Workshop on Multimedia Systems for Surveillance (MMSS) in conjunction with
8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2011

Smart Resource-aware Multimedia Sensor Network for Automatic Detection of Complex Events

Fadi Al Machot*, Carlo Tasso†, Bernhard Dieber*, Kyandoghere Kyamakya*,
Claudio Piciarelli†, Christian Micheloni†, Sabrina Londero†, Massimiliano Valotto†,
Paolo Omero†, Bernhard Rinner *

*Alpen-Adria-Universität Klagenfurt
firstname.lastname@uni-klu.ac.at

† Faculty of Sciences of the Univeristy of Udine
Department of Mathematics and Computer Science
firstname.lastname@uniud.it

Abstract

This paper presents a smart resource-aware multimedia sensor network. We illustrate a surveillance system which supports human operators, by automatically detecting the complex events and giving the possibility to recall the detected events and searching them in an intelligent search engine. Four subsystems have been implemented, the tracking and detection system, the network configuration system, the reasoning system and an advanced archiving system in an annotated multimedia database.

1. Introduction

The project Smart Resource-Aware Multi-Sensor Network (SRSnet) aims to develop a smart resource-aware multi-sensor network capable of autonomously detecting and localizing various events such as screams animal noise, tracks of persons and more complex human behaviours. The project's research areas include :

- Collaborative audio and video analysis.
- Complex event detection.
- Network reconfiguration.
- Intelligent web user interface for data warehouse.

After the detection of simple and complex events, there is a user interface and a multimedia data warehouse which stores detected simple and complex events as well as multimedia data such as images and short audio/video sequences. From the data gathered and processed within the sensor network, SRSnet filters events that are relevant to users and inserts them into the data warehouse via a web service interface. Via a convenient interface, users can query for specific events in the data warehouse.

The events we are considering are divided in two classes, simple events and complex events:

1. Simple Event: This is the simplest form of events e.g. run, walk, shot, etc.
2. Complex Event: a complex event which is the combination of the simple events e.g. groups of persons are running, group of persons are fighting, group of persons are running in different direction, etc.

So far, different papers about video surveillance systems based on audio and video features have been published. Many researchers use Bayesian networks for event recognition such as [10], [28], [2] or [12]. Others, use Support Vector Machines (SVM) [17] or hidden markov models [22] [11].

Research in context-aware computing is increasing, many researchers base their work on context computing and reasoning, see [27], [25], [29], [13], [26] or [23].

Arsic *et al.* [31] developed a system for multi-camera person tracking and left luggage detection. Their system supports human security staff to analyze the scenes algorithm based on homographic transformation and subsequent analysis of the observed object trajectories. The application of the system includes the detection of abandoned objects and loitering people.

Vu *et al.* [30] have developed an audio-video event recognition system for public transport security. Their ap-

978-1-4577-0845-9/11/$26.00

proach consists of six modules and includes: (a) object detection and tracking (b) face detection and tracking, (c) temporal multi-camera analysis (d) primitive audio event detection (e) primitive video event detection and (f) audio-video scenario recognition.

Another way to build a surveillance system for complex event detection is to use logic programming whereby several approaches have been illustrated in [24] [4] [20]. Hakeem and Shah [9] have presented a hierarchical event representation for analysing videos. The temporal relations between the sub-events of an event definition are represented using the interval algebra of Allen and Ferguson [1]. In the ASP based approach for event recognition, however, the availability of the full power of logic programming is one of the main features. It further allows activity definitions to include spatial and temporal constraints. In particular, some logical programming languages do not offer arithmetic operation built-ins and numeric constraints can affect decidability. The reason of the high performance of ASP on chip that the representation of the knowledge base and the solver size are not expensive. The Solver is 47 Kilo byte and is written in C, where most of the existed hardware platforms are able to execute it.

A well-known system for activity recognition is the Chronicle Recognition System (CRS). The language includes predicates for persistence and event absence [3]. The CRS language does however not allow mathematical operators in the constraints of the temporal variables. Consequently, CRS cannot be directly used for activity recognition in video surveillance applications. Shet et al. have presented a logic programming approach for activity recognition [19]. The temporal aspects of the definitions of Shet, Davis et al. are not well represented, there are no rules for computing the intervals in which a complex event takes a place.

