

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: 30 June 2010

Actual delivery date: 1 July 2010

Deliverable D3.1

Trial scenarios and assessment methodologies (version 1)

D3.1

Version: 1.6

Author: UNIVIE-INRIA

Contributors: IDIAP, TCF, GTT

Reviewers: MULT, RATP

Dissemination level: PU

Related document(s): D2.1 End user requirement & system objectives (v1)

Number of pages: 22

Document information

Ver.	Date	Changes	Author (partic.)
1.0	08/06/2010	First Edition of document following conference call	C. Carincotte (MULT)
1.1	11/06/2010	First draft	E. Oberzaucher (UNIVIE)
1.2	16/06/2010	Update and structure change	J.M. Odobez (IDIAP) R. Emonet (IDIAP)
1.3	18/06/2010	Integration of inputs from IDIAP & TCF	E. Oberzaucher (UNIVIE)
1.4	22/06/2010	Integration of further inputs (GTT) and formatting	E. Oberzaucher (UNIVIE)
1.5	25/06/2010	Document reorganisation	C. Carincotte (MULT)
1.6	30/06/2010	Integration of review suggestions and INRIA inputs	E. Oberzaucher (UNIVIE)

Ver.	Date	Approval/Rejection decision/comments	Author (partic.)
1.5	25/06/2010	Approved	C. Carincotte (MULT)
1.6	01/07/2010	Approved	D.T.V. Tran (RATP)

Copyright

© Copyright 2010, 2014 the VANAHEIM Consortium

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

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the VANAHEIM Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.

This document may change without notice.

1 Executive Summary

One central task of VANAHEIM is to delineate how to address the identification of usual/unusual events in an automatic audio/video surveillance system. In this document we first recapitulate the trial scenarios identified in D2.1 (End user requirement & system objectives (v1)). Then we describe how we will address the assessment of these scenario detections/recognitions. We propose two ways of evaluation of the system; technical assessment of monitoring applications based on manual annotation, and a user evaluation procedure mainly based on the acceptance of the system.

Table of contents

1	EXECUTIVE SUMMARY	4
2	INTRODUCTION.....	7
3	TRIAL SCENARIOS DEFINITION.....	8
4	ASSESSMENT METHODOLOGIES.....	9
4.1	TECHNICAL ASSESSMENT METHODOLOGY FOR AUTONOMOUS SENSOR SELECTION	10
4.2	TECHNICAL ASSESSMENT METHODOLOGY FOR HUMAN-CENTRED APPLICATIONS.....	12
4.3	TECHNICAL ASSESSMENT METHODOLOGY FOR COLLECTIVE BEHAVIOUR BUILDING	15
4.4	USER EVALUATION.....	17
4.4.1	<i>User evaluation for Automatic Sensor Selection</i>	<i>19</i>
4.4.2	<i>User evaluation for Human-Centred Applications.....</i>	<i>19</i>
4.4.3	<i>User evaluation for Collective Behaviour Building</i>	<i>19</i>
5	CONCLUSION.....	20
6	APPENDIX – INDEXES FOR COLLECTIVE BEHAVIOUR TECHNICAL ASSESSMENT	21

List of Figures

Figure 1 Evaluation process of video understanding algorithms.....12
Figure 2 Example image of coded group, with 4 group members (1 partially occluded). 12
Figure 3 Monitoring equipment behaviour coding (manipulation category). 13

2 Introduction

In this document we will outline the trial scenarios and describe how we will address the evaluation and assessment of the components in charge of the scenario detection and/or recognition.

Section 3 first recapitulates the trial scenarios described in detail in D2.1.

Section 4 is then dedicated to the specification of the assessment methodologies. We propose two ways of evaluation of the system (technical assessment and user evaluation), in order to evaluate how the events detected can be reliable and useful for the end user. We assume that all methods of evaluation proposed should be carried out, so the different aspects of validity can be covered.

The technical assessment of performance evaluation will be based on the detection of true positives versus false positives as tested against manual annotation. The technical assessment of autonomous sensor selection will use several approaches: a) comparison with manual annotation, b) testing the detection rate of unusual events registered in the logbooks of operators, c) testing the performance in detection of unusual events of operators or other people. Technical assessment of human-centred monitoring and reporting applications will solely be based on manual annotation. Last, collective behaviours are to be evaluated employing both, measures comparing obtained results with annotated data and measures indicating the correctness of the results in unsupervised manner.

