

# Performance evaluation of real-time video content analysis systems in the CANDELA project<sup>1</sup>

Xavier Desurmont<sup>a</sup>, Rob Wijnhoven<sup>b,c</sup>, Egbert Jaspers<sup>b</sup>  
Olivier Caignart<sup>d</sup>, Mike Barais<sup>e</sup>, Wouter Favoreel<sup>f</sup>, Jean-François Delaigle<sup>a</sup>

<sup>a</sup> Multitel A.S.B.L., Av Copernic, 1, B-7000 Mons, Belgium.

<sup>b</sup> Bosch Security Systems, Eindhoven, The Netherlands.

<sup>c</sup> Eindhoven University of Technology, Eindhoven, The Netherlands.

<sup>d</sup> IT-OPTICS, Mons, Belgium.

<sup>e</sup> Vrije Universiteit Brussel, Belgium. <sup>f</sup> Traficon, Belgium.

## ABSTRACT

The CANDELA project aims at realizing a system for real-time image processing in traffic and surveillance applications. The system performs segmentation, labels the extracted blobs and tracks their movements in the scene. Performance evaluation of such a system is a major challenge since no standard methods exist and the criteria for evaluation are highly subjective. This paper proposes a performance evaluation approach for video content analysis (VCA) systems and identifies the involved research areas. For these areas we give an overview of the state-of-the-art in performance evaluation and introduce a classification into different semantic levels. The proposed evaluation approach compares the results of the VCA algorithm with a ground-truth (GT) counterpart, which contains the desired results. Both the VCA results and the ground truth comprise description files that are formatted in MPEG-7. The evaluation is required to provide an objective performance measure and a mean to choose between competitive methods. In addition, it enables algorithm developers to measure the progress of their work at the different levels in the design process. From these requirements and the state-of-the-art overview we conclude that standardization is highly desirable for which many research topics still need to be addressed.

Keywords: real-time processing, computer vision, performance evaluation, MPEG-7, video content analysis.

## 1. INTRODUCTION

The number of security and traffic cameras installed in both private and public areas is increasing. Since human guards can only effectively monitor a limited number of camera monitors, automatic analysis of the video content is required. Examples of challenging applications [2] are monitoring metro stations [3] or detecting highway traffic jams, unattended object detection [1] and detecting of loitering persons. Since the last decade, many algorithms have been proposed that try to solve the problem of scene understanding. The level of understanding varies highly from only detecting moving objects and outputting their bounding boxes (e.g. the OpenSource project “Motion”<sup>2</sup>), to tracking of the objects over multiple cameras, thereby learning common paths and appearance points [28], [30] or depth maps and amount of activity in the scene [29].

Apart from functional testing, there are several other reasons for evaluating the video content analysis (VCA) systems; scientific interest, measuring the improvement during development, benchmarking with competitors, commercial purposes and finally legal/regulatory requirements. However, most literature describing VCA 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.

---

<sup>1</sup> This work is part of the European ITEA project CANDELA (Content Analysis Networked DELivery Architectures), ip02013 <http://www.extra.research.philips.com/euprojects/candela/>

<sup>2</sup> OpenSource project Motion: <http://sourceforge.net/projects/motion/>

However, for VCA 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 VCA system.

The remainder of the paper is organized as follows: Section 2 describes an overall framework for video content analysis, search & retrieval and performance evaluation, Section 3 gives an overview of earlier work related to performance evaluation, publicly available video data sets, ground truth (GT) annotation tools and evaluation metrics and Section 4 concludes and indicates future work.

## 2. CONTENT ANALYSIS AND EVALUATION FRAMEWORK

Within Europe already several other projects have addressed systems with VCA functionality. These are e.g. CAVIAR<sup>3</sup>, MODEST<sup>4</sup> [2], RETRIEVE<sup>5</sup> and ADVISOR<sup>6</sup>. The VSAM<sup>7</sup> project worked on related topics in the USA. These projects have shown the challenge of robustness of VCA algorithms, and the importance and complexity of performance evaluation of these algorithms.

