Multimedia Tools and Applications
DOI 10.1007/s11042-013-1575-9

Requirements for multimedia metadata schemes in surveillance applications for security

J. van Rest · F.A. Grootjen · M. Grootjen · R. Wijn · O. Aarts · M.L. Roelofs · G.J. Burghouts · H. Bouma · L. Alic · W. Kraaij

Abstract

Surveillance for security requires communication between systems and humans, involves behavioural and multimedia research, and demands an objective benchmarking for the performance of system components. Metadata representation schemes are extremely important to facilitate (system) interoperability and to define ground truth annotations for surveillance research and benchmarks. Surveillance places specific requirements on these metadata representation schemes. This paper offers a clear and coherent terminology, and uses this to present these requirements and to evaluate them in three ways: their fitness in breadth for surveillance design patterns, their fitness in depth for a specific surveillance scenario, and their realism on the basis of existing schemes. It is also validated that no existing metadata representation scheme fulfils all requirements. Guidelines are offered to those who wish to select or create a metadata scheme for surveillance for security.

Keywords: *surveillance, human behaviour, annotation, metadata representation scheme, event, action, multimodal, multi-sensor, ONVIF, MPEG-7, PETS.*

1. Introduction

Surveillance systems are used for a wide range of settings and purposes. They have been proven valuable in the fight against national and international terrorism, they support national security and border policy, and they help to fight crime in

[*] The online version of this article contains supplementary material.

J. van Rest, M. Grootjen, R. Wijn, O. Aarts, M. L. Roelofs, G. J. Burghouts, H. Bouma, L. Alic, W. Kraaij:
TNO, The Hague, The Netherlands.
e-mail: jeroen.vanrest@tno.nl

F.A. Grootjen:
Radboud University, Nijmegen, The Netherlands.

public places [19]. In these settings, successful operation depends on the effective use of collected data and metadata. Metadata are necessary to transfer knowledge, and to train, test and evaluate surveillance personnel and algorithms.
In surveillance systems for security, these metadata schemes are important for operational, technical, scientific and commercial reasons. First, from an operational perspective, metadata schemes are needed to efficiently create a surveillance report. Second, from a technical perspective, the modular and distributed nature of surveillance systems requires metadata schemes to allow communication between the subcomponents. Third, from a scientific and commercial perspective, they are a precondition for defining ground truth annotations and benchmarking different systems. Fourth, from a purely scientific perspective, a metadata scheme – in the form of standard test resources and a standard for semantic annotation – facilitates cooperation between supporting research domains.

The four perspectives described above place different requirements on metadata schemes. Although metadata schemes have been proposed by relevant communities (e.g. industrial [32], scientific [6], governmental [14]), no single scheme covers all four perspectives. Also, the lack of alignment between the individual schemes, complicates the ability to share information and make systems interoperable. Standardization and interoperability are important because the threat of tomorrow may be unknown at design time. So, it is likely that in order to mitigate future incidents, it is necessary to connect a surveillance system to other surveillance systems or data sources in order to share information.
A sign of a lack of communication between relevant communities is found when studying the respective terminologies. The meaning of the word "behaviour" in behavioural research means "the reaction of a cognitive agent on a stimulus". However, in multimedia research the meaning tends towards "an action of a person", thereby losing the notion of action-reaction and of cognition behind behaviour. Thus, the opportunity we miss is the reasoning about *why* someone performs an action. This reasoning reveals their intent, either malevolent or beneficent. There is no right or wrong in this regard, but a lack of awareness on such differences will hinder research and cooperation.

For a particular project, or for bringing together several communities, one may wish to select or create a (new) scheme for surveillance for security. After a project team or community reaches consensus on a set of requirements (focus of this paper), the phase of selecting or creating a (new) metadata scheme can start. This step can be supported with upper ontologies (like Cyc, SUMO and DOLCE [20,24,29]) and accompanying ontology mapping and/or database linking. That phase is out of scope of this paper. This paper offers guidelines to those who wish to select or create a (new) scheme for surveillance for security, both on the level of a (global) surveillance community, and on the level of a local project team.
The work presented here is the first step in the development of a complete metadata scheme for surveillance in security. Our contribution is that we propose a clear terminology, propose requirements, validate them in a typical, complex surveillance scenario, evaluate them against common design patterns in surveillance, and illustrate how existing schemes support or neglect these requirements.

This paper is structured in a way that puts the proposed requirements in focus. Section 2 summarizes related work. Section 3 describes the terminology. In section 4 we propose the requirements for a metadata representation scheme for surveillance for security. Section 5 evaluates the requirements extent to which the requirements are useful (for typical surveillance practice), practical (by applying them to an example) and realistic (by evaluating several existing metadata schemes against them), and section 6 discusses the findings of our evaluation. Section 7 summarizes the main conclusions.

2. Overview of Related Work

This section describes the context and related work on surveillance and existing metadata schemes. In 2005 Francois [12] focused on the importance of describing events in multimedia data, and continued by showing how relations between events and objects can be used to describe multimedia data in detail. In 2007, Westermann [46] built a case for a common event model for multimedia data. He mentions multiple new perspectives on events, e.g. he describes why causality as a type of relation between events is worthwhile to describe. In the same year, Annesley [2] started from the needs of the application domain of surveillance, and gave three types of requirements: from the scene, from the system, and non-functional requirements. Schallauer in 2009 [37] further elaborated on the notion of requirements from the surveillance domain for the design of metadata schemes, i.e. they described 11 requirements which are easily recognisable, e.g. “description of event and their properties”, “flexible in terms of modalities and sensor types” and “abstraction levels”. SanMiguel in 2009 [36] further elaborated on the requirements of the system, and particularly also in the capabilities of a surveillance system, e.g. in terms of algorithms such as tracking. Each case of this previous work introduces its own metadata scheme which distracts from the quality of the requirements that they gave. None of these papers give an evaluation of the requirements which they started with, for example, none of these papers is explicit in terms of which, or how many abstraction layers are needed.
For the evaluation of our requirements on realism, several existing metadata schemes were selected. This is done without aiming to obtain a complete list of all schemes, since the goal of our paper is to propose and evaluate requirements for such types of schemes and not to present an extensive survey. Our selection of metadata representation schemes for this paper is based on three disjunctive criteria, i.e. qualifying on one criterion is sufficient for selection. The three criteria are: (i) familiarity to audience, (ii) capability of reasoning about intent/behaviour, and (iii) footprint of current application. The rationale for selecting these criteria are explained next. As a consequence of criteria (i) and (iii) schemes are covered that are considered and used for research and also applications. This is important because surveillance is both an active research topic and it is also widely applied in practice. Due to criterion (ii) schemes are covered that are not yet known well and/or are not yet applied in practice. This category of schemes includes the ones that are relatively new and aim to address insufficiencies in the well-known schemes, i.e., including metadata beyond the sensor, beyond the static scene, and beyond the standard processing of objects and tracks. This paper covers new schemes that include metadata about intent and behaviour. These are important elements for effective surveillance, because one of the objectives of surveillance is to early detect and potentially prevent crimes. Assessment of intent and behaviour facilitates reaching this objective.

