WPSS: Watching people security services

Henri Bouma [1*], Jan Baan [1], Sander Borsboom [2], Kasper van Zon[3], Xinghan Luo [3], Ben Loke [4], Bram Stoeller [5], Hans van Kuilenburg [3], Judith Dijk [1]

[1] TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands.
[2] Cameramanager.com, Hogehilweg 19, 1101 CB Amsterdam, The Netherlands.
[3] VicarVision, Singel 160, 1015 AH Amsterdam, The Netherlands.
[4] Noldus Information Technology, Nieuwe Kanaal 5, 6709 PA Wageningen, The Netherlands.
[5] Eagle Vision, Energiestraat 16B, 1411 AT Naarden, The Netherlands.

ABSTRACT

To improve security, the number of surveillance cameras is rapidly increasing. However, the number of human operators remains limited and only a selection of the video streams are observed. Intelligent software services can help to find people quickly, evaluate their behavior and show the most relevant and deviant patterns. We present a software platform that contributes to the retrieval and observation of humans and to the analysis of their behavior. The platform consists of mono- and stereo-camera tracking, re-identification, behavioral feature computation, track analysis, behavior interpretation and visualization. This system is demonstrated in a busy shopping mall with multiple cameras and different lighting conditions.

Keywords: Surveillance, CCTV, security, tracking, behavior analysis, threat, biometry.

1. INTRODUCTION

To improve security, the number of surveillance cameras is rapidly increasing. However, the number of human operators remains limited and only a selection of the video streams are observed. Intelligent software services can help to evaluate people's behavior and show the most relevant and deviant patterns. In this paper, we present a software platform that contributes to the observation and analysis of human behavior. This platform was developed by a consortium of multiple companies in the WPSS-project. The platform consists of several components. It uses real-time tracking in multiple mono-cameras for overview, tracking in stereo-cameras for high location accuracy, and person re-identification to couple information between cameras. To enrich the person description, we used face analysis, body-part detection and 3D pose estimation. Finally, information is combined, tracks are analyzed, behavior is interpreted by a reasoning engine and directly visualized. The security services fuse all information and the results are made available in an online platform. This system has been tested and demonstrated in a crowded shopping mall with multiple cameras and different lighting conditions.

The outline of the paper is as follows. The technical system is presented in Section 2. The demonstrator setup is described in Section 3 and the results are shown in Section 4. Finally, the conclusions are presented in Section 5.

2. METHOD

The WPSS system consists of monocular and stereo cameras, video management, tracking and re-identification, measurement of behavioral features, track analysis and behavior analysis and graphical user interface to support interaction (Figure 1). Each is explained in more detail in the following subsections.

* henri.bouma@tno.nl; phone +31 888 66 4054; http://www.tno.nl

Henri Bouma , Jan Baan, Sander Borsboom, Kasper van Zon, Xinghan Luo, Ben Loke, Bram Stoeller, Hans van Kuilenburg, Judith Dijk, "WPSS: Watching people security services", Proc. SPIE, Vol. 8901, (2013). http://dx.doi.org/10.1117/12.2031639

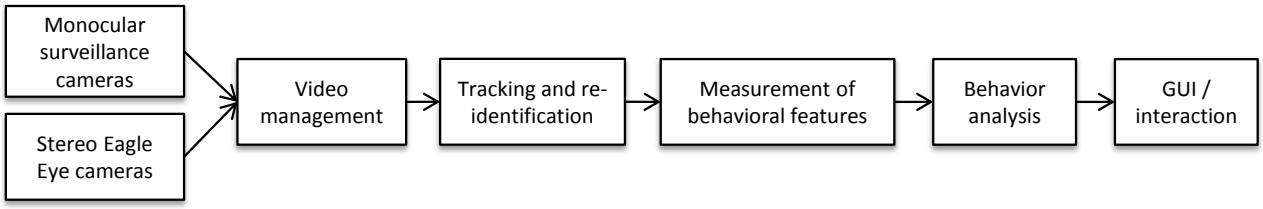


Figure 1: The WPSS system consists of monocular and stereo cameras, video management, tracking and re-identification, enrichment, behavior analysis and GUI / interaction.

2.1 Flexible distribution of video streams

The base of any video analytics platform is getting the video from their sources and distributing it to the different analytics software clients. A Video Management System (VMS) allows the video streams to be centrally managed, recorded and used by multiple component without bandwidth impact for the camera. However, several specific requirements had to be taken into account for clients using video analytics processing.

