Video/Audio Networked surveillance system enhAncement through Human-cEntered adaptIve Monitoring

Large-scale integrating project
Grant Agreement n°248907
01/02/2010 – 31/07/2013

Contractual delivery date: June 30, 2012
Actual delivery date: October 11, 2012

Deliverable D3.3
Trial scenarios and assessment methodologies (version 2)

D3.3

Version: 1.16
Author: INRIA, UNIVIE
Contributors: MULT, IDIAP
Reviewers: RATP, THALIT
Dissemination level: PU
Related document(s): Deliverable D3.1 (Trial scenarios and assessment methodologies (version 1))
Number of pages: 18

Document information

Ver.	Date	Changes	Author (partic.)
	30/06/2010	Deliverable D3.1 (Trial scenarios and assessment methodologies (version 1))	
1.0	15/06/2012	First Edition of document	H. Falciani (INRIA)
1.1	27/06/2012	Integration of input about group tracking	S. Zaidenberg (INRIA)
1.5	29/06/2012	Integration of behavioural sciences input	E. Oberzaucher (UNIVIE)
1.6	29/06/2012	Final formatting	S. Zaidenberg (INRIA)
1.9	16/07/2012	Contribution to section 3.1.1	E. Oberzaucher (UNIVIE)
1.12	17/07/2012	Contribution to sections 3.1.1 and 3.1.2	E. Jouneau (MULT)
1.13	17/07/2012	Contribution to section 3.1.1	R. Emonet (IDIAP)
1.14	18/07/2012	Contributions to section 3.1.2	C. Chen (IDIAP)
1.15	18/09/2012	Document reformatting	E. Jouneau (MULT)
1.16	10/10/2012	Final version	C. Carincotte (MULT)

Ver.	Date	Approval/Rejection decision/comments	Author (partic.)
1.16	12/10/2012	Approved	A. Grifoni (THALIT)
1.16	23/10/2012	Approved	F. Sabourin (RATP)

Copyright

Consisting of:

Coordinator:	Multitel asbl (MULT)	Belgium
Participants:	Gruppo Torinese Trasporti (GTT)	Italy
	Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP)	Switzerland
	Institut National de Recherche en Informatique et en Automatique (INRIA)	France
	Régie Autonome des Transports Parisiens (RATP)	France
	Thales Communications (TCF)	France
	Thales Italia (THALIT)	Italy
	University of Vienna (UNIVIE)	Austria

1 Executive Summary

One of the key task of VANAHEIM is to identify how to address the evaluation of usual/unusual events detection in an automatic audio/video surveillance system. In this document we build upon the generic assessment methodologies delineated in D3.1 (Trial scenarios and assessment methodologies (v1)), and describe in more details how we addressed and will continue to address the assessment of the developed detection/recognition modules. Both technical assessment of monitoring applications based on manual annotation, and user evaluation procedures are thus updated with respect to the detection modules implemented and evaluated until now.

We also propose one additional way of evaluation of the system: expert assessment of monitoring applications based on algorithms output, and an iterative expert evaluation procedure mainly based on the acceptance of selected sub datasets.

Table of Contents

List of Figures

List of Tables

2 Introduction

In this document we build upon the trial scenarios definition and generic assessment methodologies delineated in D3.1 (Trial scenarios and assessment methodologies (v1)), and describe in more details how we addressed and will continue to address the assessment of the developed detection/recognition modules.

Section 3 presents the technical assessment methodology. The technical assessment of performance evaluation is based on the detection of true positives versus false positives as tested against manual annotation. In this section, we present the metrics and evaluation methodologies used with respect to the detection modules implemented until now.

Section 0 presents the expert assessment methodology. The expert assessment of performance evaluation will be based on the selection of video sequences corresponding to the discovered activity classes. The collective behaviours are to be evaluated employing achievement based on measures comparing obtained results with annotated data and expertise in short round trip iterative collaborations between algorithm researchers and behaviourist expert to converge faster in addressing more and more complex behaviours.

Section 5 presents the user evaluation methodology. The user evaluation procedure will mainly be based on an acceptance assessment of the developed systems. The user acceptance index aims to estimate users' attitudes to and their perception of applications(s) investigated, usually based on questionnaire surveys, interviews.

All the assessment methodologies (technical, expert and user) will be used to feed-back into the development the system, first and second to make it more accurate and last to increase its usability.

3 Technical assessment methodologies

Technical assessment aims at formally evaluating the algorithms' performances at the research level. These methods mainly evaluate results conformity to a human pre-defined Ground Truth on the annotated evaluation video dataset. This form of assessment has been carried out and described in technical deliverables for each developed algorithm.

3.1 Technical assessment methodology for autonomous sensor selection

3.1.1 Evaluation of the single stream modelling

The methods presented in the different deliveries D4.1 and D4.3 model the recurrent activities occurring in the scene and affect a score to evaluate the drift between the model and the current video content.
The first evaluation is done on the recognition of the discovered activities. For this purpose, we manually label each frame of the video with one or several activities, we then infer the video with our model, vary the threshold and compare to the Ground Truth to obtain classic precision recall curve that estimates how well the activities are recognized (see **Error! Reference source not found.**a for an example).

