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 2.0 - 15/01/2008

The VISOR system

(VIdeo Surveillance Online Repository)

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

1. Introduction

This technical report describes the work carried out until M12 by the UoM team in the VIDI-VIDEO project. The conducted activities are in the area of surveillance and include two main tasks in the WP7: the creation of a usergroup for the research community on videosurveillance and the sharing of a dataset for experimentation and evaluation.

TASK 7.7 VIDEO SURVEILLANCE *(from the VidiVideo Annex)*

Surveillance video collection

An important element of the task is to create a contact with user groups of surveillance enabling a higher impact of the results of VIDI-Video. Different sources of surveillance data video will be available, such as fixed indoor and outdoor cameras, mounted at high positions with a large field of view, moving cameras with pan, tilt and zoom capabilities, fixed indoor cameras, and mobile cameras, such as those mounted on board of cars of some private surveillance companies. At UoM tools and video analysis techniques have been developed, and could be used to further provide annotation for *a posteriori* logging. It could also be used as a searching system for activities detection in case the VIDI-Video system is too general in its capabilities.

The task will perform the following activities:

- Providing a large collection of security and surveillance videos, in order to create a complete set of views of a significantly wide area, covering a 24 hours time frame, with different, also non-overlapping, views. Videos will be provided about outdoor and indoor scenes, such as roads, public parks, offices and university campus. This allows potential queries such as *find me all sequences that contain a person pushing a stretcher from 6.00am to 6.30am* or *give me all the clips of video acquired in this area containing a person with a read coat.*
- Metadata annotation in MPEG-7 to ensure interoperability with Task 6.2 and 4.1, allowing us to provide additional features and metadata to the query engine.
- Testing the capability of the concept detection techniques developed in the project, by means of a sub-set of thesaurus such as people, face, car, bicycles, all providing insight in a surveillance setting. Videos will be provided about outdoor and indoor scenes, such as roads, public parks, offices and university campus.
- Compare the results obtained with the general-purpose features extractors and invariants, as defined in the Tasks 4.1, 4.2 and 4.3 with specific surveillance techniques that take into account additional information such as camera calibration data.

Surveillance User group

This as preparation to the aim of development of a forum for the surveillance community, able to attract user groups which will have the ability of providing new requests and use cases, sharing knowledge and annotated video and testing different approaches to the video surveillance application field.

To this aim, the Unimore's Video Surveillance Online Repositoy (named ViSOR) [1] has been exploited and enriched with new video sequences. The dataset has been collected by UoM and annotated by UoM in collaboration with IIT. A unified list of surveillance concepts defined in collaboration between UoM , UVA and and UniFi of specific concepts to learn has been defined and described in the previous report [3].

In this report, instead, the Unimore Video Surveillance Online Repositoy (named ViSOR) exploited to share the surveillance footages is described.

The ViSOR interface has been presented in a national summit on Public Video Surveillance named VideoGov in September 2007 and in an International Symphosium organized by the British Machine Vision Association [2].

2. Motivation

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 [4] or the VSSN workshops of the ACM Multimedia Conference [5]. 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. For instance the PETS 2007 benchmark contains 8 video dataset, each one with 4 different views. The I-Lids dataset contains 14 sequences for stopped vehicles and abandoned luggages.

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.

In VidiVideo we exploit the ViSOR environment [1], a video surveillance online repository for annotation retrieval[1]. First aim of ViSOR is to gather and make freely available surveillance video footages for the research community on pattern recognition and multimedia retrieval. Using ViSOR, we cannot only collect videos for annotation and retrieval but also 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.

3. 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

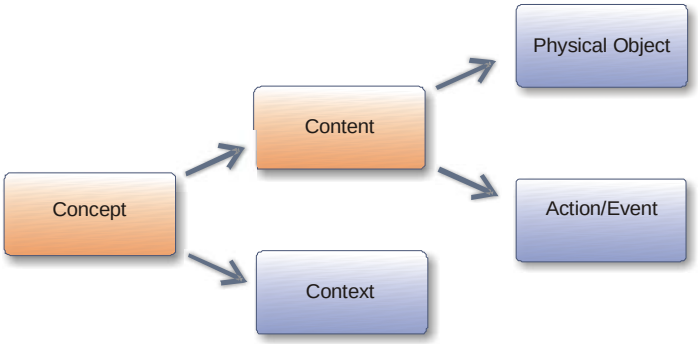

Fig. 1: Hierarchical taxonomy of the video concept categories

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.