Within the CANDELA project, content analysis is addressed in a broad application domain. Partners from medical imaging, multimedia, traffic analysis and surveillance aim to construct a general integrated VCA framework. An overview of the CANDELA project is given in [26].

A distinction can be made between those systems that will suffer a critical failure if time constraints are violated (hard real-time), and those that will not (soft real-time). The VCA system from Desurmont *et al.* [1] has soft real-time requirements since not all video frames necessarily have to be processed to detect moving objects. However, there are other constraints with respect to timing. For real-time surveillance systems, where the operator wants to receive alarms for certain events, the system is required to work with limited delay. When an alarm should be generated when a car enters the scene, the VCA system should not apply classification until an object has left the scene, but do this classification on the fly.

Within the CANDELA project, a general architecture has been developed, enabling all types of real-time VCA systems. Using this architecture, a common demonstrator system is being built by various partners of the project. The major components of the architecture are presented in Figure 1. The proposed framework has three main modules.

The first is the video content analysis module, where video acquisition is applied (frame grabbing from a video camera). Moving objects are segmented from the input video images. Object features (like speed and size) are gathered and all generated meta- and video data is stored in the overall database, which is subsequently used for search & retrieval and performance evaluation. Because video acquisition for surveillance and traffic is applied continuously (non-stop), the analysis of the content is applied in real-time.

The second module is the search and retrieval module, where users can search through all stored video by specifying certain criteria (like object type, size, speed and trajectory). The returned results are presented to the user through a graphical user interface. Since the searching can be applied off-line, this module is not real-time constrained.

The third module comprises the performance evaluation of the video content analysis algorithm(s). The metadata, generated in the analysis step, is compared with the expected results, i.e. the so-called ground truth (GT). Certain content analysis results are compared and presented to the user. Section 3 discusses the topic of performance evaluation in more detail.

Although the scope of the CANDELA system is rather broad, this paper focuses on evaluation of the real-time video content analysis. The following section will discuss the involved research areas and introduce a classification of the related work from the literature into different semantic levels.

---

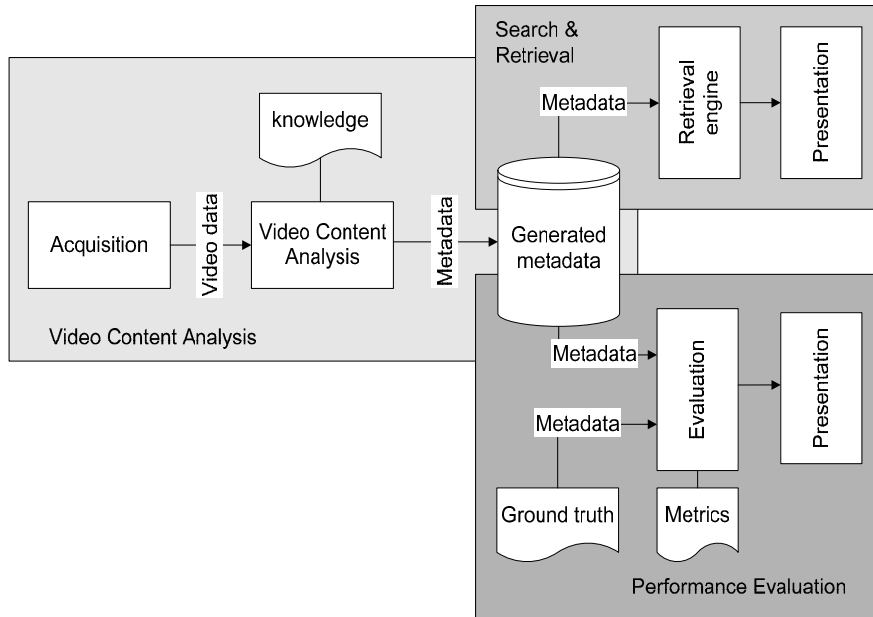
<sup>3</sup> CAVIAR: Context Aware Vision using Image-based Active Recognition: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>

<sup>4</sup> MODEST: Multimedia Object Descriptors Extraction from Surveillance Tapes: <http://www.tele.ucl.ac.be/PROJECTS/MODEST/>

