

D4.1 Deliverable

Description of the State-of-the-Art CANTATA



Project number: ITEA05010
Document version no.: 1.0
Status: final
Edited by: François Bremond, *INRIA*
Monday, 12 March 2007

ITEA Roadmap domains:

Major: Services & Software creation
Minor: Cyber Enterprise

ITEA Roadmap technology categories:

Major: Content
Minor: Data and content management

History

Document version #	Date	Remarks
v0.10	13 November 2006	Initial document start by Eefje Vandamme, Barco
V0.11	19 December 2006	Draft based on CANDELA by Eefje Vandamme, Barco Input from Xavier Desurmont, ACIC
V0.12	20 December 2006	Cedric Marchessoux, Barco, organization of the draft inclusion (outlines) and update of the draft content based on Candela publication + 1 st draft for section 3.5
V0.13	5 January 2007	State of the art on surveillance and monitoring application domain based on [72]
V1.0	25 May 2007	Document was accepted by the PMT

Contributors:

François Bremond, *INRIA*
 Nghiem Anh-Tuan, *INRIA*
 Xavier Desurmont, *ACIC*
 Cédric Marchessoux, *Barco*
 Raoul Djeutane, *Codasystem*
 Patrick Blandin, *CRP Henri Tudor*
 Brecht Vermeulen, *IBBT (IBCN)*
 Dimitrios Makris, *Kingston University*
 Neda Lazarevic-Mcmanus, *Kingston University*
 Caroline Machy, *Multitel*
 Pedro Fonseca, *Philips Research*
 Roel Tryen, *Philips Medical Systems*
 Jorma Palo, *Solid*
 Wouter Favoreel, *Traficon*
 Gunnar Holmberg, *UPF*
 Rick Koeleman, *VDG Security*

TABLE OF CONTENTS

1	Introduction	4
1.1	The Aim of the activity	4
1.2	Common state of art for the different domains of application	5
1.2.1	Requirements for validation	5
1.2.2	MCA annotation	5
1.2.3	MCA validation	6
	Pixel-level; image-frame segmentation for ICA and VCA.....	7
	Object-based evaluation per frame for VCA	8
	Object-based evaluation over an objects life time for VCA	9
	Evaluation of object features with higher semantic levels.....	10
	Metrics	10
1.2.4	Management and presentation.....	11
2	Consumer electronic applications (home and mobile multimedia) domain	12
2.1	MCA annotation.....	12
2.2	MCA validation	15
2.3	Management and presentation.....	16
3	Medical imaging applications domain	18
3.1	Evaluation of diagnostic accuracy	18
3.1.1	Aim and relevance	18
3.1.2	ROC analysis.....	18
3.1.3	User-involved study setup.....	19
3.1.4	Generalization to other domains	20
3.2	MEDical Display Simulation Chain (MEDISIC) for determining medical quality,	20
3.2.1	Context and goal	20
3.2.2	Display simulation chain	21
3.2.3	Human Visual Observer Model.....	21
3.2.4	Validation	22
4	Surveillance and monitoring applications domain	23
4.1	Qualitative evaluation of output results	23
4.2	Comparison of output results with ground truth	24
4.3	Alternative to ground truth	25
4.4	Existing Workshops and Projects.....	26
4.5	Conclusion	30
	REFERENCES	30

1 Introduction

1.1 The Aim of the activity

The activity preparing the production of the Deliverable 4.1 encompasses all the topics addressed in the WP4:

- Topic 4.1: Requirements for validation
- Topic 4.2: MCA annotation
- Topic 4.3: MCA validation
- Topic 4.4: Management and presentation
- Topic 4.5: Display chain simulation for determining medical quality

Except topic 4.5, the topics of WP4 are represented in a flowchart on figure 1. They correspond to the different elements of a complete MCA validation chain. Database of images or video is required as input. The images/videos from the database need to be annotated for building the ground truth (GT). MCA algorithms are applied and should be evaluated using the GT. The last stage is an important one corresponding to the presentation of the results.

The activity should map with the different domains that are targeted within the CANTATA project:

- Consumer electronics
- Medical imaging
- Surveillance and monitoring

Deliverable D4.1 aims to establish a complete state-of-the-art of MCA annotation, validation and presentation techniques. MCA meant Multi-Content-Analysis regrouping Image and video content analysis. Therefore, MCA could be developed for static or dynamic applications. For instance, some Video Content Analysis (VCA) validation techniques based on frame analysis or Image Content Analysis (ICA) validation techniques can be applied on videos and on images. Nevertheless, not all VCA validation techniques can be used for ICA.

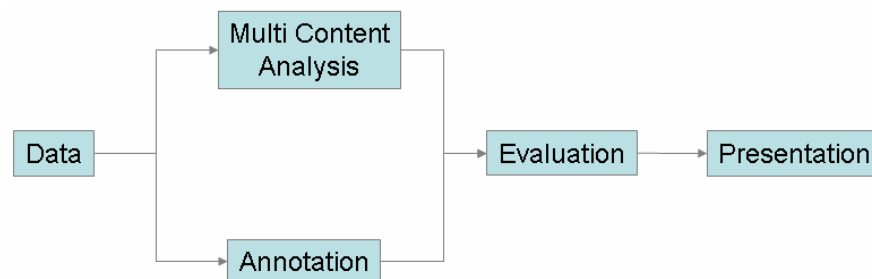


Figure 1: Validation chain.

A first a state of art of common MCA annotation, validation and presentation techniques and the validation requirements techniques is presented in the next section. Then, a state of the art for each application domain is presented in separated sections:

- Consumer electronics
- Medical imaging
- Surveillance and monitoring

The list is not exhaustive; techniques described or cited in the next section could be applied to several domains. The next section is inspired from a publication in the context of Candela project [1].

1.2 Common state of art for the different domains of application

1.2.1 Requirements for validation

Since the last decade, many algorithms have been proposed that try to solve the problem of scene understanding (content analysis). The level of understanding varies highly from only detecting moving objects and outputting their bounding boxes (e.g. the OpenSource project “Motion”¹), to tracking of the objects over multiple cameras, thereby learning common paths and appearance points [30, 32] or depth maps and amount of activity in the scene [31].

Apart from functional testing, there are several other reasons for evaluating the MCA systems; scientific interest, measuring the improvement during development, benchmarking with competitors, commercial purposes and finally legal/regulatory requirements. However, most literature describing MCA algorithms, cannot give objective measures on the quality of the results. For example, for video compression algorithms the criterion is to minimize the absolute difference between the decoded result and the original with the PSNR as standard metric. However, for MCA algorithms no standard with criteria exists. Some evaluation metrics have already been proposed in the literature, but most cover only limited part of a complete MCA system. For the validation, as shown on Figure 1, a database is required and in addition annotation methods to prepare the Ground truth.

1.2.2 MCA annotation

GT needs to be made available, describing the true properties of the images sequence. Because the level of accuracy of the GT is required to be very high, the process of creation can be quite time consuming. Several tools for annotating GT descriptions of images and video scenes have been made available. These annotation tools are sometimes referred to as ‘tagging tools’. Some of the available annotation tools are listed below.

The “Open Development environment for evaluation of Video Systems” (ODViS) is a framework that can be used to simplify the users annotation task. It allows the embedding of tracking algorithms, to create a ‘noisy’ GT description, depending on the quality of the tracking algorithm used. Users then only need to manually adjust this first GT description. Next to the annotation task, also an evaluation has been included. Jaynes *et al.* [26] explain that researchers can easily define GT data, visualize the behaviour of their surveillance system, and automatically measure and report errors in a number of different formats.

Another project, also enabling both annotation and evaluation of MCA algorithms, is the “Video Performance Evaluation Resource” (ViPER) [19, 21]. Results of evaluation can be visualized.

Collings, Zhou and The [25] propose an open source tracking test bed and evaluation website. They designed an annotation tool to be used with Matlab.

The CAVIAR project provides an annotation tool, written in Java. The source code is available from their website².

¹ OpenSource project Motion: <http://sourceforge.net/projects/motion/>
² Project IST CAVIAR (IST-2001-37540), EC Funded:
<http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>

Other proposals in literature often mention the use of graphical annotation tools, without much detail. *Nascimento and Marques* [10] describe an annotation tool that provides a tentative segmentation that needs to be adjusted by the user to avoid full manual annotation.

Above mentioned tools have not been evaluated by the authors in much detail, so no objective comparison can be provided. However, most tools use different formats to store the annotated GT metadata, causing limitations for the reuse of the GT.

The mentioned annotation tools write the GT descriptions to file. The various tools use different formatting of the metadata. Most tools use a proprietary XML description. Details on the format of the CAVIAR tool are explained in [32]32. Although a standard format is not mandatory for evaluation and benchmarking, it would be convenient. Therefore, the CANDELA project uses a limited subset of the MPEG-7 standard. Most important is that the same features are stored. Bounding boxes from proprietary XML descriptions can be converted to MPEG-7 descriptions with simple tools, as long as both definitions are known. These tools can even be included in the performance evaluation system.

For comparison purposes, the definition of these features described in the GT is very important. If the interpretation of a feature in the GT data is different from the interpretation in the MCA algorithm, evaluation is not feasible. For example, consider evaluation of object tracking using the location of a single point that describes the location of the object over time. What is the exact definition of that point describing the location? Is it the centre of the objects bounding box, the middle of the bottom line-part of the bounding box, or the median of the positions of all foreground pixels in the object?

No standards have yet been defined on what should be stored in a GT description. Because most MCA algorithms are evaluated for their segmentation or tracking performance, only segmentation masks or bounding boxes are stored. However, to evaluate higher level descriptions of the scene, more data will have to be supplied by the user during the annotation process (e.g. the real-world object height in meters). The MPEG-7 standard defines how bounding boxes and higher-level descriptions can be defined, but the total set of descriptors in the standard is too extensive for evaluating most MCA algorithms.

Another problem is the occlusion of objects. *Black, Ellis and Rosin* [29] already mention that this is a difficult issue, since the person that annotates the image or the video has to decide upon the desired behaviour of a MCA algorithm is. Should the algorithm keep tracking a (partially) occluded object?

1.2.3 MCA validation

The importance of performance evaluation of MCA algorithms has been addressed by various projects, and has led to various research publications. In 2000, the IEEE holds a yearly workshop on Performance Evaluation of Tracking and Surveillance (PETS). Discussed is the evaluation of the tracking performance of algorithms. However, these performances are mostly referring to quality of the result and not to the computational performance that is very relevant for real-time systems, i.e. hardware and software constraints like computation time and memory requirements. This indicates that algorithm development is still in its infancy, i.e. optimized implementations for industrial applicability are just being considered. Note that real-time constraints have significant impact on the chosen algorithm, the performance and timing properties such as latency and throughput. Due to the complexity of vision algorithms, the needed resources usually depend on the video content (e.g. very dynamic scenes compared to static low-motion video typically result in higher processing requirements). Other issues indirectly related to content analysis are database retrieval, network transmission and video coding. These could be evaluated with well known standard metrics like PSNR for image quality, bandwidth and maximum delay for network.

However, for the remainder of this section, we will only consider evaluation methods that describe the functional performance of MCA algorithms.

Since a complete MCA system comprises multiple semantic levels, evaluation can be done on certain levels. MCA algorithms that only perform segmentation of moving objects, and no tracking over multiple video frames, can only be evaluated to their pixel-based segmentation. Algorithms that only output alarms (e.g. car detected) can only be evaluated for proper classification, and not for their segmentation quality. Therefore, first needs to be defined what exactly should be evaluated for each semantic level. For each level, different metrics are required.

Multiple semantic levels for evaluation can be defined.