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). Detailing these three catecories, a list of surveillance classes has been defined and reported in Table 1.

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

Class	Category
Person	*PhysicalObject*
BodyPart	*PhysicalObject*
GroupOfPeople	*PhysicalObject*
FixedObject	*PhysicalObject*
MobileObject	*PhysicalObject*
ActionByAPerson	*Action/Event*
ActionByPeople	*Action/Event*
ObjectEvent	*Action/Event*
GenericEvent	*Action/Event*
Video	*Context*
Clip	*Context*
Location	*Context*

Table 1: Set of surveillance classes

"Is-a" concepts			
Name	**Definition**	**Type**	**Dynamic**
Building	Shots of an exterior of abuilding (IDLSCOMM 226)	*T/F*	*False*
House	A freestanding single familyhome. (IDLSCOMM 164)	*T/F*	*False*
Office	Office environment with desks, chairsand/or white-collar workers. (IDLSCOMM 85)	*T/F*	*False*
Windows	An opening in the wall or roof of abuilding or vehicle fitted with glass or other transparentmaterial. (IDLSCOMM 109)	*T/F*	*False*
…			
"Has-a" concepts			
Name	**Definition**	**Type**	**Dynamic**
Position_BBOX		*bbox*	*True*
PositionBar	2D Position of the gravity center	*point*	*True*
Contour	Contour of the object	*polygon*	*True*

Table 2: An excerpt of the *FixedObject* concept list.

instances of that class). Other concepts, instead, describe some characteristics or properties of the instance, in a "HASA" relation with it (e.g., the contour, the color, the position, the bounding box can be descriptive features of FixedObject instances). An excerpt of the ViSOR concept list related to the FixedObject class is reported in Table 2, but the complete list of the video surveillance concepts can be directly downloaded from the ViSOR portal.

4. Annotation format

The native annotation format supported by ViSOR is Viper [10], 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 [10]). 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

Data Type	Description and Extended Configuration
bbox	A bounding box; it is a rectangle on the image.
bvalue	A Boolean value: either “true” or “false”.
circle	A circle, in terms of center point and radius.
ellipse	An ellipse, in terms of its bounding box.
fvalue	A floating point number.
lvalue	An enumeration type. In the config part the list of allowed values must be defined.
obox	An oriented bounding box.
point	Some specific pixel in the image.
polygon	A polygon or polyline, given as a list of points.
relation	A set of object identification numbers to a certain type of descriptor.
svalue	A string value. Remember that strings must be xml-escaped.

Table 3: Viper Data types.

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 [10] or directly to the ViSOR portal [1].

4.1. Attributes data types

As beforementioned, each descriptor can be described by a set of attributes. The allowed datatypes for these attributes have been obtained from the Viper data types (See Table 3); even if the ViSOR structure is flexible enough to store user defined data types, we encourage to exploit the default types only, for example formatting user data type into a string field.

5. Web Interface

The ViSOR web interface 1 has been designed in order to share the videos and the annotation contents. Some screen shots of the web interface are shown in Fig. 2.

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.

Fig. 2: Some screenshots of the ViSOR web interface.

6. User Guide

The Visor web interface has been developed in order to share the videos and the annotation contents. A screenshot of the web interface is shown in Fig.3, together with the name convention adopted for the different regions of the page.

Fig. 3: Screenshots of the VISOR web interface

In this section the main functionality of the web interface are presented.

6.1. Video browse

The video browsing is available selecting the *Video* menu item. Videos are divided into a set of categories, as shown in Fig. 4.a. After selecting a category, a list of videos belonging to it is reported (Fig. 4.b).

As default, the categories and the video are shown using the thumbnails modality. Selecting show list from the menu on the top of the client area the same information are reported in a list style (Fig. 5.a and 5.d). The complete list of videos without selecting a particular category can be obtain selecting *All videos* as in Fig. 5.b. From the same menu it is possible to switch to a clip level details (Fig. 5.c).

a b

Fig. 4: Video browse interface. a) video categories. b) video thumbnails of the Indoor Domotic Unimore Category

