VISOR: VIDEO SURVEILLANCE ON-LINE REPOSITORY FOR ANNOTATION RETRIEVAL

Roberto Vezzani and Rita Cucchiara

Imagelab - Dipartimento di Ingegneria dell'Informazione
University of Modena and Reggio Emilia, Italy

ABSTRACT

The Imagelab Laboratory of the University of Modena and Reggio Emilia has designed a large video repository, aiming at containing annotated video surveillance footages. The web interface, named ViSOR (VIdeo Surveillance Online Repository), allows video browse, query by annotated concepts or by keywords, compressed preview, video download and upload. The repository contains metadata annotation, both manually annotated ground-truth data and automatically obtained outputs of a particular system. In such a manner, the users of the repository are able to perform validation tasks of their own algorithms as well as comparative activities.

***Index Terms*—** Video repository, video surveillance, annotation, ViSOR

1. INTRODUCTION

Video surveillance involves many central and open problems of computer vision and multimedia, including multi-camera calibration, object detection, tracking, recognition, event detection and retrieval. Regarding these problems, algorithms, implementations, complete systems have been proposed and discussed. A number of researchers have conducted careful comparisons between specific algorithms as well as complete system implementations. The community focused on performance evaluation has formed and holds annual conferences devoted to the subject, like the PETS workshop series[1] or the VSSN workshops of the ACM Multimedia Conference [2]. Perhaps the most valuable outcome of these meetings has been the introduction of a variety of testbed datasets that can then be used to study various video surveillance algorithms. As new research problems emerge, new controlled datasets should also be produced. Several dataset have been generated and published to cover a particular topic. Some examples of available datasets are reported in the table of Figure 2. The major drawbacks of these repositories are the lack of user interaction and the too specific target. For example, user cannot share their own annotation data, or grow the dataset with other videos, or comment them, and so on.

This work was supported by the project VidiVideo (Interactive semantic video search with a large thesaurus of machine-learned audio-visual concepts), funded by *EC VI Framework programme*.

This work presents the ViSOR project, a video surveillance online repository for annotation retrieval[3]. First aim of ViSOR is to gather and make freely available surveillance video footages for the research community on pattern recognition and multimedia retrieval. At the same time, our goal is to create an open forum and a interactive repository to exchange, compare and discuss results of many problems in video surveillance and retrieval. Together with the videos, ViSOR defines an ontology for metadata annotations, both manually provided as ground-truth and automatically obtained by video surveillance systems. Annotation refers to a large ontology of concepts on surveillance and security related objects and events, defined including concepts from LSCOM and MediaMill ontologies. Moreover, ViSOR provides tools for enriching the ontology, annotating new videos, searching by textual queries, composing and downloading videos.

2. VIDEO SURVEILLANCE CONCEPT LIST

To ensure interoperability between users a standard annotation format has been defined together with the structure of the knowledge base. The knowledge which could be extracted from video surveillance clips is structured as a simple "concept list": this taxonomy is a basic form of ontology where concepts are hierarchically structured and univocally defined. The concept list can by dynamically enriched by users under the supervision of the ViSOR moderator to ensure the homogeneity and the uniqueness. The goal is to create a very large concept list avoiding synonymy and polysemy drawbacks.