In this paper We are solely dealing with the detection of complex events in short term, that is, within some seconds only (or up to maximum one minute).

This paper is organized as follows: Section 2 describes the general architecture of the proposed system. Section 3 and 4 illustrate the techniques exploited for network configuration and feature extraction. Section 5 illustrates the phase of complex event detection using Answer Set Programing (ASP). Section 6 presents the multimedia data warehouse and its architecture. Finally sections 7 and 8 describe the obtained results and provide an outlook to future work.

2. A Smart Resource-Aware Multimedia Sensor Network

In the SRSnet project we propose a visual sensor network to be used in areas with limited infrastructure, i.e., areas without access to fixed communication infrastructure or power supply. SRSnet is an audio/video sensor network, that performs surveillance tasks and detects complex events while utilizing resources in a smart way in order to operate on battery and renewable energy for longer periods of time. As a multimedia sensor network, it not only captures and transmits sensor information but it also performs onboard data processing. In our case, object detection, localization and tracking are performed on audio and video data. The system consists of components for i) audio and video processing, ii) complex event detection, iii) network reconfiguration and iv) storage and presentation of multimedia and event data. Complex events are detected by aggregating simple events from the audio/video processing. Information on detected events as well as PTZ configuration are the input for the network reconfiguration subsystem which adapts the sensor configuration and node's power states according to the network's current tasks. The detected complex and simple events along with supporting multimedia data are stored in a multimedia data warehouse which also provides an advanced interface. As a use case we will deploy the SRSnet in a national park where park rangers will be the targeted users defining areas of interest, events etc.

Figure 1 shows the data flow from the sensors to the multimedia data warehouse.

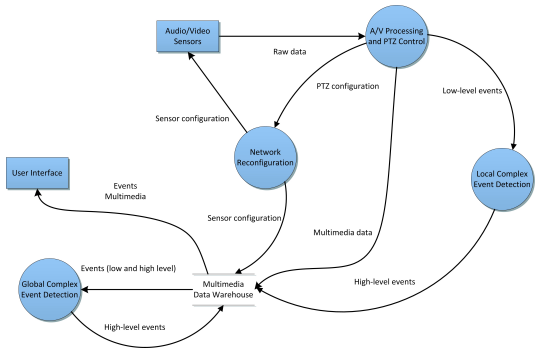

Figure 1. The data-flow from sensors through the SRSnet. Note, that the multimedia data warehouse is not deployed within the network but is a cloud service that is remotely accessed on demand.

3. Network Reconfiguration

The network reconfiguration aims at providing high surveillance quality while minimizing energy consumption to prolong the network lifetime. It also aims at selectively switching off sensors and nodes to save energy. The task is therefore a multidimenstional optimization problem. We must perform sensor selection to find a minimal set of sensors in order to perform a certain task but we also need a task assignment to find resource-minimal allocation of tasks to nodes. We model the resource and energy usage of surveillance algorithms running on our sensor nodes, then

we use multiobjective evolutionary algorithms to find assignments of tasks to cameras by taking into account tradeoffs between surveillance quality and resource usage.

4. Object tracking and detection

In order to give a semantic interpretation of the monitored scene, the system adopts a two-steps approach: first, low-level features are extracted from raw video and audio data, then the features are used as basic building blocks for the high-level reasoning system (section 5).

Moving objects are detected in video sequences by extracting zones of changes in the acquired images. Since the cameras can move, either by human operators or by being controlled by the system itself in order to optimize the visual coverage, standard background-based change detection techniques cannot be applied. We thus used a technique specifically studied for moving cameras and based on frame-by-frame comparison [33]. Detected objects are then classified using an adaptive high-order neural tree [34] applied to the normalized central moments of the binary detection image. Finally, the position of each object is projected on a map of the monitored environment and tracked by means of a combination of Kalman and Camshift filters. The map projections are computed using homographic mappings with geometric compensation for possible camera motions.