- Pixel-level; image-frame segmentation for ICA and VCA
- Object-based evaluation per frame for VCA
- Object-based evaluation over an objects life time for VCA
- Object-features like object type, speed, size (in meters)

Most algorithms operate in a bottom-up fashion. Segmentation of the video image into static background and moving foreground pixels is usually the first step. Since the performance of following steps in the MCA chain are depending on this first segmentation step, most of the proposed evaluation metrics in literature at this level, are pixel based and only consider the segmentation results. In more recent work, object-based evaluation metrics are proposed. However, there are quite some issues in this object-based evaluation. On a frame-basis, one can compare the overlap of the bounding box of the detected objects with the boxes in the GT. However, also the tracking of objects over time should be considered. Splitting of objects in two objects should also be taken into account. Recent work shows good improvements in this area [10]. We will now discuss some proposals from literature, categorized by the above-discussed semantic levels.

Pixel-level; image-frame segmentation for ICA and VCA

Zhang [16] lists multiple evaluation methods for image segmentation. Segmentation can be evaluated using analytical or empirical methods. The analytical method considers the principles, requirements and complexity of algorithms. In the latter, test video sequences are used to measure the quality of the segmentation results. Some pixel-based metrics are defined to evaluate segmentation algorithms.

Correia and Pereira [7] propose metrics for evaluation of image segmentation methods with the goal to create objective measures corresponding to evaluation by a human observer. Results are shown using the MPEG-4 video test sequences. The authors conclude that evaluation of video object segmentation is a problem, for which no satisfying solution is yet available in literature.

Erdem et al. [12] present three performance evaluation metrics that do not require segmented GT. They propose spatial differences of colour and motion and the boundary of the segmented video image and the temporal difference between the colour histogram of the object in the current frame and previous video frames. The authors show that under certain assumptions, the time-consuming annotation of GT is not necessary. However, when more than segmentation only is required, GT will have to be generated anyway.

Prati et al. [11] evaluate multiple shadow segmentation algorithms, using modified detection rate and false alarm rate metrics, called the shadow detection rate and the shadow discrimination rate. A comparison on multiple algorithms is given on a pixel-level, doing frame-by-frame comparison. However, the final result of shadow on the object-level is not considered.

Rosin and Ioannidis [13] present an evaluation of eight different threshold algorithms for change detection in a surveillance environment. Pixel-based evaluation is applied, but the authors conclude that this can sometimes give misleading rankings.

Renno et al. [9] evaluate four different shadow suppression algorithms, using video from a nightly soccer match with quite some shadow because of the lighting used. The used evaluation metrics are all based on the number of correctly detected pixels on a frame-basis. The four metrics used are the detection rate, the false positive rate, the signal-to-noise ratio and the tracking error. Finally, using an average of all values over time, all four algorithms are compared. However, this paper only focuses on the segmentation phase and other aspects like splitting and merging of multiple objects is not considered.

Oberti et al. [17] propose the use of Receiver Operating Characteristics (ROC) curves. They present pixel-based metrics for evaluation and show that the obtained ROC curves can be used to extract useful information about the system performance, when changing external parameters that describe the conditions of the scene (e.g. the number of objects in the scene). ROC curves can be used to find the optimal working point for a set of parameters. This work is extended in [18] and [20]. *Gao et al.* [23] also use ROC curves to display the performance of multiple segmentation algorithms. These curves contain the probability of a false alarm (FA) against the probability of a miss detect (MD).

Chalidabhongse et al. [22] propose Perturbation Detection Rate (PDR) analysis that has some advantage over ROC analysis. Four background subtraction algorithms are evaluated for their segmentation performance. No GT is needed for the evaluation method, but the method does not consider detection rates through the video frame or over time.

Object-based evaluation per frame for VCA

The above-discussed papers all discuss pixel-level evaluation. More recent proposals also discuss the object-based performance evaluation. In [19], *Mariano et al.* present seven metrics for evaluation of object detection algorithms by comparing properties of the objects' bounding boxes. The proposed metrics work on both pixel- and object-based comparisons between GT and results. However, the authors already mention that the proposed metrics need to be extended for algorithms that track objects over time and space.

Nascimento and Marques [10] propose new metrics that do cover the splitting into, and merging from, multiple objects. Several types of errors are considered: splits of foreground regions; merges of foreground regions; simultaneous split and merge of foreground regions; false alarms and detection failures. False alarms occur when false objects are detected. The detection failures are caused by objects that are in the GT and are not detected. The authors evaluated five different segmentation algorithms with the proposed metrics using the PETS2001 sequence. Region matching is applied using a corresponding matrix to match detected objects in the output with objects in the GT. Also the PETS 2004 sequences are evaluated using the metrics as defined in the CAVIAR project. The proposed metrics do consider splitting and merging, but only from a segmentation point of view. Tracking of objects over time is not considered, which is quite important if two objects approach towards each other and then move away again.

In recently published work, *Lazarevic-McManus et al.* [73] propose localised pixel-based and object-based metrics for the evaluation of object detection algorithms in the context of object tracking applications. Proposed metrics aim to address issues vital for effective object tracking such as: (i) the proportion of true objects located, (ii) the degree of local confusion caused by falsely detected objects, and (iii) the degree of fragmentation of true objects.

In [74], the same group of authors discusses a methodology for the evaluation of object detection algorithms in the context of end-user applications, as opposed to the evaluation of the stand-alone detection modules.

Since a good performance, in the context of a particular end-user application, inevitably requires a trade-off between two (or more) desirable metrics, a standardisation of application scenarios for the research community is proposed. (For example, algorithms which achieve high detection rates at the expense of high false positive rates are not adequate for end-user applications in which false alarms are too costly.) They illustrate utilisation of ROC and F-measure techniques for (i) system parameters selection and (ii) comparison of performances of multiple algorithms, within the trade-off constraints for a specified end-user application scenario. These recommendations can be extended to any problem involving evaluation of performance of a binary classifier with trade-off constraints.

Object-based evaluation over an objects life time for VCA

The previous proposals did not consider tracking of the objects over time. Needham and Boyle [6] present evaluation methods for positional tracking (object trajectories); how well can a tracker determine the position of the target object? They propose metrics for displacement between two trajectories, both in the spatial and the temporal domain and define a measure for the area between two trajectories. However, the authors only consider trajectories of equal length. In a system that is working with large variations on the input data (consider large, long shadows from upcoming sun or mirroring effects in rainy days), the length of the time-interval of the tracked objects might not be equal to the length of the interval in the GT.

Rossi and Bozzoli [15] presented a simple performance evaluation method for their tracking system, by comparing how many objects crossed a certain line. This method is very simple and requires very little effort for the creation of GT. However, the performance results give a limit insight in the actual performance of the tracking algorithm.

Xu and Ellis [24] present a tracking approach that can deal with partial occlusion and grouping. Two measures for the performance of the tracking system are proposed. The approach does not require GT and can not be used to compare multiple algorithms. The first measure is the tracking error between the actual and predicted (from previous video frames) position values. The next measure is the path coherence, which represents the level of agreement between the derived object trajectory and the motion smoothness constraints. These two metrics are proposed because they are the basis of most existing motion correspondence algorithms that usually assume the smoothness of motion.

Pingali and Segen [14] propose two methods to evaluate performance of object tracking algorithms. They present metrics for cardinality measures, durational accuracy measures and positional accuracy measures. The first method requires extensive availability of GT, while the second method is scalable in GT detail. Events are used to describe the location of objects at certain times. The more events annotated, the more accurate the GT. The authors used line segments to annotate these events for the reference video set. A tool with a graphical user interface is used to annotate the GT data.

Black, Ellis and Rosin [29] propose three metrics for comparing the trajectories of objects in the GT set, with the detected trajectories. The path coherence metric assumes that the trajectory should be smooth subject to direction and motion constraints. The colour coherence metric measures the average inter-frame histogram distance of a tracked object. This distance is assumed to be constant between consecutive image frames. Furthermore, the shape coherence metric gives an indication of the expected object bounding box, compared with the detected bounding box. Outlier GT tracks are removed by applying a threshold to the three error values.

Evaluation of object features with higher semantic levels.

We discussed proposals from the literature for the first three levels mentioned earlier. Literature study for the fourth and fifth level did not reveal a lot of relevant work. The fourth level seems rather straightforward at first, but the lack of standardized criteria hamper proper evaluation and benchmarking. Consider for example object classification. It seems trivial to compare the object type from the GT with the detected object type. However, for product realization, the response time might be of importance. How much of the object life time is required to output a confident classification type. No criteria have been proposed for this problem.

For the fifth level, which includes behaviour, it seems even simpler to compare the results from a MCA system with the GT. For example, the system should detect a parking car. The MCA system could decide that the object type is not car, but person. It might, however, detect that the object stops. Does this give 50% accuracy or 0% because it should be a car for sure? Hence, it is not clear how to present the performance results. Furthermore, it is not clear how time delays in detection of behaviour need to be handled.

Summarizing, most evaluation proposals are based on pixel-based segmentation metrics only ([7, [9, 11, 13, 16, 17, 20, 22,23]. *Oberti et al.* [18] also consider the object level evaluation, and Mariano *et al.* 19 consider object tracking over time. *Nascimento and Marques* [10] consider object-based evaluation and introduce splitting and merging of multiple objects. Evaluation of object trajectories is only proposed by the authors of [6, 14, 15].

Most proposals require manually annotated GT, while a limited set of proposals apply performance evaluation using metrics that do not require GT (12, 22, 24).

Metrics

Evaluation of multi content analysis can be applied on multiple semantic levels. If an algorithm only detects objects and outputs bounding boxes per processed video frame, we need different comparison rules than when an algorithm only outputs “car entered scene” and “car left scene”. The latter is more straightforward, since we only need to compare the time-interval of detection and the object-type that was detected. For each level of evaluation (see Subsection Related work) various metrics have been proposed in the literature.

There are mainly two types of metrics: binary and numeric metrics.

Binary metrics are coming from qualitative processes like detection or classification (e.g. non-detection, false alarm or misclassification). Every process leads to new decisions that should be evaluated (“is this pixel part of the background or the foreground ?”). A ground truth allows us to classify decisions as correct or not. Whenever it is correct it would be called “true” and when it is incorrect it would be called “false”. For the detection process we usually use such “binary metrics”: based on standard statistical methods of comparing two populations of values which are derived from observations with their true or expected values. In this case, the ground truth is used as the true values.

Numeric metrics are made by quantitative processes like an observation of estimation within a sequence of image. Typical errors affect the position, the shape of object, its speed or the delay of a time stamp.

To quantify these errors we use scores called “numeric scores” and are computed to quantify the accuracy of the detection or the tracking algorithms. Examples are average position and velocity errors; average number of observations before tracking is initiated; average number of frames before tracking is terminated etc.

Because various different metrics are used, the performance evaluation results from MCA algorithms currently cannot be compared. As already said, the evaluation will be done by comparing the tested output results with the GT according to relevant criterion. The criteria are precisely defined by metrics which produce an objective global score. Even if both the MCA and the GT create the same descriptions and the metrics are well defined, the matching is still problematic. For example, if an object, event or behaviour is detected, how do we know what the corresponding object, event or behaviour is that we should compare in the GT. Should the frame number in the sequence be the same? This gives problems if the detection is somewhat delayed. Should objects from the MCA and GT that are spatially closest together be compared? What if the MCA seems to detect a car with a certain speed, precisely as described in the GT, but in fact the detected object is a bicycle at another location in the scene? The evaluation tool may show a good MCA performance while its output was incorrect.

Somehow, we should define matching rules that can be used by the evaluation tool to match the descriptions from the MCA with descriptions in the GT. This implies that the evaluation results not necessarily represent the true performance. The output has certain reliability, dependent on the complexity of the matching rules. Note that human evaluation of e.g. “the colour of a passing car”, using a GT involves complex matching. Intuitively, the human brain matches the objects type, their trajectory, their sizes, shape, the time instance, and the relation with other events in the scene. Simultaneously, the human compares the colours. Defining the matching rules for performance evaluation is an open research topic to be explored.