The first criterion is familiarity with the relevant communities of multimedia research, behaviour research and surveillance industry. These schemes are therefore taken from international standards bodies (ISO, W3C [13], MPEG-7 [26,38,39]), scientific languages (VIPER [8,23,45], CVML [7,21], SAM [37], ETISEO [10,28]), scientific conferences and benchmarks (TRECVID [33,43], PETS [11,34], CAVIAR [6]) or governmental benchmarks (i-LIDS [14]), or relevant communities such as the industry collectives PSIA [35] and ONVIF [32]. The second criterion is coverage of some of the newer concepts such as reasoning about intent and behaviour (HumanML [13]). The third and final criterion is whether the scheme is (commercially) applied, i.e. has a large footprint (Milestone [25], Noldus [30], ANVIL [17,18]). Several of these languages are based on VIPER (e.g., PETS, iLIDS, TrecVID) or CVML (PETS, CAVIAR) and many are based on XML (e.g., CVML, VIPER, HumanML, ONVIF scene description interface, SAM, Milestone alert data). Together, this gives us a broad view, from relatively "open" languages that can be used for a wide range of applications, to more specific languages that were developed to serve a specific goal.

3. Terminology

To create a single reference frame for communication between relevant communities, a coherent terminology is proposed in this paper. This terminology is mainly based on terms typically used in metadata representation schemes and is enriched with frequently used terms from the surveillance and the system engineering domains. Sections 4 and 5 demonstrate that this terminology is indeed suitable for defining metadata for surveillance.

The term metadata itself may need some additional clarification [http://en.wikipedia.org/wiki/Metadata]. In this paper the definition of *descriptive metadata* is used, which is data describing instances of application data. However, in many emerging technologies data on the structure of the system are used during the runtime phase of the system, because this allows for adaptivity, autonomy, traceability and transparency. The manner in which the metadata are acquired (e.g. through manual annotation, automated sensing or through configuration files) is beyond the scope of this paper.

Surveillance is the process of watching over a person or place [22]. In general, this is done to protect someone or something and, as such, has different specific goals before, during or after a specific incident. Typical surveillance examples cover prevention (deterrent), preparation (training or creating a baseline for normal behaviour), mitigation of threats and direct after effects and investigation after the incident. In this paper, surveillance is limited to security, e.g. fighting crime and terrorists.

A *system* is a construct or collection of different elements that together produce results not obtainable by the elements alone [15] (Fig. 1). A *perceptual system* is a system that has the ability to perceive itself and/or its environment. A *surveillance system* is a perceptual system with the function of surveillance. A typical surveillance system consists of machine components (e.g., the sensors, storage and communication) and a human operator.

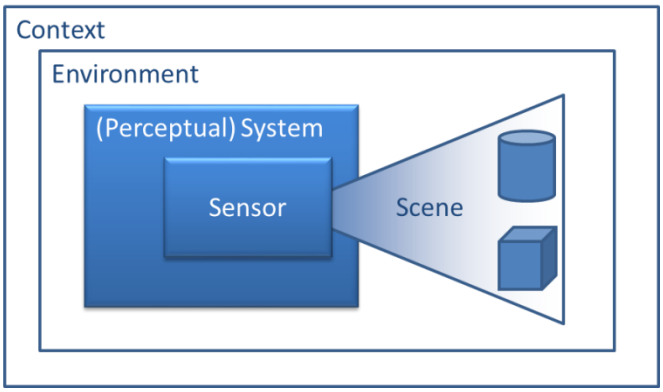

Figure 1. The relations between context, environment, (perceptual) system, sensor and scene.

The *environment* is the system's surrounding that could interact with the system. The typical environment for a surveillance system is the area under surveillance and the location(s) of the system components. A *scene* is information that flows from a physical environment into a perceptual system via sensory transduction. Within the perceptual system this information is formalized as a subset called 'image of the scene' and is at the lowest abstraction level represented by recorded footage.

The *context* of a surveillance system consists of the factors that influence the system and necessarily include the environment and the scene. Typical examples of surveillance context are the local culture, the level of terror threat, and the weather conditions. Additionally, world knowledge as prior probability, and known correlations between events and actions, are also a part of the surveillance-system context.

The data from the sensor can be interpreted at different levels of abstraction [16,40] as illustrated in Fig. 2. Using specific terminology for metadata on each level greatly facilitates scientific and engineering discussions.

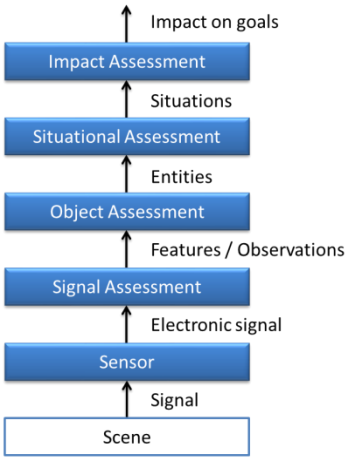

Figure 2. Multi-layered view on abstraction of information (JDL or NAIHS model [16,40]).

A sensor produces *signals*, which can be processed to form *observations* (synonym *features*). Observations are individual measurements. Multiple

observations can be made of the same phenomenon over time, through different sensors, from different viewpoints or in different modalities. For example, the colour of an object, as observed by multiple cameras under different lighting conditions, results in different pixel colour values for the same object. Even though the colour of the real object is the same, the actual internal representation from multiple cameras is different. A surveillance system needs to be accommodated to cope with conflicting observations, possibly even benefitting from it.

An *entity* is an object (or part of an object) whose presence is relevant to the surveillance system. Typical entities are cars, persons and buildings. The *relevance* of an entity to a surveillance system establishes the specificity with which it needs to described. For example, a person in a car could be describe by one entity for some applications, but could be two for other applications. There is a tension with being exhaustive in the entities in the ontology. It is common to use the part-of relationship, object inheritance and spatial relations to model complete worlds of relevant entities. The perceptual system and subcomponents themselves can also be entities. This is typically the case when the function of the surveillance system depends on being able to reason about its components, e.g. to optimise resource allocation ("where do we point this pan-tilt-zoom camera?"), or to create resilience in case of malfunction or attack against the surveillance system itself.

Security and surveillance are concerned with (among other issues) protecting against harmful actions. Such harmful actions (i.e. protecting, managing resources) are done by an *autonomous agent*. This is a subclass of the class "entity", and it can be a person or an autonomous automated system that takes actions based on perceptual stimuli. The complementary class of an agent is an *inanimate object*, e.g., buildings, locations and vehicles. Depending on the application it is relevant to make a further distinction in *immovable inanimate objects* such as buildings and *moveable inanimate objects* such as a briefcase.

An *action* is something done as a movement in contrast to doing nothing, and can only be performed by an agent with a certain level of autonomy. An action is limited in diversity, is more or less continuous, and is seen outside the scope of a stimulus-response process. Examples of actions are walking, smoking, shopping and driving a car. The *activity* is similar to an action, but considers a longer period of time. Taking a plane to travel to another country is no longer an action, but should qualify as an *activity*. An *event* is an observable occurrence of something: an *action* performed by someone at the scene, or a state change within the perceptual system. For example, a power loss on one of the surveillance cameras could be detected and classified as an event, but would not classify as an action.