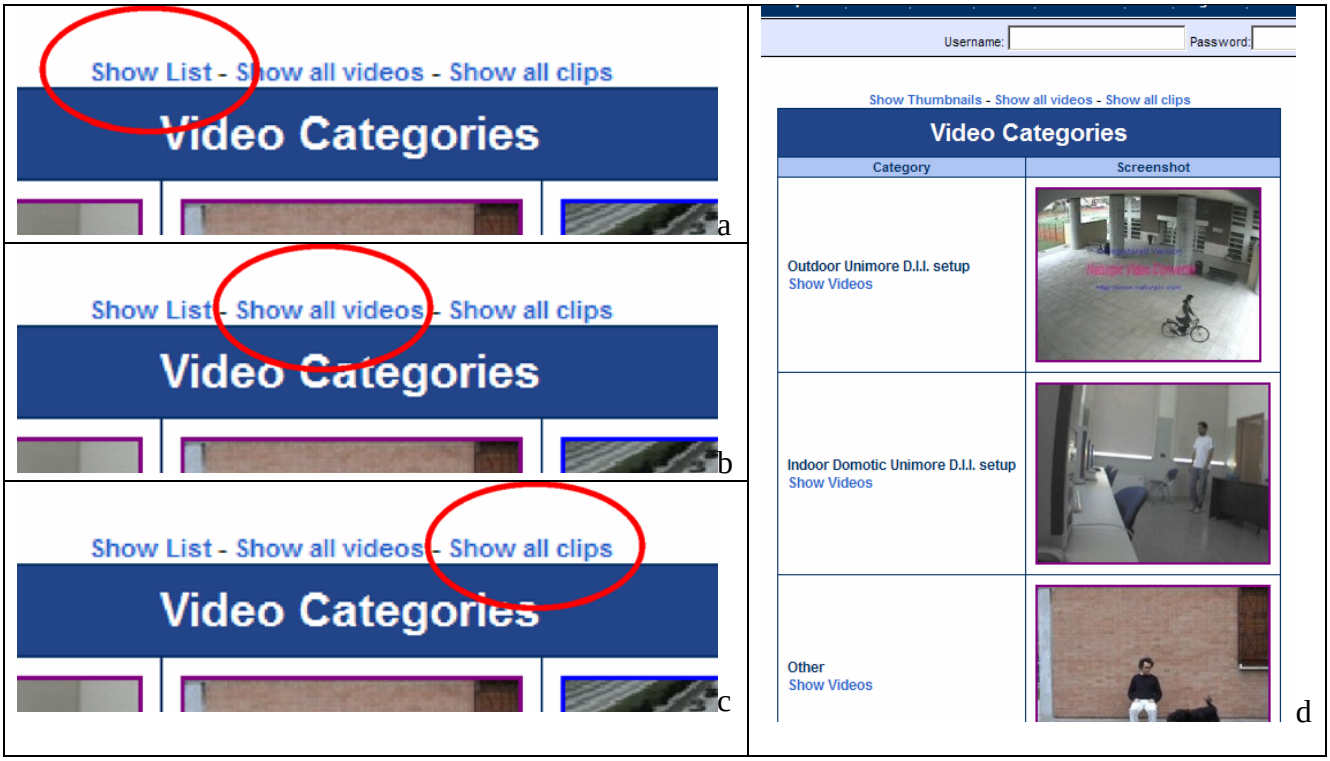

Fig. 5: Video browse interface. View selection

6.2. Video details

Three modalities have been implemented to allow video access in different ways: video preview, based on a flash compressed stream, single screenshot (a representative frame of the entire video) or a summary view, in which clip level screenshots are reported.

The video details are reported in a list as shown in Fig. 6.

Fig. 6: Video details: single screenshot view, video information, camera information annotation download

6.3. Annotation export

For each video, a set of annotations are provided, both ground truth and automatic annotations. Directly from the video details page, the interface allows to download the entire annotation as well as a subset of the annotation fields, filtering by frame number, descriptor or single attribute (See Fig. 7).

Fig. 7: Selective annotation download

6.4. Video Search

Instead of browsing the set of videos as reported in the previous section, the interface allows query of videos basing on the annotated concepts or assigned keywords (Fig. 8).

In the first case, given a concept the system will report all the video associated with an annotation containing that concept.

Fig. 8: Video Search

6.5. *Forum*

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 online 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.

The forum is Powered By: Snitz Forums 2000.

