ETISEO

EVALUATION
DU **T**RAITEMENT ET DE L'**I**NTERPRETATION
DE **S**ÉQUENCES VID**EO**.

(Evaluation for video understanding)

Introduction to evaluation and metrics

Version 1.04

Référence : E01

Introduction

The objectives of this document are:

- Firstly, to present the ETISEO project, its objectives, partners, planning and tasks in order to inform potential participants.

- Secondly, to introduce the video sequence selection criteria and the metrics expected to be used during the project. Active participants will be invited to contribute to the update of this document and to the definition of the optimal metrics and associated video sequences (see. § Video sequence database §Evaluation criteria and metrics).

Table of contents

ETISEO

1. PROJECT OBJECTIVES

ETISEO (Evaluation du Traitement et de l'Interprétation de Séquences Vidéo) is a two-year project, starting in January 2005 and sponsored by the French government in order to evaluate vision techniques for video surveillance. All participants from all over the world are welcome to participate to this project. The project is based on a voluntary basis: participants are expected to test their algorithms on their own and to send their results to the evaluator. Nevertheless, all participants will be welcome to attend freely workshops organized during the project. Each participant will naturally benefit of the resources generated by the project team, coordinator, data providers and scientific leader. At the end of the project, video database with ground truth and metrics with evaluation tools will be provided to participants.

Video surveillance is an important application of computer vision. Many years of research and experimentations has led to the development of innovative commercial applications. Nevertheless, video sequence analysis and interpretation is still a very active research area. Algorithm robustness must still be improved and dependencies between algorithms and their conditions of use must be reduced. In the meantime, the maturity of this technology favours the realization of a comparative study between existing methods. Conclusions of this study will act as a basement for future work and research in this field. Strengths and weaknesses of algorithms as well as unsolved problems will be highlighted. The commitment of industrials into the evaluation process will help us to determine performances of software currently used in applications and to identify where they should be improved. As a consequence, a first step towards a standardisation of video surveillance systems will be achieved. We list hereunder the four main project objectives:

1. To acquire precise knowledge of vision algorithms, which is essential in order to diagnose technological limits and to bring appropriate solutions. This means that we must have several comparison points throughout the processing chain: detection of physical objects of interest, classification of physical objects of interest, tracking of physical objects of interest and event recognition. During meetings, participants will define tools allowing such detailed measures together with fine annotations.

2. To create a space for discussion among participants (both research labs and companies) in order to obtain a large agreement on criteria for video reference selection and evaluation criteria used to assess vision algorithm performances. These criteria will be defined during project meetings. This large agreement is mandatory for criteria to be considered as a standard in the domain. Also, effective relations will be established with existing evaluation projects in this domain (e.g., VACE, VITAL, PETS).

3. To create two ontologies, which will ease communication between all participants in this domain: researchers, software developers and end-users (e.g., administrations, companies). The first one will describe technical concepts used in the whole video interpretation chain (e.g., a blob, a mobile object, an individual trajectory) as well as concepts associated to evaluation criteria. The second one will describe concepts of the application domain (e.g., a bank attack event). These ontologies will lead to a comparison standard for vision algorithms.

4. To conceive automatic evaluation tools for vision algorithms to allow a fair and quantitative comparison between algorithm results and reference data. A large and meaningful database of video sequences and their annotations will be created. They will represent the diversity of video surveillance applications and will be classified according to criteria corresponding to limitations of existing vision algorithms. For instance, concerning shadow detection, the database will contain video sequences illustrating several types of shadow: shadows generated under different illumination sources, shadows more or less contrasted, shadows on a textured ground... in order to determine the reactivity of an algorithm to a specific problem. In counterpart, annotations will be defined with respect to the functionality under consideration (e.g., mobile object tracking). These annotations will contain reference data as well as a detailed description of the tested video sequence to allow a fine diagnosis of encountered problems. Finally, the automatic evaluation tool will analyse accurately how a given algorithm manages a given problem. For instance, dynamic occlusions will be computed on the basis of reference data and the number of correctly managed dynamic occlusions by a given tracking algorithm will be computed.

The evaluation tool together with a large agreement on evaluation criteria will enable a clear view on current performances in this domain. Also, this project should lead to the acquisition of precise knowledge of vision algorithms and thus the development of robust and generic intelligent video surveillance systems.

2. PARTNERS

All teams from all over the world are welcome to participate at the evaluation process at any time during the ETISEO project. Currently, there are several involved teams: SILOGIC (France), INRIA (France), INRETS-LEOST (France), CEA-List (France), LASL (Laboratoire d'Analyse des Systèmes du Littoral, Université du Littoral Côte d'Opale) (France), Vision-IQ (France), University of Reading (UK), Kingston University (UK), Multitel (Belgium). There are several roles defined in ETISEO: data provider, scientific leader, evaluator, project coordinator and participant. Participants are teams that have already developed algorithms and want to test their technology on their own. The role of each team is represented in the figure hereunder. The participating roles will be defined for the first seminar.