Entities, actions, events and their attributes are observables [4]. In their context they form a *scenario*: a synoptical collage consisting of a meaningful series of actions and events. One inseparable moment of a scenario is a *situation*: the constellation of objects and their interrelations, placed in an environment and a context. Within a scenario, entities are involved in the events and actions. Sometimes as actor, sometimes as subject of an action. Therefore, the entities

have a *role* in the scenario. A scenario occurs in the environment of the surveillance system.

Behaviour is the reaction of an autonomous agent to a stimulus in relation to its environment, which includes other agents or the perceptual system itself. This reaction can be immediate or delayed, overt or covert, conscious or subconscious, voluntary or involuntary, and internal or external. Behaviour expresses the relation between stimulus and reaction, as compared to activities expressing the actions alone and, therefore, provides more insight into an agent's intentions. A stimulus can intentionally be introduced as an intervention designed to reveal deviant behaviour. In such a case, the way the subject reacts to the stimulus discloses its intent. The consequence is that behaviour cannot be described as a stand-alone activity, but must be expressed in relation to an agent, a stimulus to that agent (i.e. the history of events within the perceptual range of that agent) and his environment.

A *goal* is a desired end-state (situation) for an agent. An agent can aim for multiple, sometimes conflicting, goals at the same time. For example, an agent's goal can be to mitigate the impact of calamities, to prevent crime, or to collect evidence in the case of crime. *Intent* is the state of an agent that directs its action towards a specific goal. Intent towards an illegal goal is what is required to make an illegal action a *criminal action*. Such intents are also known as a hostile intent.

A *security risk* is the occurrence probability of a particular unwanted situation. A risk is chance times impact. If the unwanted situation actually occurs (chance equals 1) then it becomes a security incident. A risk refers to a situation that has not yet occurred, and therefore cannot yet be observed. The *impact* of a scenario is expressed as the impact on the goal of a particular agent. A *plan*, the means an agent utilizes towards its goal, consists of a set of actions resulting in a scenario that achieves the goal. A *modus operandi* is the legal term for a plan with an illegal or criminal goal [42]. A *security threat* is an indication or warning of a security incident, which can be more or less concrete.

An *ontology* is the working model operating on a set of concepts, such as entities, that are specified to create a vocabulary for exchanging information. Typical problems for an ontology include the clustering of entities, relating within a hierarchy, and subdividing entities according to similarities and differences. The terms defined in this section create an ontology; however, for the purpose of surveillance, surveillance-related concepts should also be introduced in the ontology, e.g. inanimate *object:fence*, *action:running* and *event:sensor* failure.

The next sections discuss respectively requirements for, and applications of, surveillance metadata representation schemes, using the terminology from the current section.

4. Requirements for Metadata Schemes

Nine requirements for metadata schemes for surveillance for security are identified. The purpose of this section is to be as specific as possible with regard

to these requirements. The first requirement is about *coverage of the surveillance domain*. Next follow five requirements which are related to the five abstraction levels of Figure 2. The final three requirements are related to *traceability*, *uncertainty and chance* and *observation capabilities*. In section 5, for the schemes introduced in section 2, it is determined whether they support these requirements.

4.1 Coverage of surveillance domain (requirement 1)

Concepts should be specified to cover all relevant concepts in the surveillance domain. For example, the concept immobile inanimate object will, for any useful application in surveillance, probably need to be further specified into house, office building and road. This allows to take the difference of such concepts into account when reasoning about them. Which level of detail is needed depends on the specific goal and context of a particular surveillance system. For example if the goal is merely intruder detection in a sterile zone, then the concept of a car is sufficient. If the goal is to detect suspicious activity on a parking lot with lots of cars and people, then a notion of an (perhaps yet unseen) person in a car may be necessary. In the evaluation of the requirements in section 5 it will be demonstrated which concepts and relations should be defined for a pickpocket case. In each of the next five requirements (about data abstraction) examples are given of concepts which should be included in an ontology for surveillance.

There are limits to the amount of specificity that is useful for surveillance. For example, in the surveillance domain, the relevance of objects smaller than on the level of "limbs" is usually very low. Below that level, they would simply be attributes of objects, such as "hand-palm-scan", or "face-scan". However, in the medical domain, this is typically where it starts to become relevant, and the definition of smaller objects is more than relevant, e.g. "kidney", "aorta", etc. The mapping of such an ontology to ontologies of other domains is beyond the scope of this paper. Too much freedom in covering concepts may lead to bloated ontologies and redundant concepts, which hamper usability.

4.2 Metadata about sensor (requirement 2)

Metadata about the sensor and resulting signal should be described.

The first abstraction level comprises the sensor. Its type (camera, radar, microphone), placement, orientation (i.e. internal and external calibration), platform, cardinality, resolution, internal clock, and configuration determines a lot about the signals coming from it. For active sensors the location and type of the transmitting device could also be relevant, such as the radar transmitter or light source.

These data are often crucial to interpret signals coming from sensors. E.g., without these metadata, the forensics professional is left to analyse the (video) signal itself to determine the possible contents and time stamps of data. Aspects of the resulting signal, such as spatial and temporal resolution, determine what can be inferred from the data. With technologies such as super-resolution, stitching and sensor arrays, it is possible to construct a "virtual" sensor from more than one sensor(type), which still "only" produces a new type of signal. Relevant information about the surveillance system itself can be derived from this metadata, such as covered area of the sensor plan (including white spots and

overlaps), the dependencies of resources such as communication and power and whether a component has been tampered with. Another motivation to describe metadata about the hardware/software of a surveillance system is *standardization in interoperability*. Yesterday's threat may not be the same as the threat of tomorrow. This implies that not all relevant connections for a surveillance system may be known at design time. Depending on future unknown events, it may be necessary to connect a surveillance system to other systems or data sources. This *system of systems* approach requires metadata on the system components to be available, in order to manage interoperability runtime.

4.3 Metadata about observations (requirement 3)

Metadata on observations should be described separately from metadata on objects, their attributes, events and actions.

Entities, actions, events and their attributes can be observed by multiple sensors and different modalities simultaneously and/or successively. A scream can be heard, but can also be seen by observing the face. Multiple sensors can simultaneously observe the same property or activity. Making an observation of a person in the scene does not mean that the system automatically knows to which specific internal representation of persons this observation must be attributed to. Typical solutions for signal attribution are detection, tracking and recognition technologies, assisted by a robust geometrical model of how sensors physically relate to the environment and/or to each other i.e. a calibration and a synchronisation. This requirement becomes more important when the number of sensors increases, driven by falling costs of sensors and the benefits of overlapping sensors, such as alternative viewpoints in case of occlusions, redundancy during component failure and seamless integration with neighbouring sensors. Examples of concepts which should be in an ontology for this requirement are detections (bounding boxes) of vehicle, number plate, person and face in video signals and scream- and broken-glass-detections in audio signals.

4.4 Metadata on entities, events, actions and attributes (requirement 4)

Metadata about observables should be described.