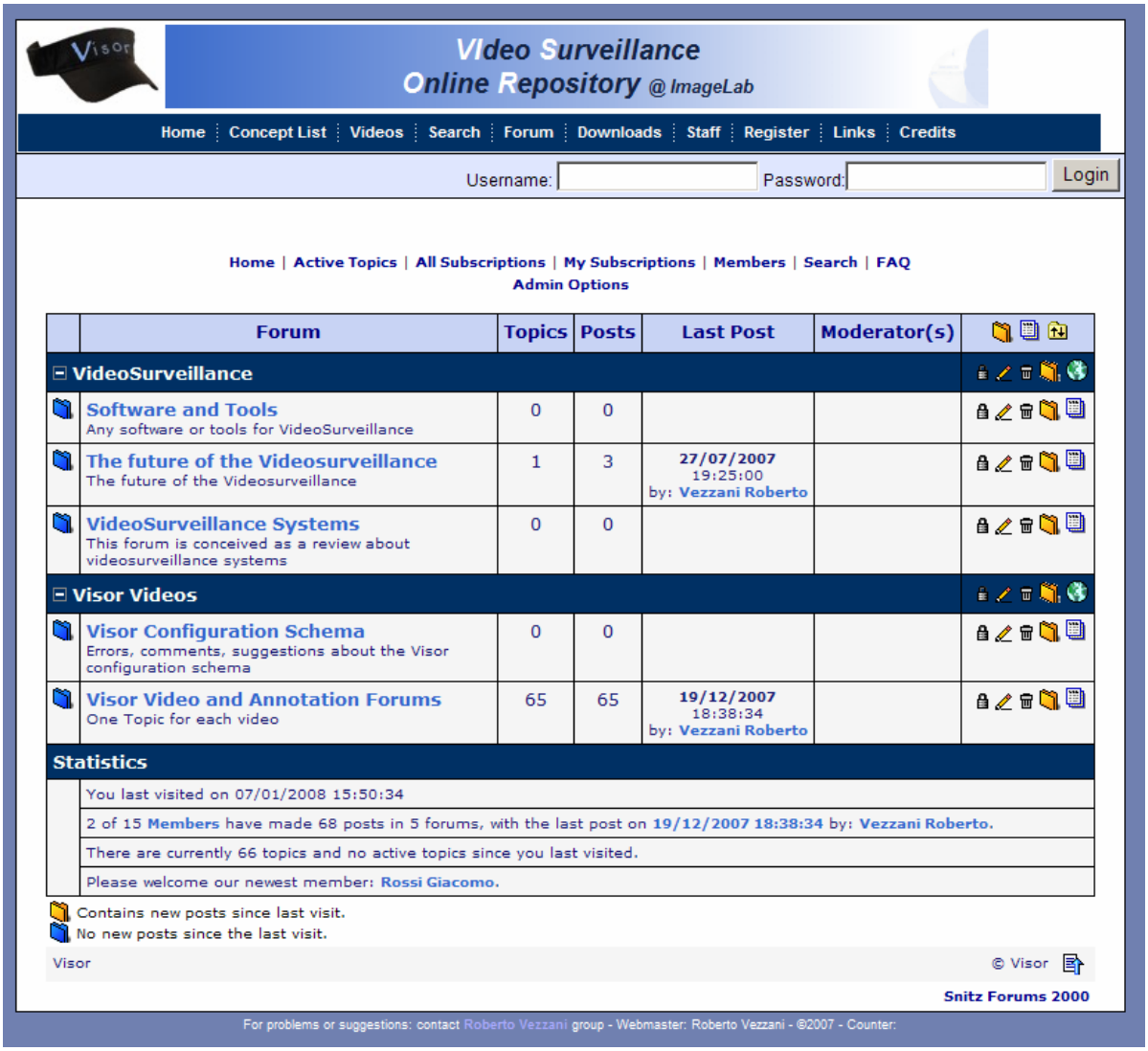

Fig. 9: Forum screenshot

6.6. User registration

Due to privacy and copyrights issues, the web access to some resources has been protected by a user level login. To this aim, a registration form is available from the *Register menu* (See Fig. 10) . The registrations are sent to the visor administrator that can allows user to login into the system. An email with the temporary password will be sent to the granted users.