<sup>5</sup> RETRIEVE: Realtime Tagging and Retrieval of Images Eligible for use as Video Evidence: <http://www.retrieve-project.org/>

<sup>6</sup> ADVISOR: Advanced Digital Video Storage Online Retrieval System: <http://advisor.matrasi-tls.fr/>

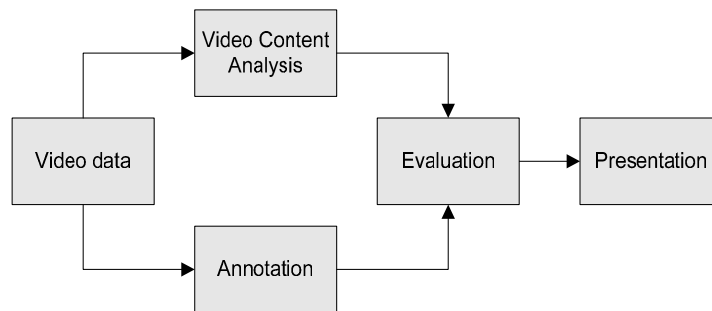
<sup>7</sup> VSAM: Video Surveillance and Monitoring: <http://www-2.cs.cmu.edu/~vsam/>



**Figure 1: Overview of the total system.**

### 3. PERFORMANCE EVALUATION

This section elaborates on the evaluation of real-time video content analysis for surveillance and traffic applications. To enable performance evaluation, multiple steps have to be taken. First, video sequences must be available. Secondly, there are results from a VCA function that needs to be evaluated. Thirdly, GT needs to be generated and stored. Then, the GT needs to be compared with the generated results, requiring unambiguous definition of the metrics. And finally, the evaluation results are combined for each video sequence for presentation to the user. Hence, we distinguish four main topics for evaluation as shown in Figure 2; creation of ground-truth data, available evaluation data sets, performance metrics, and presentation of the evaluation results.



**Figure 2: Performance Evaluation.**

Section 3.1 will first present the state-of-the-art from the literature. Subsequently, Subsection 3.2 through 3.5 will discuss each research topic separately.

#### 3.1. Related work

The importance of performance evaluation of VCA algorithms has been addressed by various projects, and has led to various research publications. From the year 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 VCA algorithms.

Since a complete VCA system comprises multiple semantic levels, evaluation can be done on certain levels. VCA 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.

- 1) Pixel-level; segmentation of the video image in moving objects and static background
- 2) Object-based evaluation per frame
- 3) Object-based evaluation over an objects life time
- 4) Object-features like object type, speed, size (in meters)
- 5) Behavior of objects; e.g. car parking, person enters door

As already mentioned, 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 VCA 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 [8]. We will now discuss some proposals from literature, categorized by the above-discussed semantic levels.

1) Pixel-level; segmentation of the video image in moving objects and static background.

Zhang [14] 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 [6] 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.* [10] present three performance evaluation metrics that do not require segmented GT. They propose spatial differences of color and motion and the boundary of the segmented video image and the temporal difference between the color 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.* [9] 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 [11] 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.* [7] 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.* [15] 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 [16] and [18]. Gao *et al.* [21] 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.* [20] 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.

### 2) Object-based evaluation per frame.

The above-discussed papers all discuss pixel-level evaluation. More recent proposals also discuss the object-based performance evaluation. In [17], 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 [8] 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 the move away again.

### 3) Object-based evaluation over an objects life time.

The previous proposals did not consider tracking of the objects over time. Needham and Boyle [5] 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 [13] 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 strictly limit insight in the actual performance of the tracking algorithm.

Xu and Ellis [22] 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 [12] 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 [27] 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 color 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.

#### 4), 5) Evaluation of object features with higher semantic levels.

We discussed proposals from the literature for the first three mentioned 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 behavior, it seems even simpler to compare the results from a VCA system with the GT. For example, they system should detect a parking car. The VCA 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 behavior need to be handled.