Audio sensors are used to localize audio sources. For audio localization, microphone arrays are used. The first step is to compute the Time Difference of Arrival (TDOA) based on the measurement of the time difference between the signals received by different microphones. We use the Multichannel Cross-Correlation Coefficient (MCCC) method [32] to calculate the TDOA, since it allows to take advantage of the redundant information provided by multiple sensors. Moreover, to improve the resolution of the peaks for the TDOA estimations and minimize the influence of noise and interferences, we apply a Phase Transform filter before calculating MCCC. From TDOA, it is possible to compute the direction of arrival of a sound for each microphone array. By using at least two separate arrays is thus possible to identify the position of a sound source by means of triangulation. If more than one source is identified, a beamformer and a spectral distance comparison provide a guide to solve the problem of associating the directions of arrival of the arrays.

5. The reasoning system

Complex event detection in audio-video sensor networks is based on three main steps. The first step is the extraction of features using object recognition and object tracking algorithms, then the detection of simple events, like walking, running or shouting, Finally, complex events are detected by combining the simple events.

Beside these main steps, we have to define the context model which describes all the information that may influence the way a scene is perceived. The state of an environment is defined as a conjunction of predicates. The environment must be modeled to retrieve the position, orientation and types of objects, as well as position, information and state of other objects from information observed in the environment.

After building the context model, we need a context interpreter which provides the context reasoning services including inferring contexts, resolving context conflicts and maintaining the consistency of context knowledge base. Different inference rules can be specified and input into the reasoning engines [27].

For the description of regions of interest in the image or to detect the coordinates of moving objects close to the important regions, geometric correction is required.

Additionally, in multi-sensor networks (e.g audio and video), the extraction of features from video-audio streams is the basis for data fusion and is needed to combine data in order to estimate or predict entity states. Data fusion techniques combine data from multiple sensors to achieve more specific inferences than what could be achieved by using a single sensor.

Answer set programming (ASP) is a declarative programming paradigm based on the answer set semantics of logic programming [8]. Logic programming can be extended to represent new options for problems in the head of the rules and to provide this by meaning of ordered disjunctions. In complex event detection systems the combination between spatial and temporal relationships among sensor nodes is needed, where this combination helps to detect different scenarios in a logic sequence of events. Moreover, ASP provides this relationships among sensor nodes to detect complex events.

For an example, in the case of the complex event ”persons are running in a forbidden area”, we have to consider the temporal and spatial contexts. This event happens when *i)*, persons are entering the forbidden area and *ii)*, when they are running inside it. Using the temporal specification and the constraints of ASP support to prevent non logical sequences of the simple events to detect the desired scenario. Firstly, the event starts when some one is inside the forbidden area and secondly, it ends when the persons is moving in the forbidden area for a specific period of time.

Using ASP, we need a knowledge base which supports the detection of complex events. In the knowledge we have to define the specification of the direction, the specification of the zones and the specification of simple events.

As an example, we illustrate a simple event (a dog is running), which happens exactly if an object is moving with a speed of more than 5 kilometers per hour.

```
run(X,S,T,Z,D,C1,C2,FI,CI,ST,SAI,SZ,OT):-
S>5,
object(X),
hasSpeed(X,S),
hasTime(X,T),
hasZone(X,Z),
hasX(X,C1),
hasY(X,C2),
hasDate(X,D),
hasFrameId(X,FI),
hasCameraId(X,CI),
hasSoundType(X,ST),
hasSoundArrayID(X,SAI),
hasSoundZone(X,SZ),
hasObjectType(X,OT),
OT=dog.
```

To detect a complex event such as a running group of persons, we need to identify at least two persons. If the distance between these two persons is less than 3 meters and both are running, then the whole group is running. The condition on the distance is specified in the predicate $near(X_1, X_2)$, where X_1 and X_2 present the observed persons:

```
near(X1,X2):-
X1!=X2,
dist(X1,X2,D),
D<3,
hasObjectType(X1,OT1),
hasObjectType(X2,OT2),
OT1=person,
OT2=person,
hasTime(X1,T1),
hasTime(X2,T2),
T1=T2.
```

The condition of running is specified in the predicate $run(X_1, OT1)$ and $run(X_1, OT2)$, where X_1 is the observed object. The two variables $OT1 and OT2$ refer to the type of the detected object. The last condition makes sure that the two observed objects are different.

```
groupPersonsRunning(X1):-
run(X1,OT1),run(X2,OT2),
near(X1,X2),
OT1=human,
OT2=human,
X1!= X2.
```

Uncertainty can not be avoided in practical visual surveillance applications. We consider now one class of uncertainty, the one called detection uncertainty. Detection uncertainty is a part of our knowledge base. We consider two types, the first one is the uncertainty of the localization and the second one is the uncertainty of the object type classification. We are getting these uncertainty values from the low level feature extraction. In the actual phase of our project we do not have to consider the logic uncertainty since our rules in the KB are always true. We use a real-value weight to represent the confidence of each rule in the KB.