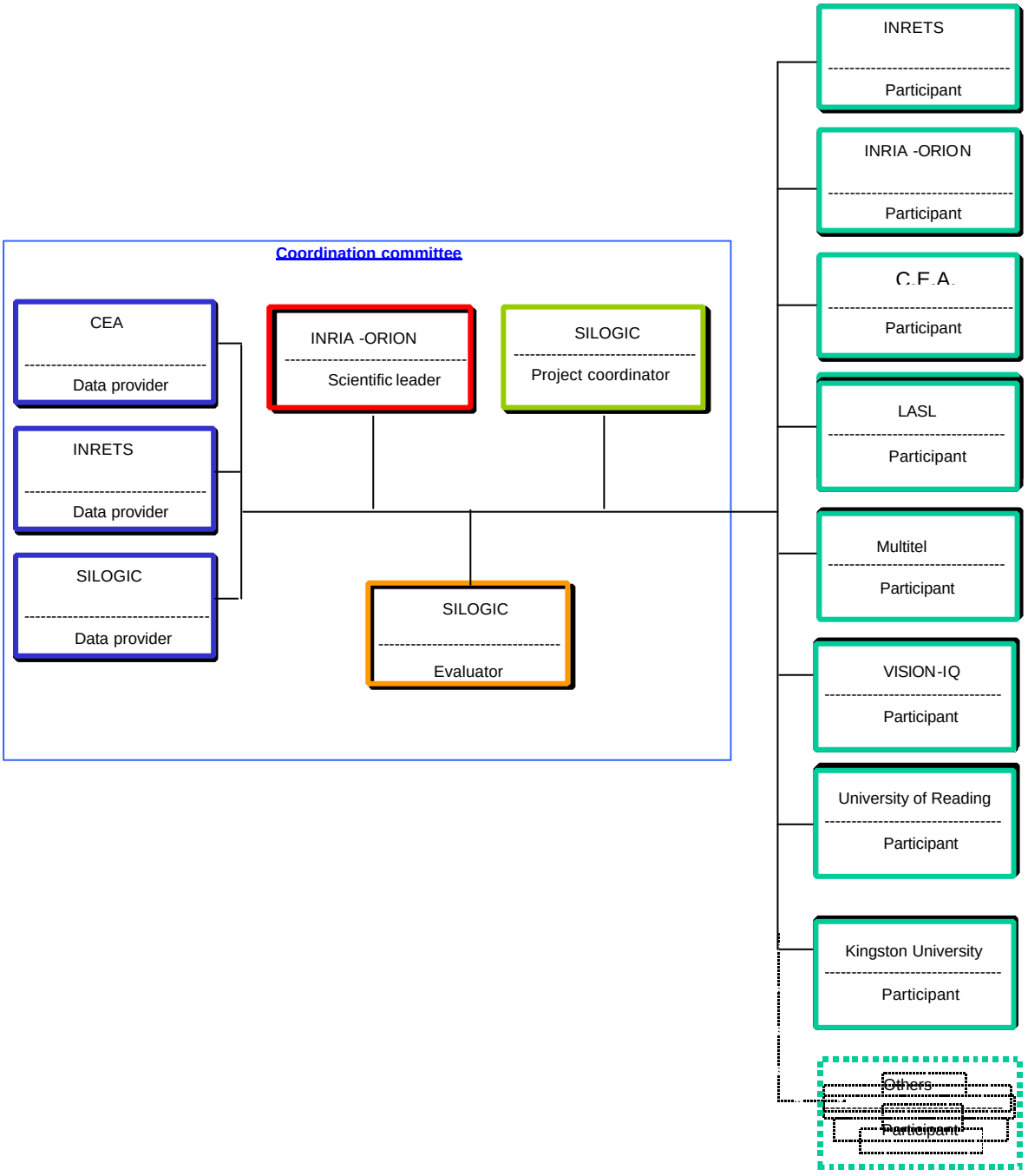

Figure 1: Partners roles

The ORION team from INRIA Sophia-Antipolis is the scientific leader of the ETISEO project. This role includes five tasks:

- To spread out largely ETISEO objectives to the scientific community in order to obtain the maximum number of participants.
- To organize ETISEO scientific meetings. This task means to create and guide discussions and to collect information and/or decisions as objectively as possible.
- To propose evaluation criteria and metrics, to discuss them with participants and to incorporate proposed modifications.
- To develop an automatic evaluation tool, which uses evaluation criteria collectively defined during meetings. This tool will be given to the evaluator to allow him to evaluate algorithms impartially. This tool is needed to be able to process the large number of video sequences collected during the project, to compare in a unique manner all algorithms from all participants.
- To write scientific reports describing the evaluation process conducted in the ETISEO project (criteria and metrics choice, evaluation results) and to disseminate results in the international scientific community.

SILOGIC is the ETISEO coordinator and the official contact for the ministry. Being also the ETISEO evaluator, SILOGIC will not act as a participant. Objectivity and impartiality are of prime importance and must be acknowledged as rules for the entire evaluation process. The evaluator (SILOGIC) will be in charge of:

- Interface development to link participant algorithms with evaluation tools.
- Data collecting and centralisation from the various providers.
- Annotations and reference data definition.
- Data and tools dissemination towards participants.
- Result collecting from participants.
- Result evaluation.
- Result synthesis together with the scientific leader.