1.2.4 Management and presentation

During evaluation, values are calculated for each metric, for a certain time interval. Combining these results over the total time-interval (of one video sequence or multiple sequences) can be applied in various ways.

Some authors show histograms of results over time, while others summarize measures over time. Other metrics require statistical analysis methods like mean or median. With low variance, the mean gives the typical working performance. One could also be interested in the maximum error rate in any of the tested cases to prove the limits of the system. From these results also other interesting measurements from the industry can be computed, like the mean time before failure (MTBF). However, different ways of presenting results from metrics make it impossible to compare various evaluation results. Therefore, besides defining standard metrics (as mentioned in the previous subsection), for each method, the way of presenting the results for a complete evaluation should be researched and standardized.

2 Consumer electronic applications (home and mobile multimedia) domain

2.1 MCA annotation

The need for good annotation

Developing technologies related to musical audio signal processing requires data. For instance, implementing algorithms for automatic instrument classification requires annotated samples of different instruments. Implementing a voice synthesis and transformation software calls for repositories of voice excerpts sung by professional singers. Testing a robust beat-tracking algorithm requires songs of different styles, instrumentation and tempo. Building models of musical content with a machine learning rationale calls for large amounts of data. Besides, running an algorithm on big amounts of diverse data is a requirement to ensure its quality and reliability.

Likewise, the development of good technologies for video signal processing requires large amounts of annotated video data. This is true both for the benchmarking of algorithms, necessary to assess the progress while developing new methods and techniques to address new applications, but also for the actual algorithm development, for an example to compile annotated training and test data sets needed to e.g. train face detectors, object recognizers or for any other unsupervised learning-based methods. Although much is common between the video and audio content analysis world, this section will focus more on MCA annotation for audio-related content and technology.

Annotation tools

Manual media annotation is a time-consuming and tedious process; many tools have been developed to make this task easier. We can cite as examples:

- MUCOSA (Music Content Semantic Annotator). MUCOSA is an environment for the annotation and generation of music metadata at different levels of abstraction. It is composed of three tiers: an annotation client that deals with microannotations (i.e. within-file annotations), a collection tagger, which deals with macro annotations (i.e. across files annotations), and a collaborative annotation subsystem, which manages large-scale annotation tasks that can be shared among different research centers. The annotation client is an enhanced version of WaveSurfer (<http://www.speech.kth.se/wavesurfer/>), a speech annotation tool. The collection tagger includes tools for automatic generation of unary descriptors, invention of new descriptors, and propagation of descriptors across sub-collections or playlists. Finally, the collaborative annotation subsystem, based on Plone (<http://plone.org/>), makes it possible to share the annotation chores and results between several research institutions.

Happiness							
Song	Set Class Class Index	Class	Tonal Descriptor: Strength	Meter	Danceability	BFM	
02 LOVE REALLY HURTS WITHOUT YOU	<input type="checkbox"/>	Low	0.646959	4	0.81174	122	
03 BLACK AS HE'S PAINTED	<input type="checkbox"/>	Medium	0.672659	4	0.81997	112	
04 WHOSE LITTLE GIRL ARE YOU?	<input type="checkbox"/>	Low	0.854263	4	0.88595	100	
01 ON THE RUN (HOLD ON BROTHER)	<input type="checkbox"/>	High	0.774791	4	0.77182	100	
(I BELIEVE IN) TRAVELLIN' LIGH	<input type="checkbox"/>	Medium	0.839752	4	0.94154	88	
(PORTRAIT OF) BOJANGLES	<input type="checkbox"/>	Low	0.858428	3	0.75373	78	
01 - PISTE 01	<input type="checkbox"/>	High	0.875541	4	1.0631	55	
05 - PISTE 05	<input type="checkbox"/>	?	0.752535	3	0.92376	105	
"HEROES"	<input type="checkbox"/>	High	0.809925	4	0.90365	52	
007	<input type="checkbox"/>	?	0.56232	4	0.83828	68	
(NOW AND THEN THERE'S) A FOOL	<input type="checkbox"/>	?	0.732852	4	0.78676	109	
(MARIE'S THE NAME) HIS LATEST	<input type="checkbox"/>	Medium	0.698993	4	0.8056	169	
(YOU'RE THE) DEVIL IN DISGUISE	<input type="checkbox"/>	High	0.764071	4	0.77728	68	
(BONUS TRACK) A LITTLE LESS CO	<input type="checkbox"/>	?	0.768793	4	0.8445	102	
(LET ME BE YOUR) TEDDY BEAR	<input type="checkbox"/>	Low	0.772971	4	0.92173	149	
...	<input type="checkbox"/>	Medium	0.482934	4	0	56	
'TAINT NOBODY'S BIZ-NESS IF I DO	<input type="checkbox"/>	High	0.826035	4	0.76768	81	
(THERE'S NO PLACE LIKE) HOME F	<input type="checkbox"/>	High	0.787399	4	0.88546	73	
01-GIDDY STRINGS.MP3	<input type="checkbox"/>	?	0.198542	4	0	88	
04-EASY ROD.MP3	<input type="checkbox"/>	?	0.819829	4	0.77328	121	

MUCOSA: A screenshot of the descriptor creator that is included in the collection tagger

- MiXa (MusicXML Annotator).** MiXa enables users to associate any element of musical content, such as a note, a lyric, and a title, with some additional information such as chords, comments, and impressions. MiXa can also handle annotations created by multiple users, and these annotations are managed separately for each annotator. This allows application systems for annotations to deal with various comments created by the different annotators. Since the content and form of annotation data depend on applications or services, the system allows application developers to define their own semantics of annotation data.

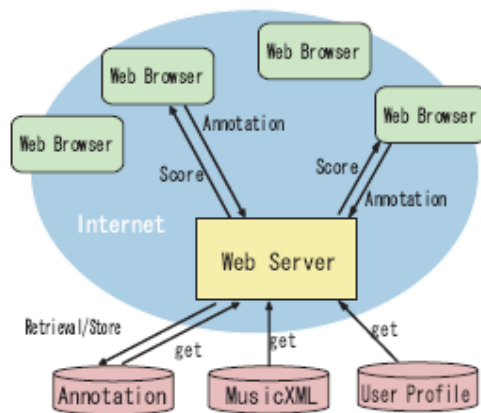
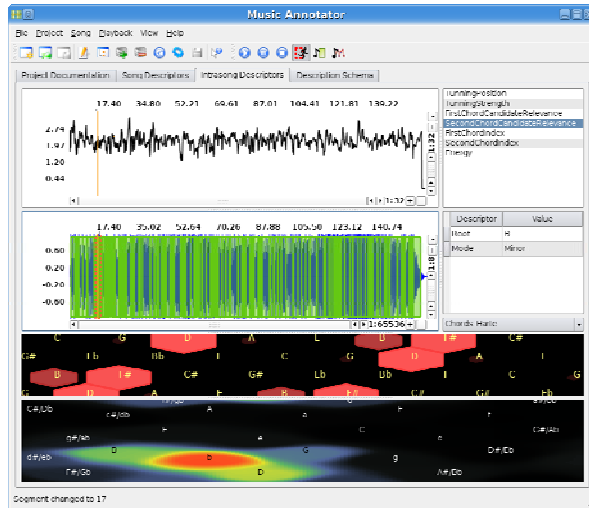


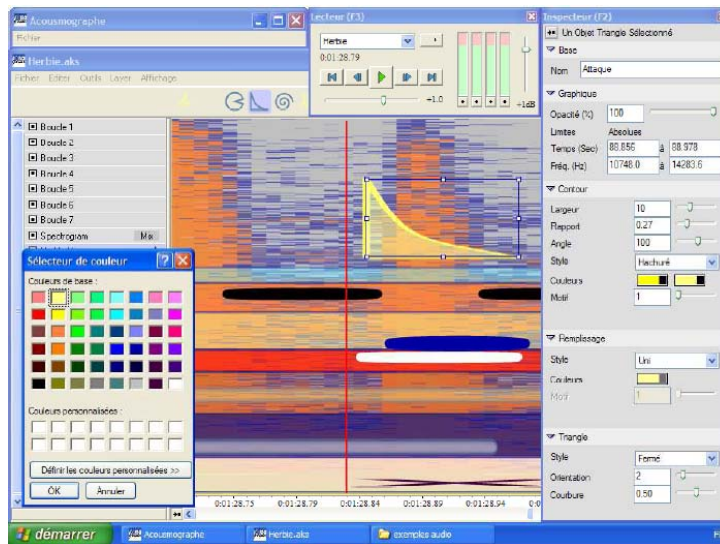
Figure 5: System architecture of MiXa

- CLAM Music Annotator** (http://iua-share.upf.edu/wikis/clam/index.php/Music_Annotator). This is a GPL tool that can be used to visualize, check and modify music information extracted from audio: low level features, note segmentation, chords, structure. The tool is intended to be useful for the music information retrieval research whenever one needs to:
 - Supervise and correct the results of automated audio feature extraction algorithms.
 - Generate manually edited annotations of audio as training examples or ground truth for those algorithms.



The CLAM Music Annotator

- **Acousmographie** (http://www.ina.fr/grm/outils_dev/index.fr.html). The Acousmographie is software for listening, analyzing and representing sound signals. The Musical Research Group began working on the project in 1991 in order to facilitate the locating, annotation and description of all forms of discourses on non-written music be they electro-acoustic or oral tradition. The third generation of the Acousmographie offers the capacity to analyze multi-tracks with multi-window demonstrations; in order to allow for almost immediate access to any part of the signal, no matter the length of the sound, the sonogram is calculated as a background task, etc..



MRG's Acousmographie

Annotation formats

Annotation tools are meant to be used in conjunction with validation tools as well as with database systems which store and help managing the metadata. Alongside generic data description and exchange formats such as XML and YAML (<http://www.yaml.org/>), there are data formats specifically designed for music and media applications:

- **MusicXML** (<http://www.recordare.com/xml.html>). Music XML, is a DTD designed primarily for a easing automated parsing and manipulations of musical scores. But as shown by the MiXa annotation tool, it also enables standardized musical annotation exchange and storage.
- **MXF** . MXF is media annotation format, it is a "container" or "wrapper" format that supports a number of different streams of coded "essence", encoded with any of a variety of codecs, together with a metadata wrapper which describes the material contained within the MXF file. MXF has been designed to address a number of problems with non-professional formats. MXF has full timecode and metadata support, and is intended as a platform-agnostic stable standard for future professional video and audio applications.

2.2 MCA validation

Validation metrics

In order to compare the calculated results to the annotated ground-truth, a standard metric is needed. Precision, recall and F-measures are standard measures used in Information Retrieval. For each query, the recall is defined as the fraction of the relevant documents which has been retrieved (as explained by Baeza 1999). Given a piece i ($i=1, \dots, N$), we define:

$$\text{Recall}_i = \frac{nFoundItems_i}{nItems_i} \quad (1)$$

where $nFoundItems_i$ is the number of found documents which are similar to the query i , and $nItems_i$ is the total number of relevant pieces to the query i .

$$\text{Precision}_i = \frac{nFoundItems_i}{n} \quad (2)$$

where n is the number of pieces returned by the algorithm. We finally compute an average of precision and recall measures, using as a query all the pieces in the collection.

$$\text{Recall} = \frac{1}{N} \sum_{i=1}^N \text{Recall}_i \quad (3)$$

$$\text{Precision} = \frac{1}{N} \sum_{i=1}^N \text{Precision}_i \quad (4)$$

Recall and precision measures, as defined originally, assume that all the documents in the answer set have been considered. We will compute recall and precision values for $n=1, \dots, M$.