As for the user evaluation procedure, it will be mainly based on the acceptance of the system from the operational point of view, for which we introduce an evaluation sheet to be used in the evaluation process.

Both, the technical assessment and the user evaluation will be used to feed-back into the development of the system, first to make it more accurate and second to increase its usability.

3 Trial scenarios definition

The aim of VANAHEIM is to study innovative surveillance components for autonomous monitoring of complex audio/video surveillance infrastructure. To do so, three main components will be developed in the VANAHEIM system:

1. Automatic sensor selection based on audio-visual scene activity modeling

Several approaches based on different paradigms (saliency detection methods, unsupervised statistical models applied independently to each camera or jointly to groups of cameras, audio-visual joint processing in well-suited physical configurations) will be elaborated for this task. Most of these approaches will be based on some saliency or abnormality index, that will be used to select the most relevant or unusual audio/video streams to be presented to operators in control rooms.

2. Investigation of behavioral cues for human-centered monitoring and reporting.

The components developed here shall allow the analysis of audio-video data in close to real-time for pre-defined scenarios and provide/report immediate information extracted from the data streams to operators. The main emphasis is on the investigation of the usefulness of precise behavioral cues (head pose, body posture) for the detection of the situations that can slow down or interfere with the everyday use of public transportation (event detection applications), as well as the reporting of useful information extracted from the data streams (situational reporting). Several scenarios investigating the above issues have been defined to this end, such as:

- *Abandoned and stolen luggage scenario;*
- *Group detection, people arguing, entering in conflict scenario;*
- *Crowd/Flows of people scenario;*
- *Monitoring equipment scenario;*
- *And situational reporting scenario.*

3. Collective behavior building and online learning from long-term analysis of passenger activities.

Collective behavior building should allow an offline analysis of audio-video data and metadata stored in the system, and provide information on trends, repeated events and relationships among events that can be useful for information retrieval and resources planning.

According to the above components, different approaches will be used to evaluate the result of the research. In a first step, algorithms will be evaluated on an individual basis, mainly by measuring their technical performance with appropriate measures (often detection and false alarm rates). In a second step, user-evaluation will be performed in operational conditions, in particular for the first component (stream selection), for which objective measures are hard to define, and the third component (collective behaviors), which requires dedicated displays and human-computer interfaces to present the learned behaviors to operators.

The next section details the evaluation procedures for each of these three main categories.

4 Assessment methodologies

The main objective of the evaluation process is to evaluate the overall performances of the information services, algorithms and tools developed by the project, drawing also possible conclusions on their transferability.

In VANAHEIM, the CONVERGE guidelines¹ have been taken as reference for methodological and formal aspects of assessment (especially for the user evaluation). In this context the term “assessment” is used for “the process of determining the performance and/or impacts of an application”. “Evaluation” is a larger process, involving also the “determination of a value of an application in comparison with alternative applications and/or to a base case and the production of recommendation for the decision makers”.

Two main categories of assessment objectives will thus be taken into account:

1. Technical assessment

This is the most basic level of assessment. It determines the technical parameters of system performance, for example the response time of the system, the capacity of the knowledge extraction process, hardware reliability etc. It would not, however, address the impacts of the system beyond its own boundaries. The particular technical parameters to be assessed will be chosen to satisfy the decision context and the decision-makers' needs.

Technical assessment will be expressed in terms of how well the knowledge extracted corresponds to the ground truth, defined ex-ante as a selection of representative data (video/audio sequences) and system performance, in terms of precision (i.e. ratio of true positives over true and false positives) and sensitivity (i.e. ratio of true positives over true positives plus false negatives). The technical assessment will be performed all along the duration of the project, to allow the assessment of the expected improvement of system operational efficiency.

In most cases, behaviour coding by UNIVIE and scientific annotations by researchers will be used as a basis for determining the scenes or events that shall be detected by the system.

The intermediate tests and verification phase at project mid-term will define the appropriate set of indicators and measurements, according to user requirements analysis, ontology definition and knowledge representation, and will provide an early feedback to further development carried out in the technical work packages.

2. User evaluation

User evaluation (acceptance assessment) aims to estimate users' attitudes to and their perception of the applications(s) investigated, usually based on questionnaire surveys, interviews, etc. Here the users may be the operators who implement and operate the systems, the passengers taking benefits of the service, or the security operators who buy and use the applications. Regarding evaluation by system operators, the results of dedicated technical assessment, impact assessment or cost-benefit analysis for the respective system is usually decisive. User-friendliness of the human-machine interface may be a further key factor influencing the system operator's acceptance.