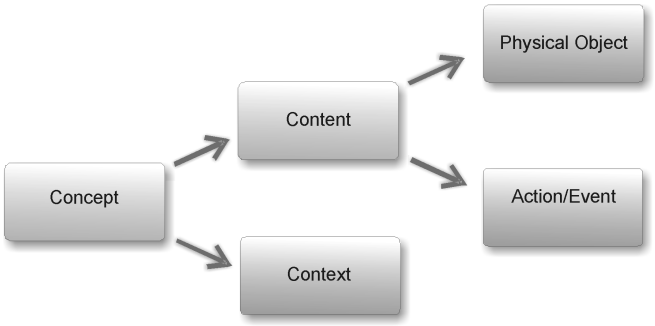

Fig. 1. Hierarchical taxonomy of the video concept categories

Dataset		Website	Topics	Ground-Truth	Size
CANDELA		http://www.multitel.be/~va/candela/	Indoor left-luggage and traffic monitoring on road intersection	no	16 indoor
Etiseo		http://www-sop.inria.fr/orion/ETISEO/	Object Detection, Object Localization, Object Tracking, Object Classification.	yes	86 video clips
i-Lids (AVSS 2007)		ftp://motinas.elec.qmul.ac.uk/pub/iLids/	Stopped vehicles and abandonded luggage	yes	14 sequences
ObjectVideo Virtual Video		http://development.objectvideo.com/	Tool to generate virtual video sequences for surveillance purposes.	yes	-
PETS	2001	http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001-dataset.html	Outdoor people and vehicle tracking	yes	5 sequences
	2002	http://www.cvg.cs.rdg.ac.uk/PETS2002/pets2002-db.html	Indoor people tracking (and counting)	yes	6 sequences
	2004	http://www-prima.inrialpes.fr/PETS04/caviar_data.html	People tracking and activity recognition	yes	28 sequences, 6 scenarios
	2006	http://pets2006.net/	Surveillance of public spaces, detection of left luggages	yes	7 datasets (4 camera views each one)
	2007	http://pets2007.net/	Multisensor sequences containing loitering, attended luggage removal (theft), and unattended luggage	yes	8 datasets (4 camera views each one)
SELCAT		http://www.multitel.be/~va/selcat/	Level crossing monitoring for stopped vehicles detection.	yes	8 sequences
SPEVI		http://www.spevi.org	Face detection and tracking	partial	10 sequences
Traffic datasets by Institut für Algorithmen und Kognitive Systeme		http://i21www.ira.uka.de/image_sequences/	Traffic surveillance in particular on road intersections	no	14 sequences
VISOR		http://imagelab.ing.unimore.it/visor	Indoor and outdoor surveillance sequences; annotation data for object detection, tracking, events, and much more.	yes	65 sequences at 12/12/2007 (in progress)
VSSN		http://imagelab.ing.unimore.it/vssn06/	background subtraction competition	no	7 sequences

Fig. 2. Available surveillance datasets

We defined a basic taxonomy to classify the video shapes, objects and highlights meaningful in a surveillance environment (see Fig. 1). A "concept" can describe either the *context* of the video (e.g., indoor, traffic surveillance, sunny day), or the *content* which can be a *physical object* characterizing or present in the scene (e.g., building, person, animal) or a detectable *action/event* occurring (e.g., falls, explosion, interaction between people).

The defined concepts can be differently related with the time space. Thus, we defined a time based taxonomy of the concepts depending on its span, e.g. the time interval during which the object is visible or the event/action is occurring. A concept can be associated to the *whole video* (e.g.: indoor, outdoor), to a *clip/temporal interval* (e.g., person in the scene), or to a *single frame/instant* (e.g., explosion, person entering the scene).

A first reference concept list has been obtained as a subset of two different predefined sets, respectively the 101-concept list of UvA[4] and LSCOM[5]. Since these lists have been defined for generic contexts, only a subset of the reported concepts have been elicited for video surveillance. Moreover, UvA and LSCOM lists are key-frame based only and are not enough to describe activities and events. An extension of the base LSCOM list have been considered (LSCOM Revised Event/Activity Annotations: video-based re-labeling of 24 LSCOM concepts [6]), but only few concepts refer to surveillance. Thus, we have collected and reported other concepts we are interesting on; most of them are defined at a very high abstraction level. Actually, a preliminary list of more than 100 surveillance concepts has been defined.

With reference to the taxonomy of Fig. 1, the video surveillance concepts can belong to three semantically different categories (*Physical Object, Action/Event, Context*). More precisely, the ViSOR ontology is structured in several classes, each of them belonging to one of the previously defined categories as reported in Table 2. A video annotation can be considered as a set of instances of these classes; for each instance a list of related concepts are assigned. Some of them directly describe the nature of the instance, i.e., they are connected to the entity with a "IS-A" relation (e.g., concepts like man, woman, baby, terrorist can be a sort of specialization of the "person" class and thus they can be use to describe instances of that class). Other concepts, instead, describe some characteristics or properties of the instance, in a "HAS-A" relation with it (e.g., the contour, the color, the position, the bounding box can be descriptive features of *FixedObject* instances).