It is usually desired to have a single value, instead of two different measures, and for this we use the F measure. The F measure can be considered to combine the information given by precision and recall, and is defined as follows:

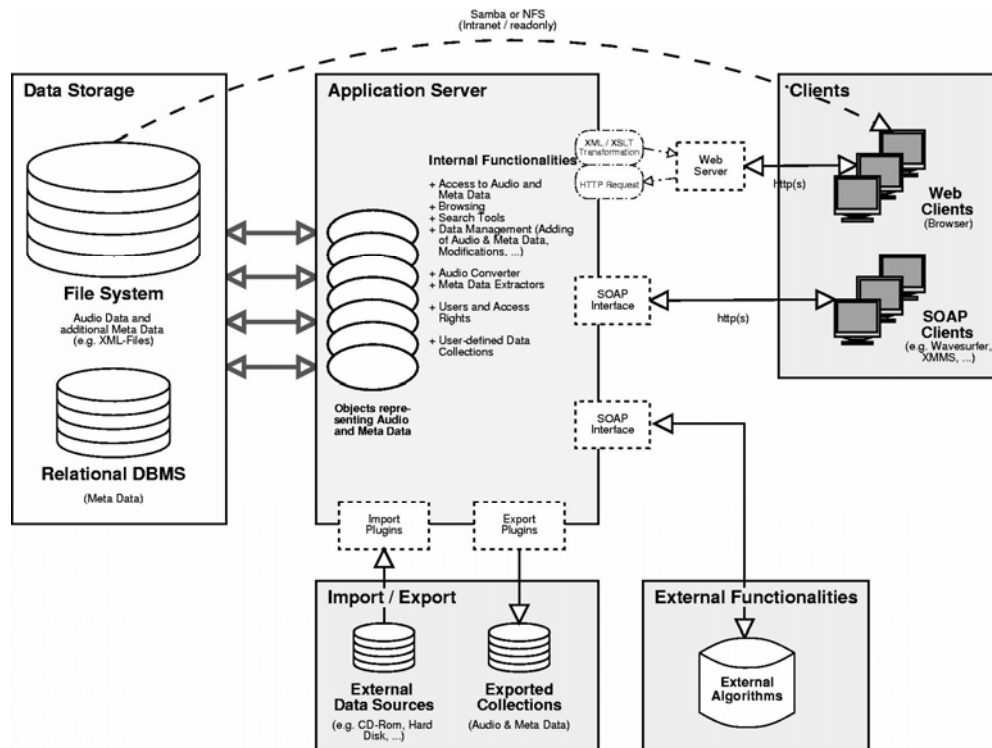
$$F = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

2.3 Management and presentation

There are many advantages in centralizing metadata together with the audio or video content. The metadata and possible descriptions linked to the content allow users to compare their algorithm with others'. Different researchers working on the same problem can increase the benchmark database size by adding more ground truth data.

The common repository can help discover synergies between automatically extracted descriptors. For instance, someone studying chord segmentation will benefit from the results of a beat tracker. Somebody studying the structure of music is likely to benefit from results from tempo evolution, chord progression and singing voice detection. A centralized and correctly labeled repository can stimulate corpus-oriented research and the use of statistical methods and learning techniques. For instance, finding correlations between certain low-level descriptors and genres, or studying the performance of a key extractor in data of different classical periods, from baroque to romanticism, is made much easier. Along with metadata, it is important to store taxonomies or ontologies. The development of a percussive instrument classifier requires the definition of a taxonomy or classification scheme of the categories of percussive instruments. Same happens in many other cases: taxonomies of musical genres, voice types, singing styles, playing modes and many more. Someone spending a Ph.D. thesis time researching on a specific topic is likely to have organized a taxonomy or an ontology that describes the concepts of that specific topic. It is imperative to preserve that knowledge. Much of the legacy metadata may refer to terms of such ontologies. If we want to reuse the metadata, we need to know exactly which restricted vocabulary was used and what was the meaning of each term. Metadata has to be well specified so that it is comprehensible both by men and computers alike. The following is an example of possibly cryptic textual metadata that describes an acoustic guitar chord: "AcousticGuitar:BbMin:Hi:Down". In order to describe this audio sample, besides specifying the instrument that produce it—an acoustic guitar—it is interesting to encode somehow that acoustic guitar is a type of string instrument. It is important to specify that the sound sample is a chord and whether it was created strumming up or down, type of strum, and so on.

Clearly a unified framework is needed. One such example is MTG-DB, developed at the Music Technology Group. MTG-DB is a common database of audio material that offers functionalities for adding audio content, browsing the database, adding metadata and dealing with taxonomies, ontology management and algorithms. It is concerned with providing a common storage for audio material and associated metadata as well as being a tool for research.



The MTG-DB architecture

Evaluation contests

In the area of text retrieval, the annual TREC evaluation contest provides a venue where evaluation can be done in an organized fashion. Following the example of TREC, both video and music retrieval communities have started to build their own evaluation contests addressing mostly content-based retrieval problem. The MIR community, that was born in the late 1990's, started discussing about the need of evaluation in the early 2000's. The evaluation task was commonly held too different from the text retrieval evaluation tasks and thus the MIR community decided to design its own evaluation contest addressing the questions that were relevant for evaluating specifically MIR methods. These discussions and workshops held in JCDL, SIGIR and ISMIR conferences lead into the birth of MIREX (Music Information Retrieval Evaluation eXchange). The first tentative step towards organized evaluation in the area of MIR was taken in 2004 during the 5th ISMIR (International Conference on Music Information Retrieval) in a form of ISMIR2004 Audio Description Contest (ADF) organized by the hosting organization of ISMIR2004, Music Technology Group - UPF, Barcelona. MIREX, is similar to the Text Retrieval Conference (TREC) approach to the evaluation of text retrieval systems. Both MIREX and TREC are built upon three basic components:

- a set of standardized collections;
- a set of standardized tasks/queries to be performed against these collections; and,
- a set of standardized evaluations of the results generated with regard to the tasks/queries.

3 Medical imaging applications domain

3.1 Evaluation of diagnostic accuracy

3.1.1 Aim and relevance

Thorough clinical validation is (or should be) an essential step in the introduction of new medical imaging applications. Before a new application will be accepted by the medical community its added value has to be demonstrated, often with respect to the existing state-of-the-art. The choice of a new imaging application should indeed be guided by a proven added value instead of the trendiness of using a new technique (which is of course no problem in a consumer application). Added value can be shown if the new application allows either a more accurate diagnosis, better patient outcome, lower cost, faster procedure times, less radiation exposure etcetera. In this project we will focus on using a more accurate diagnosis as the added value. Regulatory instances (like the US Food and Drug Administration) also require a *summary of safety and effectiveness* of a product before market introduction can take place. Effectiveness can often be shown by doing a validation.

This validation can be done by the comparison of 2 applications (or products / techniques): the one to be introduced versus the current state-of-the-art. An example of this is the introduction of digital mammography for routine screening of breast cancer, versus the conventional film mammography. Although digital mammography clearly had advantages in information management (storage, distribution, archiving), its clinical value has to be shown and compared to film mammography before the technique will be widely accepted. In this case the added value is primarily the lower cost and easier handling of digital mammography, but this should not go at the expense of lower detection sensitivity/specificity of breast lesions.

The previous example also indicates a complicating factor in the validation of medical devices. The final diagnosis will always be made by a trained medical specialist (usually the radiologist in examinations involving imaging). New techniques (being either acquisition devices, display methods or more complicated image processing algorithms) merely provide assistance to the medical specialist to allow him/her to make a better diagnosis. In that respect, the human factor should be taken into account when validating and comparing different techniques. It is only in recent years, with the advent of digital mammography and – currently – emerging techniques to apply Computer Aided Detection algorithms on medical images, that consensus is growing on the methods to use to validate and compare the performance of different systems.

Within CANTATA research will be done towards the “virtual radiologist”, i.e. how can (multi-modal) content analysis algorithms aid the medical specialist. In order to validate and compare techniques it is necessary to take one step further from standalone or technical evaluation towards clinical or user-involved evaluation. Study setups and statistics exist to perform such a user-involved evaluation and will be briefly outlined below. Before we get into the study design for a user-involved evaluation, let’s first recapture the essentials and benefits of ROC analysis.

3.1.2 ROC analysis

Recent decades have witnessed the widespread use of ROC analysis for the assessment of diagnostic imaging modalities. A ROC curve is a graphical representation of the relationship between the sensitivity (detection rate or true-positive rate) and specificity (1 - false-positive rate) as the level of aggressiveness — or reader mindset — is varied.

This is shown in Figure 1 where parameter p_3 denotes a stricter mindset than p_1 , resulting in a better detection rate (sensitivity) but at the same time being less specific (i.e. overcalling the diagnosis).

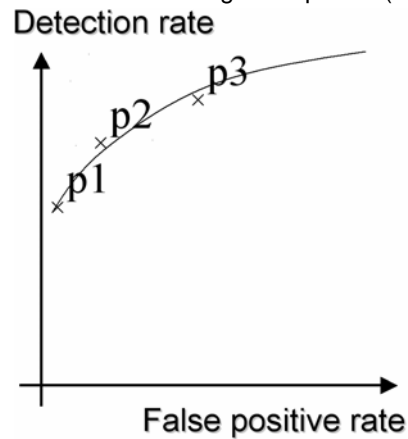


Figure 1 - ROC curve (detection rate versus false positive rate) where parameter p represents the 'mindset' of the reader.

There is often ambiguity in comparing diagnostic modalities when one has only a single sensitivity-specificity pair measured on one modality and a single sensitivity-specificity pair measured on a competing modality. There are many scenarios in which this information — without knowledge of the full ROC — is insufficient for ranking the competing systems. If we would just use sensitivity-specificity pairs (instead of the ROC curve) it is impossible to judge if the differences arose from two different underlying ROC curves (in which case the systems are different and one is better than the other), or they may arise from a common single underlying ROC curve (in which case the two systems are essentially equivalent but the 2 points just represent a different mindset).

Various summary measures of ROC performance are commonly used, including the partial area under the curve in a particular region of interest or the area under the entire ROC curve. They offer a statistical power advantage compared with using a single sensitivity-specificity pair because these summary measures in effect average over multiple “noisy” estimates of the sensitivity-specificity pairs that result from finite data sets.

Further statistical work on ROC analysis can be found in [34][35][36]

3.1.3 User-involved study setup

“In the last 2 decades major advances have been made in the field of assessment methods for medical imaging and computer-assist systems through the use of the paradigm of the receiver operating characteristic (ROC) curve. In the most recent decade this methodology was extended to embrace the complication of reader variability through advances in the multiple-reader, multiple-case (MRMC) ROC measurement and analysis paradigm.” [37]

The MRMC-ROC (“Multiple Reader Multiple Case – Receiver Operating Curve”) statistical framework is considered the “least burdensome approach” to show significant differences between different modalities that generalizes to cases and readers that were not in the study¹. Other statistics used may seemingly be simpler, but are only able to measure the performance (difference) in the particular sample studied, without allowing to generalize. The least burdensome approach means that the least number of cases should be read to reach a certain level of significance (or inversely, a higher discriminating power can be achieved with the same number of cases).

In the study setup a number of readers R should each read the same C cases using all treatment options T . Both the sample of readers R and cases C should be representative of the target groups we want to generalize to. As an example: the readers can be radiologists with sufficient training but that are not experts in the field and therefore may benefit from computer assistance. Each reader should judge each case to be either positive or negative and attach a confidence score to his judgment. This way of scoring allows to use the ROC framework that has advantages over more traditional approaches. Note that it is not always possible to apply all the different treatments T to all the patient cases C , and in those cases other statistics should be used.