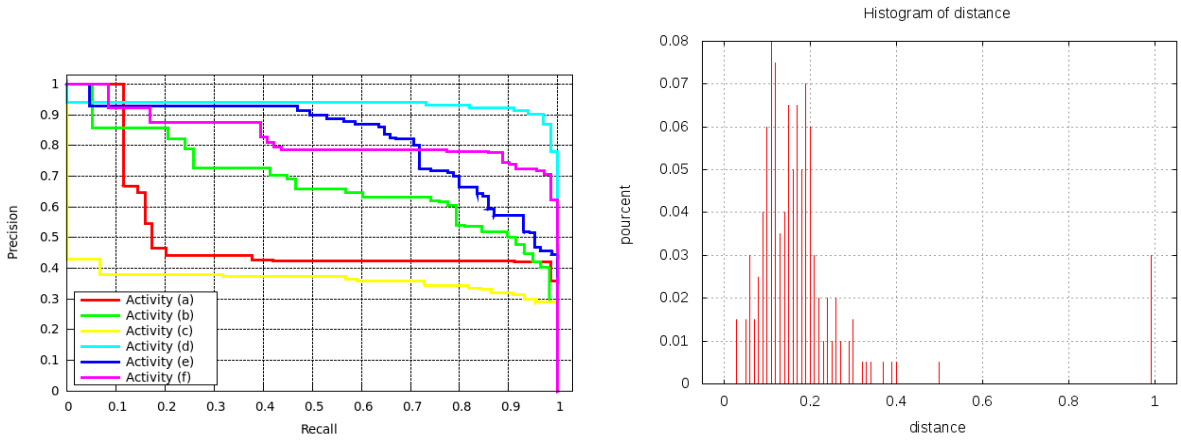

Figure 1: a) Precision recall curve (left), b) Histogram of distances (right).

A second way to evaluate this recognition is to take into account the repartition of the different activities across time. With the manually labelled frame, we can compute the Ground Truth and the observed distribution of the different activities, then compute the deviation between them with a Kullback-Leibler measure for example and finally sample this distance into 100 bins to obtain a histogram (see Figure 1.b).
The second evaluation that can be done is on the score affected to the video content. Even if the goal of the autonomous sensor selection is not to detect anomalies in the video, we can try to observe the behaviour of our model in presence of anomalies. To do that, we infer some video where there is some anomalies and inspect the main peaks of the video. We can thus compute precision recall curve on the anomaly detection case.

3.1.2 Evaluation of the stream selection task

As presented in deliverable D4.5, an experiment has been set up by UNIVIE and consisted in showing to a subject four synchronized camera views. The subject had to click on a view when it was judged interesting and the experiment was repeated on many subjects, producing a continuous Index Of Interestingness (IoI) for each video. From this experiment, we derived a Ground Truth for stream selection where we considered that when a view has a high IoI, it should be selected against the 3 other ones.
To allow for easier and better scalability, we decided to process each camera independently. Different algorithms have been explored to obtain a real valued abnormality rate from a video stream. Different algorithms and views exhibit very dissimilar distribution of abnormality rate. In particular, without any normalization it is impossible to compare the interesting rates provided by different algorithms. We thus explored how per-algorithm and per-view normalization affects the ranking results.

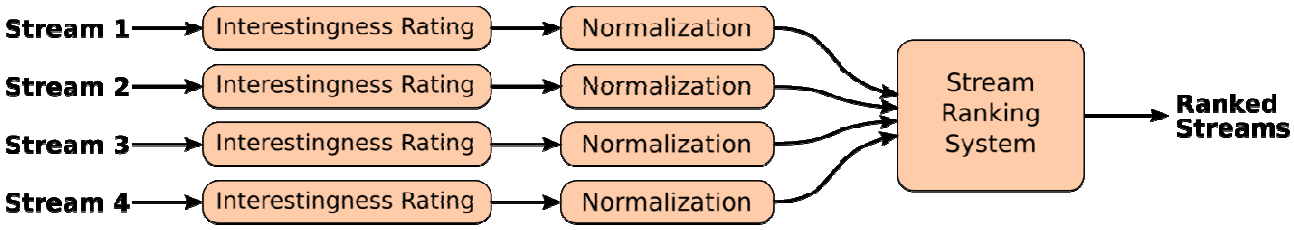

Figure 2: Architecture for the stream selection evaluation.

In a next step, UNIVIE will vary the way how interestingness is measured, to determine which methodology is suited best for the task at hand. Including more camera views in the experiments will also be considered (9 have been used until now). Also, two different kinds of assessment can be done for this task. We can evaluate the global task: the autonomous sensor selection but we can also evaluate the result produced by each stream separately.

3.2 Technical assessment methodology for human-centred applications

As described in deliverable D3.1, technical evaluation of a detection algorithm is the process of comparing algorithm results and human annotations (Ground Truth) on the same video sequence. For this, Ground Truth from UNIVIE annotations has been used, as well as annotations created in-house by partners and publically available datasets.