In Sections 4.1 to 4.3, we first describe the specific technical evaluation methodology that we will pursue for each of the VANAHEIM system component (sensor selection, real-time event and behaviour recognition, collective behaviour recognition). Then, in Section 4.4 we describe more specifically the user evaluation procedure that we will follow.

¹ 4th Framework Research, Project CONVERGE TR 1101, Transport Telematics Support and Consensus, http://cordis.lu/telematics/tap_transport/research/projects/converge.html

4.1 Technical assessment methodology for Autonomous Sensor Selection

The objective of the autonomous sensor selection algorithm is to select the audio and video streams to display at a given time. The algorithm should select the most interesting, abnormal, unusual streams. Three ways are currently envisaged to perform its technical assessment, as detailed below.

- A first assessment methodology for the sensor selection algorithm will be to use annotated audio and video sequences. The annotations can either be a binary annotation (interesting, not interesting) or a more continuous measure of the degree of abnormality. These annotations can be used as a ground truth following the scheme presented in Figure 1 (p. 12) to evaluate whether the autonomous sensor selection algorithm is efficient or not. The main drawback of this assessment approach is the requirement for long annotated audio/video streams and the need for expert annotations.

From the behaviour coding point of view, one way to tackle this issue would be to include an additional code in the behaviour observation catalogue, named “calls for attention”. This would mean that observers regard the scene as interesting without further specifying the cause for this. Since the causes (i.e. too high density, someone remaining too long at the ticketing machine, etc.) are deducible from the other behaviours coded, the overlap between scenes selected by the system and by coders can be used as a means to evaluate the accuracy of scene detection.

- Since the project preliminary dataset also contains short audio/video clips with abnormal events (GTT logbook created and regularly updated by GTT operators), a second approach will be to reuse these abnormal events already annotated in multimedia streams. From these events, we can judge whether a selection process is efficient by counting the number of times the relevant stream(s) were chosen by the autonomous system when the abnormal events occurred. As these clips are rather short and out of their context (i.e. we do not have long time before and after the event, but just the event itself), we must integrate them in longer videos and run the same evaluation process as with annotated audio/videos. This method has the advantage to reuse existing annotations but the main drawback is that we will test only over a limited set of configurations; indeed, it will allow to test normal conditions against strong abnormalities, and won't allow to evaluate the selection component on the “most relevant” criteria (i.e. how does the selection component perform when all the streams are normal; in other words, are the selected streams the less normal among all the normal ones).

In this solution based on abnormal event mixed with normal ambiance before performance evaluation, acoustic properties of assessment locations are taken into account for the audio part. We propose in the framework of VANAHEIM to broadcast abnormal audio events (with professional audio player) during the night and then collect these samples as trial's location dependant abnormal audio event. These specific audio samples will be mixed to normal ambiance before algorithm's evolution. The benefits of the method are manifold. One of the most important is the number of abnormal audio events we will be able to simulate. Using professional databases as Sound-Ideas Audio Database, we will address more audio events rather than methods using only actors to simulate abnormal behaviours.

- The last assessment methodology will be closer to user studies and will consist in putting operators or other human subjects in front of different sensor selection system; the autonomous one that is designed in the project and a random one. We could then evaluate the relevance of autonomous sensor selection based on questionnaires, task specific performance measures or measures of boredom. Autonomous sensor selection could also be compared to manual sensor selection, using the same evaluations method.

At project mid-term, we will thus test the system by putting students in an operator-like environment. On a large computer screen several videos will be shown. In the first condition the videos will be selected at random, in the second, the system will make a pre-selection for interesting scenes based on automatic audio and video analysis. The students will have the task to click on the scenes that are interesting. The evaluation will be based on the students' performance in the two conditions: scores should be higher with the preselected scenes by the system.

However, when conducted with non-operator subjects, the risk of this last assessment approach is that the subject's opinion might not reflect an operator's opinion: what is normal or usual from an expert operator might be perceived as abnormal for a non-expert observer (and vice versa). The best evaluation of the adequacy of the system can be obtained through long-term user studies: do the users use the system if they can or, do they drop it and fallback to the fully manual selection?

