University of Modena and Reggio Emilia

D.I.I. - DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

VidiVideo

Interactive semantic video search with a large thesaurus of machine-learned audio-visual concepts

Tech Rep 7.0 - 7/01/2010

ViSOR - updates

Roberto Vezzani, Rita Cucchiara

Dipartimento di Ingegneria dell'Informazione

University of Modena and Reggio Emilia

via Vignolese 905 – 41100 Modena, Italia

Tel +39-059-2056111 Fax +39-059-2056129

email: {roberto.vezzani, rita.cucchiara}@unimore.it

1. Enhanced Graphical Interface

The web interface of ViSOR has been revised in order to improve its usability.
In particular, the video browsing section has been modified and new interactive behviours have been developed.
These changes have been highlighted by a comparative study of famous multimedia portals, like youTube and LiberoVideo.

A javascript slideshow effect allows to have a look of the content of each Video Category, and of the Video Summary only moving the mouse pointer over the thumbnails.

The Video detail page has been rearranged in order to improve the page readability.

Figure 1: new interface for video details

Following the other multimedia portals, the Related video section has been enhanced and highlighted.

2. New videos

One of the registered user has uploaded a new video sequence for traffic monitoring. Traffic monitoring is a surveillance task which was not jet covered by ViSOR.
In the following figure, the video is reported.