For the statistical background we refer to reference [38]

3.1.4 Generalization to other domains

The vision of CANTATA is to develop a virtual specialist in different application domains (multimedia, medical and surveillance). Such a virtual specialist should assist the human operator to do his task faster and/or more accurate, in spite of the rapidly increasing amount of information to be processed. In CANTATA we have identified the “virtual butler”, the “virtual radiologist”, the “virtual display” and the “virtual policeman”.

However, most validation techniques available (of which a selection is described in this report) only take into account the “standalone” or technical validation. This is without any doubt a very powerful and objective way to measure and compare MCA systems, but it does not give insight into the added value of the MCA system to the ultimate goal: the virtual specialist. In other words, there is no validation of how much (if any) improvement was obtained by using the MCA system. A proposal therefore is to take into account the user(s) of the system and to measure the added value for the user. The described MRMC-ROC analysis is a possible candidate for such an evaluation, but its applicability to the different application domains should still be studied.

3.2 Medical Display Simulation Chain (MEDISIC) for determining medical quality,

3.2.1 Context and goal

In the context of a medical portable display, we want to develop a model that can predict clinical accuracy of a display based on the JNDMetrix framework (Figure 2). The medical *display simulation chain* should accurately model physical characteristics of display systems within the existing framework of our human visual system model (JNDMetrix [39]). We are developing a perception model that can predict if a modification (compression artifact, transmission error ...) in the image is clinically relevant.

Medical images that are modified using new compression algorithms in the portable wireless display environment should be clinically validated. It has to be checked if these images still contain enough image quality and if the modification is clinically relevant. In short: there is need for new metrics that describe what the possible clinical impact of a display system or an image processing algorithm is. The work will be based on literature on medical display simulation and human observer performance [40-41-42-43-44-45-46-47-48-49-50].

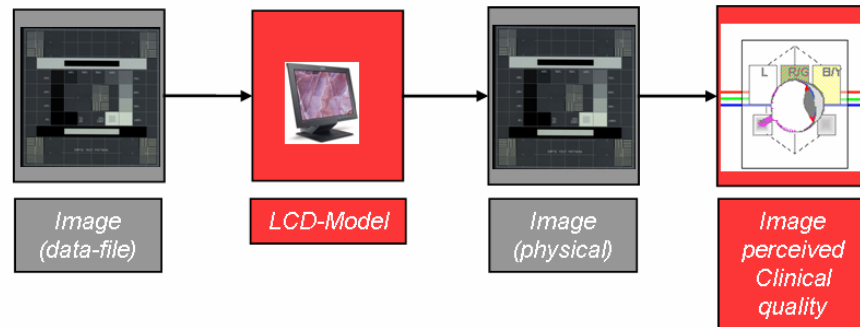


Figure 2. Medical display chain simulation

3.2.2 Display simulation chain

A complete software application will need to be developed to allow modeling of display characteristics, starting from raw images (on the disk) to the physical image displayed. The important parameters to take into account are:

- Viewing angle,
- Brightness,
- Contrast ratio,
- Shape of transfer curve, (Gray Scale Display Function)
- Bit depth,
- ...

The result of the simulation should represent the colored intensity map coming out the screen (XYZ map in cd/m^2).

3.2.3 Human Visual Observer Model

A Human Visual Observer Model should be defined to clinically interpret simulated images and predict perceived differences between them. A very important part of the display simulation chain is to be able to model clinical lesions like micro-calcifications or tumors. In order to validate the display simulation chain we can run the simulation twice: once with a medical feature (like micro-calcification) present and once when this feature is absent. The HVOM should predict if the modeled display is able to show the micro-calcification. For a detailed discussion on how a tumor JND template can be created we refer to [42-43-44-45-46-47-48-49-50].

Because of the compression artifacts it is possible that a tumor becomes more difficult to perceive than without the artifacts, which can cause the false negative fraction to increase. Inversely, compression artifacts could also be considered to be tumors and therefore also the false positive fraction can increase. By doing this type of simulations on a large number of images it is possible to come to a reliable estimate of performance using the receiver operating characteristic (ROC) curve. During the general performance evaluation, the entire portable display system (including all of its sub components) will also be tested concerning speed and functionality, including the throughput (frames per second, bandwidth requirements) of the compression algorithm. Also the accuracy of our classification algorithm (to classify medical and non-medical images) will be tested. The developed system to comply with HIPAA will also be tested in practice. Finally, also the results of our latency reduction algorithms will be evaluated.

3.2.4 Validation

Validation is extremely important. We will validate the display simulation by comparing with published results and by own psycho physical study [44].

4 Surveillance and monitoring applications domain

As said previously, the evaluation of video processing performances has received an increasing attention in the scientific community during the past few years, based on the observation that we lack standard performance metrics to compare and rank the various available algorithms. In addition, performance evaluation is a natural component of recently proposed controlled architectures for video analysis. In fact, the analysis of evaluation results can give a feedback to the controller in order to keep a maximum performance rate.

Performance evaluation has been of interest for a long time in several domains such as edge detection [51], [52] or range image segmentation [53]. However, these techniques are not key techniques for video understanding applications as they often do not offer real-time processing performances. In consequence, the associated evaluation methods are in general of little interest for our purpose as they are dedicated to the analysis of a very specific problem. Here, we thus focus on describing previous works related to the evaluation of video processing techniques intended for surveillance and monitoring applications as it is described in [72]. In this section, we first present qualitative works on the evaluation of video processing techniques and then a review of past and on-going workshops and projects which have been created to address the evaluation issue.

4.1 Qualitative evaluation of output results

Historically, first investigations on the evaluation of video processing performances were consisting in obtaining a qualitative performance evaluation of an algorithm or a comparison between different algorithms having an identical functionality. This is reasonable considering the fact that the techniques were at an early development stage. In consequence, the evaluation was mainly intended to demonstrate the feasibility of an approach. We briefly describe here such works on qualitative performance evaluation.

First, in [8], the authors propose a set of metrics for assessing segmentation result quality when no reference segmentation is available, which is called “standalone evaluation”. Their overall objective is to create objective measures corresponding to the evaluation which can be performed by a human observer. This work takes place in the context of the MPEG object-based coding and description of video sequences. They distinguish the evaluation of individual objects and the overall evaluation for a video sequence in term of a partition into a set of objects. Concerning the individual object evaluation, they define intra-object features (e.g., spatial features such as compactness, temporal features such as size stability) and inter-object features which give indication on whether the objects were correctly identified as separate entities (e.g., local contrast to neighbors). The overall evaluation consists in a weighted formula of the individual object evaluations. Experiments were conducted on few sequences of the MPEG4 test set, with 4 different segmentations. The conclusion of their work is that this kind of evaluation is able to qualitatively rank different algorithms. However, it is not clear that these metrics will still work on a larger set of test video sequences, as reported results do not show high discriminative potential. In addition, this kind of work is of no help to acquire a deep expertise on programs as the evaluation remains vague.

Following the same idea, several works propose perceptual criteria to evaluate performances [54]. The key idea is to design evaluation criteria which take into account visually desirable properties of reference (ideal) segmentations. A recent work reported in [12] proposes to use as evaluation metrics the spatial differences of color and motion and the temporal differences between the color histogram of objects in the current and previous frames. The obvious advantage of these works is that they do not require any ground truth definition, which is well-known to be a time-consuming process, especially for low-level processing such as a segmentation process (i.e., pixel-based ground truth). However, their associated drawback is that they do not enable sound quantitative comparison, which is essential to understand the behavior of video processing programs.

Moreover, in [55], the authors propose a perturbation method for comparing performances of background subtraction methods. This evaluation method does neither require to have foreground objects present in the video sequences nor to have knowledge of the foreground distribution, as it is the case for classical comparison with ground truth.

Instead, the basic assumption is that the shape of the foreground distribution is locally similar to the background one, but shifted or perturbed. In order to perform the evaluation, they first select a parameter set (for each technique) to achieve a certain and fixed false alarm rate. Then, they apply a perturbation on the entire background distribution by vectors in uniformly random directions of the RGB space, with magnitude m . Finally, they measure the detection rate as a function of the applied perturbation of magnitude m . The authors have compared four techniques differing on the background model: mixture of Gaussians, uni-modal, non-parametric and codebook-based. Detailed results are reported on four different video sequences (one indoor and three outdoors). However, unless obvious considerations (e.g., the mixture of Gaussians performs poorly on the indoor sequence), only slight performance variations are registered in average. They conclude their work by arguing that this evaluation method can be useful for qualitative comparison of sensitivity of different algorithms and can serve as a comparison guide to choose a set of parameters for a particular algorithm for a particular application.

In conclusion, this kind of evaluation only provides general and qualitative evaluation results. These general ideas may serve to confirm the assumptions which can be found in scientific papers that describe the techniques. However, such evaluation does not enable to acquire a deep expertise on programs. The reason is that almost all techniques perform as well as others, in average. Results are reported globally for a test set of video sequences. In consequence, this kind of evaluation does not provide sufficient insights to understand when and why a technique is better than another one (e.g., under which environmental conditions).

4.2 Comparison of output results with ground truth

Once the feasibility had been demonstrated, several works have investigated the issue of system reliability through a quantitative comparison of video processing results with ground truth data. Ground truth data are provided by human experts who manually annotate video sequences with the help of a graphical tool. We can mention the Viper tool coming from the University of Maryland [21], and the tool coming from the University of Edinburgh in the framework of the European project CAVIAR [33] which are both Java tools conceived to draw object bounding boxes and to assign user-defined semantic information to the objects (e.g., *occluded person*). The ODVIS (Open Development environment for evaluation of Video surveillance Systems) [26] proposes a tool for researchers to embed their tracking algorithms in the ODVIS framework. A noisy ground truth for the tracking task is automatically computed. Users have simply to refine this first noisy description to obtain the ground truth. In [10], the authors propose a semi-automatic annotation tool for pixel-based ground truth where users have only to adjust/validate a trial segmentation which is automatically computed. In consequence, depending on the level of details of the ground truth (e.g., simple list of points corresponding to target centroids up to a pixel-based description for each target for each frame), these works on quantitative comparison with ground truth present evaluation metrics to assess the performance of video processing algorithms.

In [11], the authors evaluate different shadow detection algorithms. First of all, they propose a two-layer taxonomy of shadow detection algorithms. The first classification criterion considers whether the underlying process is deterministic or statistic. Statistical approaches are then subdivided into parametric and non-parametric methods. Deterministic approaches are then subdivided into model-based or model-free methods. In a second time, the authors introduce the concept of shadow detection rate and shadow discrimination rate, which are derived from the usual detection rate and false alarm rate. Thanks to these metrics, they provide a pixel-based comparison between these four classes of methods, using representative algorithms of each class. Experiments are performed on a test set of 5 video sequences containing indoor and outdoor scenes, with shadows ranging from dark and small to light and large, and with objects of variable size and speed. Each sequence lasts approximately a thousand of frames. In conclusion, this work has demonstrated its ability to rank algorithms and to envisage a classification of algorithms in function of the environmental conditions, even if the number of tested sequences remains low. This work constitutes a first step to obtain a deep understanding of algorithm behaviors and to find explanations when and why they succeed or fail. For instance, they found out that statistical approaches are well suited for indoor environments. However, this work addresses a very specific problem and must be extended and generalized in an evaluation methodology applicable to the whole video processing chain.