Summarizing, most evaluation proposals are based on pixel-based segmentation metrics only ([6], [7], [9], [11], [14], [15], [18], [20] and [21]). Oberti et al. [16] also consider the object level evaluation, and Mariano *et al.* [17] consider object tracking over time. Nascimento and Marques [8] consider object-based evaluation and introduce splitting and merging of multiple objects. Evaluation of object trajectories is only proposed by the authors of [5], [12] and [13].

Most proposals require manually annotated GT, while a limited set of proposals apply performance evaluation using metrics that do not require GT ([10], [20] and [22]).

### 3.2. Data sets

As mentioned in the beginning of this section, the first requirement for an evaluation of a VCA algorithm is video data. To enable proper benchmarking with other algorithms, it makes sense to evaluate algorithms with standard video data. Moreover, to provide a faithful evaluation, dedicated video is required to invoke the VCA properly. It should be representative and contain both typical and worst-case scenery. Apart from the video content, also the characteristics of the image sensor, the resolution and frame rate have impact on VCA performance. For this purpose, various test video data sets have been made publicly available via the Internet. An overview of these sets is given in Table 1 and screenshots of the sequences are shown in Figure 3. Also non-public sequences are available (like the MPEG-4 Hallway sequence), but because of restricted access and usage, they are less useful for evaluation purposes. In addition to already available data sets, new video sequences were recorded in the CANDELA project. A set of scenarios has been defined for evaluation, and is described in [25]. The video sequences will be publicly available from the CANDELA website.

Each considered scenario has a certain variance. For example, a sequence of a highway will comprise multiple objects that move at large speeds, whereas a parking lot sequence contains only a small number of moving objects at limited speed. To evaluate the full variance of the input data, it is necessary to have several test sequences. The amount of data needed for an evaluation is dependent on the desired level of statistical significance of the test. The availability of data for evaluation is often limited. Considering a complete representation of the input would require time-consuming performance evaluation and large storage requirements for the video data.

Dataset	Available From	Events	Views	Annotation	Compression	Sequences/Seconds
<b>PETS 2000</b> <sup>8</sup>	2000	Outdoor cars moving on car park.	1 SC (a)	No	768x576, 25 fps, JPEG	1/62
<b>PETS 2001</b> <sup>8</sup>	2001	Out door People walking and bicycling,	3 SC (b) (c) (e), 1 OC (d)	Proprietary XML	768x576, 25 fps, JPEG	8/1370
	2001	Cars moving	2 SC in cars (f)	No	768x576, 25 fps, JPEG	2/230
<b>CAVIAR 1<sup>st</sup> set (PETS 2004)</b> <sup>9</sup>	July 2003	In door people walking, browsing, meeting, fighting, leaving objects, Collapsing.	SC (g)	CAVIAR XML	384x288, 25 fps, MPEG-2	28/1057
<b>CAVIAR 2<sup>nd</sup> set</b> <sup>9</sup>	Jan. 2004	In-door shopping center people browsing	2 SC (h) (i)	CAVIAR XML	384x288, 25 fps, MPEG-2	26/650
<b>CANDELA [25]</b>	Nov. 2004	In-door abandoned objects, people walking and interacting	SC (j) (k)	Not yet, MPEG-7	352x288, 12 fps, uncompr.	26/764
	Nov. 2004	Parking situations	SC (l)	Not yet, MPEG-7	352x288, 12 fps, uncompr.	5/232
	Nov. 2004	Intersection situations	SC (m)	Not yet, MPEG-7	352x288, 12 fps, uncompr.	3/171
<b>VS-PETS FOOTBALL INMOVE</b> <sup>10</sup>	2003	Outdoor people tracking in soccer match	3 SC (n) (o)	Yes for camera 3, XML	720x576, 25 fps, JPEG	5/380
<b>FGnet</b> <sup>11</sup> (PETS-ICVS)	2003	Smart meeting, position of face and eyes, facial expression recognition, gesture, face/head direction	2 SC (p) (q) 1 OC (r)	Text	720x576, 25 fps, JPEG	4/1814
<b>VISOR BASE</b> <sup>12</sup> (PETS 2002)	2002	People moving in front of a shop window	1 SC (s)	Not found	640x240, 25 fps, JPEG & MPEG-1	6/274
<b>PETS 2005</b> <sup>13</sup>	2004	Coastal surveillance	Thermal PTZ (t)	CAVIAR XML	720x576, 25 fps, JPEG	7/696
<b>ATON</b> <sup>14</sup>		Highway, campus & intelligent room	1 SC (u) (v) (w) (x)	Intelligent room: binary mask.	320x240, 10 fps, AVI (Cinepak codec)	4/180