The third abstraction level is the level of entities, events, actions and their attributes, i.e. the observables. Which observables are relevant is determined by the purpose of the system. For some applications a moving car is one object, for others a moving car consists by definition of two objects: the car and the driver, and yet another application even allows uncertainty about the existence of the driver until he can be seen directly, or can be inferred from the car's behaviour. The identity of a person is also covered on this level. Examples of concepts which should be in an ontology for this requirement are:

- for entities: persons, vehicles, groups of persons and their location;
- for events: sensor failure;
- for actions: giving, taking, talking, falling, hitting and running;

4.5 Metadata about situations and scenarios (requirement 5)

Situations and scenarios should be described by describing relevant relations between observables.

So far, there are sensors, signals and observations, and entities, events and actions and their attributes. The new element to be introduced is the relation between these concepts, because together they make up *situations* and, when seen over time, *scenarios*. These relations can be of different natures, such as *temporal* (before, while, after) and *spatial* (above, under, in) [27,46]. Other, less conventional relations are equally possible, such as *interaction*, including *communication* (talks with, gives to, follows) and *legal* (owns, is married to, suspected of). These relationships are essential to interpret the situation. The relationship *causation* between actions and events is also relevant for surveillance and security. Hostile agents (pickpockets, burglars, terrorists) execute a *modus operandi* because it *causes* their desired goal. Recognizing such chains of cause and effect in the environment of a surveillance system helps to predict (unwanted) situations. This makes a surveillance system less dependent (and therefore vulnerable) on pre-configured rules, and allows the system to reason ad hoc about the effects of certain situations, which increases the resilience and adaptivity of the surveillance system.

4.6 Metadata about risks and impact (requirement 6)

Goals, hypothetical scenarios, risk and impact should be described.

Typically, a security system, including the human in the loop, is configured in such a way that it has an aggregated positive impact on the goals of the security system (i.e. to prepare for the case of calamities, to prevent calamities, to mitigate the impact of calamities or to collect evidence in the case of calamities). Describing goals, hypothetical scenario's, risk and impact also lets the security system use information about goals of others, such as those of criminals or those of potential victims. This requirement then also facilitates the emergence of (semi-) autonomous surveillance systems, such as UAVs. If a goal is a desired end-situation then this requirement requires the possibility to describe a situation which does not exist yet, and to link this to a cognitive agent as one of his goals.

4.7 Traceability (requirement 7)

Metadata should describe the origins of data.

The ability to trace where a particular piece of information comes from, allows surveillance systems and their maintainers, designers and researchers to trace system errors and false hypotheses about the known world back to its origins. For forensics there is the question of authenticity: has this information been tampered with or not? Being able to reconstruct the path of how the information came to be in the first place, is then very helpful. This is only possible if metadata contains references to the abstraction level directly below it. For example, a signal should have a reference to the sensor (or human) that it came from, an entity should have references to the observations in signals that its instantiation was based on, and a situation should have references to the entities it is comprised of.

4.8 Uncertainty and chance (requirement 8)

The metadata representation scheme should facilitate that signals, observations, observables and situations are accompanied by the level of certainty that the responsible subcomponent has with regard to the accuracy of it and any alternatives.

Uncertainties are part of every step of data and information processing and can arise from all sorts of angles. The precision of the calibration of sensors is limited. Memory and bandwidth constraints in humans and machines lead to the rounding of numbers, and there may be multiple conflicting hypotheses about the real world, all in some degree supported by the available data points. The purpose of a surveillance system is not to have the perfect situational awareness, but to have a situational awareness that is good enough to get the job done. Typically this leads to the goal of *information dominance*, i.e. the situation that your information position is significantly better than all realistically possible opponents. When one has information about the confidence level of a piece of data (e.g., of detections or matches), one will typically also want to have alternatives to that data. In many ways of describing certainty or chance, alternatives are intrinsically given, e.g. a binominal distribution for the height of a person.

4.9 Observation capabilities (requirement 9)

The metadata representation scheme should facilitate the description of available (technical) capabilities in relation to possible data requests.

A surveillance system should be able to reason about alternative ways to get the required type of information, which requires knowledge about the available technical capabilities in relation to the data request. For example: is there a capability of zooming in to get more resolution and contrast on the face of that person, or is there another camera available? This requirement is not limited to the sensor, but also includes processing capabilities such as: “is face recognition available?”, or “do we have tracking available?”, and “can we combine audio with video to determine who started shouting in a crowd?” It is also useful to describe (the capabilities) of human operators: “”how many operators do we need to monitor these video feeds live?”

4.10 Other requirements

When actually designing a metadata scheme, more requirements than these will have to be taken into account: *streaming metadata*, *openness* and *ease-of-use* are some that come to mind. However, these appear to be requirements that are generally addressed with either tooling, methodology or legal actions and, therefore, do not depend on the metadata scheme as such.

5. Evaluation

This section evaluates the requirements in three ways. First, the benefits of the four requirements related to abstraction levels are evaluated in breadth in terms of common and emerging surveillance design patterns (Sec. 5.1). Second, the need for the requirements is illustrated in depth in a complex surveillance example of

pickpocketing (Sec. 5.2). Third, the requirements are evaluated against existing metadata schemes to identify which scheme fulfils which requirements (Sec. 5.3).

5.1 Evaluation on common surveillance design patterns

This section evaluates the four requirements related to abstraction levels (requirements 2-5) against good practices for surveillance systems. In our experience with the police and commercial surveillance companies, we found that surveillance is commonly done with a limited set of basic design patterns. Design pattern are general reusable solutions to commonly occurring problems, which are often used in architecture [1] and object-oriented software development [5]. We define these 'good practices' in the surveillance domain now explicitly as design patterns, or *surveillance patterns*. These patterns have similar purposes (create *situational awareness*), and similar input (e.g., video, sound, tweets) and output (hit/alarm or no-hit). We identify the following surveillance patterns: "threshold alarm", "profiling", "concentric circles of protection", "bag of observations" and "scenario view". These surveillance patterns can be applied by both machine and human, but a human professional can shift seamlessly between these patterns, while machines must be explicitly designed to apply them. Each pattern has its own strengths and weaknesses, so there is no perfect surveillance pattern: each pattern has to fit requirements such as efficiency, efficacy and lack of invasiveness, all of which depend on the local situation. The surveillance patterns are described in order of increasing complexity.

The surveillance pattern *"Threshold Alarm"* works on the basis of putting a threshold on the attributes of a single observable. Typical and often implicit reasoning in this pattern is to consider only the presence (or absence) of an entity, e.g. *smoke* or *a person*. This pattern is used in situations where a very specific risk is present (i.e. *fire* or *burglary*) in an environment with very little distractions, which allows for highly specialized, yet relatively simple observation. A metadata representation scheme that accommodates this pattern must be able to separate observables (smoke or a person) (req. 4) from situation assessment (*fire in the building*, or *a burglar on the premises*) (req. 5) to apply the threshold in between.

The surveillance pattern *"Profiling"* extrapolates information about an entity from multiple other data points. In security and surveillance this surveillance pattern is often implemented in border control and object security. E.g. behaviour profiling uses data about behaviour of a person to make an assessment about the intent of that person. The data representation scheme must facilitate linking multiple aspects, actions and behaviour to a single entity [3]. In the case of behavioural profiling, as all of these actions and behaviour will (generally) not happen at the same time. This could require functionality such as person tracking and recognition, which in turn may require additional support from the representation scheme, such as the description of an *identity* (req. 4), or some locally unique (biometric) feature.

