in such videos and on the subset of its aspects that are to be effectively registered (detected, identified and interpreted). In an approximate manner, in the context of the annotation of a multi-view capture of the action within a store, where one intends to mark the image segments (and identify the entities present in them), that correspond to a maximum of ten people (clients and staff), the file storing the annotations will have a size of 5 MB per minute, for each view (which roughly corresponds to a 14 MB footprint of RAM memory when loaded). This represents a greater memory consumption than that of other tools (such as CVML as employed by VIPER software [16]), however it also enables a much more dense annotation of media, and a far more efficient manipulation of that annotation.

The loading of the annotations data from file to memory is done at an approximate rate of 34 MB/s. Once loaded onto memory, the sequential playback of any of the videos, with the annotating information (the part of it which is graphically depictable), superimposed over it, works smoothly. This is all the more remarkable as this process was tested on an hour long real HD video, and not over, for instance, an animated sequence of low resolution photos for less than a minute, as is the case for the VIPER project's tools.

The obtained results demonstrate the flexibility of the model. They show how it enables an agile and dense annotation of video, across multiple views, and an easy alteration and manipulation of such annotations, be it at the detection (frame segment definition) or interpretation (entity or event identification) levels. The results also demonstrate that the model is perfectly adequate to support a real time access to its contents for their rendering over the video. Overall, the attained results thus validate the adequacy of the defined metadata format.

5 Technical contributions

The COG language constitutes a comprehensive tool for the expression of computer vision processing results as well as for the definition of ground-truth data for the training of CV applications from various different areas of application. It thus fills a void in this area, regarding the universal, polyvalent and coherent expression of such cognitions.

In a summary manner COG's contributions are the following:

a) the structure of the COG format, which is loosely inspired by the process of natural cognition, is more logically correct than existing alternatives. These randomly intermix the expression of information pertaining to different levels of cognition, while COG performs

an adequate separation and isolation of such different levels. This enables the isolation of visual markings (frame ROIs), which are bound to their specific frames (and thus in a temporally sequential manner), from logically higher aspects, such as the identifiers of the realities enclosed in such markings. This isolation enables an easier alteration, manipulation and interconnection of the overall extracted information, and avoids unnecessary informational redundancies (characteristic of existing alternatives), that lead to added work at the time of metadata manipulation/alteration.

b) stemming from a) the COG format enables several functionalities, such as an absolutely expeditious correction of the identity or characteristics of an entity or event, without any need to operate over the detection metadata (image segment defining metadata). For instance, and considering the example depicted in Fig. 6, one may easily change an entity's description (e.g. g1:entity1), by altering the information inside its declaring element, and without the need for any further alteration to the metadata. Also, one may dissociate or associate an identification (e.g. 880:ent1) to a detection (box 880:1:1) without changing anything at the detection level. These examples illustrate the minimization of the impact that changes have on the overall metadata file.

c) the isolation of the ***Interpretation*** information in its own level also enables and agilizes the realization of high level searches for specific events and entities in the overall metadata and, consequently in the visual content as well. For example, the searching for all instances of the occurrence of events of type "*Client Conversation*", or for a specific "*Client Conversation*" event, as well as for all the intervals where a specific "*Client*" entity is visible.

The possibility to perform this type of searches over the COG structure paves the way for the expeditious realization of a broad set of studies and calculations pertaining to patterns of events, or of entity behaviours, observable within a specific video object.

d) in an innovative manner, the COG format also supports the coherently interrelated expression of meta-information extracted from multi-view video objects, and facilitates the playback of the visual aspects of that information, together with the video content, to a video player.

e) additionally, the developed COG metadata manipulation module enables the employment and exploitation of that metadata in a variety of contexts, and under stringent temporal performance demands.

The development of the COG format, and its manipulation module, thus constitutes a notable contribution for the CV field as it provides it with a tool for the standard and coherent expression of its processing results, paving the way for the sharing of such results and the development of large scale cooperative CV initiatives. This format will thus ease the work of computer interpretation of reality facilitating its large scale employment in such fields of application as the retail sector or sports. It may be used in those contexts, for example, to define the ground-truth for CV mechanisms training and calibration, as well as for the expression of their processing results and their sharing between different CV provisions for workload distribution.

6 Conclusions

The work, presented in this paper, comprised the development of a format for the expression of the information extracted/inferred from video by computer vision applications. Its usefulness, for CV tools lies in the fact that it provides a comprehensive and logically correct infrastructure for such provisions to express/register the results of their work, i.e., to express their visual detections and their interpretations of such detections.