A last important issue, which has to be addressed concerning autonomous sensor selection, is the way to be used to present/display continuously the algorithm's results. In Turin, security operators use 28 video monitors to control more than 700 video sensors. The way they use to switch is based on heuristics and also on their own knowledge of the passenger's behaviour (day of the week, time of the day, particular events in the city...). To correctly evaluate the proposed sensor selection method, we will have to suggest adapted result's display methods in order not to perturb operator in their regular way of working. The following issues will be addressed during technical evaluation:

- How to display the results: one or several dedicated monitor(s) in the video wall?
- Where the results should be displayed in the video wall?

Depending on the security operators' feedbacks, we will adapt the display methods following their requirements. Questions to users related to these issues will also be included in User evaluation datasheet presented in Section 4.4.

4.2 Technical assessment methodology for Human-Centred Applications

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 1.

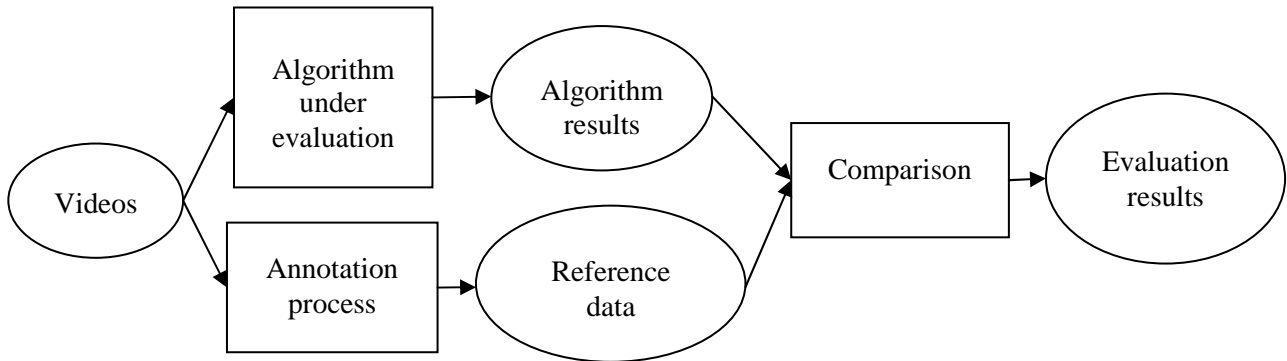


Figure 1 Evaluation process of video understanding algorithms.

This evaluation process thus requires taking a decision concerning two topics: audio/video database and related annotation, and evaluation criteria and metrics.

The prerequisite to obtain such an automatic evaluation is to first collect and define **reference data** through the **annotation** process (often called ground-truth). In general, behaviour coding by UNIVIE will be used as a basis for determining the scenes or the events that shall be used for this technical assessment. For example, groups can be manually marked in the images as a reference for the recognition system evaluation (Figure 2). In this case, videos have been manually annotated for the presence and size of groups, exploiting features such as synchronized locomotion, proximity and social interaction (e.g. communication), etc.

Other examples of frequent behaviour related to scenarios of interest such as manipulation of equipment (mostly turnstiles and ticket machines) will also be coded in the project material (Figure 3).



Figure 2 Example image of coded group, with 4 group members (1 partially occluded).



Figure 3 Monitoring equipment behaviour coding (manipulation category).

Currently, four hours of video material are annotated with the behaviour catalogue described in D2.2 (Ontology-based behaviour modelling specifications), i.e. all behaviours of all people visible on the current video material. In total, 341 people were observed for 21.7 seconds on average, resulting in 124 person-minutes. In the future, an annotated database of behaviour events, which can be selected automatically from the video-streams via the ANVIL coding software, will be provided. In addition to the reliability of the behaviours, we will use benchmark data generated for machine recognition (metadata annotations generated by researchers for conducting their studies), that will complement the behaviour coding.

Regarding the reference data themselves, for most of the scenarios, the number of events to be detected will be large enough in normal captured data. This is the case for the “Group identification” or “Monitoring equipment scenarios”. However, for other events, such a situation may not be encountered. This is the case for instance for the “abandoned luggage scenario”, or the “People arguing, entering in conflict scenarios”, which are rather rare events in the material. In this case acted scenarios might be needed to collect enough evaluation data. In addition, as mentioned in DoW, public benchmarks related to the concerned scenario will also be used, such as PETS or AVSS 2007 datasets for unattended luggage or attended luggage removal (theft), PETS 2009 or BEHAVE datasets for crowd activities monitoring.