3. PLANNING

WP 1 (**Evaluation initialisation**) defines the environment and procedures for the entire project. The coordination committee sets up rules and initialises specification documents, which will be discussed during the first meeting. Each participant will have to follow defined procedures in order to guarantee a good achievement of the project goals.
WP 2 (**Metrics and tools creation**) defines and validates criteria and metrics, creates evaluation tools, which compare algorithm results with ground truth data. The evaluation tool will use criteria, which have been discussed and defined collectively by all participants during meetings. Tools will be disseminated at the end of the project. During this work package, interface software will be developed and distributed to participants to help them to use data and to allow their algorithms to be evaluated (e.g., to output their results in the appropriate format required by the evaluator).
WP 3 (**Data collecting**) is realized by the various data providers. Data will be collected in order to satisfy the established constraints (e.g., diversity, significance). They will be centralized and managed by the evaluator, which will be responsible of their dissemination to participants. Video sequence data will be organized in three distinct data sets: work, test and evaluation data set. In addition, annotations will be defined by SILOGIC.
WP 4 (**Evaluation cycle validation**) is planned before the final evaluation of all participant algorithms. It is intended to verify protocols, technical details as well as the pertinence of criteria and metrics. It is based on a specific data set. A first comparison of algorithm performances is achieved at the end of this work package.
WP 5 (**ETISEO evaluation**) is the evaluation of the analysis and interpretation of video sequences. All data collected during the whole project will be provided to the participants for the evaluation of their work in video-surveillance. Finally, results will be published.

W7P No	WP Description	Result Description	T0+
WP 1	**Evaluation initialisation**		
WP 1.1	Internal initialisation	Procedures, rules and plan	2
WP 1.2	Criteria and reference data study	Study on criteria and reference data (draft version)	4
WP 1.3	Definition seminar	Study on criteria and reference data (approved version)	6
WP 2	**Metrics and tools creation**		
WP 2.1	Comparison tools and metrics	Comparison tools and metrics	12/16
WP 2.2	Interface software for participants	Interface software for participants	12
WP 3	**Data collecting**		
WP 3.1	Collect & diffusion of work data set	Work video data set	9
WP 3.2	Collect & diffusion of test data set	Test video data set	12
WP 3.3	Collect & diffusion of evaluation data set	Evaluation video data set	18
WP 3.4	Annotation definition	Annotations	12/18
WP 4	**Evaluation cycle validation**		
WP 4.1	Seminar to launch the evaluation process	Intermediate report	12
WP 4.2	Tests by participants	Test report	15
WP 4.3	Result collecting by the evaluator	Algorithm results from participants	15
WP 4.4	Analysis of the results and the evaluation cycle	Intermediate result evaluation Evaluation cycle report	18
WP 5	**ETISEO evaluation**		
WP 5.1	Tests by participants	Test report	21
WP 5.2	Result collecting by the evaluator	Algorithm results from participants	21
WP 5.3	Result analysis and evaluation report	Result evaluation from participants	22
WP 5.4	Workshop for evaluation result	Final evaluation report	23
WP 5.5	Result dissemination	Coordination committee	24
WP 6	**Coordination and management**	Activity report	

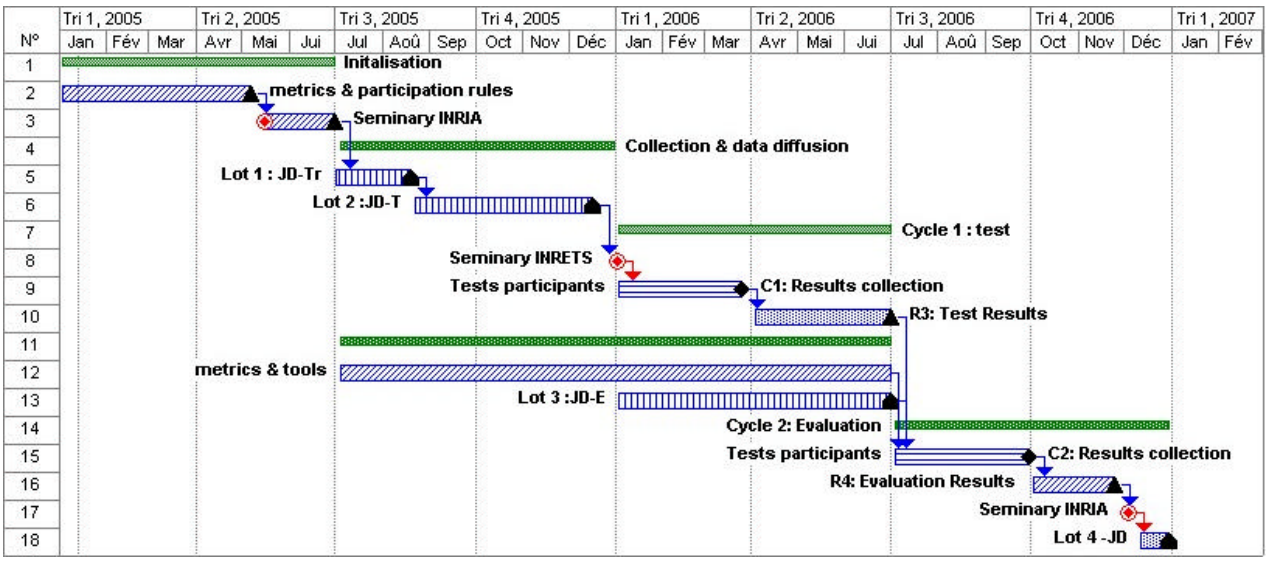

Figure 2: ETISEO planning

T0 is set to January 2005, the first and the general ETISEO workshop to define video criteria and evaluation metrics will take place in Sophia-Antipolis and is planned on May 10 and 11th, 2005.