3.2.1 Evaluation of the multi-human tracking task

As described in deliverable D5.1, multi-human tracking using a CRF model has been evaluated using the public CAVIAR dataset, VANAHEIM data from Torino and a laboratory sequence from EPFL. Performances were measured using tracker purity and object purity metrics (defined in D5.1) on the CAVIAR dataset, for which Ground Truth annotations are available.
To evaluate the multi-object tracking algorithm presented in deliverable D6.1, different measures have also been used presented in details in [2]:

- MOT metrics evaluates the quality of the bounding box alignment between tracks and ground truth:
  - MOTA: Multi Object Tracking Accuracy.
  - MOTP: Multi Object Tracking Precision.
- Configuration error:
  - FP: number of false positive.
  - FN: number of false negative.
  - FA: number of false alarm that correspond to the sum of FN and FP.
  - CD: configuration distance that is the difference between the number of detection and the number of object in the ground truth.
- Multi-detection errors that evaluates the association between objects and trackers:
  - MT: Multi Tracking.
  - MO: Multi-Object.
- Purity measures evaluate the quality of the matching between one object and one tracker:
  - FIT: Falsely Identified Tracker.
  - FIO: Falsely Identified Object.
  - TP: Tracker Purity.
  - OP: Object Purity.

On Table below, an example of the result is reported for illustration.

Table 1: Example of result for a given configuration.

MOTP	MOTA	FP	FN	MT	MO	CD	FIT	FIO	TP	OP	FA
0.53	0.49	0.1	0.36	0.04	0.05	0.34	0.19	0.21	0.64	0.47	0.47

An example of the results is shown in Figure 3.

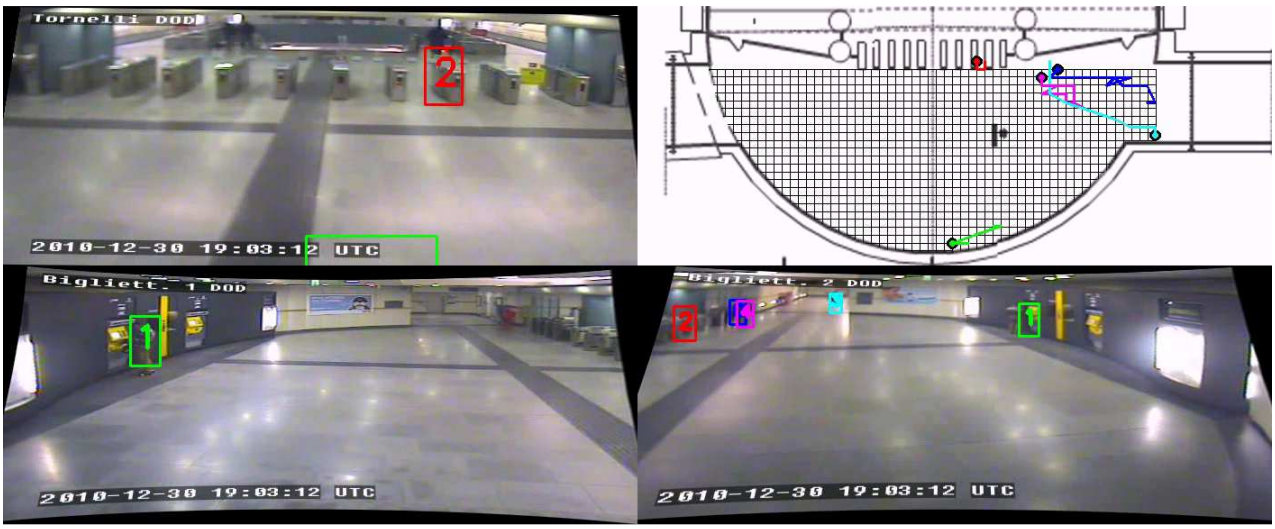

Figure 3: Example of multi-object tracking results.

3.2.2 Evaluation of the joint head and body orientation task

Joint head and body orientation has been evaluated using standard datasets (TUD body orientation dataset and the CHIL/CLEAR dataset). VANAHEIM data has been used for head localization numerical evaluation and head and body orientation estimation qualitative evaluation. For the head localization task, annotations were created in-house by partners.

Figure 4: Example of the output of head localization.

Head localization is evaluated by comparing the localized head bounding box on some (~1200) static body images. These body images are cropped based on the human detection output on Torino data, and the ground-truth head bounding boxes are annotated by hand. For each static body image, the head localization accuracy is computed as the IOU (Intersection over Union) between the ground-truth bounding box and the localized head bounding box. **Error! Reference source not found.** shows some examples of the head localization output.

Joint body and head pose estimation is evaluated by computing the angular error between the estimated body/head pose angle and the ground-truth body/head pose angle. For ground-truth, the head pose angle for CHIL is provided by the dataset itself. All other ground-truth data are annotated by hand.

Two errors are computed separately for body pose and head pose. The evaluation is conducted on four different datasets as listed in Table 2, and the evaluation results are shown in Table 3.

Dataset	# tracks (human id)	# frames	# time duration
CHIL dataset	4	~2500	~8 min
Torino dataset	3	~2000	~7 min
Indoor dataset	2	~2000	~5 min
TownCentre dataset	15	~4500	~3 min

Table 2: Summary of the four datasets used for evaluating body/head pose estimation

Dataset	CHIL	Torino	Indoor	TownCentre
Body pose error	35.3	29.4	23.6	17.4
Head pose error	36.0	30.0	23.6	18.4

Table 3: Evaluation results of body/head pose estimation.