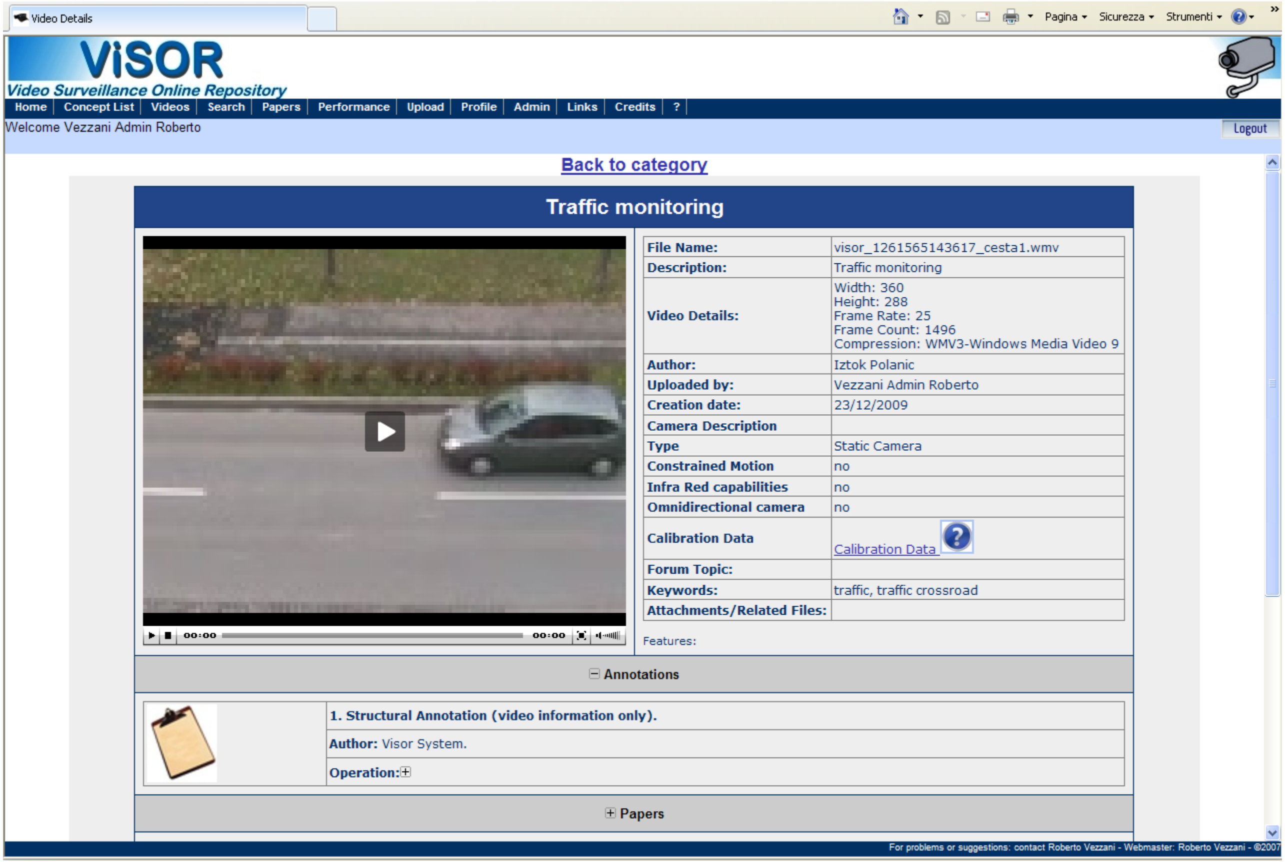


Figure 2: new videos for traffic monitoring

3. Multi Dimensional Annotation Space

Different types of annotation can be generated depending on the drill-down depth used to annotate the video and on the application goal. To this aim we defined three dimensions over which an annotation can be differently detailed: Spatial, Temporal, and Conceptual (STC space). In the following graph (Figure 3) these three dimensions are associated to the Cartesian axes and each point corresponds to a particular annotation type. For each dimension we have identified some significant values.

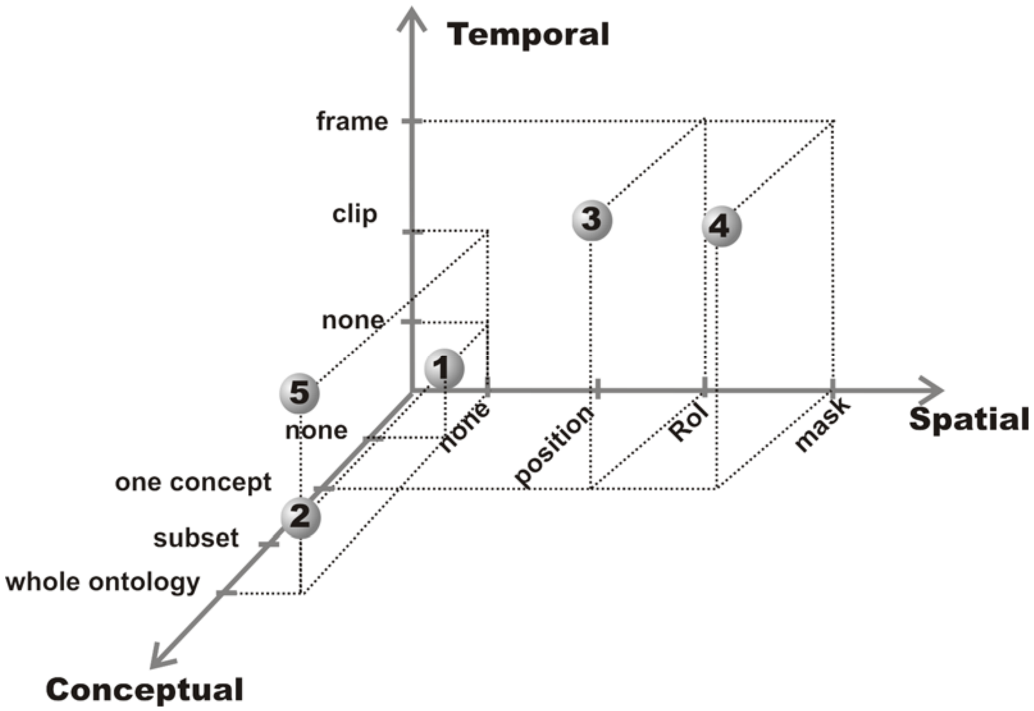


Figure 3: STC space of annotation. 1: (n,n,n) - video with metadata only; 2: (n,n,w) - video for CBVR; 3: (r,f,o) - video for people detection and localization; 4: (m,f,o) - low level segmentation; 5: (n,c,w) - TRECVid

Temporal dimension. The ontology concepts can be differently related with the time space, depending on the time interval during which the object is visible or the event/action is occurring. Some concepts can be associated to the whole video (e.g.: indoor, outdoor), others to a clip/temporal interval (e.g., person in the scene), and others to a single frame/instant (e.g., explosion, person entering the scene). But, even if an object is visible in a temporal fragment of the video only, the corresponding annotation can specify the presence of the concept without giving temporal details.