6. The design of the multimedia database

The MultiMedia Data Warehouse (MM-DW) is devoted to archive the video and audio files received from sensors. Its main goal is to provide to several classes of users a set of access features designed in order to support them in various tasks. The design of the MM-DW has started with a specific analysis performed for identifying and specifying the most suitable requirements for the user interface to the archive. We have thus analyzed the state of the art of the development tools that can be used for implementing such data base. The results of this analysis are the following: i) all the examined tools feature standard general-purpose archive & retrieval functions, not specifically specialized for the multimedia case; ii)only a few of the analyzed tools offer specific extension modules designed to handle multimedia information. In most cases, these are devoted to generic recognition and processing functions, that can be exploited in order to detect specific audio/video patterns; iii) no more specific access and retrieval functions are generally provided, adequately fitting the SRS-net scenario. Considered such situation, we have decided to adopt a specific two-layer solution for the SRS-net MM-DW, constituted by (i) a basic archiving layer and (ii) by an advanced access & retrieval & knowledge-discovery layer. Such last layer directly supports the users in various access modalities, ranging from simple retrieval to more complex data mining and knowledge discovery operations. This organization takes into account the general architecture of the smart resource-aware multimedia sensor network adopted in SRS-net, which includes a specific processing pipeline, starting from the raw files acquired by the sensors, followed by two processing phases devoted respectively (i) to identify specific low-level domain-specific features (typically specific öbjects̈) and (ii) to recognize, starting from the detected features, high level (simple or complex) domain-specific events. The output of these two processing steps is then filtered: only the parts of raw audio/video files where events have been discovered are sent to the MM-DW, which can later be accessed by the final users. Within the general approach illustrated above, the specific data fields recorded in the MM-DW are the following (with the exception of the multimedia data, all the other fields are structured): fragments of the audio/video files acquired by sensors, each one associated to a successful event detection; clusters of sensors, each one including localization information useful to infer where an event has taken place; sensor information; objects, i.e. entities (person, an-

imal, dog, shot, car, scream, etc.) detected by sensors, recognized in the first processing phase, such field including also all the information useful to recognize the case where objects detected by different sensors are actually the same object; event information, referring to both simple and complex situations (eg. running, fighting, screaming, etc.), associated with the low-level features that allowed the recognition; feature information, including data such as where, when, from which sensor was detected, and so on; properties, referring to specific settings of the sensors when a feature has been detected; zone,e.g. the subarea of the park where the detection has taken place. In the current prototype the archiving layer has been implemented by means of the open source DBMS MySQL. Java programs have been used for implementing the advanced access & retrieval & knowledge-discovery layer, to be accessed through a Web interface. The following classes of users have been considered: Network Configurator, System Administrator, Security Operator, Analyst, Generic Visitor. The first version of the prototype has been developed with reference to a specific kind of users, namely the analyst, which uses the MM-DW in order to perform generic retrieval operations over the collected data and advanced statistical analysis (by means of data mining, machine learning and knowledge discovery tools) to get knowledge from the archive.

The basic access metaphor utilized for querying the archive is a what/where/when three dimensional space, whereas the search results are visualized and can be navigated following an event/place/network three dimensional approach. A specific Analytics Module (AM) has been planned for advanced analysis of the data: it will operate on the structured parts of the archive, exploiting high level reasoning and processing functions whose expected result is the discovery of knowledge patterns of specific interest for the user e.g.(new complex events). After the ongoing implementation activity, a systematic experimental evaluation has been planned.