First, VMS-es in the past used Motion Jpeg because of its simplicity for the cameras, but since more and more cameras support highly compressed H.264, the market is moving to H.264. Most video analytics algorithms are heavily CPU bound and have IO resources to spare, so for the WPSS project the simplicity to decode, implement and skip frames of Motion JPEG was chosen.

Furthermore, the different components need to be able to uniquely identify video frames to be able to communicate about them. Normally, a VMS does have this information and it focusses on smooth playback, not on the communication of perfect timestamps to clients. This can be seen in many of the approaches to transport video, as Motion JPEG over HTTP does not have a way to include this information and most Real Time streaming Protocol (RTSP) libraries don't support it even though the RTSP standard includes it. In the end, Motion JPEG over HTTP was chosen as a transport, using an extra custom header to include the timestamp information. This was done due to its ease to implement client-side, and compatibility with firewalls and proxies.

To ensure stability and ease of maintenance, the metadata platform was divided in multiple services, each handling a single part of the data flow such as recorders, rule engines, multi-protocol metadata receivers and services to relay data using e-mail, sms and mobile push notifications. The multi-protocol receivers where used to communicate with the different analytics components and needed to support a variety of transports including udp packets, tcp sockets and HTTP POST requests. Each of these approaches was implemented as separate plug-ins to allow easy updating when the protocols used by the analytics components changed.

Communication of metadata between these services and the client in a production environment has a lot of requirements: it has to be fast, use low bandwidth, be easy to debug, but also able to get through a variety of firewalls and proxies. The Apache Thrift framework, first developed for inter-server communication at Facebook, was chosen for this task since it allows a plethora of languages, including Java, C++ and C#, to communicate data or write it to disk.

Communication of multiple metadata packets can be done using a single TCP connection while with a REST based approach, each packet would have to be sent using a separate connection, lowering the number of packets and thus bandwidth dramatically. Furthermore, the framework allows multiple types of well-defined data serialization, from compact binary representations to human readable JSON, allowing high efficiency while keeping debug capabilities and transport over text-only firewalls and proxies.

2.2 Tracking people with stereo cameras

Detecting people with a camera is a challenging computer task. Firstly it can be hard to distinguish foreground (the person to track) from background (the environment). And secondly, by the effect of occlusion, a camera does not always have a clear view of a person. Occlusion occurs when the line of sight from the camera to a person is blocked, either by an object or by another person. To overcome these issues, a system is developed to detect and track people using ceiling-mounted stereo cameras, with built-in processing units, called *Eagle Eyes*.

A configuration involving multiple Eagle Eyes is called an *Eagle Grid*. The Eagle Grid is designed to combine the information of multiple Eagle Eyes in a single map of the environment. To cover an area of 100 square meters, around 4 to 12 Eagle Eyes are needed, depending on the height of the ceiling (Figure 2).

Figure 2: Scanning a public area in shopping mall “Kanaleneiland” using Eagle Eyes.

A stereo camera typically has two lenses, each with its own image sensor. The lenses have the same direction and are shifted a dozen of centimeters in the plane perpendicular to the viewing axis. The camera records two images at once, a main image obtained by the first sensor and an auxiliary image obtained by the second sensor. Using these two images, the Eagle Eye can compute the real-world distance to each pixel in the images. Knowing this distance makes it possible to create a three-dimensional representation (i.e. a point cloud) of the scene (Figure 3).

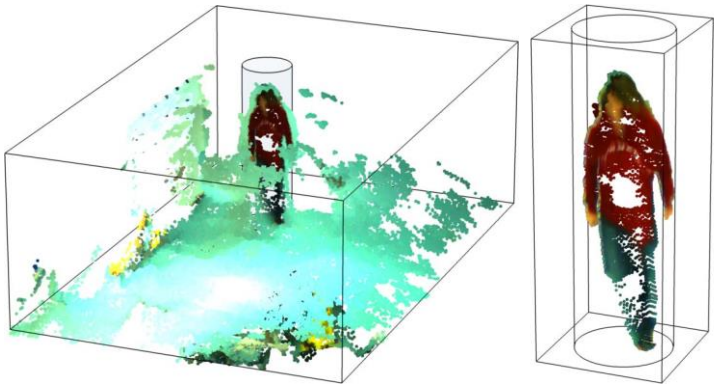


Figure 3: The cylinder containing a person is extracted from the point cloud of the entire scene.