In [6], the authors introduce metrics to evaluate the tracking task, especially the accuracy of the position estimation over time (i.e., trajectories). They propose two metrics to measure the displacement between trajectories: one considering a spatial shift and one considering a temporal shift. The credit of this work is to address the issue of video processing programs which are at an intermediate level of the processing chain. Indeed, this kind of evaluation is more complicated than simply counting the number of good/wrong detected pixels. Evaluation results are reported on three trajectories on a video sequence which last 200 frames. However, the authors only consider trajectories of equal length. This is a hard limitation regarding the fact that systems working in real conditions are likely to provide trajectories with different initial or ending time compared to ground truth (e.g., a false object termination due to a ghost, an object which is detected too late due to a lack of contrast). In addition, the evaluation of centroid trajectories alone is not sufficient. For instance, a careful evaluation must also consider splitting and merging of objects, their spatial extent or their class label.

The European project CAVIAR [56] is providing a set of reference video sequences for evaluation. These sequences are associated with ground truth bounding boxes. Moreover, this project provides a tool and a website for on-line evaluation for CAVIAR partners. Similarly, in [31], the authors propose another evaluation website which is dedicated to the evaluation of the tracking process. A main advantage of these works is that they provide data to be used by the scientific community for benchmarking tracking algorithms. Several articles have been published using these data [57], [58]. Unfortunately, a comparison of results from one article to another one remains a difficult task.

In the European project AVITRACK (Aircraft surroundings, categorized Vehicle and Individual Tracking for apRon's Activity model interpretation and Check) on apron monitoring [59], [60], partners have produced a detailed evaluation. They have used 5 sequences, each containing about 2000 images associated with ground truth bounding boxes. The longest sequence contains 18 physical objects and the most complex scene involves 12 objects globally crossing each other. In this experience, among a total of 56 physical objects evolving in the scene, 33 objects have been completely tracked without any errors. The remaining 23 tracked objects have been lost one or two times by the tracking process. In conclusion, the evaluation process presented in this work is significant and shows the effectiveness of the video analysis. However, little information on the video tracking processes can be extracted from the evaluation. This is due to the fact that the evaluation has been performed globally on simple and complex scenes mixing different types of difficulties at a time.

4.3 Alternative to ground truth

Being aware of the issues in defining ground truth (e.g., subjectivity, time-consumption), several alternatives have emerged in the literature. We briefly present here two interesting research directions to help the process of evaluation in particular cases.

First, pseudo-synthetic video sequences have been proposed as a way to automatically create test sequences with various perceptual complexities [29]. The idea is first to use an object detection (or tracking) algorithm on video sequences containing a single object. Due to the reduced complexity of this task, the output results have more chance to be considered as ground truth data. Of course, a filtering stage is applied to remove the tracks which have poor quality. In a second step, these ground truth tracks are randomly selected to compose a more complex scene. The small images corresponding to the objects present in the track are pasted into an empty background, using the calibration information to handle correctly dynamic occlusions (i.e., which object is on the foreground layer compared to the other one). Even if this technique is far from simulating the real world complexity, it can be a useful tool to test and address specific issues. For instance, the authors have been able to test the robustness of their algorithms against dynamic occlusions on long test sequences. Unfortunately, the performances of the tracker rapidly decrease with the number of objects in the scene.

Second, based on recent advances in the creation of 3D animated movies for modeling various behaviors (humans walking in museums, fishes interacting with swimmers in the sea,...) [61], some works attempted to create pure synthetic movies for the purpose of evaluating video processing programs [62]. The authors proposed a tool to generate 3D animations of humans evolving in a scene. To enhance the impression of reality, shadows and their projections against walls are modeled.

Nevertheless, these techniques are very far from being enough realistic and require a lot of processing resources as soon as we want more reality. Again, even if these techniques are not able to recreate complex real issues, they can be advised to study deeply a particular problem such as the shadow management in video processing programs.

Finally, we want to note that new types of sensors which are emerging today can be used to provide ground truth data, namely GPS (Global Positioning System) and RFID (Radio Frequency Identification). For instance, GPS is used in the European project AVITRACK [63] to provide the localization of the objects on a 3D ground of the scene viewed by cameras. They are thus able to assess the accuracy of the localization video processing task without many efforts. Nevertheless, these new solutions still have drawbacks. For instance, the GPS localization does not work inside buildings or when the objects are close to buildings in outdoor scenes. In addition, these ground truth data do not provide any information about the shape or the posture of objects.

4.4 Existing Workshops and Projects

The previous subsection has presented particular works which study the evaluation problem and which are the most relevant in the literature. Nevertheless, a common characteristic of these works is that they are isolated and thus they often have little influence on the scientific community (e.g., for becoming a standard). On the opposite, we present here a review of existing workshops and projects which address the problem of performance evaluation. This is a recent topic in the video understanding community. However, performance evaluation has already been active in other domains for a long time. For instance, in the natural language domain, the TREC (Text REtrieval Conference) competition has been created to encourage research in information retrieval from large text collections. In the multimedia community, a similar competition has been set up for broadcasting video retrieval (TRECVID, [64]). Thanks to these efforts, retrieval techniques start becoming very efficient [65]. The proposed description hereunder recalls the historical evolution of evaluation campaigns in video understanding, from qualitative ones until recent ones which involve the international scientific community and which aim at defining standard evaluation protocols, metrics and reference video sequences. After assessing the reliability of algorithms, the ultimate goal is to be able to design a certification protocol for video processing algorithms.

PETS workshops

The PETS (Performance Evaluation of Tracking and Surveillance) workshop has been created in the year 2000 to answer the need, among the scientific community, of being able to assess the robustness of video processing techniques and compare different alternatives. The first PETS workshop was very attractive and gathered a wide community of researchers. In fact, it was the first time that people were able to test their algorithms on a publicly available set of video sequences. Following this first experience, workshops are now organized on a regular basis (i.e., once or twice a year) and with a particular topic (e.g., counting people, shop monitoring, car park monitoring, face tracking in a conference meeting). In order to prepare the workshop, each participant can retrieve from a web server a set of video sequences and their associated annotations. Most of the time, these annotations are ground truth at the end-user level (e.g., number of persons, presence or absence of an event). The video database is usually composed of two sets: a training set and a testing set. The training set can be used by participants to tune and adapt their system with respect to the application under consideration (e.g., to learn a set of optimized parameter values). On the opposite, the testing set is used to produce evaluation results. In this configuration, systems cannot be modified.

In conclusion, even if evaluations are performed on the same set of sequences for all participants, the comparison of the techniques is most of the time qualitative. This is partly due to the fact that the database is large and participants can freely choose the sequence(s) that they want to produce results (e.g., participant 1 tests with sequence A and participant 2 tests with sequences B and C). There is thus a large variation of the testing conditions. Another reason is that there is no agreement on evaluation metrics. Each participant is free to use its own evaluation metrics. In consequence, it is very difficult to take advantage of these evaluation results for acquiring a deep expertise on programs. However, the important point we want to underline is that the scientific community has become aware of the utmost importance of the evaluation since the creation of PETS. Thanks to this oldest workshop, several other research projects have been created and funded to further investigate this issue.

CAVIAR project

The CAVIAR (Context Aware Vision using Image-based Active Recognition) project [56] is a research program funded under the FP5 of the European Union. This project has two main research interests. First, to investigate the use of contextual knowledge for improving the extraction and grouping of features during a recognition process. Second, to investigate the use of knowledge-based control strategies for improving the recognition speed and accuracy. On the point of view of the evaluation of performances, this project follows the general idea of PETS, though it exhibits two important differences. First, the focus of the project is restricted to the study of two applications: public space and shopping mall surveillance. Second, a huge annotation work has been realized: a total of 100 000 frames have been annotated. This tedious task has been shared between 10 different labelers. In addition, a few video sequences have been annotated by several labelers in order to study the variability introduced by the labeling task (e.g., either due to imprecision or subjectivity) [66]. This variability may be quite large if no accurate annotation rules are given to labelers before they start the annotation process.

In conclusion, this project is a first step towards obtaining a statistically significant evaluation. However, even if annotations are publicly available to other researchers, they are only used to test the CAVIAR system. There is no benchmarking taking place between the different systems from different research centers.

VERAAE workshop

The VERAAE (Video Event Recognition Algorithm Assessment and Evaluation) workshop has been launched in 2005. This workshop aims at the evaluation of performances of event recognition programs, specifically. The objective is to compare existing high-level techniques on a same set of inputs. These inputs are composed of two parts. The first one is a set of ideal inputs which are manually created. The second one is a set of real inputs which are created with a unique tracker. This second set is intended to assess the performances and the robustness of event recognition techniques in presence of degraded inputs.

In conclusion, we argue that this kind of evaluation is useful to acquire expertise on high-level techniques. However, this is of no help to understand the low-level video processing issues which are currently the most critical ones.

VACE project

The VACE (Video Analysis and Content Extraction) project [71] is a research program funded by the US government under the ARDA (Advanced Research and Development Activity) program. The research consortium currently involves 14 participants with both industrials and academics. It is funded in 3 two-year phases and they are currently in the middle of the second phase (i.e., VACE-II). The objective of the year 2005 is to perform the evaluation of the detection and tracking tasks while the evaluation of the event detection task will be performed in 2006. More precisely, they currently study the performances of four video analysis tasks: face detection and tracking, hand detection and tracking, people detection and tracking, vehicle detection and tracking, and across four domains: meeting room, broadcast news, UAV (Unmanned Airborne Vehicle) and surveillance. For each task/domain pair, they have recorded and annotated 100 clips with an average duration of 2.5 minutes each. In order to annotate the video sequences, they have written a reference document in collaboration with the NIST (National Institute for Standard and Technology). This document proposes a set of detailed instructions so that multiple annotators are able to produce reliable and accurate annotation data. The time taken for the annotation process is approximately 2000 hours.

In conclusion, this kind of figures really highlights the tremendous amount of efforts which are needed to perform a statistically significant supervised evaluation (both in time and human resources). In addition, we can underline the efforts made by this project (and especially NIST) to create a standard for the performance evaluation of video understanding systems. Finally, we can note that NIST is also involved in the ETISEO project and has already accepted to share the annotation documents with the project.

ETISEO project

The problem of other projects is that they do not help to discover working conditions of the algorithms. ETISEO [68], one of the latest evaluation programs, has tried to address this issue by characterizing for instance the video input. One of the main objectives of ETISEO is to "acquire precise knowledge of vision algorithms". In other words, ETISEO tries to underline the "dependencies between algorithms and their conditions of use". At the end of the project "strengths and weaknesses of algorithms as well as unsolved problems should be highlighted".

ETISEO tries to address each video processing problem separately, by defining accurately the problem. For instance, they handle shadows within at least three different problems: (1) shadows at different intensity levels (i.e. weakly or strongly contrasted shadows) with uniform non color background, (2) shadows at the same intensity level with different types of background images in terms of color and texture and (3) shadows with different illumination sources in terms of source position and wavelengths.

Firstly, for each problem, the video sequences have been collected to illustrate only the current problem. The video sequences should illustrate the problem at different difficulty levels. For instance, for the problem of shadows and intensity levels, they select video sequences containing shadows at different intensity levels (more or less contrasted). On these selected sequences, the appropriate part of the ground truth is filtered and extracted to isolate video processing problems. For instance, for the detection task, they evaluate the algorithm performance relatively to the problem of handling occluded objects by considering only the ground truth related to the occluded objects.

Secondly, for a given task (object detection, tracking, object classification and event recognition) ETISEO defines a sufficient number of metrics to measure and characterize the algorithm performance on various aspects. For instance, in ETISEO there are 7 metrics for the task of object detection.

Thirdly, ETISEO computes the reference data which corresponds to the expected output of the algorithm to be evaluated relatively to a given video processing task. The reference data are computed from the ground truth provided by human operators and can be improved to better correspond to the expected results. For instance, instead of evaluating the mobile object positions from the ground truth (2D-points), they can use 3D-point reference data to measure the computation of 3D object position.