7. Results

Implementation is ongoing but some results have already been achieved. We measured the execution time of the ASP solver on an embedded platform (pITX-SP 1.6 by Kontron[1]). We see that Answer Set Programming is far more suited for embedded operation than other existing paradigms. Due to the fast execution (0.4 s in average) of the iClingo[2] ASP solver, the complex event detection needs no considerable amount of resources. In our project, we can thus run the complex event detection once or twice a second which enables the audio/video subsystem to collect sufficient data for detecting complex events.

[1]http://www.kontron.com

[2]http://potassco.sourceforge.net

The evaluation of our system is done for different scenes. The test environment is a park place, equipped with several cameras and audio arrays. The events we defined are divided in two groups: simple and complex events. After conducting experiments on a benchmark dataset, we realized that, whenever the accuracy of the detection is high, then our detection ratio is over 94% for all complex events, see table 1.

The complex event	Specificity	Sensitivity
A group of persons are running	100%	98.4%
A group of persons are fighting	89%	94.4%
A group of persons are running in different directions	92.1%	96.2%

Table 1. The performance of the reasoning system

8. Conclusion

This paper has described an audio-video surveillance system able to automatically recognize high level human and animals behaviors using audio and video features. Different algorithms and technologies have been used and developed for the combined audio and video analysis, complex event detection, network reconfiguration and the advanced web user interface to the data warehouse. In our future work, we plan a systematic implementation and we will focus on complex behaviors understanding based on supervised an unsupervised learning algorithms to detect unusual behavior from learning the abnormality of video and audio features in a dynamic environment.

Acknowledgments

This work is supported by Lakeside Labs GmbH, Klagenfurt, Austria and funded by the European Regional Development Fund (ERDF) and the Carinthian Economic Promotion Fund (KWF) under grant KWF 20214/18354/27107.v

References

[1] J. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531, 1994. 2

[2] D. Arsic, F. Wallhoff, B. Schuller, and G. Rigoll. Video based online behavior detection using probabilistic multi stream fusion. *IEEE International Conference on Image Processing*, pages 606–609, 2005. 1

[3] A. Artikis and G. Paliouras. Behaviour recognition using the event calculus. *Artificial Intelligence Applications and Innovations III*, pages 469–478, 2009. 2

[4] A. Artikis, M. Sergot, and G. Paliouras. A logic programming approach to activity recognition. In *Proceedings of the 2nd ACM international workshop on Events in multimedia*, pages 3–8. ACM, 2010. 2

[5] G. Brewka. Logic programming with ordered Function. In *Proceedings of the National Conference on Artificial Intelligence*, pages 100–105. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002.

[6] J. Crowley and Y. Demazeau. Principles and techniques for sensor data fusion. *Signal processing*, 32(1-2):5–27, 1993.

[7] T. Eiter and A. Polleres. Towards automated integration of guess and check programs in answer set programming: a meta-interpreter and applications. *Theory and Practice of Logic Programming*, 6(1-2):23–60, 2006.

[8] M. Gebser, R. Kaminiski, B. Kaufmann, M. Ostrowsky, T. Schaub, and S. Thiele. Using gringo, clingo and iclingo. September 2008. 3

[9] A. Hakeem and M. Shah. Learning, detection and representation of multi-agent events in videos. *Artificial Intelligence*, 171(8-9):586–605, 2007. 2

[10] R. Howarth. Interpreting a dynamic and uncertain world: task-based control. *Artificial Intelligence*, pages 5–85, 1998. 1

[11] Y.-P. Huang, C.-L. Chiou, and F. Sandnes. An intelligent strategy for the automatic detection of highlights in tennis video recordings. *Expert Systems with Applications*, 36:9907–9918, 2009. 1

[12] X. Jiang, D. Neill, and G. Cooper. A bayesian network model for spatial event surveillance. *International Journal of Approximate Reasoning*, 2009. 1

[13] K.-E. Ko and K.-B. Sim. Development of context aware system based on bayesian network driven context reasoning method and ontology context modeling. *International Conference on Control, Automation and Systems*, pages 2309–2313, October 2008. 1

[14] M. Maroti, G. Simon, A. Ledecz, and J. Sztipanovits. Shooter Localization in Urbain Terrain. *Computer*, 37(8):60–61, August 2004.