The Eagle Eye extracts people from the point clouds and reports their position and appearance at a rate of 10 Hz to the Eagle Grid. The Eagle Grid combines the measurements of all Eagle Eyes to form trajectories (see Figure 4). A trajectory is a sequence of detected positions of a single person. Each person gets a unique numerical id which remains fixed while the person is seen by different Eagle Eyes. The Eagle Grid sends information about all currently observed people to the vision systems of our partners Camera Manager, Noldus, Vicar Vision and TNO. For each person a message describing its position is sent 10 times per second and an appearance message once per second.

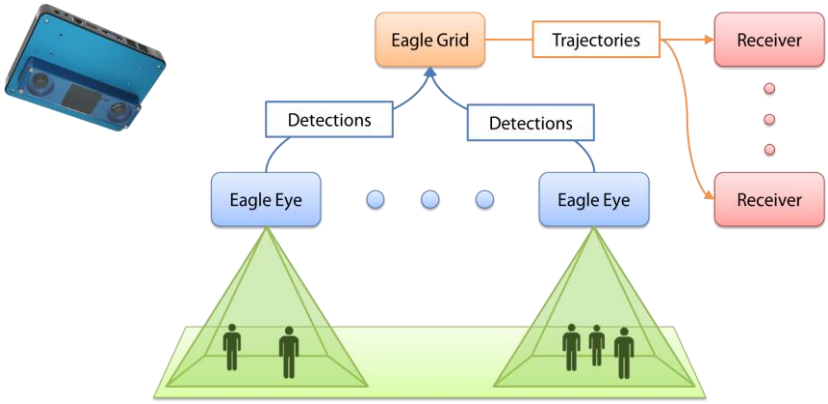


Figure 4: Schematic representation of the Eagle Eyes, the Eagle Grid and the receiving parties.

The appearance message is a description of the appearance of a person. This appearance model is based on a colored point cloud and the detected position of a person in this cloud (Figure 3). From the point cloud, a person's height is calculated as well as a color based appearance model. This model is a representation of the color of the person's clothes, face and hair. The most common method to create a color based appearance model is to construct a histogram of the distribution of the colors of the points representing a person. For the WPSS project a RGB (red, green and blue) histogram is created at different heights. This means that a person's point cloud is split up into several disks at different heights (6 disks in this case). For each disk a $n \times n \times n$ ($10 \times 10 \times 10$) RGB histogram is created. This information is sent to our partners' vision systems, who can use it to recognize the person in their own captured frames.

Using the RGB space for these histograms is not trivial. For instance, RGB suffers greatly from lighting influences. Over the years many color spaces have been proposed. Some of them are very straightforward and have been used by many (like RGB and Normalized RGB), others are complex, unintuitive and have only been used in research settings. Gevers describes 14 color spaces [15].

Although general RGB histograms are used for the WPSS project, we will explain the radial hue-saturation histograms briefly as a suggestion for future improvement. All colors in RGB space can be converted to the more intuitive HSV space (hue, saturation and value), where the tone (hue) of a color is defined by an angle on a circle, the saturation is defined as the radius in that circle and the brightness/darkness, called the value, is defined on the axis perpendicular to the hue-saturation space (see Figure 5). By ignoring the value (which is mostly influenced by light sources and shadows) a circular hue-saturation space is created, which is insensitive to changes in the intensity of the light source.

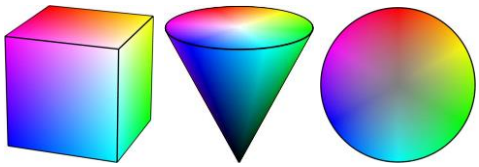


Figure 5: Illustration of the RGB, HSV and hue-saturation color space.

This hue-saturation space is well known in computer vision applications. Plotting a person's point cloud in the hue-saturation space results in Gaussian-like scatter patterns. Figure 6 shows a scatterplot of the isolated point cloud of Figure 3. The color of the jeans, shirt, face and hair can easily be distinguished as four normally distributed clusters. To capture this distribution in a histogram, one would like to partition the space in equally sized bins. The classic way is to partition the hue-saturation space in n even rings and m equally sized wedges, resulting in $n \times m$ bins, where m is around $3n$, so $3n^2$ bins in total. These bins are relatively small around the center and large around the edge. An alternative method divides the circle in $3n^2$ equally sized bins. To achieve this, each i^{th} ring is divided into $6i$-3 bins [25]. For large n the outer bins will approach a square-like form.