An evaluation, which uses dedicated evaluation criteria and metrics, will then be used to estimate the algorithms performances and highlight their pros and cons (e.g. capabilities to manage shadow, occlusions, object crossings). Results will be presented in two manners:

1. **Detailed analysis:** to obtain a clear and precise view on performances of the various algorithms at several points in the audio/video processing chain. For instance, we will use metrics that can output:
 - A defined value like 2D localisation error, time interval between object apparition in the ground truth and its detection in the system... In these cases the metrics can compute statistics on these values (mean, deviation, min & max values...), the statistics being for example computed along the video sequence.
 - A number of successes or failures of the algorithm faces to situations. In that case, we use the standard definitions:
 - The True Positive (TP): the system has detected a real situation (exists in reference data and algorithm results).
 - The False Negative (FN): a real situation has been missed by the system (exists only in reference data).
 - The False Positive (FP): the system has detected a situation that is not real (exists only in algorithm results).
 - The True Negative (TN): entity that does not fit neither with a reference data, nor an algorithm result.

These data may provide also more information that will characterize algorithms as:

- The precision = $TP / (TP + FP)$
- The sensitivity = $TP / (TP + FN)$
- The specificity = $TN / (FP + TN)$,
- The F-score = $2 * \text{precision} * \text{sensitivity} / (\text{precision} + \text{sensitivity})$: harmonic mean of the precision and sensitivity.

2. **Global analysis:** to compare globally all algorithms and to obtain a meaningful measure of their performances from the end-user point of view. For instance, it can be the minimum between the sensitivity and the precision.

In general, the number and type of indicators will be defined during the algorithms development stage in related work package, taking into account inputs/specifications coming from WP2 and WP3. Two basic requirements will be taken into account when defining indicators: they must be able to reflect clearly the related performance or impact; and they must be capable of reliable assessment using the experimental tools and measurement methods chosen.

4.3 Technical assessment methodology for Collective Behaviour building

Collective behaviour building consists mainly in 2 tasks: a) computing the most frequent activities in a subway and b) detecting abnormal activities. To analyse the frequent activities, a first stage consists in clustering all the different behaviours, dynamics or events occurring in the scene into clusters of activity corresponding to the main trends of passengers travelling in the subway. For this first stage a large amount of videos will to be collected in T3.3.

Assessment with unsupervised indexes

The technical assessment of collective behaviour building consists in evaluating the quality of these clusters using general metrics verifying the constancy of the clusters (e.g. high density, low intra-distance and high inter-distance). Several criteria (called also validity indexes) can be used such as Silhouette, Dunn, Davies-Buldin). Note that these indexes evaluate the good structuring of the clustering results rather than any correctness from the user point of view.

From the clustering results, a direct mapping can be established between the clusters and the activities observed from the data. When talking about a good cluster separability, we are basically measuring a clear distinction between the activities, which is given by the inter-cluster separability. On the same sense, when talking about a high intra-cluster homogeneity, we mean this time a good cluster cohesion and in terms of activity, this means we are able from the cluster to describe the mean activity in a concise way and without too much deviations. The Silhouette, Dunn, Davies-Buldin indexes differ between themselves only in the way the distance between clusters or the cluster's scatter is calculated. Actually some more indexes do exist in the literature but in VANAHEIM, we are concentrating in the most employed-ones; for concision matter, these indexes are described in Appendix 6 (p. 21).

Assessment by activity cluster visual inspection

In order to get closer to an end-user evaluation point of view, a second technique to assess the process of collective behaviour building is to visualise the videos corresponding to the same cluster to verify whether the people behaviours occurring in the video correspond effectively to the same activity. In other words, we are visually assessing the intra-cluster homogeneity. If we succeed to observe the same behaviour when randomly drawing from the cluster several different records, then the frequently activities for the collective behaviour building process are validated. Note that if from visual evaluation we observe a mixture of behaviours grouped into one cluster, the clustering algorithm needs certainly to be tuned as the behaviour mixture is generally indicative of an under-fitted space, or in other terms a too low number of clusters employed to describe the whole spectrum of activities. Each cluster carries thus a given degree of confusion. On the other hand, if it is found that elements belonging to different clusters actually correspond to the same behaviour we are then in the frame of an over-clustered space and the number of clusters can be diminished. This kind of evaluation is thus particularly important. Even when the algorithm is totally unsupervised and the number of clusters does not need to be specified in advance of the clustering process, this evaluation could be employed to tune the sensitivity parameters which allow adding/creating new clusters into the systems.