When combined with physical barriers, the surveillance pattern *"Concentric circles of protection"* has the function of containing the threat in compartments. This pattern is typically used in border security, object security and VIP protection. In general, this patterns allows the surveillance system to designate the

relation between one entity and another entity that needs protection. At a minimum, this requires a general idea of the asset to protect (req. 4) and the notion of relative locations (req. 5). With regard to the perceiving components of a surveillance system (i.e. sensors and surveillance personnel), their location should be known in relation to the compartments in order to know to which compartment their output should be attributed.

New surveillance patterns are also emerging. The *"Bag of Observations"* pattern works by indiscriminately combining multiple observations to estimate the situation. In the surveillance industry this pattern is experimentally used for situations with very difficult reasoning about an individual person: e.g., crowd management, or urban security. In such contexts, much data are available and can be mined for general behaviour patterns. Combining multiple observables facilitates more robust performance than by putting a threshold on a single observable. On the other hand, this surveillance pattern misses information on the interaction between different entities (e.g., communication or movement patterns), or on the relation between events and actions (e.g., successive actions are performed by the same person, versus by different persons), so it does not use requirement 5.

A second emerging surveillance pattern is the *"Scenario View"*. With the surveillance patterns described above, it is not possible to describe all aspects of scenarios with complex behaviour, i.e. behaviour which involves changing relations between multiple persons. In the surveillance pattern "Profiling" it is difficult to describe the communication between collaborating pickpockets, the interactions of a drugs deal in a city square, early signs of a rip-deal in an airport or simply the relation *ownership* in a lost-luggage scenario [41]. This illustrates the need for a pattern that includes dynamic relations between entities such as *X can see Y*, *X speaks with Y*, *X follows Y* and *X owns Z*. This surveillance pattern needs -in fact exploits- requirement 5.

These five surveillance patterns have the purpose of creating situational awareness by analysing from low level sensor data all the way up to situational awareness. They require support in the metadata scheme –either in their output, their input or in their internal working- for four levels of information abstraction, which are addressed by four of the requirements. Table 1 compares these four requirements against these common and emerging surveillance patterns. The next section 5.2 illustrates *how* the surveillance patterns use this information.

Table 1: Comparison of surveillance design patterns on four of the abstraction level requirements. A cross in a cell means that this pattern needs this requirement to be addressed in order to be able to work.

	Common			Emerging	
Requirement (number)	Threshold alarm	Profiling	Concentric circles of protection	Bag of observations	Scenario view
Metadata about the sensor (2)	X	X	X	X	X
Metadata about features/observations (3)	X	X	X	X	X
Metadata about entities, events, actions and their attributes (4)		X	X	X	X
Metadata about situations and scenarios (relations between observables) (5)			X		X

Several conclusions can be drawn from this table. First, these four requirements (abstraction levels) are all actually needed by these surveillance patterns. Second, no additional requirements are needed (with regard to the metadata scheme) to support these surveillance patterns. Third, this table shows, not surprisingly, that these surveillance patterns all require metadata about the sensor and about features. And fourth, there also seems to be a correlation between the complexity of a surveillance pattern, and both the number and abstraction level of requirements. Together these conclusions support the first main conclusion that these are the right requirements given these common and emerging surveillance patterns.

5.2 Evaluation on a pickpocket case study

Current state-of-the-art automatic video analysis tools focus on surveillance tasks that are simple for humans, such as person tracking and action recognition. For an in depth evaluation of all proposed requirements, we focus on the complex task of detecting a pickpocket scenario (see online video [31]), which contains many of the simpler tasks.

The *modus operandi* was constructed by the Amsterdam police, based on actual observations of real pickpocket scenarios. The scene on the video is observed through a colour pan-tilt-zoom camera. Six agents are involved in the scenario: potential victims (V1, V2, V3) and pickpockets (P, P2, P3) (Figure 3). The relevant inanimate objects are the tram, the tram stop, and the road next to the tram stop. Relevant actions are standing still, walking and snatching for the victims and pickpockets, and moving, standing still and open doors for the tram.

Figure 3. Cropped frame from the pickpocket video. The pickpockets are looking for suitable victims.

Together, the persons in this scenario step through different phases: The pickpockets are loitering at the tram stop to find a victim. When a tram approaches in the distance, one pickpocket (P3) selects a candidate victim (V3). The victims are queuing to enter the tram (Fig. 4, left) and P1 pushes himself in front of V3, P2 is close to V3 in the queue, and P3 moves closer to the queue. P2 takes something from V3 and gives the loot to P3 under cover of the jackets over their arms (Fig. 4, right). Finally, P1 and P2 travel with the tram and P3 (with the loot) stays behind at the tram stop. This example is used to illustrate the relevance of requirements.

Figure 4: Cropped frames from the pickpocket video. The victims enter the tram just before P1 pushes himself in the queue (left), and while P2 gives the loot to P3 (right).

Table 2 presents the importance of requirements 1-5 through the lens of the surveillance patterns on this scenario. It shows how having requirement 5 enables the use of more complex surveillance patterns, which allow for more precise and complete descriptions of the scenario as can be seen in the column "Summary". The "Alarm Threshold" surveillance pattern selects one type of event in this pickpocket scenario that is most indicative, e.g. the *snatching*, or the event that *someone stands near the queue, but is not actually boarding the tram*. The next pattern "Bag of Observations" would select multiple events based on a general vicinity of e.g. the tram platform, e.g. the actions of *someone observing the*

platform from a distance, *someone jumping the queue* and *someone not-boarding after standing near a queue*. When seen together (i.e. within a certain short time span and at one location), these observables would be indicative for a pickpocket scenario on the tram platform. The surveillance pattern of "Behaviour Profiling" allows us to reason that P1 performed several suspicious actions within a short time span. The fourth surveillance pattern "Concentric circles of protection" allows to take the pure perspective of the asset to be protected against stealing, such as wallets. For each person on the tram platform, we would take a number of circles, e.g. taking the idea of "personal space" [http://en.wikipedia.org/wiki/Personal_space] as starting point. This would allow us to reason about events taking place in e.g. the intimate space of V3, which would normally not take place there. The fifth and final surveillance pattern "Scenario view" allows us to take into account the interaction between the pickpockets and the collaboration around the theft. We can reason about P1 having interacted with P2 and P3, then moving away from each other, and then coming together again on one tram entrance.

Table 2 – Five surveillance patterns for the events and actors that are involved in the pickpocket scenario.