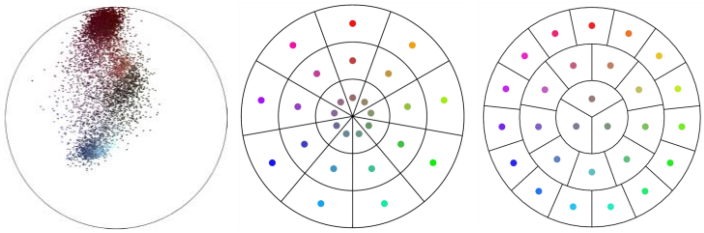


Figure 6: A scatter plot of a person's point cloud in hue-saturation space and two possible binning patterns.

2.3 Tracking people with mono cameras

The TNO architecture for tracking and re-identification is shown in Figure 7. The main components are tracklet generation and the re-identification engine and a graphical man-machine interface. Tracklet generation is an activity that continuously processes the incoming video streams to detect persons and track them within a single camera. The resulting tracklets are shared with the other partners in the WPSS project and stored in a tracklet database. This database allows our system to quickly retrieve similar candidates after human interaction without computational intensive video processing. In order to track a person in a large environment over multiple non-overlapping cameras, the separate tracklets of a certain person from different cameras need to be combined. The re-identification engine compares the query with tracklets in the database and presents the most likely candidates. This engine consists of two components:

appearance-based matching [7] and space-time localization. The combination of both is used to present the best matching candidate [8]. The human-machine interface enables the operator to interact with the system, by selecting queries and candidates. Coupling of tracks between cameras can be done automatically or semi-automatically, with user interaction. The generated tracks are send to VicarVision, Noldus and Cameramanager.

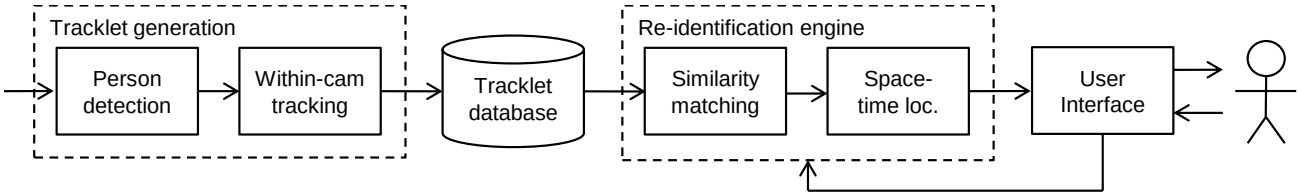


Figure 7: The system consists of tracklet generation, a re-identification engine and a graphical user interface.

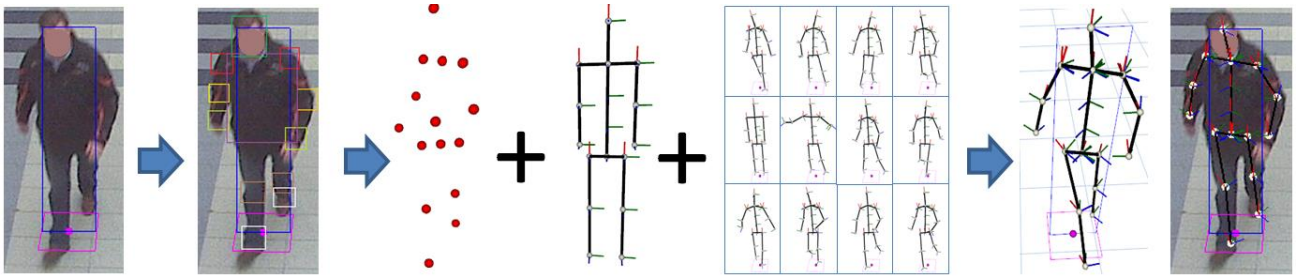
2.4 Measurement of behavioral features

To facilitate the person re-identification process, a number of different video processing components are employed to enrich the person descriptions provided by the person-tracking module. The first component uses a telephoto lens and a custom version of the FaceReader software to extract information from a person's face. The second module employs a body-part detector which detects the heads, torsos and legs of all the persons in a camera's view. These body-parts are thereafter used to automatically estimate properties as a person's height and clothing color. In addition, the detected heads are used as an input for tracking human attention by estimating the gaze direction of each person in the camera's view.

