On Building Decentralized Wide-Area Surveillance Networks based on ONVIF

Tobias Senst, Michael Pätzold, Rubén Heras Evangelio, Volker Eiselein, Ivo Keller and Thomas Sikora
Communication Systems Group, Technische Universität Berlin
Einsteinufer 17, 10587 Berlin, Germany
senst,paetzold,heras,eiselein,keller,sikora@nue.tu-berlin.de

Abstract

In this paper we present a decentralized surveillance network composed of IP video cameras, analysis devices and a central node which collects information and displays it in a 3D model of the complete area. The exchange of information between all components in the surveillance network takes place according to the ONVIF specification, therefore ensuring interoperability between products complying with the specification and flexibility regarding the integration of new devices and services. The collected information is displayed in a 3D model of the surveilled area, therefore providing a comfortable overview of the activity in large environments and offering the user an intuitive way to eventually interact with network devices.

1. Introduction

The interest on automated video surveillance systems has notably increased in the last decades as safety and security have become critical issues in many public areas, and the number of video surveillance cameras is rapidly growing. This has led to a large amount of wide area camera networks adopting different architectural choices, algorithms for automatically analyzing the huge amount of video information generated by these devices, and communication, storage and display solutions aiming to assist human operators in the task of monitoring activity in large environments [11, 7].

While the literature concerning particular video surveillance issues and specific systems is large [10, 3, 2], there is a lack of information on systems based on a global standard. In this paper we report a use case of the design of a wide area surveillance system which complies with the ONVIF specification. The ONVIF specification [5] defines a common protocol for the exchange of information between IP-based video devices. ONVIF was founded by Axis Communications, Bosch Security Systems and Sony Corporation and currently accounts with 295 member companies, including 17 full members, 22 contributing members and 256 user members. The specification includes automatic device discovery, video streaming and intelligence metadata. ONVIF makes use of web services and provides a formal conformance process, therefore assuring interoperability between products regardless of their brand. We believe that the compliance with a global standard will be a key of success of any surveillance product in the near future.

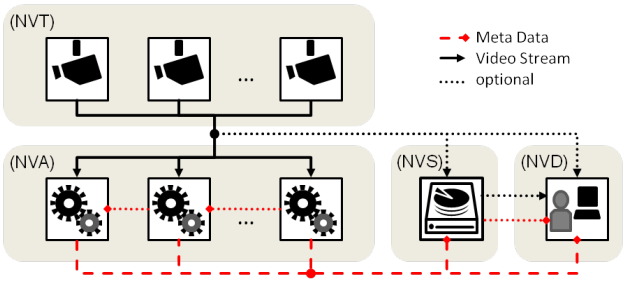

Figure 1. Architecture of an ONVIF based network.

The system we present in this paper is composed of IP video cameras whose captured video sequences are analyzed in order to send information of the activity observed at their respective locations to a central node. This information is collected at the central node and made available to the user interface. As user interface we use a 3D model of the complete area, therefore providing a comfortable overview of the activity in large environments and offering the user an intuitive way to eventually interact with network devices, thus improving the efficiency. Figure 1 depicts the proposed system.

The rest of this paper is organized as follows: In Section 2 we present the algorithms that we use in order to obtain information out of the video sequences. In Section 3 we show how this information can be sent according to the ONVIF specification. The user interface is described in Section 4. Section 5 concludes our paper.

2. Video Analysis

The base of the proposed system are the video streams and the metadata generated by the analysis devices, which set the requirements for the communication. We have currently integrated four algorithms aiming to extract information of interest in a general public space (like a railway station or an airport), namely, lost baggage, people tracks, people carrying objects, and statistics of crowd motion.

The video analysis devices provide the extracted information as objects with an associated 3D position, a timestamp and analysis specific attributes. We assume that camera position and calibration are known by the analysis devices. The 3D position of the considered objects is then computed by assuming that their base is on the ground plane. The timestamps are extracted from the video streams, which are synchronized to a global time server.

The lost baggage detector uses a dual-background model and a finite-state machine as presented in [4]. The specific information generated by this algorithm is the bounding box of the detected object.

The person tracking module detects the shape of upper body regions by means of a Histogram of Oriented Gradients algorithm. Furthermore, these detections are supported by a uniform motion analysis aiming at the exclusion of ambiguous regions [6]. The resulting detections are tracked by applying a Kalman-filter.