	Events		**Observing platform from a distance**	**Waiting on platform**	**Communicating to someone**	**Moving in front of someone else**	**Stalling line**	**Snatch something**	**Handing something over**	**Board tram**	**Remain on platform**	**Summary**
Surveillance patterns	**Hypothetical detection rate**		High	High	Low	Medium	Medium	Low	Low	High	High	
	Indicativeness of scenario		Low	Low	Low	Low	Low	High	Low	Low	High	**Summary**
Alarm	Pickpocket							P2				**P2 snatched**
Bag of words	Pickpocket		P1	P2, P3, V1-3	P3 tot P2 and P3 to P1, V1-3	P1	P1	P2	P2 to P3	P1, P2, V1-3	P3	**All events occurred**
Behaviour Profiling	P1		Yes		Yes	Yes	Yes			Yes		**P1 behaved rude**
	P2			Yes	Yes			Yes	Yes	Yes		**P2 stole something**
	P3			Yes	Yes						Yes	**P3 received something**
	V1			Yes	Yes					Yes		**V1 acted normal**
	V2			Yes	Yes					Yes		**V2 acted normal**
	V3			Yes	Yes					Yes		**V3 acted normal**
Concentric circles of protection	V3: personal space			V2, V3	V1-3	P1	P1	P2	P2	P1, P2, V1-3		**Close to V3: some unusual events with several different people**
	V3: within observation distance		P1	P1-3	P1-3						P3	**Nothing unusual around V3**
Scenario view	Pickpocket scenario	Waiting for easy victim	P1	P2, P3, V1-3								**A specific pickpocket scenario took place, it involved P1-3 as pickpockets and V3 as victim.**
		Selecting victim			P3 tot P2 and P3 to P1, V1-3							
		Position relative to victim				P1						
		Distracting intended victim					P1					
		Snatch valuable						P2				
		Hide Loot							P2 to P3			
		Leave location								P1, P2, V1-3	P3	

We conclude that storing not just the personal data itself, but also metadata about where (Req. 2) and why (Req. 6) it was recorded explains that a surveillance system is used near a tram stop to prevent or mitigate the threat of pickpockets, which may be beneficial for privacy and legal aspects. Describing how it was captured and processed (Req. 2-6) explains that a surveillance system is, e.g., comprised of human operators with a particular training, or an automated processing system that acts on the location of people. Describing the relation between raw sensor data, personal data and the use of personal data (Req. 7) gives transparency and allows for the correction of errors in the system. For example, in the pickpocket case, the alternative hypothesis can also be described that it is not P2 who is the thief cooperating with P3, but that P3 is the thief and P2 is actually a victim. The description of runtime choices and potential alternatives (Req. 8) gives insight into why certain errors or decisions were made. Finally, the description of additional observation requests (Req. 2 and 9) gives more insight into the amount of invasiveness of the surveillance system, in this case the fact that the video zoomed in on these people boarding the tram.

Based on these conclusions, we draw the second main conclusion that these requirements are indeed sufficiently rich to describe also the more complex situations in the right amount of detail.

5.3 Evaluation of requirements on existing metadata schemes

This section describes how current metadata schemes (Sec. 2) address the nine requirements (Sec. 4) already. Based on this comparison we can both learn whether the requirements are feasible and get an impression of the coverage of these requirements is by current schemes. Publicly available data were used to perform the analysis. TRECVID and PETS were also included in the analysis, although they are benchmark conferences that use CVML and VIPER as schemes. However, in their use of these schemes they have created labels and attributes, which alters their fulfilment of several requirements and therefore, they are discussed separately. The result of our evaluation is presented in Table 3.

	HUMAN ML	ETISEO	I-LIDS	Noldus	Milestone	PSIA	MPEG7	CVML	ONVIF	SAM	ANVIL	VIPER	TRECVID	PETS
1. Coverage of relevant domain	Yes [d]	Yes	Yes	Yes	Yes	Yes	No [c]	Yes	Yes	Yes	No	No [c]	No [c]	Yes
2. Metadata about the sensor	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
3. Metadata about features/observations	No	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No
4. Metadata about entities, events, actions and their attributes.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Fixed in SIN, KIS	Fixed
5. Metadata about situations and scenarios (relations between observables)	Yes	Fixed	Fixed	Yes	No	No	Yes	Fixed	No	Yes	No	No	Fixed in SED, MED	Fixed
6. Describe goals, hypothetical situations and scenarios	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No
7. Traceability	No	No	No	No	Yes	Yes	Yes [b]	No	Yes	Yes	No	Yes [b]	No	No
8. Uncertainty and alternatives	No	No	No	No	No	No	Yes [b]	No	No[a]	Yes	No	No [a]	Yes [f]	No
9. Observation capabilities	No	Yes [b]	Yes [b]	No	Yes [e]	Yes	Yes [b]	No	Yes	Yes[b]	No	Yes [b]	No	No

Table 3: Comparison of metadata schemes on the 9 defined requirements. Notes: (a) Only probabilities, no alternatives. (b) Not explicitly, but can be modelled. (c) Only some general low level signal-related concepts. (d) Scope of human communication. (e) Not explicitly, but can be implemented using software API. (f) Each shot is given a separate probability per concept which could be seen as alternative hypotheses. (Fixed) Not possible in general, but available for a fixed set of predefined events/scenarios.

There is no scheme that fulfils all requirements (i.e. there are no rows completely filled with Yes). Some widely used schemes in different scientific communities score on about half of the requirements. These metadata representation schemes (like CVML and ETISEO) are designed to evaluate specialized research on new techniques e.g. in Video Analytics. To get beyond the state of the art these schemas facilitate the more abstract requirements 5 and 6, but only in a limited form for fixed events and/or scenarios. In an academic setting of researching surveillance systems, i.e. where a specific research focus determines the constraints, it is not strange to see that requirements 7 and 9 are seldom met.

There are representation schemes that have an *open* character (e.g. MPEG7, VIPER and in extreme XML). Their openness allows them to describe almost anything. However, this also has a drawback: the more open a specification, the less specific it is. E.g. one might argue to use natural language English for video annotation, claiming it has sufficient descriptive power; however, choosing a natural language involves considerable loss in specificity and causes ambiguity and interpretational issues. For example, MPEG-7 potentially scores very high. The only missing requirement seems to be the inclusion of a relevant ontology specific for surveillance (Req. 1). Other schemes are also flexible or extensible, but none as much as MPEG-7. The great flexibility of MPEG-7 can however also negatively impact the ease of use for novices. CVML has even made a design choice that goes against requirements 3: three levels of abstraction are described in one XML-tag: "*Reporting tracked entities in a scene is done with the <entity> tag, where data such as the bounding box and orientation, but also high-level information such as role and scenario is provided. Groups of entities have their own bounding box, role and scenario, and this is output for each frame in a video sequence.*" [7]

Some of the representation schemes contain technical detail (e.g. MileStone, ONVIF and PSIA). This is understandable since they describe working systems and reflect the current state of the technology. From these representation schemes we can deduce what is needed to actually *implement* a complete scheme. However, the respective companies and consortia have found no reason (yet) to accommodate for requirements 5 and 6 in these formats.

Two of the requirements are seldom met in metadata schemes: (6) *goals and hypothetical scenarios*, and (8) *uncertainty and alternatives*. This is not surprising, as they enable emerging functionality, such as influencing behaviour, multi-hypothesis reasoning and autonomous surveillance systems. They might be addressed with classical AI frameworks. Since *goal* is expressed in terms that were already needed for other requirements, such as *scenario* and *agent*, requirement 6 about *goals and hypothetical scenarios* would require a relatively simple extension of the metadata representation scheme. It could even be expressed by using the link we already need for *traceability* to express that this scenario does not originate from sensor data, but rather from the mind of an agent.