Finally, ETISEO provides a unique automatic evaluation tool to accurately analyze how a given algorithm addresses a given problem.

As a summary, in ETISEO, for each video sequence, there are three types of associate data. The first one is the ground truth (e.g. object bounding box, object class, event etc.) given by human operators at each level of the four video processing tasks. The second one is the general annotation on the video sequences concerning video processing problems (e.g. weak shadows) or concerning recording conditions (e.g. weather conditions such as sunny day). The final information is the camera calibration and contextual information about the empty scene describing the topology of the scene (e.g. zone of interest).

All the video sequences of ETISEO (about 40 sequences) are selected and classified according to the problems they illustrate. These sequences have been processed by 16 international teams participating to the evaluation program in two phases.

The main limitation of ETISEO is that it does not define quantitative methods to measure the difficulty level of the videos illustrating a given video processing problem. For instance, ETISEO uses the terms "normal" or "dark" to describe the intensity levels of video sequences. Therefore, the selection of video sequences in ETISEO according to their difficulty levels is subjective and not precise enough. Furthermore, this subjective judgment also makes the prediction ability of the evaluation difficult. To overcome this issue, [69] has proposed a new evaluation approach that tries to quantify the difficulty levels of each video processing problem. With the definition of difficulty level, they can measure the capacity of algorithms on solving each video processing problem separately. The authors have applied this approach to two problems: detection of weakly contrasted mobile object and detection of shadow. The preliminary results show that with this approach, they can determine the upper-bound of algorithm capacity on solving these problems.

In conclusion, despite its limitation, this recent project is a real step towards the definition of standard evaluation metrics and of an appropriate evaluation protocol. In addition, the collaboration of all participants (e.g., NIST) has guaranteed a large dissemination of the project results and will favour the definition of a standard.

RATP contest

Finally, to close this presentation of past and on-going research projects on performance evaluation, we briefly describe a contest which has been launched by the RATP [67], the French company of public transport in Paris. They have organized a call for real-time event detection solutions for enhanced security and safety in public transportation beside the AVSS (Advanced Video and Signal-Based Surveillance) international conference [70]. They have recorded many video sequences of several minutes from metro stations. The diversity of sequences is rather large: different metro stations, different cameras (analog, digital, black and white, infrared), different moments in a day (morning, peak hour, evening, night) and different events (nothing, abandoned baggage). The winner is the system which has demonstrated the best results on the whole set of video sequences. The end-users were a little bit disappointed because results were below what they were expecting. Moreover, the metrics were not really convincing in discriminating the competitive solutions.

We can draw two conclusions from this contest. First, there is an increasing need coming from end-users to compare between the various existing video understanding systems. However, on a scientific point of view, such comparisons are of no help neither to identify current issues in the domain nor to acquire expertise on techniques. Second, end-users become suspicious about existing systems. This is due to the fact that current systems are not enough robust and adaptable to the diversity of situations. In consequence, end-users become rapidly unsatisfied of the results of video understanding systems (e.g., a too high number of false alarms). As a consequence, they want to test and validate themselves the systems, before deciding whether to use them or not.

VIVID project

VIVID - Video Verification of Identity tracking evaluation website [85] offers an online evaluation tool for object tracking algorithms. It provides nine publicly accessible ground-truthed datasets, visible and thermal IR video sequences, for tracking of ground vehicles from airborne sensor platforms. Also available are an open-source tracking testbed in C++ and mechanism for uploading tracking results for automated scoring. The evaluation methodology and metrics are an usual combination of temporal and spatial correspondence of the ground truth and detected objects represented by their bounding boxes, with the additional distance metrics produced for a bitmap version of the matched objects [84]. Although it suffers from same weaknesses as other similar evaluation initiatives, a good aspect of this website is that it provides a simple to use online tool with clear metrics definitions and free annotated datasets.

i-Lids project

The i-LIDS [87], provided by the Home Office Scientific Development Branch (HOSDB), is the UK Government's benchmark dataset for video-based detection systems with the intention to help in policing and counter-terrorist operations. The video sequences are real world CCTV footages based on four scenarios: abandoned baggage detection, parked vehicle detection, doorway surveillance and monitoring, and sterile zone or perimeter monitoring. Each scenario comprises video datasets of 24 hours of footage. The high-level ground truth annotation is provided in XML format. The datasets include a wide range of weather and lighting conditions. Users can filter the data according to weather conditions or alarm events. From the summer of 2007, HOSDB is planning to carry out evaluation of participating algorithms. The scoring is based on the F-Measure to take into account the relative influence of detection rate and false alarm rate on the outcome of the evaluation. This choice of evaluation metrics represents a significant shift in understanding of a 'good' performance of a video-based algorithm. Although the dependency of the evaluation criteria on the end-user application has been in principle acknowledged by the previous evaluation initiatives, this notion has not been supported by their design of the evaluation methodology and the selection of metrics.

This is probably due to the fact that both algorithm design and evaluation were essentially performed by the researchers without sufficient input from the end-users. This indicates the importance of involving all members of the visual surveillance chain in the process of evaluation of existing technologies. For a discussion on the impact of end-user applications on the outcome of the evaluation and the use of the F-Measure in the performance evaluation of video understanding algorithms see [86]

4.5 Conclusion

In parallel, several workshops and projects have been created to answer to the increasing need of having effective evaluations. The overall trend is to propose automatic tools allowing comparing results with reference data. These automatic tools enable to produce significant quantitative results on standard test video sequences. However, a standard database does not exist yet and the ground truth definition is a hard task. In addition, there is a large variability of end-user needs. In consequence, we need a general evaluation framework as proposed in ETISEO to get better insights on the video understanding approaches and on current technical issues. This framework should contain for instance a precise characterization of the video input to be used for the evaluation. In the same way, we need a flexible evaluation tool as provided by ETISEO which enables to acquire expertise on the video processing programs composing a video understanding platform. The evaluation tool should be parameterized in function of the task to evaluate, the target application and the environmental description. An example of flexibility which is required is the ability to select the part of the ground truth which is relevant to evaluate a given task. For instance, to evaluate a face tracker, we do not need a bounding box but an ellipse delineating the face. On the opposite, in a people counting application, the bounding box or just the center of gravity of the objects is perhaps sufficient.

REFERENCES

1. X. Desurmont, R. Wijnhoven, E. Jaspert, O. Caignard, M. Barais, W. Favoreel and J.F. Delaigle, "Performance evaluation of real-time video content analysis systems in the CANDELA project", conference on Real-Time Imaging IX, part of the IS&T/SPIE Symposium on Electronic Imaging 2005, 16-20 January 2005 in San Jose, CA USA
2. X. Desurmont, A. Bastide, J.F. Delaigle, B. Macq, "A seamless modular approach for real-time video analysis for surveillance", *5th International Workshop on Image Analysis for Multimedia Interactive Services*, Instituto Superior Tecnico, Lisboa, Portugal, April 21-23, 2004.
3. A.Cavallaro, D. Douxchamps, T. Ebrahimi and B. Macq, "Segmenting moving objects: the MODEST video object kernel", *Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2001*, Tampere, Finland, May 16-17, 2001.
4. F. Cupillard, F.Brémont and M. Thonnat, "Tracking groups of people for video surveillance", *2nd European Workshop on AVBS Systems (AVBS2002)*, University of Kingston, London, UK, Sept. 2001.
5. T. Ellis, "Performance Metrics and Methods for Tracking in Surveillance", *Proc. of the Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2002)*, Copenhagen, Denmark, June 2002.
6. C.J. Needham and D. Boyle, "Performance Evaluation Metrics and Statistics for Positional Tracker Evaluation", *Proc. of the Computer Vision Systems: Third International Conference, ICVS 2003*, vol. 2626, pp 278—289, Graz, Austria, April 2003.
7. P.L. Correia and F. Pereira, "Objective evaluation of video segmentation quality", *Image Processing, IEEE Transactions on*, Vol. 12, Num. 2, pp 186—200, Feb. 2003.
8. Correia, P. and Pereira, F. (2001). "Standalone objective evaluation of segmentation quality. In Proceedings of the International Workshop for Image Analysis for Multimedia Interactive Services (WIAMIS'01), Tampere, Finland.
9. J.R. Renno, J. Orwell and G.A. Jones, "Evaluation of shadow classification techniques for object detection and tracking", *IEEE International Conference on Image Processing*, Suntec City, Singapore, Oct. 2004.
10. J. Nascimento and J.S. Marques, "New performance evaluation metrics for object detection algorithms", *6th International Workshop on Performance Evaluation for Tracking and Surveillance (PETS 2004)*, ECCV, Prague, Czech Republic, May 2004.

11. Prati, I. Mikic, M.M. Trivedi and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 27, Num. 7, pp 918—923, July, 2003.
12. C.E. Erdem, B. Sankur and A.M. Tekalp, "Performance measures for video object segmentation and tracking", *Image Processing, IEEE Transactions on*, Vol. 13, Num. 7, pp 937—951, July, 2004.
13. P.L. Rosin and E. Ioannidis, "Evaluation of global image thresholding for change detection", *Pattern Recognition Letters*, Vol. 24, Num. 14, pp 2345—2356, Oct. 2003.
14. S. Pingali and J. Segen, "Performance evaluation of people tracking systems", *Applications of Computer Vision, 1996. WACV '96.*, Proceedings 3rd IEEE Workshop on, pp 33—38, Sarasota, FL, USA, Dec. 1996.
15. M. Rossi and A. Bozzoli, "Tracking and counting moving people", *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, Vol. 3, pp 212—216, Austin, TX, USA, Nov. 13-16, 1994.
16. Y. J. Zhang, "A survey on evaluation methods for image segmentation", *Pattern Recognition*, Vol. 29, Num. 8, pp 1335—1346, Aug. 1996.
17. F. Oberti, A. Teschioni and C.S. Regazzoni, "ROC curves for performance evaluation of video sequences processing systems for surveillance applications", *Image Processing, 1999. ICIP 99. Proc. 1999 Int. Conf. on*, Vol. 2, pp 949—953, Kobe, Japan, Oct. 1999.
18. F. Oberti, E. Stringa and G. Vernazza, "Performance evaluation criterion for characterizing video-surveillance systems", *Real-Time Imaging*, Vol. 7, Num. 5, pp 457—471, Oct. 2001.
19. V.Y. Mariano et al., "Performance evaluation of object detection algorithms", *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, Vol. 3, pp 965—969, Aug. 2002.
20. F. Oberto, F. Granelli and C.S. Regazzoni, "Minimax based regulation of change detection threshold in video-surveillance systems", in *Multimedia video-based surveillance systems*, G.L. Foresti, P. Mähönen, C.S. Regazzoni, pp 210—233, Kluwer Academic Publishers, 2000.
21. Doermann, D. and Mihalcik, D. (2000). "Tools and techniques for video performance evaluation." In Proceedings of the International Conference on Pattern Recognition (ICPR'00), pages 167--170, Barcelona, Spain.
22. T.H. Chalidabhongse, K. Kim, D. Harwood and L. Davis, "A perturbation method for evaluating background subtraction algorithms", *Proc. Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 2003)*, Nice, France, Oct. 2003.
23. X. Gao, T.E. Boult, F. Coetsee and V. Ramesh, "Error analysis of background adaption", *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, Vol. 1, pp 503—510, Hilton Head Island, SC, USA, June 2000.
24. M. Xu and T. Ellis, "Partial observation vs. blind tracking through occlusion", *British Machine Vision Conference 2002 (BMVC2002)*, University of Cardiff, UK, Sept. 2-5, 2002.
25. R. Collins, X. Zhou, S. K. The, "An Open Source Tracking Testbed and Evaluation Web Site", *Proc. of 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Breckenridge, Colorado, January 7 2005.
26. C. Jaynes, S. Webb, R. Matt Steele and Q. Xiong, "An Open Development Environment for Evaluation of Video Surveillance Systems", *Proc. of 3rd IEEE Int. Workshop on Performance Evaluation and Tracking and Surveillance (PETS)*, pp 32—39, June 1 2002.
27. R.G.J. Wijnhoven, "Scenario Description - Technical Document v.0.6", CANDELA Project, Bosch Security Systems B.V., Eindhoven, The Netherlands, 2004.³
28. P. Merkus, et. al, "CANDELA – Integrated Storage, Analysis and Distribution of Video Content for Intelligent Information Systems", *Proceeding of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, EWIMT*, London, UK, Nov. 25—26, 2004.
29. J. Black, T. Ellis and P. Rosin, "A novel method for video tracking performance evaluation", *Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pp 125—132, Nice, France, Oct. 2003.
30. D. Makris, T.J. Ellis and J. Black, "Learning scene semantics", *ECOVISION 2004, Early Cognitive Vision Workshop*, Isle of Skye, Scotland, UK, May 2004.