4. DATA TERMINOLOGY

This section enumerates and defines all the vocabulary describing data used in a video understanding process.

Image: array of pixel generated at a time step by a video camera (e.g., composite, CCD, CMOS, PTZ, omni directional). An image is characterized by a timestamp (year, month, day, hour, minute, second, millisecond) and can correspond to a frame interleaved or not. An image can be of the following type: colour, black and white, infrared and with different compression levels.

Video sequence: temporal sequence of images which are generated by a video camera. A video sequence can be represented as a live stream (e.g., composite signal, MJPEG stream), as a file (e.g., a MPEG4 encoded file) or as a sequence of files (e.g., a sequence of JPEG files).

Video clip: a part of a video sequence, which corresponds to a particular situation to be evaluated.

Scene: the physical space where a real world event occurs and which can be observed by one or several video cameras. A scene without any physical object of interest is called an **empty scene**.

Blob: 2D image region that has been segmented based on regions (e.g., homogeneous in motion, colour, energy or texture information) or contours (e.g., using a shape model). This region can be defined as a set of pixels (not necessarily connected) or as a polygon delimiting its contour. It can be characterized by 2D features such as a colour histogram, a density, a 2D width and height.

Moving region: a blob that has been created following a motion criteria (e.g., either optical flow or reference image subtraction).

Physical object: a real world object in the scene. There are two types of physical objects: **physical object of interest** and **contextual object**.

Physical object of interest: a physical object evolving in the scene whose class (e.g., person, group, vehicle) has been predefined as interesting by end-users and whose motion cannot be foreseen using a priori information. It is usually characterized by a semantic class label, 2D or 3D features (e.g., 3D location, width and height, a posture, a trajectory, a direction, a speed), a list of blobs, an initial tracking time, a camera number for the camera which is the best seeing the object (in a multi camera configuration), an identifier. An identifier can either be defined locally to the current image, globally on the video sequence or globally on a scene (in a multi camera configuration).

Contextual object: a physical object attached to the scene. The contextual object is usually static and whenever in motion, its motion can be foreseen using a priori information. For instance, it can be in motion such as a door, an elevator, a fountain, a tree or displaceable (by a human being) such as a chair, a luggage.

Event: generic term to describe any event, action or activity happening in the scene and visually observable by cameras. Events of interest can either be predefined by end users or learned by the system. Events are characterized by the involved objects of interest (including contextual objects and zones of interest), their starting and ending time and by the cameras observing the events. Examples of events are “intrusion in a forbidden zone”, “detection of an abandoned bag”, “detection of a fighting situation”, ”a meeting between two people”...

5. VIDEO UNDERSTANDING FUNCTIONALITIES

Video understanding is a process recognising user events in a scene observed by video cameras. The whole processing chain goes from pixel analysis up to alarm generation and is composed of four main video processing tasks:

- Task 1: detection of physical objects of interest: decomposition of the image into blobs (2D regions) corresponding to potential physical objects of interest. A typical approach consists in separating first moving pixels from non-moving pixels and then clustering moving pixels into blobs. A moving pixel is a pixel whose intensity is sufficiently different from the corresponding pixel in a reference image (e.g., background or previous image). Advanced functionalities consist in being able to distinguish interesting moving pixels generated by human activities from those corresponding to noise generated by contextual objects (e.g., moving trees), shadows or reflections. These advanced functionalities may require the use of contextual information (e.g., 3D geometry of the empty scene), a sophisticated reference image or chromatic information about pixels. The output of task 1 is a grey level image (0 = background, n = identifier of the physical object of interest).

- Task 2: classification of physical objects of interest: classification of blobs into labels corresponding to classes of physical objects of interest, with respect to a predefined semantic: person, vehicle, group of persons, etc. Advanced functionalities consist in refining object classes (e.g., motorcycle, cycle, car, truck, airplane, for the vehicle class), in splitting objects (e.g., two separate persons are better than a group), in computing a posture and an orientation for objects, in computing their 3D parameters while taking into account static occlusions by contextual objects. The output of task 2 is a list of physical objects with their properties.

- Task 3: tracking of physical objects of interest: process which consists in matching objects detected at image time t-1 with those detected at image time t and maintaining a unique identifier for each object over the whole video clip. Advanced functionalities consist in tracking separately rather than globally physical objects in case of dynamic occlusion, in tracking accurately objects even in case of static occlusion, in tracking objects in a network of cameras with overlapping or distant field of view. The output of task 3 is a list of physical objects with their properties (the camera having the best viewing point to observe the physical objects, trajectory, kinematics, time-filtered properties) and their links to previous objects.

- Task 4: event recognition: the goal of this step is to recognize any event predefined by the user (abandoned bag, forbidden zone access, attack) from descriptors given by preceding tasks (e.g., shape, speed, position and trajectory). An event is characterized by involved objects, the event recognition initial time and the camera having the best viewing point (in a multi camera configuration).

There are many ways to implement these tasks. Some systems only address task 1, some systems combine the first three tasks into a single task while others merely skip task 2. Video understanding systems may address globally these four main tasks. However, when it is possible, we propose to evaluate them at the end of each task. The output of task 4 is a list of events together with their involved physical objects.