Requirement 8 *uncertainty and alternatives* may require some additional work. Describing alternatives does not require large changes in schemes, merely the flexibility to pass more than one instance of data instead of one. Uncertainty may

be slightly more complex. On one hand, it seems easy to pass along one extra variable about chance or a distribution of values. On the other hand, the way uncertainty should be modelled on data can depend on the data type and on the way the data were obtained. This may require a more elaborate description of uncertainties.

The third main conclusion is that there is no fundamental problem with implementing these requirements, because for every requirement there is at least one existing metadata scheme which supports it. In addition, for the requirements which have relatively little support, we can give practical guidelines on how they could be implemented.

6. Discussion

This discussion focusses on suggestions for further development of a metadata scheme. We foresee three approaches for further development of metadata schemes where the requirements identified in this paper could be applied. The first approach is, for example, after a large security incident where a reconstruction of the incident must be made from multiple sets of metadata. Such a reconstruction could use database linking and ontology alignment; however, these are complex processes that require making many choices which impact on the usability of the resulting reconstruction. These requirements could serve as a guideline for these choices. The second approach is the construction of a new metadata scheme that should serve as a unifying metadata scheme among surveillance systems of all kinds. This scenario would typically start with a thorough stakeholder analysis in order to validate these requirements, and to determine the support in the relevant communities for each requirement. Several options for fulfilling each requirement may be considered, and multiple metadata schemes should be assessed for their fitness as a starting point for the new scheme. For example, based on the evaluation of requirements on existing schemes, MPEG-7 could be a basis on which to build one coherent scheme, or PSIA and ONVIF could extend this (e.g. with MPEG-7), or one of the scientific schemes could be gradually extended. Home Office CAST appears to have initiated the creation of a new scheme [44]. Ontology linking may be part of such a procedure. The third approach is the construction of an ad hoc metadata scheme for a local project or software tool. In this scenario the engineers simply go through the list of requirements and select those that are required for their own goal. In this scenario the requirements are simply a checklist.

The requirements collected here are not meant as requirements for all research and development on the areas of surveillance or multimedia. Especially people in highly specific disciplines will find them too broad for their own purposes. However, when re-using datasets and annotations that were originally created for different purposes, these requirements will help to identify where the potential pitfalls lie. The actual uptake and use of a metadata scheme involves more than can be accomplished with a set of requirements as described in one paper. Also, the support for a (new) metadata representation scheme does not only depend on the quality of the scheme itself, as we can see on the lack of support for SAM, which addresses almost all requirements. In the actual application of these requirements, whether for the design of a new scheme or for the slight adaptation

of an existing one, the hands on involvement of relevant communities is required. Such a process requires workshops with stakeholders, trial-and-error, demonstration versions and acceptation tests, and is stimulated with (tools that support) backwards compatibility. The authors hope to provide a starting point for such a process.

7. Summary

This paper describes the requirements for metadata representation schemes for scientific research and operational systems in the area of surveillance for security, metadata on: the sensor, features and observation, events, entities, their actions and attributes, situations and scenario, goals and intent, hypothetical scenarios, traceability, alternatives and chance, and additional observation actions. A terminology is proposed to define the metadata requirements. We identify and describe five common and emerging surveillance design patterns, and evaluate the fitness of the requirements on these surveillance patterns, both in breadth as in depth with a pickpocket case. We conclude that these are the right requirements given these common and emerging surveillance patterns, and these requirements are indeed sufficiently rich to describe also the more complex situations in the right amount of detail. Finally, we conclude that there is no fundamental problem with implementing these requirements, because for every requirement there is at least one existing metadata scheme which supports it. In addition, for the requirements which have relatively little support, we can give practical guidelines on how they could be implemented. There is however no metadata representation scheme that fulfils all the requirements. Especially the requirements of relations between entities, goals, intents, uncertainty and alternatives are rarely met. Of the examined schemes, MPEG7 and SAM fulfil most of the requirements but MPEG7 might be too expressive in its general form, and SAM has not found much support yet.

Acknowledgements

This work was performed as independent research of the applied research programme Dutch Top sector High Tech Systems & Materials: Roadmap Security, Passive Sensors. [9]. The authors thank Aart Beukers (Eye-D Security Experts) and the Amsterdam police for kindly providing the instruction video.

References

1. Alexander C (1977) A Pattern Language: Towns, Buildings, Construction
2. Annesley J, Colombo A, Orwell J, Velastin S (2007) A profile of MPEG-7 for visual surveillance, IEEE Int. Conf. AVSS, 482–487
3. Bouma H, Vogels J, Aarts O, Kruszynski C, Wijn R, Burghouts G (2013) Behavioral profiling in CCTV cameras by combining multiple subtle suspicious observations of different surveillance operators, Proc. SPIE 8745
4. Burghouts GJ, Marck J (2011) Reasoning about threats: from observables to situation assessment. IEEE Trans Syst Man Cybern 41(5):608–616
5. Buschmann F, Meunier R, Rohnert H, Sommerlad P (1996) Pattern-Oriented Software Architecture, Volume 1: A System of Patterns. John Wiley & Sons
6. CAVIAR: Context aware vision using image-based active recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/
7. CVML: Computer vision markup language. http://homepages.inf.ed.ac.uk/tlist/cvml/spec.html