Thus, three temporal levels of description are defined :

• *none or video-level: no temporal information are given;*

• *clip: the video is partitioned into clips and each of them are described by the set of descriptor instances;*

• *frame: the annotation is given frame by frame.*

Let us consider an indoor video V1 stored in ViSOR, during which a child is sitting on a chair. Using the previously defined ontology, we can annotate the video with four descriptor instances I = {I1, I2, I3, I4}. I1, of type Location, annotates the type of video and thus the instance is detailed with the concept Indoor. The others instances are Person, StaticObject, and ActionByAPerson, which can be used to describe the concepts child, chair, and sitting action respectively.

An annotation at None or Video Level does not contain any temporal information, but this does not mean that the child is visible all time long. This can be used for testing the retrieval engine to access the video repository and search, for instance, a video containing a child. A Clip level annotation requires instead a partitioning of the Video into Clips defined by temporal boundaries. An instance In is then referred to a clip, but again it is not required persistence of the instance during all the clip. This is important in order to select only the part of the video where the action occurs, avoiding to download the entire video sequence. Finally, a frame level annotation specifies frame temporal boundaries for each entity.

Spatial dimension. Similar considerations can be done for the spatial level. Using the previous example, it is possible to specify where the child is in the scene by means of a single position (e.g., the centroid), of a Region of Interest (e.g., the bounding box) or giving the complete pixel-level mask. Summarizing, we can have the following four spatial levels:
• *none (image level : no spatial information are given and the concept is referred to the whole frame;*
• *position: the location of the concepts is specified by a single point, e.g. the centroid;*
• *ROI : the region of the frame containing the concept is reported, for example using the bounding box;*
• *mask: a pixel level mask is reported for each concept instance.*

Conceptual dimension An annotation can be done taking into account all the ViSOR concepts, in order to give a description of the video content as detailed as possible. In some cases, instead, only a few concepts are considered.
Some applications are specifically devoted to a single concept detection (e.g., smoke detection) and includes details on this particular concept only. Finally, each video can be described by a set of metadata (such as the file name, the frame rate, the frame size, and so on) related to the file itself and not to the semantic content of the video. In this case we talk of syntactical annotation only in opposition to the semantic annotation previously defined. Specifying the type of annotation by means of a conceptual level is important in order to infer if the lack of a concept in an annotation implies its real absence in the video or not. Summarizing, we have defined these four conceptual levels:
• *none (Syntactical level): no semantic information are provided; free-text keywords and title can be provided.*
• *one concept : only one particular concept is considered and annotated; other concepts can be added but they are not the focus of the annotation itself;*
• *subset: only a subset of the ViSOR surveillance concepts are considered and the subset adopted should be indicated;*
• *whole ontology: all the ViSOR surveillance concepts are considered.*

In Fig. 3 the STC space is illustrated with some example of possible annotation types. The example (1) is a typical video stored with metadata only (i.e., syntactical annotation) provided by the owner which uploaded the video. Point (2) represents instead a typical video annotation which takes into account the whole ontology but without any spatial or temporal information, i.e. the annotation is a simple list of the concepts included in the video. CBIR systems typically require this kind of annotation to perform queries by concept.
Points (3) and (4) correspond to annotations which indicate for each frame the ROI (eg. for face) and the pixel mask (eg., for object tracking) respectively. Finally, point (5) is similar to (2) but it represents a keyframe-annotation such as the one used in TRECVid, where the annotation content is referred to clips identified with keyframes. To better explain the ViSOR ontology and the STC classification we have reported three case studies related to common video surveillance problems:

object tracking, action recognition, smoke detection. For each of them, some sample frames from the ViSOR dataset are shown in Fig. 4.
Since occlusions are a still challenging problem for tracking systems, ViSOR has been enriched by videos with different occlusion types, captured in both indoor and outdoor environments (first row of Fig. 4). In the second row of Fig. 4 we have reported some frames of the ViSOR action recognition dataset; in particular, frames from the jumping, drinking, taking of the jacket, sitting and abandoning object actions are reported. Finally, a computer vision smoke detector should identify the presence of smoke by means of video analysis and issue an alarm to alert nearby people that there is a potential fire. Samples of some videos for smoke detection are presented in the thirs row: these videos contain smoke generated by a smoke bomb.

Figure 4: Case studies: smoke detection, people tracking with occlusions, action recognition

ViSOR implements the STC annotation classification: each annotation is described by a set of icons which indicates its spatial, temporal and conceptual level (Figure 5).

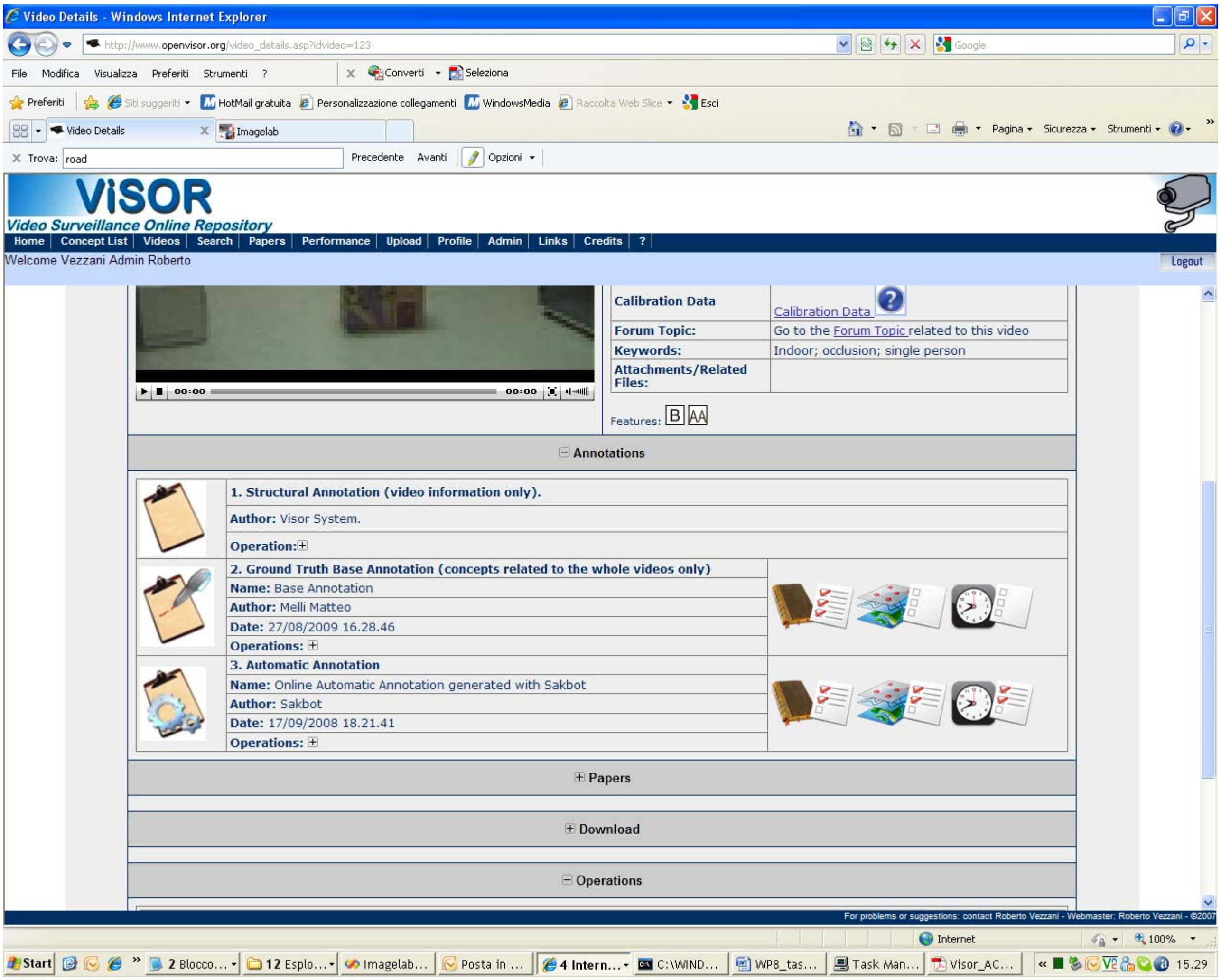
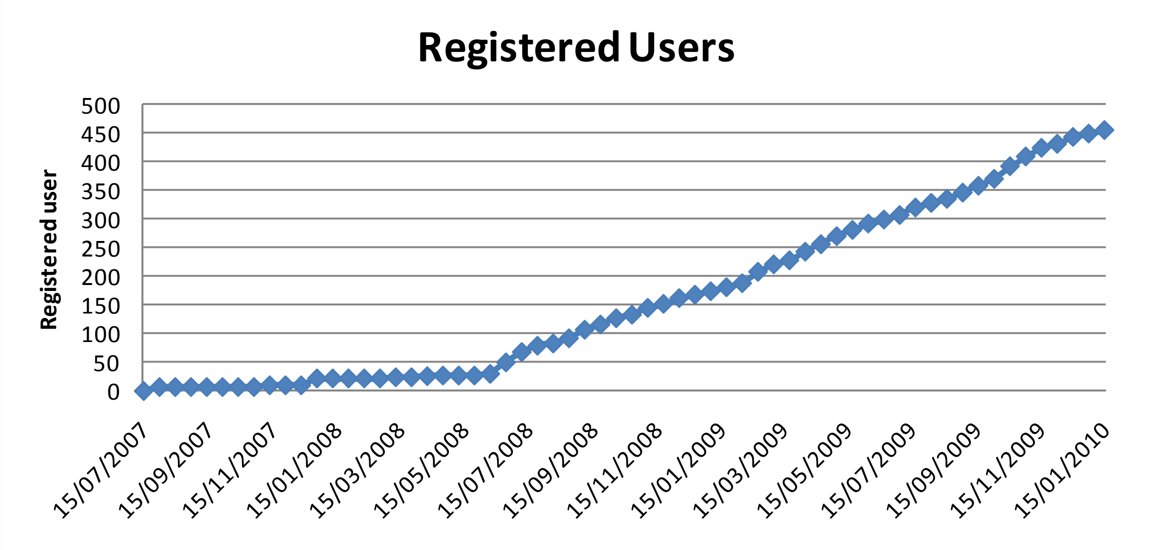
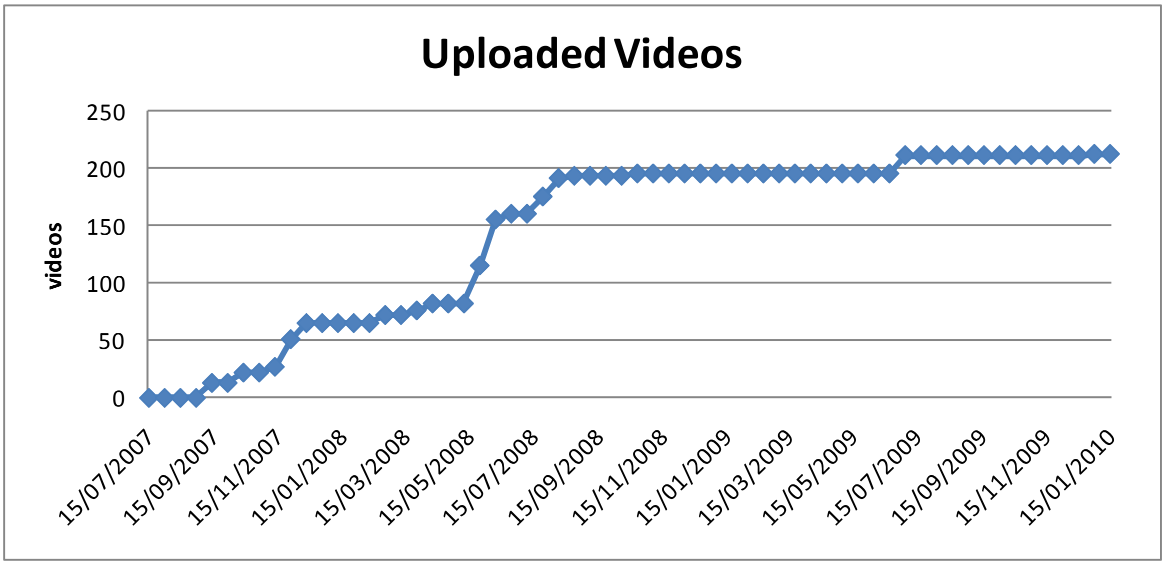
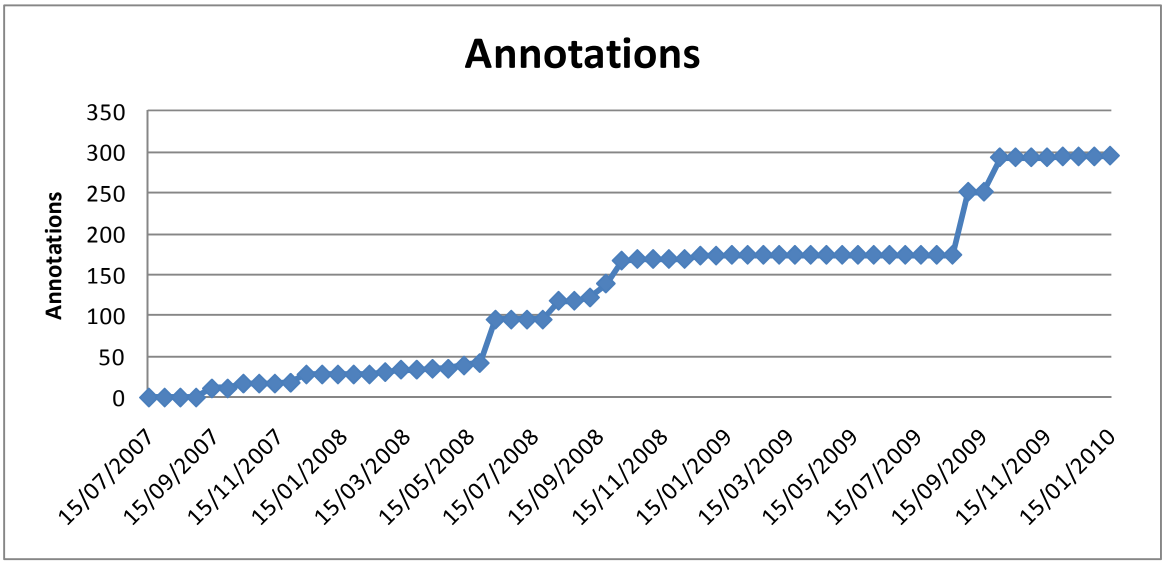


Figure 5: STC classification used in ViSOR

4. Updated user statistics (7/01/2010)







5. Events

We have organized a workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis (PRAIxHBA). During the event, endorsed by VidiVideo, we have advertise ViSOR.

6. Recent publications

R. Vezzani, R. Cucchiara, "Video Surveillance Online Repository (ViSOR): an integrated framework" (available online) in Multimedia Tools and Applications, DOI 10.1007/s11042-009-0402-9, 2009

The paper " R. Vezzani, R. Cucchiara, ViSOR: Video Surveillance Online Repository" has been ACCEPTED in the BMVA Annals, Special Issue on Surveillance and Security Performance Evaluation.