The COG format presents a more logically correct structure than existing alternatives. It performs a separate registration of the information that pertains to the different levels of cognition enabling a more adequate manipulation and interconnection of that information. The format in scope, unlike most existing ones, also supports the coherent expression of the information extracted from multi-view video objects, as well as it eases the playback of that information together with the annotated video.

The work here described comprised also a software tool developed for the manipulation of computer vision metadata registries (including those associated to multi-view content), structured in accordance with the developed format. This tool enabled the validation of the usefulness and adequateness of the devised format, as it demonstrated how simply and efficiently the metadata may be produced and manipulated. It also proved the ease of access to the computer vision information expressed in the devised format, as it performed a real-time loading and playback of meta-information over the associated video content (single or multi-view), in scenarios characterized by high density metadata and HD video [1].

Our work therefore contributes, very relevantly, to the development of a common language for the description of CV experiences and for the communication of such experiences between collaborative and distributed CV provisions.

Nevertheless, there is room for improvement, both for the conceived format as well as for the implemented software tool. The COG format would benefit from a reduction of its verbosity leading to a lesser volume of data and making its manipulation easier. This may possibly be attained by eliminating empty *DetectionSets* and *IdentificationSets* or by writing the COG metadata in a specifically devised binary format, instead of text.

Furthermore, the manner in which the COG metadata is stored to disk, read from it, serialized and deserialized, needs to be altered so that the performance of the Metadata Manager, pertaining to the loading and saving speed of annotation data, may be improved. This may be attained by: changing COG so that it enables the splitting-up of the registered information (particularly that of detection and interpretation), into multiple independent files that may be loaded, altered and saved in a piecemeal manner (this however, may present some file management and handover issues); though the employment of a dedicated xml binding tool; or through the employment of a database for metadata storage.

Another desirable improvement is to enable the frame segments defined in COG to have shapes other than quadrangular ones, which presently is not supported.

Acknowledgments The Work was largely developed in the context of: project Media Arts and Technologies (MAT), NORTE-07-0124-FEDER-000061, financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT); Project QREN 23277 RETAIL PRO, a co-promotion R&D project funded by European Regional Development Fund (ERDF) through ON2 as part of the National Strategic Reference Framework (NSRF), and managed by Agência de Inovação (ADI); Project QREN 33910 ARENA, a R&D

project funded by European Regional Development Fund (ERDF) through ON2 as part of the National Strategic Reference Framework (NSRF), and managed by IAPMEI - Agência para a Competitividade e Inovação, I.P.