Assessment with synthetic data

A third technique is to simulate a large number of trajectories corresponding to people activities, to cluster these trajectories and to evaluate the obtained clusters. Several advantages can be attached to working with simulated data:

1. **Data quality.** Because we control the way data is generated, 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 set in advance
3. **Data labelling.** While generating the simulated data, it is possible to set a tag for each activity. The objective for this is to later compare the grouping established by the clustering algorithm with that artificially set at the beginning while generating the simulated data.

Evaluating the obtained clusters in this case can be done employing supervised indexes because we have a ground-truth to compare to (the grouping established artificially at the beginning while generating the simulated data). The supervised measure we are employing in the project for this purpose is the Jaccard index described in Appendix 6, Section 6.4 (p. 22).

Assessment with video ground-truth

The activity clusters can also be evaluated compared to video ground-truth when end-users are able to provide such information (i.e. the main activities occurring in the subway) on the processed data. At this point, videos of benchmarking data could be used. However there is no such database currently available corresponding to VANAHEIM scenes. The contingency plan here is to employ the video material that evaluates the human-centred applications. Because, it mainly consists of abnormal activities, we can evaluate our system by testing that these activities are not included in the frequent activity clusters for behaviour building, which actually describe the normal and most frequent trends of behaviour in the station.

Assessment of abnormal activity detection

In the collective behaviour building, when all frequent activities have been identified, the detection of abnormal activities is similar to the autonomous sensor selection part and consists in selecting the scene activities deviating from normal behaviour and of interest for the security operators. Therefore this process can be assessed as the autonomous sensor selection described before.

4.4 User evaluation

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, etc. Here the users may be the operators who implement and operate the systems, or the security operators who buy and use the applications. Regarding acceptance by system operators, the results of dedicated technical assessment, impact assessment or cost-benefit analysis for the respective system is usually decisive. User-friendliness of the human-machine interface may be a further key factor influencing the system operator’s acceptance.

Another user evaluation index that could be used is the impact assessment. Impact assessment is the measurement or estimation of the impacts (effects) of an application, e.g. those on safety, security, environmental conditions or efficiency of the transport system, for the particular target groups (system operators, , decision makers etc.) likely to be affected. Since this means determining how the indicators of assessment may have changed, a comparison is implied: either a “before-and-after” measurement, or comparison with alternative(s). The decision context should determine which impacts to assess and the target groups to be considered.

For VANAHEIM, the main target group will be the security operators. Thus the main acceptance and impact assessment will consist in evaluating how efficiently the VANAHEIM system can be usable and useful, mainly to identify “Scenes of Interest” (SOI) for the different parts of the system. To do so, we should compare how many SOIs could be processed by a security operator before and after VANAHEIM utilisation.

Within the user acceptance assessment, as proposed in the tale below, users’ opinions/evaluations will be gathered towards three main fields: HMI, accessibility to information, and ease of use. Within the impact assessment, it will evaluate the impact of knowledge extracted in the fields of security operations, on criteria such as response time, credibility of information, usefulness of information, and level of interoperability.

<i>Assessment category</i>	<i>Assessment objective</i>	<i>Indicators</i>	<i>Group involved in evaluation</i>
User evaluation	1. Acceptance assessment	1. HMI	End Users
		2. Accessibility of the information	End Users
		3. Ease of use	End Users
	2. Impact acceptance	4. Response time	End Users
		5. Credibility / Reliability	End Users
		6. Usefulness	End Users
		7. Level of interoperability	End Users

The user acceptance will be evaluated by end users through demonstrations of algorithms usage firstly with test data and then in a real life operations, where possible, and through simulations in all the other cases. The values used for this assessment will be qualitative, using a scale from 1 to 5, where the scale corresponds to the following assessed performances: “Very poor”; “Poor”; “Fair”; “Good”; “Excellent”. Table below presents a generic evaluation datasheet integrating these different user criterions, which will be updated and refined at the mid-term and final evaluation stages in WP7, so as to exactly fit the developed components.

Next subsections now present in more details the foreseen user evaluation procedures that will be applied for each part of the VANAHEIM system, i.e. the autonomous sensor selection, the human-centred applications and the collective behaviour building.