Based on people tracks and the corresponding video stream, people carrying objects are detected by analyzing the foreground mask of the tracked persons as proposed in [1]. Alternatively, tracked persons can be analyzed by means of an optical flow motion model as recently proposed in [9], which is independent from foreground segmentation results although it imposes a higher computational complexity. This method processes the metadata generated by the people tracking algorithm and mark the tracks corresponding to people carrying objects.

The crowd motion is measured by generating long-term motion trajectories based on a robust local optical flow [8]. We divide the ground floor into equally sized tiles and compute statistics of the motion observed. Each tile represents an object in the metadata and contains the mean direction of the motion.

3. Communication

The Open Network Video Interface Forum (ONVIF) has developed an open standard aiming to achieve the interoperability between networked video surveillance devices and components. The ONVIF specification defines a network layer of IP security devices described by web services based on the Organisation for the Advancement of Structured Information Standards (OASIS). The definition of a client is provided by the Web Services Description Language (WSDL). A key advantage of this approach is to assure the client integration regardless of its brand. Data structure and exchange between web service requester and provider is based on the Simple Object Access Protocol (SOAP). SOAP is a message exchange protocol which can be implemented over HTTP, HTTPS, RTSP...

```
<wstop:TopicNamespace name="NUE" targetNamespace="...www.nue.de/...">
    <wstop:Topic name="LostBaggage"/>
    <wstop:Topic name="PeopleTracking"/>
    <wstop:Topic name="MotionStatistic"/>
    <wstop:Topic name="DetectionPeopleCarryingBaggage"/>
</wstop:TopicNamespace>

<xs:complexType name="WsntNotificationFrameExtension"
  xmlns:tt="http://www.onvif.org/ver10/schema"
  xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:wsnt="http://docs.oasis-open.org/wsn/b-2">
  <xs:complexContent>
    <xs:extension base="tt:FrameExtension">
      <xs:sequence>
      <xs:element ref="wsnt:Topic" use="required"/>
      <xs:element name="Source" type="tt:ItemList" use="required"/>
      ...
      </xs:sequence>
    </xs:extension>
  </xs:complexContent>
</xs:complexType>
```

Figure 2. **Required Extension of the WS-Topics definition:** *Frame* class was derived to provide analysis specific topics and camera specific data.

Every device integrated in an ONVIF compliant network has to implement a device service (Figure 1 depicts the architecture of an ONVIF based network). Device services can be categorised into four classes:

- Network Video Transmitter (NVT) provide one or more video streams, such as a camera or a box combining the video streams from several video cameras.
- Network Video Analytics (NVA) are devices used to analyze video, audio or metadata and provide additional information not delivered by the input stream.
- Network Video Display (NVD) provide the representation of media stream and the interface between system and human operators.
- Network Video Storage (NVS) provide the meaning for recording streamed media and metadata as well as the capability of accessing this data in a structured manner.

According to the ONVIF specification the analysis functionalities offered by the proposed video surveillance system can be considered as NVAs. In order to exchange the data between NVA, NVD and NVS, ONVIF specifies a *MetaDataStream* class[1]. ONVIF bears with two different kinds of analysis results, events and a more comprehensive scene description. Therefore events and analytics schemata are defined. The proposed system is based on analytic schemes. In order to receive an analytic metadata stream the requester has to subscribe to it on the NVA device

[1] http://www.onvif.org/onvif/ver10/schema/onvif.xsd

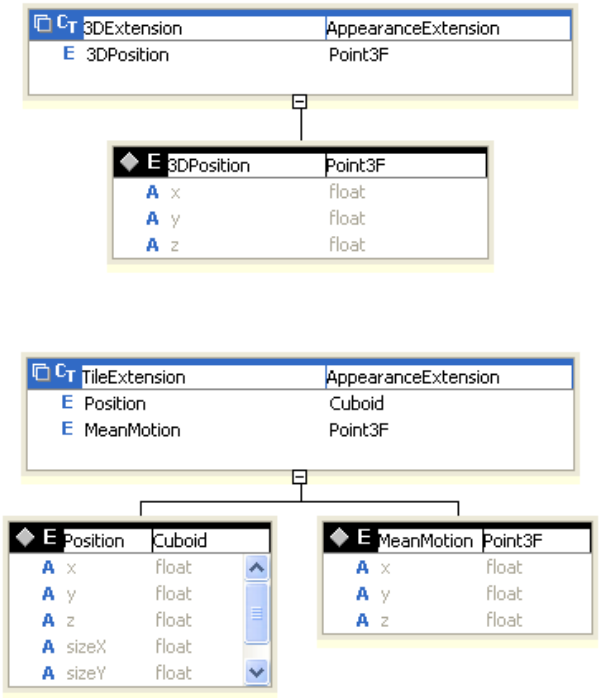

Figure 3. **Required extensions of the ONVIF specification:** *Appearance* class was provided with additonial elements in order to describe system specific data.