Specialization relations are always static, i.e, they do not change during time; for example, a person can be a man or a woman, but reasonably it cannot switch between them during the video clip. Differently, some "HAS-A" relation can be dynamic; for example, the position and the color of the person can be different frame by frame. Thus, we have distinguished the "HAS-A" concepts in "static" and "dynamic" concepts. In Table 1 an excerpt of the ViSOR concept list related to the Person class only is reported. A complete list of the video surveillance concepts can be directly downloaded from the ViSOR portal.

"Is-a" Concepts			
Name	Definition	Type	Dynamic
Adult	Shots showing a person over the age of 18 (LSCOM #181)	True/False	-
Aggressor	(LSCOM #461)	True/False	-
Baby	images of babies (children that are too young to walk) (LSCOM #247)	True/False	-
Boy	One or more male children. (LSCOM #183)	True/False	-
Child	images of children (LSCOM #273)	True/False	-
Civilian_Person	One or more persons not in the armed services or police force. (LSCOM #105)	True/False	-
Female	(LSCOM #21)	True/False	-
Girl	One or more female children. (LSCOM #184)	True/False	-
Male	(LSCOM #17)	True/False	-
Person	Shots depicting a person. The face may be partially visible (LSCOM #217)	True/False	-
Police/security	(LSCOM #42)	True/False	-
"Has-a" Concepts			
Position_BBOX		rectangle	True
PositionBar	2D Position of the gravity center	point	True
Contour	Contour of the object	polygon	True
IDPerson	Application defined ID	integer	False
RealHeight	Real height of the person	float	False
PersonName	Name of the person	string	False

Table 1. An excerpt of the Person concept list.

3. ANNOTATION FORMAT

The native annotation format supported by ViSOR is Viper[7], developed at the University of Maryland. The choice of this annotation format has been made due to several requirements that Viper satisfies: it is flexible, the list of concepts is customizable; it is widespread avoiding the difficulties to share a new custom format (e.g., it is used by *Pets* and *Etiseo*); it is clear and easy to use, self containing since the description of the annotation data is included together with the data. An annotation tool has already been developed by the same authors of the standard (namely, ViPER-GT [8]). Finally, it is possible to achieve a frame level annotation that is more appropriate than the clip level annotation adopted by other tools.

The annotation data is stored as a set of records. Each record, called *descriptor*, annotates an associated range of frames with a set of attributes. To inform applications of the types of descriptors used to create the data file and the datatypes of the associated attributes, users must provide configuration information at the beginning of file. To this aim, Viper files are basically composed by two sections; the first one is called *config part* and explicitly outlines all possible descriptors in the viper file. It defines each descriptor type by name and lists all attributes for each descriptor. From the ViSOR portal a predefined *config file* containing the video surveillance concept list described in the previous section can be obtained. The second section of each Viper file, namely *data part*, contains instances of the descriptors defined in the *config part*. For each instance, the frame span (i.e., range) of the descriptor visibility together with a list of attributes values are reported. For more details refer to the Viper manual [8] or directly to the ViSOR portal [3].

Class	Category
1. Person	PhysicalObject
2. BodyPart	PhysicalObject
3. GroupOfPeople	PhysicalObject
4. FixedObject	PhysicalObject
5. MobileObject	PhysicalObject
6. ActionByAPerson	Action/Event
7. ActionByPeople	Action/Event
8. ObjectEvent	Action/Event
9. GenericEvent	Action/Event
10. Video	Context
11. Clip	Context
12. Location	Context

Table 2. Set of surveillance classes

4. WEB INTERFACE

The ViSOR web interface has been designed in order to share the videos and the annotation contents. Some screen shots of the web interface are shown in Fig. 3. ViSOR supports multiple video formats, search by keywords, by video meta-data (e.g., author, creation date, ...), by camera information and parameters (e.g., camera type, motion, IR, omni-directional, calibration). Until now about 60 videos belonging to different scenarios, like indoor, outdoor, smoke detection, have been added to ViSOR but the number of video is growing day by day. A screenshot of the video representative of the indoor category is shown in Fig. 4.