3.2.3 Evaluation of the group detection task

For group tracking and the underlying people detection, in-house annotations have also been carried out. Group bounding boxes have been annotated in 3 video sequences (128, 1373 and 17992 frames) and people – in 1 sequence of 2 hours. UNIVIE annotations have been used for evaluation as well. Section 4.2 describes this expert evaluation in more details.

As described in deliverable D6.1, group event detection has been qualitatively evaluated on the Torino data, as well as data recorded at the Eindhoven airport. A quantitative evaluation has been made impossible by the lack of interesting annotated events (the only event relative to groups is “change in group size” which is not quite relevant to the subway security operators). This lack is explained by the difficulty to screen huge amounts of data in order to find interesting events (such as vandalism/conflict). The huge amount of data to screen and annotate is a limitation for the creation of relevant Ground Truth. Nevertheless, a quantitative evaluation on CAVIAR data has been carried out and the results will be presented at the AVSS 2012 conference in September 2012 [1].

3.2.4 Evaluation of the People flow/crowd monitoring task

People flow/crowd monitoring has been evaluated on in-house annotated sequences as well. 3 metrics were used: relative error on cumulative counting, mean and variance on the flow error and correlation on flow (defined in D5.1). Experiments have been conducted on data from Torino, both from the CARATAKER and VANAHEIM projects.
The assessment of the crowd analysis algorithm presented in deliverable D5.1 is made using 2 different methods. The first method is to analyse the person count across time and compare it to the Ground Truth. For better readability, results smoothed on one minute window are presented in addition to the raw results for each configuration. The Ground Truth is composed of the number of people present in the scene per frame.
Figure 7 illustrate the curve we can obtain in both cases.

Figure 5: Illustration of detection result for crowd analysis.

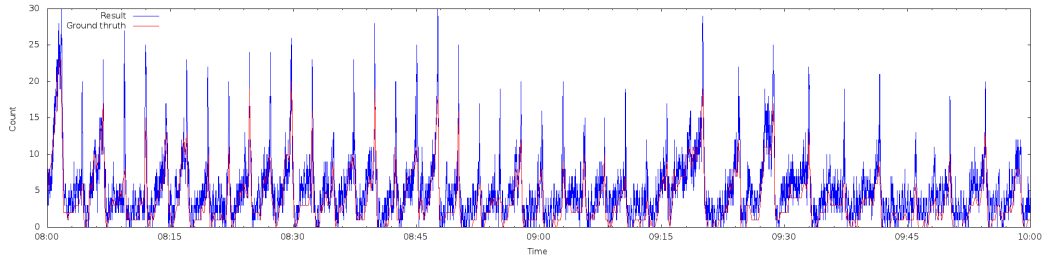

Figure 6: Counting per frame curve.

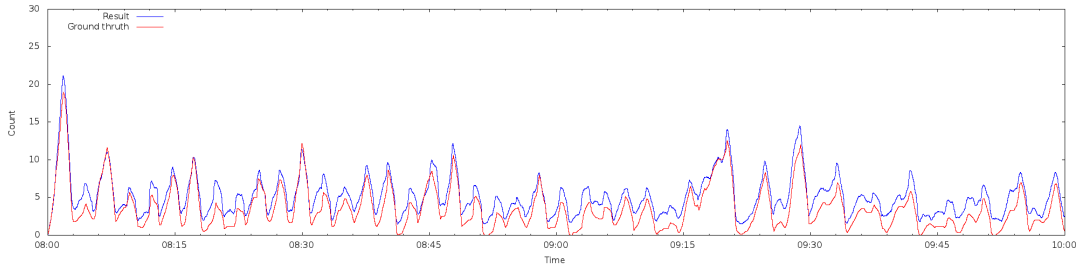

Figure 7: Smoothed counting.

The second method used to evaluate the performance of the algorithms is based on 2 metrics: the mean error, which is computed as the mean of the absolute value of error between the result and the Ground Truth, and the correlation between result and Ground Truth. These metrics are computed on the smoothed curves.

View	Mean number of people	Number of frames	Error	Correlation
Via1B	3,6	721	1.2	0.95
Via2A	2.4	721	0.88	0.68

Table 4: Performance of crowd counting algorithm.

3.2.5 Technical assessment methodology for the left luggage detection task

Event detection algorithms have been evaluated using VANAHEIM data as well. Left luggage detection has been evaluated on manually in-house annotated VANAHEIM data. Precision and recall metrics allow quantifying results.

3.3 Technical assessment methodology for collective behaviour building

The technical assessment for collective behaviour building has been reported in Deliverable D6.2.
The collective behaviour building has the following objectives:

- Investigate the modelling of context:
  - Automatically detect activity zones
  - Discover how equipment and activity zones are employed.
- Find relationships between the behaviours observed at different time periods:
  - People density in the station during weekdays, weekends,
  - Queuing times; use of vending machines in the morning, at rush hour, ...
- Find relationships between the behaviours observed at different locations:
  - People density in the subway platform or hall, or different stations
  - Different people category densities (with-without baggage; adult-children)

To achieve these objectives, the collective behaviour building employs trajectory-based analysis of detected mobiles. This analysis involves the detection of zones, simple events and activities. There are two kinds of simple events: mobiles stay in a zone or move from one zone to another. The activities such as buying a ticket are then characterized by a group of simple events.