generating that metadata, which is sent in form of *Frame* elements. Each *Frame* contains the information extracted for a single video frame, but not which kind of information the NVA generates, e.g. do the 3D position sent by a NVA correspond to a static object or a tracked person.

In order to incorporate this information we extend the *Frame* base class based on the OASIS WS-BaseNotification and WS-Topic specifications. Therefore we can sort the received frames attending to their topic. The extension of the frame class is shown in fig. 2. Beside WS-Topics, we also provide information about the video sources by defining the *Source* element. With this extension we can associate each *Frame* to an exact network camera and time, which is included in the *UtcTime* element of the base class. In order to incorporate this information we extend the *Frame* base class based on the OASIS WS-BaseNotification and WS-Topic specifications. Therefore we can sort the received frames with respect to their topic. The extension of the frame class is shown in Figure 2. Beside WS-Topics, we also provide information about the video sources by defining the *Source* element. With this extension we can associate each *Frame* to an exact network camera and time, which is included in the *UtcTime* element of the base class.

To transmit the extracted information of a NVA each *Frame* owns a vector of elements of the type *Object*. The *Object* type is specified by ONVIF and consists of elements of type *Appearance* and *Behaviour*. The *Appearance* type includes a set of default descriptors. For example the bounding box generated by the lost baggage method is described by the *ShapeDecriptor* type. We extend these descriptors to provide the capability to transmit 3D data by specifying a *3DExtension* type based on the *Appearance* class. Additionally, we also define a 3D tile type *TileExtension*, which is derived from the *Appearance* base class, in order to describe a region on the ground plane. We need the 3D tile type to represent statistics of crowd motion. Figure 3 shows a scheme of these extensions.

4. User Interface

Nowadays, a large amount of static surveillance cameras is installed inside and outside of buildings which are under security control. The aforementioned analysis-methods are able to extract a variety of metadata from each camera stream, which can only be reviewed efficiently by a single operator, if it is put into an ergonomic interface. In the case of existence of calibration data for the cameras of a building, the metadata can not only be related to a camera identifier, but in addition, metadata can be displayed as 3D objects in the correct position within the 3D building model, as depicted for the person-tracking-, lost-baggage- and crowd-motion- analysis in figures 4 to 6, respectively. A 3D abstract representation provides the operator a fast and comfortable overview. The position of cameras is plainly visible and relations between camera views are easily comprehensible for the operator, who is able to freely navigate like a humming bird in all directions through the scene.

Furthermore, detailed information (live images, camera identifiers, event timestamps, statistics of events...) can be overlaid on interaction with the 3D model. For example, by clicking on one of the person models in Figure 4 further information of the selected person like position, speed, appearance or a full-length trajectory can be shown.

Additionally, a collection of all important events is presented in one chronologically ordered event list. Figure 5 depicts this list using the example of detected static objects. This list contains an overview over all events and the appropriate information. By clicking on events in the list the point of view of the 3D-model is redirected so as to optimally show the area where the event is located. Thereby, the operator obtains a fast orientation within the scene.

Internally, individual models of buildings and objects are arranged hierarchically in a tree structure. This enables the integration of geographically dispersed areas. This easy extension of the interface to additional areas and their flexible placement provides a comfortable monitoring for the operator and is indispensable for an ergonomic wide area surveillance system.

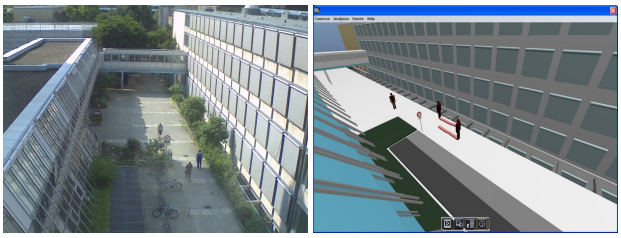

Figure 4. **Model-based representation of the person-tracking-analysis:** The video streams are analyzed and resulting people tracks are represented as objects in a 3D-model. By interacting with the objects further information is overlaid.