**Table 1: Publicly available video datasets (SC = static camera, OC = omni camera, PTZ = pan-tilt-zoom camera). The letters between brackets in the “Views” column refer to Figure 3.**

Furthermore, a distinction between the amount of data used for development and evaluation has to be made, since evaluation on the development set would obviously give unfaithful results. Moreover, to provide viable evaluation and a short algorithm development cycle, the evaluation data set is typically larger than the training set.

<sup>8</sup> PETS test sequences: <http://www.visualsurveillance.org/>

<sup>9</sup> Project IST CAVIAR (IST-2001-37540), EC Funded: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>

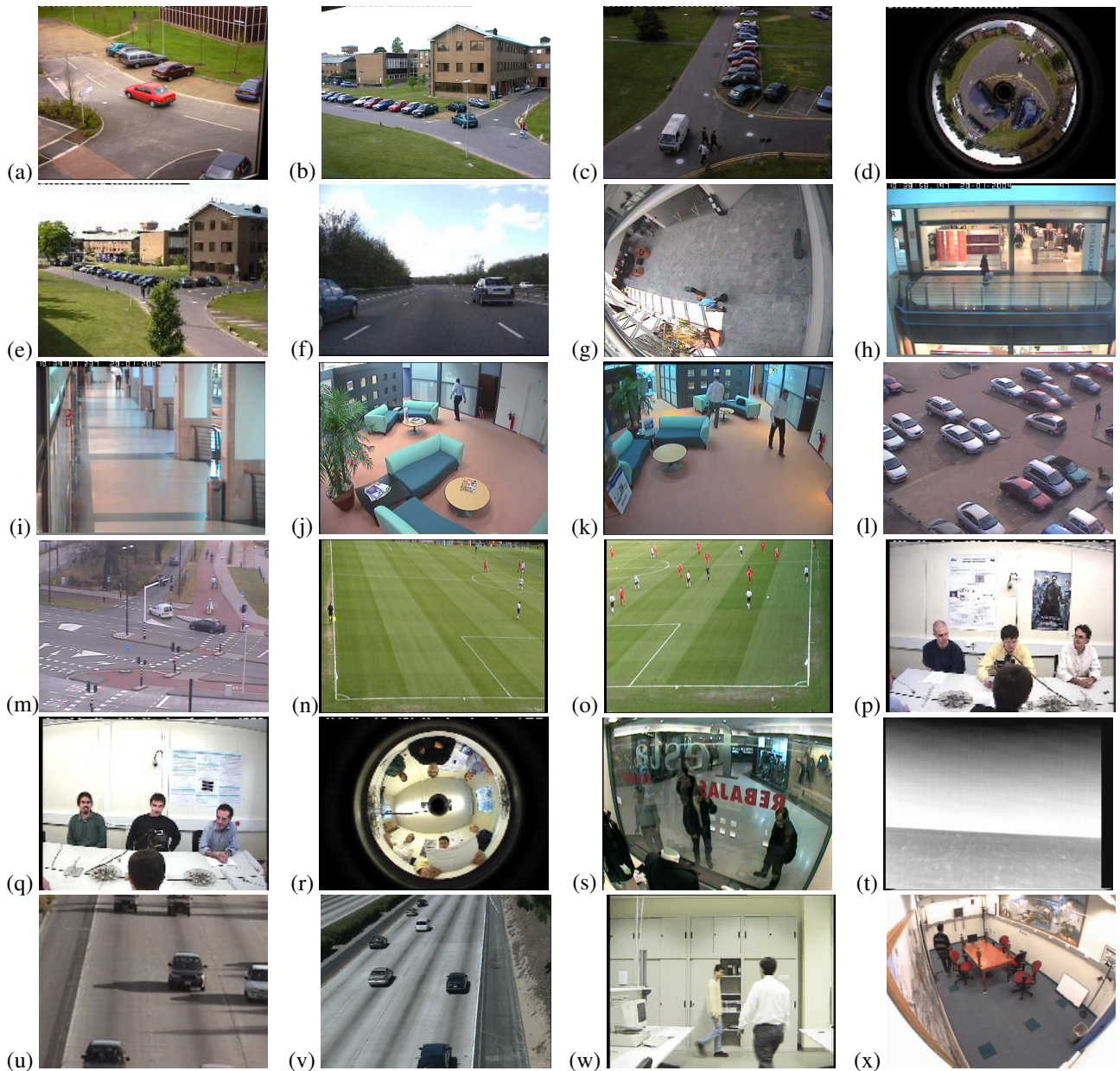
<sup>10</sup> Project IST INMOVE (IST-2001-37422): <http://www.inmove.org>