6. VIDEO SEQUENCE DATABASE

From a general point of view, collected data will correspond to views of the real world thus they will represent real difficulties encountered during software and application development. They will illustrate video surveillance applications from indoor and outdoor scenes including persons and vehicles. The configuration can be mono or multi cameras and sensors can be black and white, colour or infrared ones. It should be noted that calibration information will be available (in a format still to be defined).

The diversity of the ETISEO database is essential to obtain a meaningful algorithm evaluation. To reach this goal, different providers will create data. Each video clip will last several minutes to record a full event occurrence. The database will contain instances of increasing levels of difficulties in several categories. For instance, if the studied problem is the management of crossings between people, the database will contain sequences ranging from crossings with 2 persons to crossings implying at least 5 persons. The current video sequence categorization is as follows:

V1) Acquisition information:

V1.1) Camera configuration: mono or multi cameras,

V1.2) Camera type: CCD, CMOS, large field of view, thermic cameras (infrared),

V1.3) Compression ratio: no compression up to high compression,

V1.4) Camera motion: static, oscillations (e.g., camera on a pillar agitated by the wind), relative motion (e.g., camera looking outside a train), vibrations (e.g., camera looking inside a train),

V1.5) Camera position: top view, side view, close view, far view,

V1.6) Camera frame rate: from 25 down to 1 frame per second,

V1.7) Image resolution: from low to high resolution,

V2) Scene information:

V2.1) Classes of physical objects of interest: people, vehicles, crowd, mix of people and vehicles,

V2.2) Scene type: indoor, outdoor or both,

V2.3) Scene location: parking, tarmac of airport, office, road, bus, a park,

V2.4) Weather conditions: night, sun, clouds, rain (falling and settled), fog, snow, sunset, sunrise,

V2.5) Clutter: empty scenes up to scenes containing many contextual objects (e.g., desk, chair),

V2.6) Illumination conditions: artificial versus natural light, both artificial and natural light,

V2.7) Illumination strength: from dark to bright scenes,

V3) Technical issues:

V3.1) Illumination changes: none, slow or fast variations,

V3.2) Reflections: reflections due to windows, reflections in pools of standing water, reflections due to bright floors,

V3.3) Shadows: scenes containing weak shadows up to scenes containing contrasted shadows (with textured or coloured background),

V3.4) Moving Contextual objects: displacement of a chair, escalator management, oscillation of trees and bushes, curtains,

V3.5) Static occlusion: no occlusion up to partial and full occlusion due to static contextual objects,

V3.6) Dynamic occlusion: none up to a person occluded by a car, a person occluded by another person,

V3.7) Crossings of physical objects: none up to high frequency of crossings and high number of implied objects,

V3.8) Distance between the camera and physical objects of interest: close up to far,

V3.9) Speed of physical objects of interest: stopped, slow or fast objects,

V3.10) Posture/orientation of physical objects of interest: lying, crouching, sitting, standing,

V3.11) Calibration issues: little or large perspective distortion,

V4) Application type:

V4.1) Primitive events: enter/exit zone, change zone, walking, running, following someone, getting close,

V4.2) Suspicious behaviour detection: violence, fraud, tagging, loitering, vandalism, stealing,

V4.3) Intrusion detection: person in a sterile perimeter zone, car in no parking zones,

V4.4) Monitoring: traffic jam detection, counter flow detection, home surveillance, abandoned bag,

V4.5) Statistical estimation: people counting, car speed estimation,

Other video sequence characteristics can be investigated (such as PTZ camera, omni directional cameras, stereo vision, aerial view) however; only above characteristics will be selected due to time limitation.

The video sequence database will be classified into three sets: work, test and evaluation set. The **work data set** is representative of the various sequences contained in the database. It is distributed to participants to allow them to run, modify and adjust their algorithms the way they want. In order to give participants the maximum amount of time, a non-exhaustive data sub set will be distributed at the beginning of the collecting phase of the first data set. The **test data set** is created and distributed at the beginning of the evaluation cycle validation. It is representative of the next coming evaluation set. It contains various sequences illustrating several cases predefined in the evaluation process. The analysis of the comparison results will enable to qualify the pertinence of the chosen video sequences. The **evaluation set** is intended to assess performances of participant algorithms. It will contain the same variety of sequences as the test set but with more video clips in order to obtain sound statistical evaluation results.

Several characteristics of the tested algorithms cannot be automatically evaluated by a comparison with reference data. For instance, we can mention the processing time, the memory space usage, the amount of interactivity required, and the need of a learning phase. A questionnaire will be established during the first seminar. It will be distributed to participants along with data and they will send it back with their algorithm results. This information will enable a more detailed analysis. Finally, a live demonstration could be organized during the last workshop to run a participant algorithm on the data set and in presence of other participants and the organising committee. This opportunity will be investigated during the first seminar.

7. VIDEO UNDERSTANDING EVALUATION

This section enumerates and defines all the vocabulary used for the evaluation of a video understanding process as well as the considered process of automatic supervised evaluation.