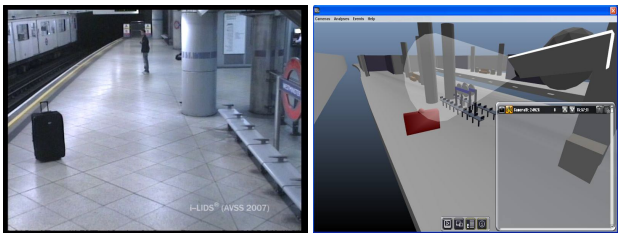

Figure 5. **Model-based representation of the lost-baggage-analysis:** The video stream is analyzed and new static regions are detected and represented in the 3D-model and in a system event list which contains the detection results for all cameras in the network.

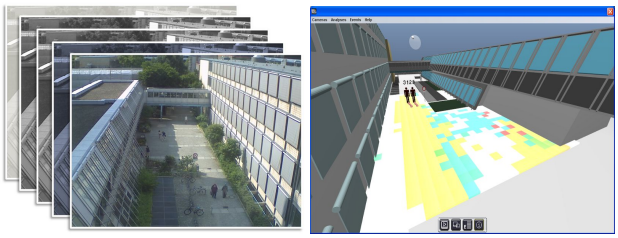

Figure 6. **Model-based representation of the crowd-motion-analysis:** The video stream is analyzed and, based on the information provided by long-term trajectories of persons, a statistic of the motion observed at equally sized image regions is computed, therefore obtaining information of the predominant motion directions in crowded environments. This information can be displayed as coloured tiles in the ground plane.

5. Conclusions

In this paper we presented a decentralized video surveillance system complying with the ONVIF specification. The system is composed of ONVIF conform Network Video Transmitters (NVT), Analytics (NVA), Displays (NVD) and Storage (NVS), which are integrated as web services. Video streams are provided by the NVTs and analysed by several NVAs. The overall extracted information is then collected by the NVD and displayed in a 3D model of the complete area. To transmit the analysed data in a 3D spatio temporal space we extend the data exchange specified by ONVIF and define the corresponding topics. Designing complying with standards ensures interoperability between products and flexibility regarding the integration of new devices and services, thus allowing to reduce costs and facilitate maintainability of the systems.

The representation of the generated metadata in a 3D model provides a comfortable overview of the surveillance area and brings the possibility of integrating interaction of the user with network devices. Furthermore, a model representation can also be considered regarding privacy protection and ethic concerns.

References

[1] C. B. Abdelkader and L. Davis. Detection of people carrying objects: A motion-based recognition approach. In *Automatic Face and Gesture Recognition*, pages 378–383, 2002. 2

[2] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, P. B. Osamu Hasegawa, and L. Wixson. A system for video surveillance and monitoring. Technical report, VSAM Final Report, May 2000. 1

[3] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000. 1

[4] R. Heras Evangelio, T. Senst, and T. Sikora. Detection of static objects for the task of video surveillance. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 534–540, January 2011. 2

[5] Open Network Video Interface Forum. Onvif core specification ver 2.0, November 2010. 1

[6] M. Pätzold, R. Heras Evangelio, and T. Sikora. Counting people in crowded environments by fusion of shape and motion information. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, (PETS Workshop 2010)*, pages 157–164, Boston, USA, 2010. 2

[7] T. D. Räty. Survey on contemporary remote surveillance systems for public safety. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(5):493 –515, sept. 2010. 1

[8] T. Senst, V. Eiselein, R. Heras Evangelio, and T. Sikora. Robust modified L2 local optical flow estimation and feature tracking. In *IEEE Workshop on Motion and Video Computing (WMVC 11)*, pages 685–690, 2011. 2

[9] T. Senst, R. Heras Evangelio, and T. Sikora. Detecting people carrying objects based on an optical flow motion model. In *IEEE Workshop on Applications of Computer Vision (WACV 11)*, pages 301–306, 2011. 2

[10] Y.-l. Tian, L. Brown, A. Hampapur, M. Lu, A. Senior, and C.-f. Shu. Ibm smart surveillance system (s3): event based video surveillance system with an open and extensible framework. *Mach. Vision Appl.*, 19:315–327, September 2008. 1

[11] M. Valera and S. Velastin. Intelligent distributed surveillance systems: a review. *Vision, Image and Signal Processing, IEE Proceedings* -, 152(2):192 – 204, april 2005. 1