FaceReader is the world's first tool that is capable of automatically analyzing facial expressions. For the automatic evaluation of human behavior this tool provides valuable information. Facial expressions like anger, fear or disgust can potentially be used as an indicator for interesting events or threatening situations. In addition to the facial expressions, the face analysis module estimates a list of characteristics of the faces which are analyzed. These characteristics include a person's gender, age, ethnicity, the presence of facial hair (beard and moustache) and glasses. The key components of the face analysis module consist of a face detector [28] and an Active Appearance Model (AAM) [19] of the face which is classified by numerous neural networks. The model classifications are averaged over multiple video frames by tracking each person with a Kalman Filter and a HSV histogram of a person's clothing color. The face analysis module operates on the same area of view as the person tracker. To enable matching between face data and person tracking data. The bounding boxes provided by the person tracker are re-projected onto the view of face analysis camera.

The body part detection module is based on a cascade of classifiers using GentleBoost [14]. The features used in this module are Histogram of Orientated Gradients (HOG) descriptors [13] and Region Covariance descriptors [27]. Using the integral image approach of [28] these features can be calculated rapidly on different scales and different locations in the image. Detections are only performed on small regions of the image, in the neighborhood of the bounding boxes of the detected persons provided by the person tracking module. This does not only benefits the detection speed, it also reduces the number false body part detections in the image.

The body part detection estimates the 2D coordinates of the body parts based on image evidences which are results of perspective projection of the 3D human subject on the 2D camera image plane. Due to the absence of depth information, proper 3D interpretation of a given 2D body pose is a non-trivial task and is a highly ambiguous problem. Additional knowledge and constraints are required to restrict the size of the solution space. Although the latest research like [23] and [22] on 3D pose from a single view demonstrated promising results, these complex approaches are still far from robust real-time real-life application, due to the condition assumptions, the subject dependency and the slow processing speed because of the enormous computing. Based on the work of [2] and inspired by the state-of-the-art, we developed a simplified yet more practical and flexible solution for 3D pose estimation from a typical single surveillance view. See the pipeline in Figure 8.



(a) (b) (c) (d) (e) (f) (g)

Figure 8: Pose estimation pipeline: (a) Locate the person in the 2D image; (b) Locate the person's body parts; (c) 3D goal positions mapped from 2D pose to drive the skeleton; (d) The 3D kinematic skeleton; (e) Example training poses from the UMPM benchmark [2] for the pose solution space; (f) Estimated pose in 3D; (g) 3D pose superimposed on 2D image.

Unlike [23] in which the human body is modeled simply by the 3D coordinates of the joints, we use the more refined kinematically constrained human skeleton model [2] (Figure 8(d)), driven by 3D goal positions (Figure 8(c)) via inverse kinematics (IK). Any human pose can be defined as a set of 30 constrained joint angles regardless the size variations between subjects. The CCD-based IK iterations [29] drive the skeleton as a whole, the angular and length constraints restrict any abnormal position of the body parts caused by erroneous goal positions. In addition, by modeling common pedestrians' repetitive pose sequence like walking, a pose solution space is setup to further constrain the pose configuration and guarantee the resulting poses are always realistic. The motion capture dataset from the UMPM benchmark trained the pose solution space (Figure 8(e)). The angle sets of the training poses are converted to ranked principal components by PCA methods to build the solution space.

Having the 2D body part detection as the input, similar to [23] the pose ambiguous pose hypothesis set is setup by stochastic exploration of all possible 3D poses mapped from this 2D pose. The 3D pose candidates are used as goal positions to drive the skeleton by IK. Each of the resulting angle set which defines a skeleton pose, is converted to the PCA solution space. Normally a direct match in the solution space is not expected, therefore the Euclidean distance of each candidate to all solution poses is measured to find the closest. If the distance is below a threshold, the closest pose solution is selected to represent the candidate. Otherwise the candidate is regarded as abnormal pose and discarded. Filtering out the false positives we get a subset of the solution poses represents the candidate poses. Finally, within the subset the pose with minimum sum of squared distance to its candidate and the previous time step pose is determined as the estimated 3D pose. The 3D skeleton is scaled by the estimated height of the subject, and the current 3D pose estimation is limited to head and torso due to the challenging limb detection, see the example results in Figure 18(c)(d).