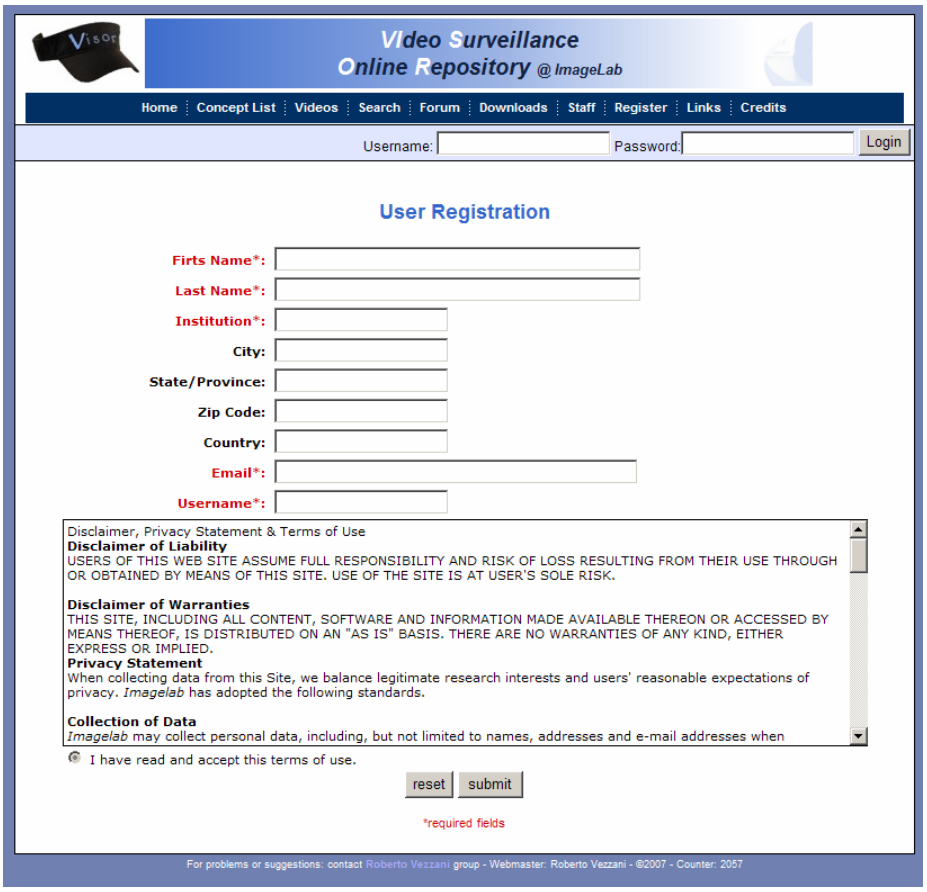

Fig. 10: User registration

7. System features

7.1. Automatic video management

When a video is uploaded into the ViSOR system, the Moderator can change the provided description and attributes and then he can approve the media.

Thus, the file is automatically processed in order to generate:

- a compressed flash version for video preview
- an MPEG1 version principally used for VIPER-GT annotations
- a JPEG screenshot (from the first frame)
- a clip segementation (fixing the number of clips or the number of frames per clip) and a clip screenshot (first frame of the clip)

All these tasks are performed exploiting the command line version of FFMPEG [11]. Thus, the video formats supported by the ViSOR system are the same of the FFMPEG library (a list is reported in Table 5).

7.2. ViSOR features and requirements

ViSOR features and requirements	
Webserver	IIS 5
Database	Microsoft Access 2003
Server Side Script Language	ASP 1.0 - VB script
Client Side Script Language	Java
Additional Server side CGIs	Written in C++ and compiled under Microsoft Windows.
FFMPEG	FFmpeg version SVN-r10461, Copyright (c) 2000-2007 Fabrice Bellard, et al. libavutil version: 49.5.0 libavcodec version: 51.43.0 libavformat version: 51.12.2 built on Sep 11 2007 01:20:02, gcc: 4.2.1-sjlj (mingw32-2)
Forum	Powered by Snitz Forums 2000 Version 3.4.06