References

1. Barrett D (2013) One surveillance camera for every 11 people in Britain, says CCTV survey. The Telegraph. http://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in- Britain-says-CCTV-survey.html
2. Carvalho P, Cardoso JS, Corte-Real e L (2012) Filling the gap in quality assessment of video object tracking. Image Vis Comput 30(9):630–640
3. Carvalho P, Oliveira T, Ciobanu L, Gaspar F, Teixeira LF, Bastos R, Dias MS, Cardoso JS, Côrte-Real e L (2013) Analysis of object description methods in a video object tracking environment. Mach Vis Appl 24(6): 1149–1165
4. Castro H, Alves AP (2009) Cognitive object format, international conference on knowledge engineering and ontology development. Funchal. doi:10.5220/0002263103510358.
5. Doherty AR, Hodges SE, King AC, Smeaton AF, Berry E, Moulin CJA, Lindley S, Kelly P, Foster C (2013) Wearable cameras in health: the state of the art and future possibilities. Am J Prev Med 44(3):320–323. doi: 10.1016/j.amepre.2012.11.008
6. Drost B, Ulrich M, Navab N, Ilic S (2010) Model globally, match locally: efficient and robust 3D object recognition. In CVPR
7. Francescani C, NYPD (2013) expands surveillance net to fight crime as well as terrorism. Reuters, http://www.reuters.com/article/2013/06/21/us-usa-ny-surveillance-idUSBRE95K0T520130621
8. Information technology - multimedia content description interface - part 9: Profiles and levels, amendment 1: extensions to profiles and levels ISO/IEC 15938-9:2005/Amd.1:2012 (2012)
9. Kojima A, Tamura T, Fukunaga K (2002) Natural language description of human activities from video images based on concept hierarchy of actions. Int J Comput Vis 50(2):171–184
10. List T, Fisher RB (2004) CVML – An XML-based computer vision markup language. Proceedings of the 17th international conference on pattern recognition ICPR
11. Manjunath BS, Salembier P, Sikora T (2002) Introduction to mpeg-7: multimedia content description interface. ISBN: 978–0-471-48678-7
12. Marr D (2010) Vision. A computational investigation into the human representation and processing of visual information. The MIT Press, Cambridge. ISBN 978-0262514620
13. Newcombe RA, Davison AJ (2010) Live dense reconstruction with a single moving camera. In proceedings of the ieee conference on computer vision and pattern recognition (CvPR) 1:2.2
14. Pereira F, Koenen R (2001) MPEG-7: a standard for multimedia content description. Intern J Imag Grap 1(3): 527–547
15. Project CAVIAR website, http://homepages.inf.ed.ac.uk/rbf/CAVIAR
16. Project ViPER website, http://viper-toolkit.sourceforge.net
17. Reisslein M, Rinner B, Roy-Chowdhury A (2014) Smart camera networks [guest editors' introduction]. Computer 47(5):23–25. doi:10.1109/MC.2014.134
18. Saligrama V, Konrad J, Jodoin P (2010) Video anomaly identification: a statistical approach. IEEE Signal Process Mag 27(5):18–33
19. Sanes DH, Reh TA, Harris WA (2006) Development of the nervous system. Elsevier Academic Press, London
20. Sano M, Bailer W, Messina A, Evain J-P, Matton M (2013) The MPEG-7 audiovisual description profile (avdp) and its application to multi-view video IVMSP Workshop. 2013 IEEE 11th, pp 1–4, 2013.
21. Schallauer P, Bailer W, Hofmann A, Mörzinger R (2009) SAM – an interoperable metadata model for multimodal surveillance applications. In proceedings of spie defense, security, and sensing 2009. Orlando
22. Vezzani R, Cucchiara R (2010) Video surveillance online repository (ViSOR): an integrated framework. Multimedia Tools Appli 50(2):359–380
23. Volkmer T, Smith JR, Natsev A (2005) A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. Proceedings of the 13th annual ACM international conference on multimedia, pp 892–901

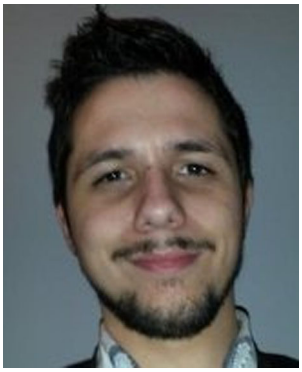
24. Wines M (2011) China: chongqing will Add 200,000 surveillance cameras. The New York Times. http://www.nytimes.com/2011/03/11/world/asia/11webbrfs-Cameras.html?_r=0
25. Yan Y, Ricci E, Subramanian R, Lanz O, Sebe N (2013) No matter where you are: flexible graph-guided multi-task learning for multi-view head pose classification under target motion. International conference on computer vision
26. Yan Y, Ricci E, Subramanian R, Liu G, Sebe N (2014) Multi-task linear discriminant analysis for multi-view action recognition. IEEE Trans Image Process 23(12):5599–5611

Helder Castro holds a PhD in Electrical and Computers Engineering awarded, in 2013, by the University of Porto. He has, for the last ten 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 and sustainable media content delivery.

João P. Monteiro holds an MSc degree in Biomedical Engineering from the University of Porto. He is currently doing his PhD and working at the Visual Computing and Machine Intelligence Group within INESC TEC in Porto. His PhD topic is personal health systems for assessment of upper extremity impairments. His main research interests are computer vision, machine learning and medical decision support systems.

Américo Pereira holds an MSc in Computer Science awarded, in 2013, by the University of Porto. He is currently doing research within the context of nationally funded research projects. His main research areas are computer vision and image processing.

Diogo Silva holds an MSc degree in Informatics and Computing Engineering finished in October of 2014, with a Thesis in Artificial Intelligence. Working as a researcher at INESC Porto, in Computer Vision and Machine Learning, since October 2014.



António Gil Coelho holds an MSc in Electrical and Computers Engineer awarded, in 2014, by the University of Porto. For the past 9 months he has been doing research work at INESC TEC within the context of research projects. His main research areas are computer vision and image processing.

Pedro Carvalho got his PhD in Electrical and Computer Engineering by the University of Porto in 2012. He is a Senior Researcher at INESC TEC where he has managed projects or research teams. He has developed work in the field of Multimedia with a greater focus on Computer Vision in the last six years. His main research interests are Computer Vision, with a particular focus on Object Detection and Tracking, and Multimedia Systems.