The analysis of simple events shows some interesting results. For example, in Table below, we can easily see the people density at different zones as well as the tendency to move from one zone to another. We can also see that the percentages of these simple events does not change much in 3 week days, Wednesday 20/10/2010, Thursday 21/10/2010, and Saturday 23/10/2010.

Table 5: Discovered simple events after online analysis of 10 hours of video for different days.

Wednesday			Thursday			Saturday		
rank	(%)	Event	rank	(%)	Event	rank	(%)	Event
1	31.46	at zone Turnstiles	1	29.74	at zone Turnstiles	1	28.33	at zone Turnstiles
12	1.79	at zone Vending machine1	12	1.57	zone Vending machine1 to zone Turnstiles	12	1.65	at zone Vending machine1

Similarly, the analysis of activity can provide end-user some useful information as depicted in Figure 13. For example, there are many people at Zone 1 at one certain period of time.

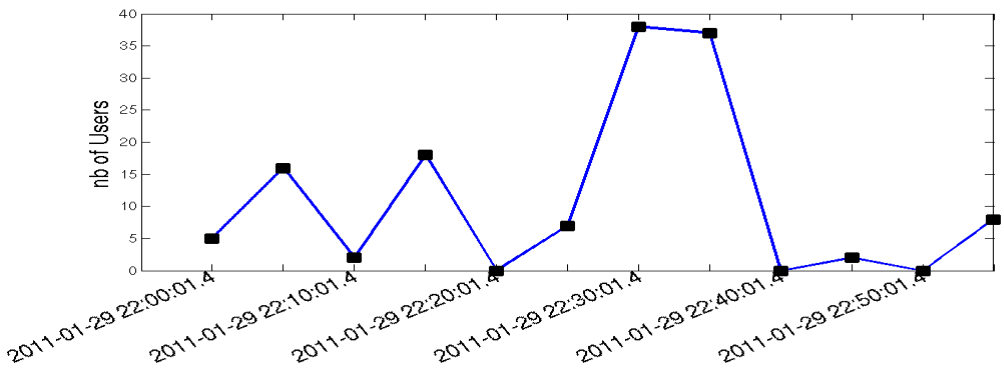

Figure 8: Activity evolution at Zone 1.

The collective behaviour building has two implementations in Matlab and in PL/SQL (relational database). They are a little bit different in how the simple events are detected.
Zone discovery is evaluated by comparing detected zones with zones predefined by human. An example is given in Table 6.

Table 6: Zone overlapping results.

Database implementation	Matlab implementation	GT zone names
76%	71%	zoneEntrance1
76%	52%	zoneHall
47%	84%	zoneVending1
0%	0%	zoneMap1

Activity discovery is evaluated by comparing detected activity with activity defined by human.

Table 7: Activity evaluation results.

Database True Positive	Matlab True Positive	Ground Truth activities	Activity class
40	40	40	zoneEntrance2_to_zoneTurnstiles
189	186	195	All GT activities matching

The processing time for both implementations is also evaluating using a 48 hours long dataset. Among these two implementations, again, the Matlab implementation is too slow to process the video data whereas the database version can process 48 hours video data in 7 minutes.

The drawback of the evaluation is that we lack the ground truth for large video data to correctly estimate the system performance in the discovery of zones, events and activities. In the near future, we will try to overcome this problem with the use of the ANVIL Ground-Truth files defined for the turnstile video camera analysed and manually annotated by UNIVIE.

4 Expert assessment methodology

The main objective of the evaluation process is to evaluate the overall performances of the systems, algorithms and tools developed by the project, drawing also possible conclusions on their transferability. But in an early stage, the behaviourist experts' assessments substituting to the user assessment offers the opportunity to fine-tune the algorithms in a shorter iterative way than with the user assessment only. It will be an addition to the user assessment.

Evaluating algorithm results using Ground Truth annotations created by behavioural experts may be tricky. Defining with precision the expected results and the nature of information to annotate is not an easy task. The main difficulty comes from the fact that scientists such as computer vision experts and behavioural experts do not speak the same language. There is a semantic gap between concepts handled by each party. While in behaviour research the *group* term is limited to people knowing and interacting with each other, computer vision might perceive people with similar trajectories as members of one group. Additionally, identifying groups as such is not a trivial task for the human observer, either. In fact, groups showed the lowest inter-observer-agreement.

In addition to the two main categories of assessment objectives (technical and user) will thus be taken into account: *expert assessment*. This additional form of assessment aims at bridging the gap by improving communication between cross-domain experts. To reach this goal, both sides need to gain better representation of the other domain's concepts and increase the communication rate.

The proposed assessment method is based on an iterative process inciting better communication and resulting in a convergence toward shared concepts and common representations.

Beyond response time of the system, the capacity of the knowledge extraction process, hardware reliability etc., this assessment would address the understanding of the system parameters. The particular technical parameters to be assessed will be chosen to satisfy the decision context and the behaviourists' analysis of resulting video selections.

The following sub-sections explain the iterative process and the way this method is intended to be applied on two algorithms within the VANAHEIM system: group behaviour detection and long-term offline analysis.