Table 4: ViSOR features and requirements

Supported Codec	Encoding	Decoding	Comments
MPEG-1 video	X	X	
MPEG-2 video	X	X	
MPEG-4	X	X	
MSMPEG4 V1	X	X	
MSMPEG4 V2	X	X	
MSMPEG4 V3	X	X	
WMV7	X	X	
WMV8	X	X	not completely working
WMV9		X	not completely working
VC1		X	
H.261	X	X	
H.263(+)	X	X	also known as RealVideo 1.0
H.264		X	
RealVideo 1.0	X	X	
RealVideo 2.0	X	X	
MJPEG	X	X	
lossless MJPEG	X	X	
JPEG-LS	X	X	fourcc: MJLS, lossless and near-lossless is supported
Apple MJPEG-B		X	
Sunplus MJPEG		X	fourcc: SP5X
DV	X	X	
HuffYUV	X	X	
FFmpeg Video 1	X	X	experimental lossless codec (fourcc: FFV1)
FFmpeg Snow	X	X	experimental wavelet codec (fourcc: SNOW)
Asus v1	X	X	fourcc: ASV1
Asus v2	X	X	fourcc: ASV2
Creative YUV		X	fourcc: CYUV
Sorenson Video 1	X	X	fourcc: SVQ1
Sorenson Video 3		X	fourcc: SVQ3
On2 VP3		X	still experimental
On2 VP5		X	fourcc: VP50
On2 VP6		X	fourcc: VP60,VP61,VP62
Theora	X	X	still experimental
Intel Indeo 3		X	
FLV	X	X	Sorenson H.263 used in Flash
Flash Screen Video	X	X	fourcc: FSV1
ATI VCR1		X	fourcc: VCR1
ATI VCR2		X	fourcc: VCR2
Cirrus Logic AccuPak		X	fourcc: CLJR

4X Video		X	Used in certain computer games.
Sony Playstation MDEC		X	
Id RoQ	X	X	Used in Quake III, Jedi Knight 2, other computer games.
Xan/WC3		X	Used in Wing Commander III .MVE files.
Interplay Video		X	Used in Interplay .MVE files.
Apple Animation	X	X	fourcc: 'rle '
Apple Graphics		X	fourcc: 'smc '
Apple Video		X	fourcc: rpza
Apple QuickDraw		X	fourcc: qdrw
Cinepak		X	
Microsoft RLE		X	
Microsoft Video-1		X	
Westwood VQA		X	
Id Cinematic Video		X	Used in Quake II.
Planar RGB		X	fourcc: 8BPS
FLIC video		X	
Duck TrueMotion v1		X	fourcc: DUCK
Duck TrueMotion v2		X	fourcc: TM20
VMD Video		X	Used in Sierra VMD files.
MSZH		X	Part of LCL
ZLIB	X	X	Part of LCL, encoder experimental
TechSmith Camtasia		X	fourcc: TSCC
IBM Ultimotion		X	fourcc: ULTI
Miro VideoXL		X	fourcc: VIXL
QPEG		X	fourccs: QPEG, Q1.0, Q1.1
LOCO		X	
Winnov WNV1		X	
Autodesk Animator Studio Codec		X	fourcc: AASC
Fraps FPS1		X	
CamStudio		X	fourcc: CSCD
American Laser Games Video		X	Used in games like Mad Dog McCree
ZMBV	X	X	Encoder works only on PAL8
AVS Video		X	Video encoding used by the Creature Shock game.
Smacker Video		X	Video encoding used in Smacker.
RTjpeg		X	Video encoding used in NuppelVideo files.
KMVC		X	Codec used in Worms games.
VMware Video		X	Codec used in videos captured by VMware.

Cin Video		X	Codec used in Delphine Software games.
Tiertex Seq Video		X	Codec used in DOS CDROM FlashBack game.
DXA Video		X	Codec originally used in Feeble Files game.
AVID DNxHD	X	X	aka SMPTE VC3
C93 Video		X	Codec used in Cyberia game.
THP		X	Used on the Nintendo GameCube.
Bethsoft VID		X	Used in some games from Bethesda Softworks.
Renderware TXD		X	Texture dictionaries used by the Renderware Engine.
AMV		X	Used in Chinese MP3 players.

Table 5: Supported video formats

A table reporting some ViSOR statistics is depiceted in Fig. 11. They refer to the period of nov-dec 2007. The number of video and contact is growing.

Users	
Registered	23
Staff	8
Videos	
Sequences	65
Annotated	28
Clips	430
Concepts	
Total Videosurveillance IS-A concepts	96
IS-A concepts used	60
HAS-A concepts	36
Counters	
Web Accesses	8092
Video Downloads	629

Fig. 11: system statistics at 31/12/2007

References

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

[2] R. Vezzani, R. Cucchiara, "Visor: Video Surveillance Online Repository", Proceedings of BMVA symposium on "Security and surveillance: performance evaluation", London, 2007

[3] R. Vezzani, R. Cucchiara, "Video surveillance concepts and the VISOR system (VIdeo Surveillance Online Repository)", Tech Rep VV1.0, 20 Apr 2007.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] "FFMPEG at sourceforge," Website, 2007, http://ffmpeg.sourceforge.net/index.php