The torso and the leg regions provided by the body part detector are used to provide a semantic description of the clothing color. The purpose of this module is to provide a human understandable description of the main colors of a person's clothes. A pre-defined binary mask is used to discard the pixels in the region which are most likely not part of the body. The remaining pixels are then converted to the HSV color space and put into a histogram. For a single person these histograms are aggregated over a track of multiple frames. At the end of the track the histogram is normalized and classified by a decision tree. By using the detected head and leg regions of a person a course estimate can be made of the height of a person. If the height of one or multiple reference objects in the scene are known, the height of a person can be estimated from the bounding boxes of the detected body parts [12]. Attention of people in a scene can be used as an indicator of interesting areas and events. Any single object or person receiving attention from a large number of people is likely to be worthy of further investigation. Inspired by the work Benfold and Reid [6], our system tracks human attention by making a rough estimate of the gaze direction for each person in the scene.

Our gaze estimation module operates on head images which are provided by the body part detector. These images are first scaled down to a fixed size of 20 by 25 pixels and then divided into a grid of overlapping rectangles. From each rectangle the HOGs and the covariance descriptors are computed and the results are put in one large feature vector. This vector is then fed into a multi-class support vector machine to classify the extracted features into one of the eight classes. As illustrated in Figure 9, each of these classes consists of a range of 45 degrees in yaw angle. The resulting gaze estimates are projected on a 2D map of the scene and the results can be employed to create heat map of human attention.

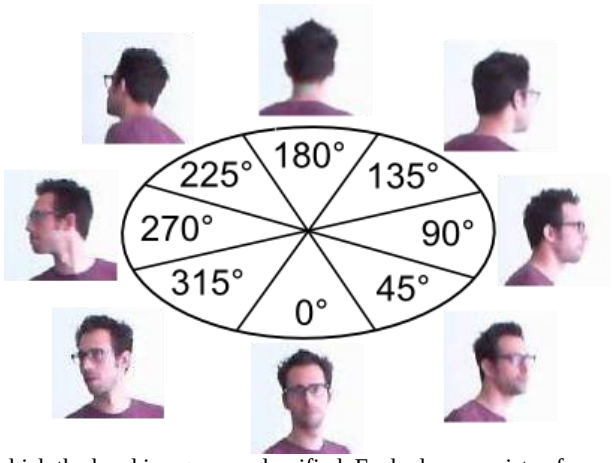


Figure 9: The eight classes in which the head images are classified. Each class consists of a range of 45 degrees in yaw angle.

2.5 Track analysis and behavior analysis

TrackLab [26] is a new tool of Noldus for measurement, recognition and analysis of spatial behavior. TrackLab is designed as a very open and flexible system, so that the software works with a wide variety of tracking technologies. It supports a wide variety of indoor tracking solutions, including TNO people tracking system in combination with CCTV or mono cameras [8], Ubisense™ ultra-wideband sensors and tags [30], EagleEye™ stereo cameras [17], and Noldus' video-based PeopleTracker™ [1]. But it can also be used with GPS systems for outdoor tracking, or in a retail environment where satellite reception is possible, for instance a shopping mall with a glass roof. For live import of real-time data, TrackLab uses the Noldus Communication Framework (NCF), which is a thin wrapper around RabbitMQ, an implementation of the Advanced Message Queuing Protocol (AMQP) [21]. As such NCF enables the setup of a flexible and scalable distributed system, i.e. a location based acquisition system just has to interface with the NCF to stream the position data into TrackLab. Multiple, often complementary, location based tracking systems can stream their data into TrackLab at the same time. For offline import of tracks, TrackLab support GPX, CSV and JSON formats.

Once the location data is in the TrackLab software it can be visualized in a variety of ways, the track data can also be edited to remove outliers, and a statistical analysis report is generated. The analysis variables are based on established parameters for quantification of behavior based on location. The analysis helps you to gain insight into the spatial and temporal behavior of customers. For real-time applications of the system, the analysis variables can be used to control external software, for example presentation of information on a display when a person has followed a particular path through the shop. TrackLab calculates a large number of statistics related to the location and movement of the subjects. The user can define zones (such as area around shopping mall entrance, the ATM area, etc.) and the statistics are then calculated both for the entire track and for the individual zones. This enables calculations such as the dwell time in the region of the ATM, the average speed in walkways, which regions of the shopping mall the visitors visit infrequently, and how much time the visitors were standing still in the shopping mall. Furthermore there are a number of statistics which can be used to quantify searching behavior of the visitors, based on an analysis of the path shape of the visitors. It is also possible to calculate group statistics across all the tracks in a study, for instance the average time that all the visitors spent in the zone near to a new advertising hoarding.