Three modalities have been implemented to allow video access: video preview, based on a compressed stream, single screen shot (a representative frame of the entire video) or a summary view, in which clip level screen shots are reported. Currently ViSOR contains about 450 clips.

Fig. 3. Some screenshots of the ViSOR web interface

For each video a set of annotations can be shared and downloaded, both ground truth and automatic annotations. The web interface allows to download the entire annotation as well as a subset of the annotation fields, filtering by frame number, descriptor or single attribute. The annotation can be exported in the VIPER format; an MPEG7 format export module is under development.

4.1. User Group

Another important aspect for a research community is the information exchange and the opportunity to share opinions, requests, comments about the videos and the annotations, and so on. Thus, the web portal of ViSOR includes a forum in which one topic for each video, generic topics on video surveillance, and topics on ViSOR (e.g., call for videos) are already active. In addition, each registered user can create his own topics.

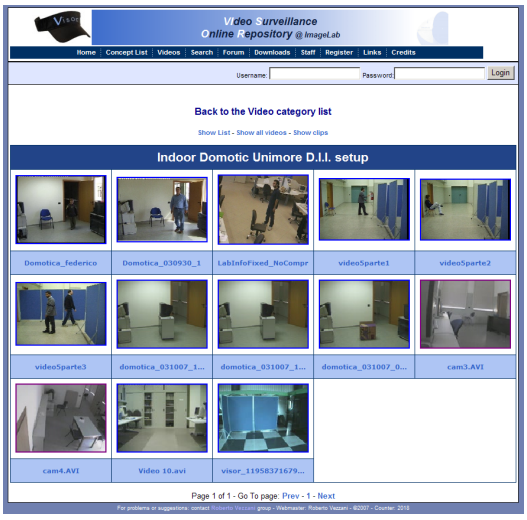

Fig. 4. Thumbnails of the ViSOR videos belonging to the Indoor category

5. CONCLUSION AND FUTURE WORK

ViSOR is a dynamic repository of annotated video sequences related to surveillance applications. A suitable ontology for surveillance domains has been defined in order to assure a better and easier interoperability among users.

This project (funded by VidiVideo EU project) is recently started and even if the interface and the database structure have been developed, the population of the database is just on an initial stage. Nonetheless, its interactive interface and the free available tool set are key points to become a reference repository of surveillance and security videos for many multimedia applications.

6. REFERENCES

[1] "Pets: Performance evaluation of tracking and surveillance," Website, 2000–2007, `http://www.cvg.cs.rdg.ac.uk/slides/pets.html`.

[2] *VSSN '06: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, New York, NY, USA, 2006. ACM, General Chair-Jake K. Aggarwal and General Chair-Rita Cucchiara and Program Chair-Andrea Prati.

[3] "Visor portal," Website, 2007, `http://imagelab.ing.unimore.it/visor`.

[4] C.G.M. Snoek, M. Worring, J.C. Van Gemert, J.M. Geusebroek, and A.W.M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proceedings of the 14th ACM Int'l Conference on Multimedia*, New York, NY, USA, 2006, pp. 421–430, ACM.

[5] M.R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, Smith J. R., P. Over, and A. Hauptmann, "A light scale concept ontology for multimedia understanding for trecvid 2005," Tech. Rep., IBM Research, 2005.

[6] L. Kennedy, "Revision of lscom event/activity annotations, dto challenge workshop on large scale concept ontology for multimedia," Tech. Rep., Columbia University ADVENT, 2006.

[7] D. Doermann and D. Mihalcik, "Tools and techniques for video performance evaluation," *Proc. of Int'l Conference on Pattern Recognition*, vol. 04, pp. 4167, 2000.

[8] "Viper toolkit at sourceforge," Website, 2005, `http://viper-toolkit.sourceforge.net/`.