Evaluation criterion: an evaluation criterion is an evaluation functionality to compare video understanding algorithm results with reference data. For instance, for the task “detection of physical objects of interest”, a criterion can evaluate the accuracy of the 2D or 3D location of objects, another one can evaluate the quality of the object shape. For the task “classification of physical objects of interest”, a criterion can evaluate the quality of the assigned class labels. In addition, these criteria can be detailed with regard to video clip categories. In the previous example, the assignment of class labels under static occlusion situations could be qualified, for instance.

Evaluation metric: a distance between video understanding algorithm results and reference data implementing an evaluation criterion. A way of displaying evaluation results is to use a ROC (Receiver Operating Characteristic) curve defined as a plot of the true positive rate against the false positive rate.

Ground truth data: data given by a human operator and which describe real world expected results (e.g., physical objects, events) at the output of a video understanding algorithm. These data are supposed to be unique and corresponding to end user requirements even if in many cases, this information can contains errors (annotation bias). These data can be written in a XML or MPEG7 format.

Annotation: information associated to a video clip including ground truth data plus other types of information about technical difficulties (e.g., shadows) and recording conditions (e.g., weather conditions) of the video clip under consideration. These annotations can provide several types for false or incorrect results (e.g., wrong classification, wrong detection).

Reference data: data supposed to be constant and unique, corresponding to a functionality of a video understanding task and used to evaluate the output of a video understanding algorithm at a given task level. Reference data include ground truth data, data given by a video expert and data computed from all annotation and contextual information. For instance, the 3D position of a person is a reference data computed from the bounding box given by a video expert and the calibration matrix. In addition, rules should be given to video experts in order to define as objectively as possible particular data. For instance, for a partially occluded person, one can choose to draw the bounding box for the visible part only or for the full object (including its hidden part).

Automatic supervised evaluation: process assessing algorithm performances in an automatic manner by comparing algorithm outputs with reference data. Automatic evaluation means that all the comparison is done without human interaction. In the automatic case, criteria and metrics are predefined and encoded into the evaluation system. Supervised evaluation means that video experts and human operators provide reference data used for the comparison.

In ETISEO, we are interested in an automatic supervised evaluation. The overall description of this evaluation process is depicted in figure 1.

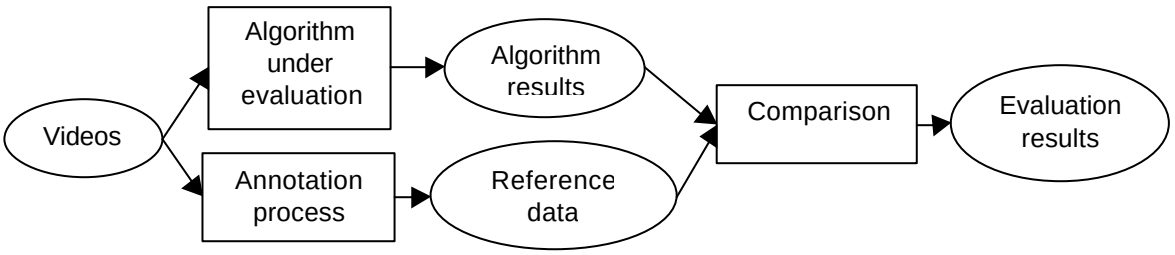

Figure 1: evaluation process of video understanding algorithms

The prerequisite to obtain an automatic evaluation based on a reference data comparison is to define reference data through the annotation process. Then, a comparator, which uses evaluation criteria and metrics, must be designed. Finally, this comparator (called evaluation tool) can be run on a video understanding algorithm to produce evaluation results quantifying the adequacy between algorithm outputs and reference data. This evaluation is run at each step (i.e., at each image for a mono camera processing) and its goal is to highlight algorithm capability to solve a set of current problems (e.g., shadow management, occlusions, object crossings) for each task of the video understanding chain. The evaluation process requires taking a decision concerning three topics: video database, annotation, evaluation criteria and metrics.

Results will be presented in two manners:

1. Detailed: to obtain a clear and precise view on performances of the various algorithms at several points in the video interpretation chain. Results will be synthesized into three arrays:
   - True Positives: results matching reference data.
   - False Positives: results not matching reference data.
   - False Negatives: results missing with respect to reference data.

   For each array, a column will present results according to a category of video sequences (see section 5) while a row will present results according to a task or a sub task (see section 4).

   We describe hereunder a synthetic view of these concepts and their relationships, as well as evaluation qualificators: precision, sensitivity and specificity.

	Reference Data (RD)	Noise (N)	
Computed	True Positive (TP)	False Positive (FP)	Precision = TP/(TP+FP)
Not Computed	False Negative (FN)	True Negative (TN)	
	Sensitivity = TP/(TP+FN)	Specificity = TN/(FP+TN)	

2. Global: to compare globally all algorithms and to obtain a meaningful measure of their performances. For instance, it can be the minimum between the sensitivity and the precision.

8. EVALUATION CRITERIA AND METRICS

This section describes criteria and metrics, which will be used by the evaluation tool, for each task.

The elements are defined in collaboration with participants.