31. D. Greenhill, J. Renno, J. Orwell and G.A. Jones, "Learning the semantic landscape: embedding scene knowledge in object tracking", *Real-Time Imaging, Special Issue on Video Object Processing for Surveillance Applications*, Jan. 2004.
32. T.J. Ellis, D. Makris and J. Black, "Learning a Multi-camera Topology", Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), ICCV 2002, pp. 165—171, Nice, France, 2003.
33. T. List and R.B. Fisher, "CVML – An XML-based Computer Vision Markup Language", International Conference for Pattern Recognition, Cambridge, UK, Aug. 2004.
34. Metz CE. "Basic principles of ROC analysis." *Semin Nucl Med* 1978; 8:283–298.
35. Metz CE. "ROC methodology in radiologic imaging." *Invest Radiol* 1986; 21:720–733.
36. Metz CE. "Some practical issues of experimental design and data analysis in radiological ROC studies." *Invest Radiol* 1989; 24:234–245.
37. R.F.Wagner et al, "Assessment of Medical Imaging and Computer-assist Systems:Lessons from Recent Experience", *Acad Radiol* 2002; 9:1264–1277
38. Dorfman DD, Berbaum KS, Metz CE. "Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method." *Invest Radiol* 1992; 27:723–731. 8:328–334.
39. www.sarnoff.com/products_services/video_vision/jndmetrix/visual_quality.asp
40. J. Jung, R. Geelen, and D. Vandenbroucke, "The virtual image chain: a powerful tool for the evaluation of the perceived image quality as a function of imaging system parameters," in IS&T/PICS Conference Proceedings, pp. pp. 128–131, 2001.
41. C. Marchessoux and J. Jung. A virtual image chain for perceived image quality of medical display. In *SPIE Medical Imaging*, February 2006.
42. E. A. Krupinski, "Medical Image Perception: Influence of Monitor Quality on Observer Performance", *Society of Computer Applications in Radiology*, www.scarnet.org
43. E. A. Krupinski, J. P. Johnson, H. Roehrig, J. Nafziger, J. Fan, J. Lubin, "Use of a Human Visual System Model to Predict Observer Performance with CRT vs LCD Display of Images", *Journal of Digital Imaging*, Volume 17, Issue 4, Dec 2004, Pages 258-263
44. E. A. Krupinski, J. P. Johnson, H. Roehrig, J. Lubin, "Optimizing soft-copy mammography displays using a human visual system model: influence of display phosphor", *Academic Radiology*, Feb 2003, 10(2), p. 161-6
45. E. A. Krupinski, J. P. Johnson, H. Roehrig, M. Engstrom, J. Fan, J. Nafziger, J. Lubin, W. J. Dallas, "Using a human visual system model to optimize soft-copy mammography display: influence of MTF compensation", *Academic Radiology*, Sep 2003, 10(9), p. 1030-5
46. E. A. Krupinski, J. P. Johnson, H. Roehrig, J. S. Nafziger, J. Lubin, "Viewing images on and off axis with CRT and LCD monitors: effects on observer and model performance", *Proceedings of SPIE Medical Imaging 2005*, Vol. 5749, pp. 281-287
47. J. P. Johnson, J. Lubin, J. S. Nafziger, E. A. Krupinski, H. Roehrig, "Channelized model observer using a visual discrimination model", *SPIE Medical Imaging Conference*, 13-17 Feb, San Diego, CA (2005).
48. A. Badano, B.D. Gallas and D. H. Fidara. "Visual detection with non-Lambertian displays: model and human observer results"
49. W.B. Jackson, M.R. Said, D.A. Jared, J.O. Larimer, J.L. Gille and J. Lubin. "Evaluation of human vision models for predicting human-observer performance". *Image Perception conference*, 1997, ISBN 0-8194-2447-1
50. P. G. Engeldrum, "Image quality modeling : Where are we ?," in IS&T/PICS Conference Proceedings, pp. pp. 251–255, 1999.
51. Heath, M., Sarkar, S., Sanocki, T., and Bowyer, K. (1997) "A robust visual method for assessing the relative performance of edge-detection algorithms". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1338—1358.
52. Bowyer, K., Kranenburg, C., and Dougherty, S. (1999). "Edge detector evaluation using empirical roc curves". *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99)* , pages 359--366, Fort Collins, USA. IEEE Computer Society Press.
53. Min, J., Powell, M., and Bowyer, K. (2004). "Automated performance evaluation of range image segmentation algorithms" *IEEE Transactions on Systems, Man and Cybernetics*, 34(1):263—271.
54. Cavallaro, A., Gelasce, E., and Ebrahimi, T. (2002). "Objective evaluation of segmentation quality using spatio-temporal context". In *Proceedings of the IEEE Conference on Image Processing (ICIP'02)*, pages 301--304, Rochester, NY, USA.

55. Horprasert, T., Kim, K., Harwood, D., and Davis, L. (2003). "A perturbation method for evaluating background subtraction algorithms." In Ferryman, J., editor, Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS) , pages 110--116, Nice, France.
56. Caviar (2004) <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>
57. Dibos, F., Pelletier, S., and Koep, G. (2005). "Real-time video segmentation." In Tubaro, S., Sarti, A., and Lupica, F., editors, Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance , pages 382--387, Como, Italy. IEEE Computer Society Press.
58. Hall, D. (2005). "Automatic parameter regulation for a tracking system with an auto-critical function." In Proceedings of the IEEE International Workshop on Computer Architecture for Machine Perception (CAMP05) , pages 39--45, Palermo, Italy.
59. Borg, M., Thirde, D., Ferryman, J., Fusier, F., Bremond, F., and Thonnat, M. (2005). "An integrated vision system for aircraft activity monitoring." In Ferryman, J., editor, Proceedings of the 7th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS05) , Breckenridge, Colorado, USA. IEEE Computer Society Press.
60. Aguilera, J., Wildenauer, H., Kempel, M., Borg, M., Thirde, D., and Ferryman, J. (2005). "Evaluation of motion segmentation quality for aircraft activity surveillance." In The Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 2005) , Beijing, China.
61. Terzopoulos, D. (1999). "Artificial life for computer graphics." Communication of the ACM, 42(8):32—42.
62. Santuari, A., Lanz, O., and Brunelli, R. (2003). "Synthetic movies for computer vision applications." In Hamza, M., editor, Proceedings of the 3rd IASTED International Conference on Visualization, Imaging and Image Processing (VIIP'03), pages 1--6, Benalmadera, Spain. Acta press.
63. Avitrack (2004). <http://www.avitrack.net>.
64. Trecvid (2001). <http://www-nlpir.nist.gov/projects/trecvid/>.
65. Quénot, G., Moraru, D., Ayache, S., Charhad, M., Guironnet, M., Carminati, L., Mulhem, P., Gensel, J., Pellerin, D., and Besacier, L. (2004). "Clips-lis-lsr-labri experiments at trecvid 2004." In Proceedings of the TREC Video Retrieval Evaluation, New-York, NY, USA.
66. List, T., Bins, J., Vasquez, J., and Tweed, D. (2005). "Performance evaluating the evaluator". In The Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 2005), Beijing, China.
67. Ratp (2005). <http://www.ratp.com>.
68. Etiseo (2005). <http://www.etiseo.net>.
69. A.T Nghiem, F. Bremond, M. Thonnat, R. Ma "A new Evaluation Approach for Video Processing Algorithms" accepted paper – IEEE Workshop on Motion and Video Computing – WMVC Feb. 23-24, Austin, Texas, 2007.
70. Creds (2005). <http://www-dsp.elet.polimi.it/avss2005/>
71. Vace (2004). <http://www.ic-arda.org/InfoExploit/vace/index.html>.
72. Benoît Georis "Program Supervision Techniques for Easy Configuration of Video Understanding Systems", PhD Thesis 2006
73. N. Lazarevic-McManus et al (2006). "Designing Evaluation Methodologies: The Case of Motion Detection". In Proceedings of 9th IEEE International Workshop on PETS, pages 23-30, New York, June 18, 2006
74. N. Lazarevic-McManus et al (2006). "Performance Evaluation in Visual Surveillance using the F-measure" In Proceedings of the ACM Multimedia Workshop on VSSN, Santa Barbara, California, October 27, 2006
75. Lessaffre, M., Leman, M., De Baets, B. and Martens, J.-P. "Methodological considerations concerning manual annotation of musical audio in function of algorithm development". Proceedings of the 4th International Conference on Music Information Retrieval, Barcelona, 2004, 64-71.
76. Adams, W.H., Lin, C.-Y., Iyengar, G., Tseng, B. L. and Smith, J. R.. "IBM multimodal annotation tool", IBM Alphaworks, August 2002.
77. Kaji, K. and Nagao, K. "MiXA: A Musical Annotation System". Proceedings of the 3rd International Semantic Web Conference, Hiroshima, Japan, 2004.
78. Tzanetakis, G. and Cook, P. R. "Experiments in computer-assisted annotation of audio", Proceedings of the ICAD, Atlanta, GE, 2000.

79. Amatriain, X., Massaguer, J., García, D. and Mosquera, I. "The CLAM Annotator: A crossplatform audio descriptors editing tool". Proceedings of the 6th International Conference on Music Information Retrieval, London, UK, 2005.
80. Cano, P., Koppenberger, M., Ferradans, S., Martínez, A., Gouyon, V., Sandvold, V., Tarasov, V. and Wack, N. "MTG-DB: A repository for music audio processing". Proceedings of the 4th Intl. Conf. on Web Delivering of Music, Barcelona, 2004
81. Gouyon, F., Wack, N. and Dixon, S. "An open source tool for semi-automatic rhythmic annotation". Proceedings of the 7th Intl. Conference on Digital Audio Effects, Naples, 2004.
82. Notess, M. and Swann, M. B. "Timeliner: Building a Learning Tool into a Digital Music Library", Proceedings of ED-MEDIA, Lugano, Switzerland, 2004.
83. Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Retrieval evaluation. In Modern Information Retrieval, Series in Cognition and Perception, book chapter 3, pages 73–99. ACM Press, Pearson Addison Wesley, first edition
84. R. Collins and X. Zhou and S.K. Teh "An Open Source Tracking Testbed and Evaluation Web Site". IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005)
85. VIVID <http://www.vividevaluation.ri.cmu.edu/main.html>
86. N. Lazarevic-McManus et al (2006). "Performance Evaluation in Visual Surveillance using the F-measure" In Proceedings of the ACM Multimedia Workshop on VSSN, Santa Barbara, California, October 27, 2006
87. i-Lids <http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/>

ⁱ Dorfman DD, Berbaum KS, Metz CE. "Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method." Invest Radiol 1992; 27:723–731. 8:328–334.