4.1 Expert assessment methodology for group detection and collective behaviour building

In VANAHEIM, many of the algorithms aiming at recognizing human-centred events are modelled in a supervised manner. This is especially true for scenarios such as abandoned luggage, equipment monitoring or group detection. In such cases, for the technical assessment, an automatic supervised evaluation will be conducted in the related work package, following the general evaluation process depicted in Figure 9.

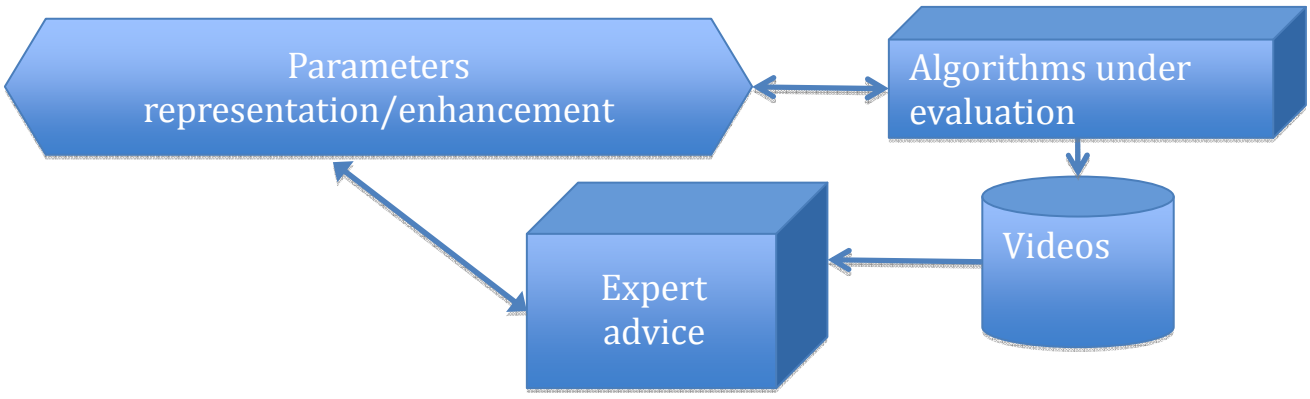

Figure 9: Evaluation process of video understanding algorithms.

One limitation to obtaining relevant Ground Truth is the large amount of data in which it is difficult for the human annotator to find relevant events. The iterative process proposed has as first goal to *filter* the available data.

The iterative process starts with applying the algorithm on a video dataset and obtaining results. These results point out potential targets (depending on the algorithm: groups, specific group behaviours, abandoned luggage, etc.) and are then evaluated by the behavioural expert, who decides which detections are correct/relevant and which are mistakes of the algorithm. This a posteriori annotation is accompanied by behavioural expert advice explaining the reasons why detection is wrong. The computer vision expert, based on the given feedback, adjusts the parameters of his algorithm in a way to obtain new results, more in adequacy with the expert opinion. These new results are, in turn, sent back to the behavioural experts, starting the next iteration step of the assessment. This loop is pursued until sufficient agreement of the algorithm output and the expert opinion.

The automatic nature of this method allows processing large amounts of data and to point out for the behavioural expert only potentially interesting video samples, which are the base for the final Ground Truth being built.

Assessment with roughly filtered data and visual inspection

Collective behaviour building and mostly offline analysis based on filtered data is to concentrate on an as large as possible number of tracked people trajectories. Filtering these trajectories with explicit criteria offers several advantages:

1. ***Data quality***. Because we control the way data is selected, it is thus possible to control the signal to noise ratio (SNR) levels in the data. The behaviour of the system can then be controlled for several levels of complexity in the input data from the SNR point of view.
2. ***Data partition***. The number of normal and abnormal behaviours to be found by the system can be bound hierarchically modulating the selection criteria sensitivity and thus the number of analysed trajectories.

Evaluating the obtained activity clusters in this case can be done comparing unsupervised activity indexes because we have a prior index to compare to (the first prior being the results obtained during GTT's videos based work).

4.2 Expert assessment methodology for detection of groups activity

Evaluation of the group behaviour detection algorithm is based on the evaluation of the group tracking algorithm. As proposed in deliverable D3.1, the formal, quantitative, evaluation uses Ground Truth annotations and counts True Positives (TP), False Negatives (FN) and False Positives (FP) by comparing detection and Ground Truth at each frame. Tracking metrics (Fragmentation, Tracking Time, etc.) are computed given the same detection vs. Ground Truth comparison.
In terms of group tracking evaluation, 5 video chunks of 2 hours each of the Tornelli view at GTT have been annotated by UNIVIE. In those videos, dense annotation has been performed around groups, shows an example. In the images below, the empty blue bounding boxes represent automatically detected and tracked groups, the striped blue bounding boxes represent UNIVIE's Ground Truth group annotations and the striped red bounding boxes represent people annotations, which are irrelevant for this application.

Figure 10: Example of Ground Truth annotation provided by UNIVIE.

The group detection algorithm has been executed and a formal evaluation was carried out. The results are shown in Table 8.

		Tornelli-2011-01-28T07_00_01	Tornelli-2011-01-28T16_00_01
Detection	TP	1789	699
	FP	1597	1141
	FN	2254	3503
	Precision (global)	0,53	0.38
	Sensitivity (global)	0,44	0.17
Tracking	Fragmentation	0,93	0.88
	Tracking time	0,06	0.04
	Purity	0,77	0.80

Table 8: Formal evaluation of the group tracking algorithm on two video sequences of 2 hours using UNIVIE's Ground Truth annotations.

