**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

CODING OF MOVING PICTURES AND ASSOCIATED AUDIO INFORMATION

#### **ISO/IEC/JTC1/SC29/WG11/N14509**

**Valencia, Spain, April 2014**

|  |  |
| --- | --- |
| **Source:** | **Requirements** |
| **Status:** | **Public** |
| **Title:** | **Compact Descriptors for Video Analysis (CDVA)**  |

# MPEG Vision

Current industry systems implement the “Compress than Analyse (CtA)” paradigm, where a video is compressed using traditional video encoding methods, for further analysis.

Guided by the requests and requirements from different industries, MPEG recently started preparations to develop tools for video processing and analysis. They will address the challenge to reverse CtA into the “Analyse than Compress (AtC) paradigm, shown in figure 1, to reduce the data amount for their transmission or storage and to achieve interoperability of implementations, applications and databases.



Figure 1 Analyse than Compress (on the top) and Compress than Analyse pipeline (on the bottom)

These tools will produce standardised descriptors called Compact Descriptors for Video Analysis (CDVA). The block diagram (Fig. 2) shows a general concept where video from the camera is analysed by the MPEG CDVA analyser (encoder) to produce a compact representation of the video content that can be easily interpreted by the CDVA interpretation engine (decoder).



Figure 2 A generic video analysis system employing CDVA descriptors

MPEG identified several functionalities that are of interest to industry and started collecting relevant requirements. The two areas of particular interest are: (1) media&entertainment and (2) automotive driver assistance. This work may be of interest to other industry domains, such as surveillance or augmented reality. MPEG invites further inputs with corresponding requirements.

MPEG is considering which descriptors and other elements of the processing system need to be standardised to ensure interoperability of various applications, components and services.

# Preliminary requirements

The initial functionalities and requirements for the automotive and media&entertainment domains are summarized below in order to stimulate discussion with the stakeholders. The functionalities specify the type of video analysis to be supported (e.g. object instance recognition, object detection & classification, scene classification) and the classes of visual objects that are of interest (e.g., persons, cars) and their properties (e.g. rigid/deformable, reflective, etc.). The requirements may also specify other constraints, such as: extraction complexity, maximum processing delay, accuracy or minimum performance levels required. A preliminary vision of the intended architecture is also provided as an illustration.

## Automotive Domain

Cameras are used for the automotive safety applications to support the drivers to avoid traffic accidents. The ability to detect specific object classes, such as pedestrians and cars, is essential to enable an improved level of safety on automotive platforms. The major camera applications are Forward/Backward Collision Warning, Cross Traffic Warning and Blind Spot warning. These applications need to detect the presence of certain objects and calculate the likelihood of the collision, and then warn the driver when the collision probability is high.

The functionality required is detection and classification of objects in the incoming video stream.

The most important object classes in this domain are people (e.g. pedestrians) and vehicles. Thus analysis has to cope with deformable objects (people) and reflective objects (e.g. cars).



Figure 3 Example of a pipeline implementing CDVA in the automotive domain

Figure 3 shows an example of a pipeline implementing CDVA in the Automotive domain. At the client side an image sensor captures raw video, which is then analysed/processed by the CDVA encoder. The encoder may apply compression for transmission over bandwidth constrained networks. At the opposite side of the network, where typically an application processor is located, the received CDVA descriptor is used for object detection and classification to determine if the detected objects belong to the classes of interest (pedestrian and vehicles in this case). The detected objects are localized in terms of temporal and spatial positions and used by collision warning system. The typical requirements for this application will include high detection rate, low probability of false alarms and very low latency.

## Media and Entertainment Domain

The current media and entertainment domain is characterised by increasing volumes of content, large number of delivery channels and an ever-growing need for the relevant content to be accessed on demand or published quickly. This is very difficult without appropriate tools to manage content items, including search for object instances in video, categorisation of scenes and content grouping. Thus, the ability to generate and exchange compact and standardized descriptors in an interoperable and efficient way is considered a key enabler in this domain.

Two example applications of particular relevance in this domain are:

1. Scalable instance search to find a specific object instance in a very large video database (e.g. like a broadcaster might have). The industry needs video descriptors that enable performing this task with smaller descriptor size and shorter runtimes as compared to using CDVS [3] in video domain. Rigid objects (e.g. textured and reflective buildings) are of highest relevance, however objects instances with more challenging properties (deformable, reflective and inherently 3D) are also of interest.
2. Object classification, to determine if a certain picture or video segment contains objects belonging to a certain class, using a database of class models.

The classes of objects of interest are numerous in this domain; an example of concept taxonomy is represented in Figure 4. It includes a sub-set of the concepts adopted for the TRECVID 2010 Semantic Indexing task [4]. The description of each concept and its mapping to the LSCOM taxonomy can be found in [5]. This includes persons, animals, and vehicles, these classes are further decomposed in sub-classes. Moreover scene-related concepts are introduced like outdoor and indoor; events, for example, can be sports. The set of classes of interest to MPEG will be defined by the relevant industry stakeholders.



Figure 4 Conceptual object taxonomy for Media & Entertainment

Figure 5 shows an example of a pipeline implementing CDVA in the Media and Entertainment domain. The video owner analyses the available content with a CDVA encoder and stores resulting descriptor in a CDVA database. Also, live video streams can be analysed and annotated in real time. This metadata can be then used by the user to search for specific object instances, object classes or in scene classification tasks. The key requirements for this application include ultra-fast matching in very large video archives (hundreds of thousands of hours), high detection rate, and low probability of false alarms. The application also requires object temporal and spatial localisation.



Figure 5 Example of a pipeline implementing CDVA in the Media and Entertainment domain

## Conclusion

Guided by the requests from stakeholders industries, MPEG recently started preparations to develop tools for video processing and analysis. These tools are intended to produce standardised descriptors called Compact Descriptors for Video Analysis (CDVA). MPEG has started defining specific requirements and evaluation conditions and is seeking further inputs from the stakeholders on that matter.

# References

[1] Draft Requirements for Compact Descriptor for Video Search, ISO/IEC JTC1/SC29/WG11/14095, Geneva, Switzerland, October 2013.

[2] Revised Automotive Requirements for Compact Descriptor for Video Search, ISO/IEC/JTC1/SC29/WG11/33075, Valencia, Spain, April 2014

[3] CDVS (2014). ISO/IEC DIS 15938-13 Compact Descriptor for Visual Search

[4]  <http://www-nlpir.nist.gov/projects/tv2010/tv2010.html#sin>
[5]  <http://www-nlpir.nist.gov/projects/tv2010/TV10-concepts-130_UPDATED.xlsx>