<sup>11</sup> Project IST FGnet (IST-2000-26434): <http://www.fg-net.org>

<sup>12</sup> Project IST VISOR BASE (IST-1999-10808): <http://www.vtools.es/visorbse/index.html>

<sup>13</sup> The sequences are Copyright (2004) T. Boulton

<sup>14</sup> ATON test sequences: <http://cvrr.ucsd.edu/aton/shadow/>



**Figure 3: Screenshots of the available video data sets.**

Next to real-life video data, also synthetic video data can be used to evaluate VCA algorithms. Typically, the amount of realism of these sequences is limited. However, the process of GT annotation is extremely simplified (see also the next subsection).

Black, Ellis and Rosin [27] present an evaluation framework for using pseudo-synthetic video, which employs video data that has been captured and is stored in a database. The authors evaluated their system using three hundred thousand video frames that are created without any user intervention. The resulting video data is more realistic than completely synthetic data, but it is difficult to reach the amount of realism of a real video sequence. For some algorithms, however, semi-synthetic video sets are useful for evaluation purposes as shown by the authors of [27].

### 3.3. Ground truth and annotation tools

In the previous subsection, video data sets were discussed. To use these video sequences for evaluating VCA algorithms, GT needs to be made available, describing the true properties of the 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 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.* [24] explain that researchers can easily define GT data, visualize the behavior 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 VCA algorithms, is the “Video Performance Evaluation Resource” (ViPER) [17], [19]. Results of evaluation can be visualized.

Collings, Zhou and The [23] 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 website<sup>15</sup>.

Other proposals in literature often mention the use of graphical annotation tools, without much detail. Nascimento and Marques [8] 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 [31]. 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 VCA 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 center 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 VCA 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 VCA algorithms.

Another problem is the occlusion of objects. Black, Ellis and Rosin [27] already mention that this is a difficult issue, since the person that annotates the video has to decide upon the desired behavior of a VCA algorithm is. Should the algorithm keep tracking a (partially) occluded object?

### 3.4. Metrics

Evaluation of video 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

---

<sup>15</sup> Project IST CAVIAR (IST-2001-37540), EC Funded: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>

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 3.1) various metrics have been proposed in the literature. There are mainly two types of metrics: binary and numeric. Binary metrics are used for detection or classification purposes, like false alarm or misclassification. Typical numerical metrics reflect errors in the position, shape or speed of an object, or in the time-delay.

However, because various different metrics are used, the performance evaluation results from VCA 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 criterions. The criterions are precisely defined by metrics which produce an objective global score.

Even if both the VCA 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 behavior is detected, how do we know what the corresponding object, event or behavior 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 VCA and GT that are spatially closest together be compared? What if the VCA 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 VCA 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 VCA 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 color 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 colors. Defining the matching rules for performance evaluation is an open research topic to be explored.

### **3.5. Presentation of results**

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.

## **4. CONCLUSIONS AND FUTURE WORK**

In this paper, we have shown a general framework for video content analysis (VCA) algorithms and the storage of the generated metadata, enabling search and retrieval and performance evaluation.