One may notice that the quantitative results shown in are quite poor. This may be partly explained by the semantic gap mentioned above. Indeed, the understanding of the concept of "group" is different for experts in two different domains. Even though a common definition has been established, the need for an iterative annotation process is clearly revealed by these results. This process will allow both experts to get a deeper understanding of the other domain's vision of the concepts at hand and finally to obtain more relevant tools for evaluating the algorithms.
The following examples illustrate various cases where the algorithm and the Ground Truth annotation disagree and show that, given this close look at the data, it is hard to decide which is correct.

Figure 11: Illustration of a detected but not annotated group.

Figure 12: The same group as in Figure 11 still tracked a few frames later.

Figure 13: Illustration of three detected groups in case of a crowd.

Figure 14: Illustration of a detected but not annotated group in the background of the scene.

Figures 11 to 14 show typical examples of cases where a non-annotated group was detected, and one can wonder if considering this group as a False Positive is correct. and shows a group of two people who probably do not know each other (and were not annotate as a group for that reason) but who have the trajectory of a group. Based on their relative trajectory, which is the only information that the algorithm uses, it is logical that they would be considered as a group by the algorithm. It is now the expert's role to decide whether this detected group is a False or a True Positive. shows a case of crowd. In case many people are close to each other and have all similar trajectories, the algorithm considers them to be groups. Moreover, in

this situation many occlusions occur and people are badly detected as individuals by computer vision algorithms used. shows an example of a group detected in the background of the image, which was probably ignored by the annotators. This case could be filtered out in quantitative evaluation, or validated a posterior by an expert.

Figure 15: Example of a True Positive, a case where annotation and automatic detection agree.

Figure 16: Illustration of a False Negative (Ground Truth group with ID 101 is not detected), and a False Positive (group with ID 172 is detected but not annotated).

Figure 15 shows an example of a straight forward True Positive and shows a failure of the algorithm to detect a real group because one person continuously occludes the other. Moreover, Figure 16 shows again an example of an ambiguous detection: a priori a False Positive but maybe a True Positive a posteriori.
An iterative a posteriori annotation process will disambiguate such cases and allow for a more relevant quantitative evaluation, *i.e.* groups detected by the algorithm will be presented to the experts, who will then verify whether they are groups.

4.3 Expert assessment methodology for Offline Analysis

Offline Analysis heavily depends on behavioural Activity definition. Precise parameter sets are available for building behaviours for real-time analysis. For instance, collective behaviour building for group activity provides metadata which are available as input for long term analysis.
Then, finding a way to ease interaction between behaviour and data-mining experts could be addressed through causal networks or such intuitive knowledge representation models. Such models could be updated by application domain experts and thus enrich the knowledge extraction process.
Expert assessment will be expressed in terms of how well the knowledge extracted corresponds to the expert's expectations and in new parameters or new settings updates proposal. Those are defined ex-ante as a validation of representative data (video sequences) and system parameters represented as a causal graph or any other relevant format (*i.e.* SQL, semantic language, etc.). Ideally the assessment will be performed until the end of the project, to allow the assessment of the expected improvement of the system operational efficiency. Knowledge representation in terms of system parameter prevalence will provide with an early precise feedback leading to further development carried out in the technical work packages.

5 User assessment methodology

User assessment aims at evaluating the system's acceptance by end-users: security, quality or maintenance operators. The following audio/video analysis algorithms are integrated in the prototype and have been used for the first evaluation:

- Abandoned Object Detection Module
- Group Detection Module
- Escalator Flow Counting Module
- Occupancy Rate on Platform Module
- Abnormality Measure on Camera Module
- Abnormality Measure on Audio Source Module
- Automatic Camera Detection Module
- Human Detector Module
- Situational Reporting Module
- Offline Analysis Tool

As detailed in D7.2 (Tests and evaluation report (v1)), the chosen test and evaluation methodology for this intermediate test was intended to have both a qualitative and a quantitative assessment of the project results. In order to do so, the evaluation has been split in 2 different steps:
- the first step consisted in interviews during which the VANAHEIM staff demonstrated in detail the whole system, asking specific questions to GTT persons with different roles to get their detailed opinion
- the second step consisted in letting operators play with the prototype autonomously with the only help of a user manual written on purpose, and then fill a questionnaire.

The interviews were taken before starting to distribute the questionnaires and the people to be interviewed were accurately selected per role and guided to the use of VANAHEIM prototype. Since from the very beginning of the interviews seemed that the automatic stream selection has a particular interest for both the operators and the managers, a huge part of the interviews was related to the concept of abnormality, that is what is abnormality for a human being and, mostly what is abnormality and what has to be shown as abnormal depending on the different roles of interviewed people.