[15] C. Matheus, K. Baclawski, M. Kokar, and J. Letkowski. Using swrl and owl to capture domain knowledge for a situation awareness application applied to a supply logistics scenario. In A. Adi, S. Stoutenburg, and S. Tabet, editors, *Rules and Rule Markup Languages for the Semantic Web*, volume 3791 of *Lecture Notes in Computer Science*, pages 130–144. Springer Berlin / Heidelberg, 2005.

[16] N. Pham, W. Huang, and S. Ong. Probability hypothesis density approach for multi-camera multi-object tracking. In *Proceedings of the 8th Asian conference on Computer vision-Volume Part I*, pages 875–884. Springer-Verlag, 2007.

[17] C. Piciarelli, C. Micheloni, and G. Foresti. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11), November 2008. 1

[18] D. Sadlier and N. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Transaction on Circuits and Systems for Video Technology*, 15(10):1225–1233, October 2005.

[19] V. Shet, D. Harwood, and L. Davis. Vidmap: video monitoring of activity with prolog. In *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, pages 224–229. IEEE, 2005. 2

[20] V. Shet, D. Harwood, and L. Davis. Multivalued default logic for identity maintenance in visual surveillance. *Computer Vision–ECCV 2006*, pages 119–132, 2006. 2

[21] L. Snidaro, M. Belluz, and G. Foresti. Representing and recognizing complex events in surveillance applications. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 493 –498, 2007.

[22] J. Snoek, J. Hoey, L. Stewart, R. Zemel, and A. Mihailidis. Automated detection of unusual events on stairs. *Image and Vision Computing*, 27(1-2):153–166, 2009. 1

[23] T. Strang. A context modeling survey. *In Workshop on Advanced Context Modelling, Reasoning and Management associated with the Sixth International Conference on Ubiquitous Computing*, 2004. 1

[24] S. Tran and L. Davis. Event modeling and recognition using markov logic networks. *Computer Vision–ECCV 2008*, pages 610–623, 2008. 2

[25] B. Truong, Y.-K. Lee, and S.-Y. Lee. Modeling and reasoning about uncertainty in context-aware systems. *Proceedings of the 2005 IEEE International Conference on e-Business Engineering*, 2005. 1

[26] G. Wang, J. Jiang, and M. Shi. A context model for collaborative environment. *Proceedings of the 10th International Converence on Computer Supported Cooperative Work in Design*, 2006. 1

[27] X. Wang, D. Zhang, T. Gu, and H. Pung. Ontology based context modeling and reasoning using owl. *In Workshop Proceedings of the 2nd IEEE Conference on Pervasive Computing and Communications*, pages 18–22, March 2004. 1, 3

[28] S. Wasserkrug, A. Gal, O. Etzion, and Y. Turchin. Complex event processing over uncertain data. *Complex Event Processing Over Uncertain Data*, pages 253–264, 2008. 1

[29] X. Ying and X. Fu-yuan. Research on context modeling based on ontology. *International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce*, 2006. 1

[30] V.-T. Vu and F. Bremond and G. Davini and M. Thonnat and Quoc-Cuong Pham and N. Allezard and P. Sayd and J.-L. Rouas and S. Ambellouis and A. Flancquart. Audio-video event recognition system for public transport security. *IET Seminar Digests, pp 414-419*, 2006. 1

[31] Arsic D. and Hofmann M. and Schuller B. and Rigoll G. Multi-Camera Person Tracking and Left Luggage Detection Applying Homographic Transformation. *PETS 2007, Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*,october 2007. 1

[32] J. Chen, J. Benesty and Y. Huang Robust time delay estimation exploiting redundancy among multiple microphonesy. *IEEE Transactions on Speech and Audio Processing*, pages 549–557, vol 11, November 2003. 3

[33] G.L. Foresti, C. Micheloni and C. Piciarelli Detecting Moving People in Video Streams. *Pattern Recognition Letters*, pages 2232–2243, vol 26, 2005. 3

[34] C. Micheloni, C. Piciarelli and G.L. Foresti How a Visual Surveillance System Hypothesizes How You Behave. *Behavior Research Methods*, pages 447–455, vol 38, 2006. 3