Four main topics for evaluation of a VCA system have been defined; creation of ground-truth data, available evaluation data sets, performance metrics and presentation of the evaluation results.

We have given an overview of state-of-the-art methodologies for performance evaluation of real-time VCA systems on several semantic levels. Most proposals in literature apply evaluation on pixel-based segmentation level. Some perform evaluation on object-level and only few consider the tracking of unique objects over time. However, a literature study for higher semantic level evaluation did not reveal a lot of relevant work. We addressed the need for further research for this higher-level evaluation, evolving towards evaluation of human-understandable video descriptions.

To enable objective evaluation, standard data sets have to be considered. We have listed several publicly available video data sets. In addition, new video data has been created within the CANDELA project, dedicated to surveillance and traffic applications. To support the performance evaluation needs, this set will be made publicly available.

In order to evaluate results from VCA algorithms, they need to be compared with the true description, the so-called ground truth (GT). To create this GT, annotation tools have been developed. However, various GT description formats are used, causing incompatibility and limit reuse across the available performance evaluation tools. Working towards a common GT format, we have addressed the need for standardization of the features stored in these GT formats and propose the use of the MPEG-7 metadata description format.

The actual comparison of the VCA results with the GT is done by applying metrics. In literature, various metrics have been proposed that all address certain behavior of VCA algorithms. However, there is no standard set of metrics that is used in every performance evaluation, which prevents the objective comparison of multiple algorithmic evaluations available in literature.

Furthermore, the representation of the results for each metric needs to be defined, in order to combine intermediate results over time. Some proposals in literature use histograms, others use statistics (minimum, maximum, average or median) or summarize the intermediate results.

Summarizing, in the four main topics required for an objective evaluation of VCA algorithms, quite some standardization effort is required. This yields for the available evaluation data sets, the GT description format, comparison metrics and their representations.

## 5. REFERENCES

- [1] 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.
- [2] A.Cavallaro, D. Douchamps, 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.
- [3] F. Cupillard, F.Br mond and M. Thonnat, "Tracking groups of people for video surveillance", *2<sup>nd</sup> European Workshop on AVBS Systems (AVBS2002)*, University of Kingston, London, UK, Sept. 2001.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] A. 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.
- [10] 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.
- [11] 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.

- [12] 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.
- [13] 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.
- [14] Y. J. Zhang, "A survey on evaluation methods for image segmentation", *Pattern Recognition*, Vol. 29, Num. 8, pp 1335—1346, Aug. 1996.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] D. Doennann and D. Mihalcik, "Tools and techniques for video performance evaluation", *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, Vol. 4, pp 167—170, Barcelona, Spain, Sept. 2000.
- [20] 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.
- [21] X. Gao, T.E. Boult, F. Coetzee 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.
- [22] 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.
- [23] R. Collins, X. Zhou, S. K. The, "An Open Source Tracking Testbed and Evaluation Web Site", *Proc. of 6<sup>th</sup> IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Breckenridge, Colorado, January 7 2005.
- [24] C. Jaynes, S. Webb, R. Matt Steele and Q. Xiong, "An Open Development Environment for Evaluation of Video Surveillance Systems", *Proc. of 3<sup>rd</sup> IEEE Int. Workshop on Performance Evaluation and Tracking and Surveillance (PETS)*, pp 32—39, June 1 2002.
- [25] R.G.J. Wijnhoven, "Scenario Description - Technical Document v.0.6", CANDELA Project, Bosch Security Systems B.V., Eindhoven, The Netherlands, 2004.<sup>16</sup>
- [26] 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.
- [27] 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.
- [28] D. Makris, T.J. Ellis and J. Black, "Learning scene semantics", *ECOVISION 2004, Early Cognitive Vision Workshop*, Isle of Skye, Scotland, UK, May 2004.
- [29] 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.
- [30] 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.
- [31] T. List and R.B. Fisher, "CVML – An XML-based Computer Vision Markup Language", *International Conference for Pattern Recognition*, Cambridge, UK, Aug. 2004.

---

<sup>16</sup> Scenario description document available from <http://www.extra.research.philips.com/euprojects/candela/>