8. Doermann D, Mihalcik D (2000) Tools and techniques for video performance evaluation. ICPR 4:167–170
9. Dutch top sector high tech systems & materials: Roadmap security, passive sensors. http://www.htsm.nl/Roadmaps/Security
10. ETISIO: Video understanding evaluation. http://www-sop.inria.fr/orion/ETISEO/
11. Fisher RB (2004) The PETS04 surveillance ground-truth data sets. Proc. 6th IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, pp 1–5
12. Francois A, Nevatia R, Hobbs J, Bolles R, Smith JR (2005) VERL: an ontology framework for representing and annotating video events. IEEE Multimedia 12(4):76–86
13. HUMAN ML: Human markup language. https://www.oasis-open.org/committees/download.php/60/HM.Primary-Base-Spec-1.0.html
14. I-LIDS: Imagery library for intelligent detection systems. Home Office, UK
15. INCOSE, a consensus of the INCOSE fellows. http://www.incose.org
16. Kester LJHM (2008) Designing networked adaptive interactive hybrid systems. IEEE Multisensor Fusion and Integration for Intelligent Systems, 2008, MFI 2008, pp 516–521
17. Kipp M (2013) Anvil: the video research annotation tool. http://www.anvil-software.org/ accessed January 4th 2013
18. Kipp M (2013) Anvil 4.0 Annotation of video and spoken language
19. La Vigne NG, Lowry SS, Markman JA, Dwyer AM (2011) Evaluating the Use of Public Surveillance Cameras for Crime Control and Prevention. Urban Institute, Justice Policy Centre.
20. Lenat DB, Guha RV (1990) Building large knowledge-based systems: representation and inference in the CYC project. Addison–Wesley, Reading
21. List T, Fisher RB (2004) CVML-an XML-based computer vision markup language. Int Conf Pattern Recog (ICPR) 1:789–792
22. Lyon D (2007) Surveillance studies: an overview. Polity Press, Cambridge
23. Mariano VY, Min J, Park J-H, Kasturi R, Mihalcik D, Li H et al (2002) Performance evaluation of object detection algorithms. ICPR 3:965–969
24. Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A (2003) Ontology library (final). IST Project 2001–33052 WonderWeb Deliverable D18
25. Milestone. http://www.milestonesys.com/
26. MPEG-7: Moving pictures expert group
27. Neely H (2010) Modeling Threat Behaviors in Surveillance Video Metadata for Detection using an Analogical Reasoner, IEEE Aerospace conference
28. Nghiem AT, Bremond F, Thonnat M, Valentin V (2007) ETISEO, performance evaluation for video surveillance systems. IEEE Conference On Advanced Video and Signal Based Surveillance, AVSS 2007, pp 476–481
29. Niles I, Pease A (2001) Towards a Standard Upper Ontology. In: Welty C, Smith B (eds) Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine, October 17–19, 2001
30. Noldus. www.noldus.com
31. Online resource pickpocket video
32. ONVIF: Open network video interface forum. http://www.onvif.org/Home.aspx
33. Over P, Awad G, Fiscus J, Antonishek B, Michel M, Smeaton AF, et al (2011) Proceedings of TRECVID 2010—An overview of the goals, tasks, data, evaluation mechanisms, and metrics, Gaithersburg, Md., USA
34. PETS: Performance evaluation of tracking and surveillance. http://pets2012.net
35. PSIA: Physical security interoperability alliance. http://www.psialliance.org/
36. SanMiguel JC, Martinez JM, Garcia A (2009) An ontology for event detection and its application in surveillance video, IEEE Int. Conf. AVSS, pp 220–225
37. Schallauer P, Bailer W, Hofmann A, Mörzinger R (2009) SAM: An interoperable metadata model for multimodal surveillance applications. Proc. SPIE, 7344
38. Sowa JF (1976) Conceptual graphs for a database interface. IBM J Res Dev 20(4):336–357
39. Sowa JF (1984) Conceptual graphs. Information Processing in Mind and Machine, 39–44
40. Steinberg AN, Bowman CL, White FE (1999) Revisions to the JDL data fusion model. Environmental Research Institute of Michigan Arlington VA
41. Surveillance of Unattended Baggage and the Identification and Tracking of the Owner (SUBITO) consortium (December 2011), SUBITO Deliverable D100.2: Final Report
42. Suzić R (2005) A generic model of tactical plan recognition for threat assessment. Proc. SPIE
43. TRECVID: TREC video retrieval evaluation. http://trecvid.nist.gov/

44. UK Home Office, Invitation to Tender Efficient Archive Retrieval & Auto Searching (EARS) CONTEST Project http://www.homeoffice.gov.uk/publications/science-research-statistics/research-statistics/home-office-science/eoi-ears-hos, Accessed June 2012
45. VIPER. http://viper-toolkit.sourceforge.net/
46. Westermann U, Jain R (2007) Toward a common event model for multimedia applications. IEEE Multimedia 14(1):19–29

Author biographies

Jeroen van Rest is a lead consultant at TNO (Networked Organisations group), the Netherlands. He studied computer science at the University Leiden (MSc degree 2002) with specialisation in Multimedia. His research interests cover broad scope of multimedia, cognitive systems and privacy for security and surveillance.

Franc Grootjen is an associate professor Artificial Intelligence at the Radboud University (Artificial Intelligence department), the Netherlands. He studied Mathematics, Physics and Computing Science at the Radboud University Nijmegen, the Netherlands (MSc degree 1992). In 2005 he received his PhD degree in Physics, Mathematics and Computer Science at the Radboud University Nijmegen, the Netherlands. His main interest is the broad application of mathematical and cognitive techniques in a linguistic environment such as Information Retrieval and Knowledge Representation.

Marc Grootjen is a researcher at TNO (Perceptual and Cognitive Systems group), the Netherlands. He is the founder of EagleScience, a human factors research oriented company. As a navy officer, he studied Mechanical Engineering at the Delft University of Technology (MSc degree 2002), the Netherlands. His main interests are: user-centered design, video annotation and process control rooms.

Remco Wijn is research scientist at TNO (Human Behaviour & Organisational Innovations group), the Netherlands. He studied social psychology at the University of Groningen after which he obtained his PhD from Utrecht University on the antecedents and consequences of being treated fairly or not. His research topics include (but not limited to) behavioural processes that lead to suspicious behaviours, radicalization, and terrorism. The central aim of these projects is determining how behaviours linked to these subjects occur and evolve, and recognizing them in an early stadium.

Olav Aarts is a scientist sociology and statistics at TNO (Human Behaviour & Organisational Innovations group), the Netherlands. He studied sociology and statistics at the Radboud University Nijmegen (MSc degree 2004) where he was involved in the Interuniversity Centre for Social Science and Methodology. In 2010 he received his PhD at Nijmegen University, the Netherlands. Currently, he is conducting research in the field of social networks, interaction between culture and technology, and statistics.

Maaike Roelofs is a scientist in human behaviour at TNO (Human Behaviour & Organisational Innovations group), the Netherlands. She studied Leisure Management at INHOLLAND Diemen, Amsterdam (BBA degree 2009), and Management, Economics and Consumer studies at Wageningen University (MSc degree 2011). Her specialisations is in consumer behaviour and deviant behaviour. Currently, her interests concentrate round behavioural and societal sciences.

Gertjan J. Burghouts is a lead scientist in visual pattern recognition at TNO (Intelligent Imaging group) , the Netherlands. He studied artificial intelligence at the University of Twente (MSc degree 2002) with a specialization in pattern analysis and human-machine interaction. In 2007 he received his PhD from the University of Amsterdam on the topic of visual recognition of objects and their motion, in realistic scenes with varying conditions. His research interests cover recognition of events and behaviours in multimedia data. Currently, he is mainly occupied with DARPA project named CORTEX.

Henri Bouma is a research scientist and project manager at TNO in the field of computer vision and pattern recognition. He received his MSc in Electrical Engineering at the University of Twente and his PhD in Biomedical Image Analysis at the TU Eindhoven, the Netherlands. His publications include articles in multiple high-impact IEEE journals. Dr. Bouma participated in projects about ship recognition for the Dutch Ministry of Defence, in the Mind's Eye program of DARPA for human action and behaviour recognition, and he leaded several projects about the automatic tracking and behaviour analysis in surveillance video for the Dutch Police.

Lejla Alic is a scientist in pattern recognition at TNO (Intelligent Imaging group), the Netherlands. She studied Electrical Engineering at Delft University of Technology (MSc degree 2001) with a specialization in signal and image processing. In 2013 she received her PhD from the Erasmus MC in Rotterdam on the topic of pattern recognition in cancer treatment prediction and response using multi-modality images. Her research interests include many aspects of imaging (processing, analysis, pattern recognition) and data annotation in multi-modality data.

Wessel Kraaij is senior research scientist at TNO (Media and Network Services) and professor at Radboud University Nijmegen (Institute for Computing and Information Sciences / Intelligent Systems / Information Foraging Lab), the Netherlands. He studied electrical engineering at the Eindhoven university of Technology (MSc degree 1988). In 2004 he received his PhD from the university of Twente on the topic of language modelling for information retrieval. He is joint coordinator of the NIST TRECVID benchmark since 2003. His research interests include multimedia information retrieval and information aggregation.