Concerning the questionnaire, the different questions to be rated from 1 to 5t were:

1. Ease of use
2. Usefulness of the information provided
3. Operational usefulness of the information provided for your specific job
4. Response time of the system to present information
5. Credibility of information provided
6. Simplicity in understanding the information provided
7. Impact of information on the safety management of the Metro
8. Impact of information provided by the system on cooperation with other entities responsible for Metro Security (e.g. Police)
9. Impact of information on operations of the Metro
10. How acceptable is to have a high percentage of false alarms
11. How acceptable is to have a high percentage of alarms that are not detected

The possible answers have been classified regarding their interest from 1 to 5 corresponding to very little to very much. Due to the nature of the area monitored, the evaluation must be done along the day to observe all different kind of behaviours/situation in the metro station. Figure 17 presents the average score per question obtained over the whole questioned user., while the evaluation of the all algorithms is fully reported in the deliverable D7.2 – tests and evaluations report (v1).

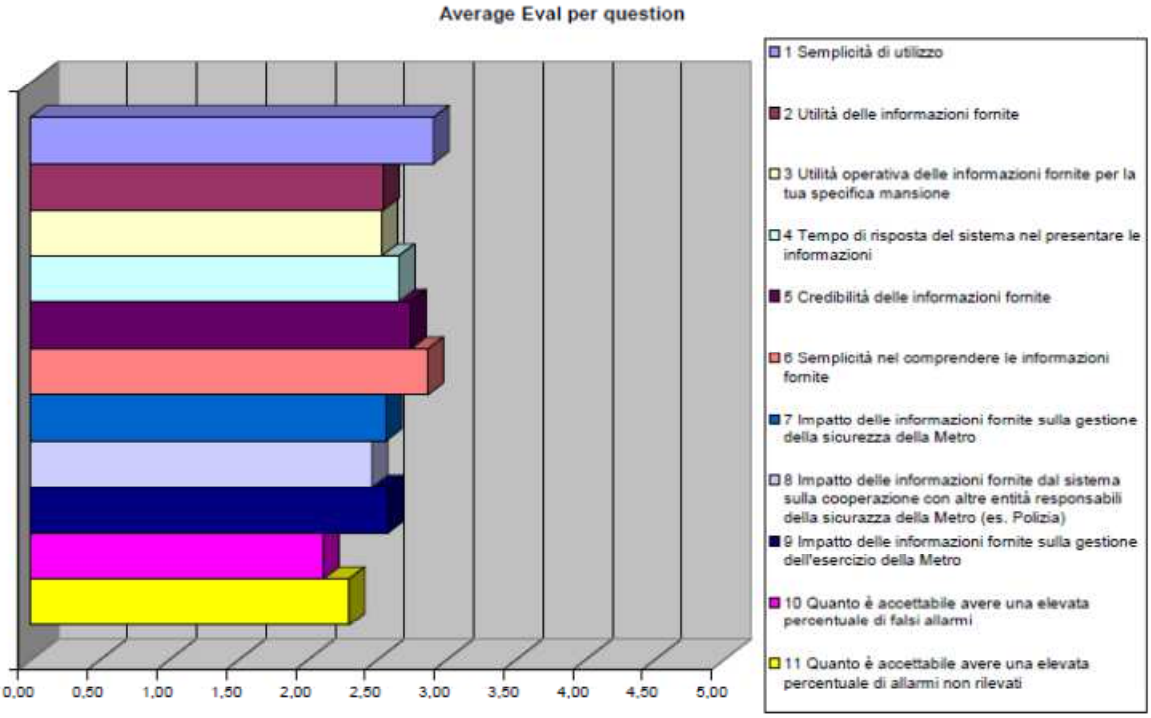

Figure 17: Average score per question.

6 Conclusion

As for the technical assessment, as initially planned, it is mainly be based on the quantitative metrics (true positives versus false positives, precision versus recall, etc) tested against manual annotation provided by behaviours coding and scientific ground truth. The technical assessment for each main part of the system has been presented; evaluation of autonomous sensor selection using comparison with manual annotation, technical assessment of human-centred monitoring and reporting applications solely based on manual annotation. Basically, these technical evaluation metrics has been defined in details in the related work package (WP4 to WP6), and related evaluations has been reported in the corresponding technical deliverables (D4.* to D6.*).

In addition, this deliverable introduces an expert assessment, which objective is to initiate a tighter collaboration between computer vision and human behaviour experts in order to obtain more relevant assessment methods and tools. The proposed method is an iterative process shortening the round trip between experts from different domains and to bridge the semantic gap due to specificity of concepts. It will also contribute to understanding which behaviours are actually of relevance – while methodological specifications pose certain limitations for what can be done, this should not be our guide for what is relevant. As mentioned in the group detection example, interpreting people with similar trajectories as belonging to the same group might be gravely misleading. Thus the iterative process aims to optimize the tasks as well as the methodologies.

Last, the user assessment and evaluation methodology of the preliminary system deployed at GTT was presented; main points were dealing with interviews, followed by a free questionnaire. With respect to the concerned deliverable (D7.2), all people involved in this user evaluation think that day by day operation will benefit from the deployment of VANAHEIM after the needed setup and tuning of the system.

7 Bibliography

[1] S. Zaidenberg, B. Boulay and B. François, “A generic framework for video understanding applied to group behavior recognition,” in *9th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Beijing, 2012.

[2] K. Smith, D. Galica-Perez, J.-M. Odobez and S. Ba, “Evaluating Multi-Object Tracking,” in *Workshop on Empirical Evaluation Methods in Computer Vision*, 2005.