Criteria and metrics list:

- **T1) Detection of physical objects of interest:**
    - C1.1) Criterion qualifying the number of blobs corresponding to physical objects of interest,
    - C1.2) Criterion qualifying the area corresponding to physical objects of interest, globally in an image.
- **T2) Classification of physical objects of interest:**
    - C2.1) Criterion qualifying the classification of physical objects of interest,
    - C2.2) Criterion qualifying the properties of physical objects of interest in case of isolated physical objects correctly detected.
- **T3) Tracking of physical objects of interest:**
    - C3.1) Criterion qualifying the frame-to-frame tracker,
    - C3.2) Criterion qualifying the long-term tracker,
    - C3.3) Criterion qualifying the long-term tracker properties,
- **T4) Event recognition:**
    - C4.1) Criterion qualifying the recognition of events globally on a video clip or on a scene in a multi cameras configuration,
    - C4.2) Criterion qualifying event properties,

T1) Detection of physical objects of interest:

	Reference Data (RD)	**Noise (N)**	
Object Detection (OD)	True Positive (TP) **Good Detection (GD)**	False Positive (FP) **False Detection (FD)**	**Precision = GD/(GD+FD)**
No Detection (ND)	False Negative (FN) **Miss Detection (MD)**	True Negative (TN) **False Detection Rejection (FDR)**	
	Sensitivity = GD/(GD+MD)	**Specificity = FDR/(FD+FDR)**	

C1.1) Criterion qualifying the number of blobs corresponding to physical objects of interest:

- M1.1.1: number of detected blobs compared to reference data using their bounding box. The matching between blobs and reference data can be done by checking if the overlapping area is above a given threshold.

  - GD = reference data having a sufficient overlap with blobs.

  - FD = blobs having no sufficient overlap with reference data.

  - MD = reference data having no sufficient overlap with blobs.

  - OD = GD + FD: all detections.

  - RD = GD + MD: all reference data.

  - **Precision**: number of GD / number of OD.

  - **Sensitivity**: number of GD / number of RD.

- M1.1.2: same as metric M1.1.1 but using the physical object shape instead of their bounding box for the comparison between blobs and reference data.

C1.2) Criterion qualifying the area corresponding to physical objects of interest, globally in an image:

- M1.2.1: computation between bounding boxes of detected blobs and reference data regions, for each image. Percentages for a video clip are computed as the sum of percentages per image divided by the number of images containing at least one reference data.

  - GD = pixels belonging to both the reference data set and the blob set.

  - FD = pixels belonging to the blob set but not to the reference data set.

  - MD = pixels belonging to the reference data set but not to the blob set.

  - FAR = pixels belonging neither to the reference data set nor the blob set.

  - OD = GD + FD: all detections.

  - RD = GD + MD: all reference data.

  - **Precision**: number of GD / number of OD.

  - **Sensitivity**: number of GD / number of RD.

  - **Specificity**: number of FDR / number of N.

- M1.2.2: same as metric M1.2.1 but using the shape of physical objects of interest instead of the bounding box.

In addition, other metrics can characterize the detection accuracy: dominant colours, texture...

T2) Classification of physical objects of interest:

	Reference Data (RD)	**Noise (N)**	
Object Classification (OC)	True Positive (TP) **Good Classification (GC)**	False Positive (FP) **False Classification (FC)**	**Precision = GC/(GC+FC)**
No Classification (NC)	False Negative (FN) **Miss Classification (MC)**	True Negative (TN) **False Classification Rejection (FCR)**	
	Sensitivity = GC/(GC+MC)	**Specificity = FCR/(FC+FCR)**	

C2.1) Criterion qualifying the classification of physical objects of interest. It can be computed by counting the number of physical objects having a correct type in order to obtain GD, FA, MD in situations where physical objects are correctly detected:

- M2.1.1: number of correctly classified physical objects of interest.

  - GC = physical objects of interest correctly classified.

  - FC = physical objects of interest wrongly classified.

  - MC1 = physical objects not classified due to classification shortcomings (e.g., unknown).

  - MC2 = physical objects not classified due to detection or classification shortcomings (e.g., lack of contrast).

  - OC = GC + FC: all detections.

  - RD1 = GC + MC1: correctly detected reference data.

  - RD2 = GC + MC2: reference data.

  - **Precision**: number of GC / number of OC.

  - **Sensitivity1**: number of GC / number of RD1.

  - **Sensitivity2**: number of GC / number of RD2.

- M2.1.2: same as metric M2.1.1. but compare the physicial objects of interest at different levels of the class hierarchy. For instance, a motorcycle classified as vehicle can be a GC at a higher level of the class hierarchy even if it is a FC at a lower level.

- M2.1.3: same as metric M2.1.1. but for reference data with a composed type (e.g., to check the ability of classifying groups of people).

- M2.1.4: same as metric M2.1.1. but for reference data with only elementary types in case of dynamic occlusions. This metric checks the ability of separating an object of composed type (e.g., group of people) in two or more objects of elementary type (e.g., person) when the video clip contains several physical objects side by side.

C2.2) Criterion qualifying the properties of physical objects of interest in case of isolated physical objects correctly detected (the metric will be computed separately for each type of physical objects):

- M2.2.1: average of the 3D distance between centres of gravity of physical objects and reference data.

- M2.2.2: average of the 3D distance between bottom point of physical objects and reference data.

- M2.2.3: average of the 2D distance between bottom point of physical objects and reference data bounding boxes.