VANAHEIM – User evaluation datasheet

Test Site:
 Day:
 Time:
 Operator:
 Test number:



Algorithm ID	EG: COMPONENT-1	Rate		Successful (+) Unsuccessful (-)
		Peak-hour	Off-peak hour	
Ref.	Questions			
	Test			
1.1	How would you rate the user friendliness of interface (HMI)?			
1.2	How would you rate the accessibility of the information provided?			
1.3	How would you rate the overall easiness of use of the tool?			
2.4	How would you rate the responses time in which the results have been provided?			
2.5	How would you rate the credibility/reliability of the information provided?			
2.6	How would you rate the (operational) usefulness of the information provided?			
2.7	How would you rate the interoperability of the service?			
	<i>How would you evaluate the impact of the information provided on the management of security?</i>			
	<i>How would you evaluate the impact of the information provided on the cooperation among other relevant bodies in charge of security?</i>			
	...			
Comments				

Scale	from 2.1.1 to 3.1.2	for 3.2.3 and 3.2.4	
1	Very Poor	1	Very Little
2	Poor	2	Little
3	Fair	3	Somewhat
4	Good	4	Much
5	Excellent	5	Very Much

4.4.1 User evaluation for Automatic Sensor Selection

The objective of the autonomous sensor selection algorithm is to select the audio and video streams to be played/displayed at a given time in the control room. The algorithm should select the most interesting, abnormal, unusual streams.

The user evaluation of this selection will be based mainly on the above-mentioned table and the related datasheet provided to the operators. Automatic sensor selection will be evaluated mainly by security operators during their day-by-day operations. The opportunity to evaluate the correct selection will be given by authentic abnormal events during operational hours that are usually reported by customers to the control room: since it is mandatory to log these events, the log itself will report if the algorithm will select efficiently the correct stream.

In addition, many interesting, abnormal and unusual streams happen during non-operational hours and are, in fact, already known by the operators since they are unusual events related to extraordinary maintenance or operation activity, such as exercitations of operators or agent. The ability to know in advance these events will be used as a ground- truth to evaluate whether the autonomous sensor selection algorithm is efficient or not.

4.4.2 User evaluation for Human-Centred Applications

The objective of human-centred applications is to report immediate information extracted from the data-streams to operators in control room. All the scenarios, even if usually not affecting regular operations of public transport, happen on a regular basis in the day-by-day operations.

Since all of them are currently reported by customers or agents immediately or at the end of the day to the control room, together with the quality assessment and evaluation that will be done by operators on single instance of the information reported by the applications, a statistic will be compiled on the efficiency of those application in comparison with non-automatic reporting of the described scenarios.

4.4.3 User evaluation for Collective Behaviour Building

The objective of collective behaviour building application is to provide information on trends, repeated events and relationships among events that can be useful for information retrieval and resources planning. For this reason the end user evaluations will be done mainly by security managers and operations planners, rather than by control room operators.

The information and trends that will emerge from VANAHEIM applications evaluation will be correlated with the information the operator already has on trend and utilization of the metro, such as the one that comes from the gates or the ticketing vending machine, in order to mutually acknowledge the two information sources. Also in this evaluation, the above-mentioned table and the related datasheet provided to the operators will be used to report the evaluation.

5 Conclusion

Taking into account the assessment of a system such as VANAHEIM one is crucial on several levels (technical performance of components and also user evaluation from operational point of view), we proposed in this deliverable an evaluation and verification framework that will be applied in VANAHEIM for both technical assessment and user evaluation of the system.

As for the technical assessment, it will 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 will use several approaches: a) comparison with manual annotation, b) testing the detection rate of unusual events registered in the logbooks of operators, c) testing the performance in detection of unusual events of operators or other people. Technical assessment of human-centred monitoring and reporting applications will solely be based on manual annotation, by comparing human observation (manual annotation or operators' observations) and outputs coming from the components themselves, so as to compute meaningful statistics about component technical performances. Last, collective behaviours will be evaluated employing both, measures comparing obtained results with annotated data and several indexes indicating the correctness of the results in unsupervised manner. Basically, these technical evaluation metrics will be refined in the related work package (WP4 to WP6), and related evaluations will be reported in the corresponding technical deliverables (D4.* to D6.*).