- M2.2.4: average of the density difference between shapes of physical objects and reference data.
- M2.2.5: average of the 3D width, height and ratio difference between physical objects and reference data.
- M2.2.6: average of the 3D posture difference between physical objects and reference data. Distances for postures will be defined accordingly.
- M2.2.7: average of the 3D orientation difference between physical objects and reference data.
- M2.2.8: average of the difference of the number of physical objects for a specific type (e.g., person) and reference data.

T3) Tracking of physical objects of interest:

	Reference Data (RD)	**Noise (N)**	
Object Tracking (OT)	True Positive (TP) **Good Tracking (GT)**	False Positive (FP) **False Tracking (FT)**	**Precision = GT/(GT+FT)**
No Tracking (NT)	False Negative (FN) **Miss Tracking (MT)**	True Negative (TN) **False Tracking Rejection (FTR)**	
	Sensitivity = GT/(GT+MT)	**Specificity = FTR/(FT+FTR)**	

C3.1) Criterion qualifying the frame-to-frame tracker: estimate whether the link between two physical objects detected at two consecutive time instants is correctly computed or not. This criterion depends on three types of information:

- The overlap between bounding boxes of detected physical objects and reference data,

- The type of detected physical objects,

- The presence of one or several links between bounding boxes of physical objects at time t and t+1 corresponding to an identical real object,

- M3.1.1: number of links between physical objects compared to reference data links.

  - GT = reference data link matching a link between two physical objects.

  - FT = a link between two physical objects not matching any reference data.

  - MT1 = reference data link not found due to frame-to-frame tracking shortcomings.

  - MT2 = reference data link not found due to detection, classification or frame-to-frame tracking shortcomings.

  - OT = GT + FT: all detected links.

  - RD1 = GT + MT1: reference data links between correctly classified physical objects.

  - RD2 = GT + MT2: all reference data links.

  - **Precision**: number of GT / number of OT.

  - **Sensitivity1**: number of GT / number of RD1.

  - **Sensitivity2**: number of GT / number of RD2.

- M3.1.2: same as metric M3.1.1. but for reference data links between physical objects of elementary types (e.g., isolated person) and physical objects of composed types (e.g., group of people) to check the ability of splitting/merging objects.

C3.2) Criterion qualifying the long-term tracker: estimate whether trajectories of physical objects are correctly detected over the duration of their presence in the scene or not.

- M3.2.1: number of detected trajectories compared to reference data trajectories.

  - GT = reference data trajectories matching physical object trajectories.

  - FT = physical object trajectories not matching any reference data trajectories.

  - MT = reference data trajectories not found.

  - OT = GT + FT: all detected trajectories.

  - RD = GT + MT: reference data trajectories.

  - **Precision**: number of GT / number of OT.

  - **Sensitivity**: number of GT / number of RD.

C3.3) Criterion qualifying the long-term tracker properties:

- M3.3.1: average number (together with variance, min and max) of detected trajectories per reference data trajectory in case of short trajectories (e.g., less than 20 frames).
- M3.3.2: average number of detected trajectories per reference data trajectory in case of long trajectories (e.g., more than 20 frames).
- M3.3.3: average of the 3D speed difference between physical objects and reference data.
- M3.3.4: average of the 3D direction difference between physical objects and reference data.
- M3.3.5: average of the initial and ending tracking time difference between physical objects and reference data.
- M3.3.6: average of the camera number difference having the best viewpoint between physical object and reference data.

These metrics can be refined by taking into account zones where events occur (e.g., zones close or far from the camera, detection in a dark or noisy zone) and characteristics of physical objects of interest (e.g., low speed, important size, numerous interactions, strong contrast).

T4) Event recognition:

	Reference Data (RD)	**Noise (N)**	
Event Recognition (ER)	True Positive (TP) **Good Recognition (GR)**	False Positive (FP) **False Recognition (FR)**	**Precision = GR/(GR+FR)**
No Recognition (NR)	False Negative (FN) **Miss Recognition (MR)**	True Negative (TN) **False Recognition Rejection (FRR)**	
	Sensitivity = GR/(GR+MR)	**Specificity = FRR/(FR+FRR)**	

C4.1) Criterion qualifying the recognition of events globally on a video clip or on a scene in a multi cameras configuration:

- M4.1.1: number of correctly recognized events compared to reference data, for each event type.

  - GR = reference data events matching recognized events.

- FR = recognized events not matching reference data events.

- MD1 = reference data events not found due to event recognition shortcomings.

- MD2 = reference data events not found due to detection, classification, tracking or event recognition shortcomings.

- ER = GR + FR: all recognized events.

- RD1 = GR + MD1: all reference data events involving correctly tracked physical objects.

- RD2 = GR + MD2: all reference data events.

- **Precision**: number of GR / number of ER.

- **Sensitivity1**: number of GR / number of RD1.

- **Sensitivity2**: number of GR / number of RD2.

- M4.1.2: same as metric M4.1.1 but the recognition is less strict: if more general events (with smaller priority) are recognized, the process is said correct.

C4.2) Criterion qualifying event properties:

- M4.2.1: average of the initial and ending event time difference between detected event and reference data event.

- M4.2.2: average of the camera number difference having the best viewpoint for observing events.

These metrics can be refined by taking into account zones where events occur (e.g., zones close or far from the camera, detection in a dark or noisy zone) and characteristics of physical objects of interest (e.g., low speed, important size, numerous interactions, strong contrast, large density of people).