As for the user evaluation procedure, it will be based on the acceptance of the system from the operational point of view, for which we introduce an evaluation sheet to be used in the evaluation process. This evaluation datasheet will be refined and used at project mid-term evaluations stage (WP7/T7.3, D7.2), so as to identify the pros and cons of currently deployed components.

Last, both, the technical assessment and the user evaluation will be used to feed-back into the development of the system until the project end, first to make it more accurate from the technical point of view and second to increase its usability from the operational point of view.

6 Appendix – Indexes for collective behaviour technical assessment

6.1 Silhouette index

The Silhouette index is defined as follows: Consider a data object $v_j \in \{1, 2, \dots, N\}$ belonging to cluster cl_i $i \in \{1, \dots, c\}$. This means that object v_j is closer to the prototype of cluster cl_i than to any other prototype. Let the average distance of this object to all objects belonging to cluster cl_i be denoted by a_{ij} . Also, let the average distance of this object to all objects belonging to another cluster $i' \neq i$ be called d_{ij} . Finally let b_{ij} be the minimum d_{ij} computed over $i' = 1, \dots, c$ which represents the dissimilarity of object j to its closest neighbouring cluster. The Silhouette index is then:

$$S = \frac{1}{N} \sum_{j=1}^N s_j \quad \text{where} \quad s_j = \frac{b_{ij} - a_{ij}}{\max\{a_{ij}, b_{ij}\}}$$

This way, the best partition is achieved when S is maximized, which implies minimizing the intra-cluster distance (a_{ij}) while maximizing the inter-cluster distance (b_{ij}).

6.2 Dunn index

The Dunn index is defined as follows: Let cl_i and $cl_{i'}$ be two different clusters of the input dataset. Then, the

diameter Δ of cl_i is defined as $\Delta(cl_i) = \max_{v_j, v_{j'} \in cl_i} \{d(v_j, v_{j'})\}$

Let δ be the distance between cl_i and $cl_{i'}$. Then δ is defined as

$$\delta(cl_i, cl_{i'}) = \min_{v_j \in cl_i, v_{j'} \in cl_{i'}} \{d(v_j, v_{j'})\},$$

and, $d(x, y)$ indicates the distance between points x and y .

For any partition, the Dunn index is

$$V_D = \min_i \left\{ \frac{\min_{i'} \left\{ \frac{\delta(cl_i, cl_{i'})}{\max_i \{cl_i\}} \right\} \right\} \quad \text{and} \quad i, i' \in \{1, \dots, N\}, \quad i' \neq i$$

Larger values of V_D correspond to a good clustering partition.

6.3 Davis Bouldin index

The Davis Bouldin index is defined as follows: This index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within cluster, cl_i , is computed as

$$S_i = \frac{1}{|cl_i|} \sum_{v_j \in cl_i} \{ \|v_j - m_i\| \}$$

m_i is the prototype for cluster cl_i . The distance δ between clusters cl_i and $cl_{i'}$ is defined as

$$\delta(cl_i, cl_{i'}) = \|m_{i'} - m_i\|$$

The Davies-Bouldin (DB) index is then defined as

$$DB = \frac{1}{N} \sum_{i=1}^N R_i \text{ with } R_i = \max_{i,i'} R_{ii'}, \quad i, i' \in \{1, \dots, N\}, \quad i' \neq i \text{ and}$$

$$R_{ii'} = \frac{S_i + S_{i'}}{\delta(cl_i, cl_{i'})}$$

Low values of the DB index are associated with a proper clustering.

6.4 Jaccard index

Consider $C = \{CL1 \dots CLm\}$ is a clustering structure of a data set $X = \{v_1, v_2, \dots, v_N\}$; and $P = \{P1 \dots Ps\}$ is a defined partition of the data. We refer to a pair of points (v_i, v_j) from the data set using the following terms:

- SS: if both points belong to the same cluster of the clustering structure C and to the same group of partition P .
- SD: if points belong to the same cluster of C and to different groups of P .
- DS: if points belong to different clusters of C and to the same group of P .
- DD: if both points belong to different clusters of C and to different groups of P .

Assuming now that a , b , c and d are the number of SS, SD, DS and DD pairs respectively, then $a+b+c+d = M$ which is the maximum number of all pairs in the data set (meaning, $M = N(N - 1)/2$ where N is the total number of points in the data set).

Now we can define the following indices to measure the degree of similarity between C and P as:

$$J = a/(a + b + c).$$