Although a quantification of the tracking data will often be necessary, in order to gain good insight into shoppers' behavior a visual presentation of that data is often invaluable. TrackLab allows visual presentation of data in three different ways: the track plot, heat map, and a graph of speed. The Track plot gives an overlay of all the tracks in your experiment on top of a floor plan of for instance the shopping mall. The software enables you to import a digital floor plan (in a bitmap format) and calibrate it so that it is the correct scale and position in relation to the tracks. You can pan and zoom the view, draw regions and points of interest (which are used in the analysis), and play the tracks back in a variety of ways. The heat map shows the difference in sample density for each location using different colors. Increased sample density is caused either by a subject spending longer at a given location, or by more subjects visiting the location during the recording. It is thus an excellent way to visualize interest in a region. For instance, if a new product (or an existing product with new packaging) is placed in several locations in a store, the heat map will illustrate how much interest there was in that product. The speed of the subjects over time can also be visualized as a graph. This is useful

information for determining both the interest in consumers in particular products as well as quantifying bottlenecks such as entry/exit passages.

It is also possible to configure TrackLab so that it provides input to other software, dependent on the path that the visitors have taken in a shopping mall, so called spatial events, or other behaviors such as standing still or walking fast, so called movement events. These events are communicated over the NCF for use by other systems, for example an event processing system. The rule based event processing system, so called RT-MMC a software prototype from Noldus, aggregates events from various event generating systems, like the spatial and movement events from TrackLab, into complex behavior events. The event processing logic can be defined by composing conditions with logical operators into a set of rules in design mode of the event processing system. A graphical user interface has been created to provide the user an intuitive and easy method to define the rules. In execution mode the event processing system will show the actual state of the rules depending on the incoming events and generates the according complex events when a rule executes successfully. The event processing system interfaces with NCF for flexible event communication with other systems.

2.6 Visualization

In the WPSS project, the Cameramanager.com user interface visualizes two types of results from the analytics components: continuously created metadata and detected events. The metadata is the data about the people in the video stream, including their location in the images, but also detected specifics such as clothing colors, length, age and gender. This data is visualized in two steps: first, the location of persons in the images is shown, allowing the user to click the persons, after which the extra information about the person is shown on a second screen with a video loop of the person over the whole time he was seen in the image.

Examples of detected events are large groups forming or loitering people in sensitive areas. In high level VMS systems, this is taken care by separate incident management systems, but the Cameramanager.com interface is designed for small business and home users and thus these events have to be communicated simpler and in an easier to understand form when the user is using the live view or by sending notifications using email or mobile push notifications. The different approaches that were implemented and tested in this project are described in the results section.

3. DEMONSTRATOR SETUP

3.1 Demonstrator at reception desk

A demonstrator setup was created at the entrance and hallway of the Noldus office in Wageningen (Figure 10). Six stereo cameras from Eagle Vision are mounted in the ceiling to track people, over an area of 12 x 3 meter, entering and leaving the office or their movement behavior around the reception desk and internet corner. The Eagle Vision tracking software streams the location data with use of the communication framework (NCF) into the movement analysis software TrackLab for real-time visualization and real-time spatial- and movement-based event detection. These behavior events are streamed to the event processing engine to detect time based behavior (for instance dwell time at the internet counters) or to a dedicated visualization application displaying the number of people in the office since the start of the day.

3.2 Demonstrator in a shopping mall

In the demonstrator setup in a shopping mall, approximately 20 PCs, 1 server, 9 stereoscopic cameras and 15 monoscopic cameras were installed to test our security services (Figure 11). There are different types of surveillance cameras, including AXIS 211M network cameras with 1280x1024 resolution. All services were tested in this environment.

In an experiment to measure the operator efficiency for real-time tracking and re-identification, only six of the cameras were used to create more blind spots. The PCs for this experiment are Dell Optiplex 9010 small form, with Intel Core i7-3770, 3.4 GHz, and 8GB DDR3 memory. The objective of this tracking and re-identification experiment is to obtain the increased efficiency by using the TNO-system for finding people and a comparison was made between people searches with and without the system.

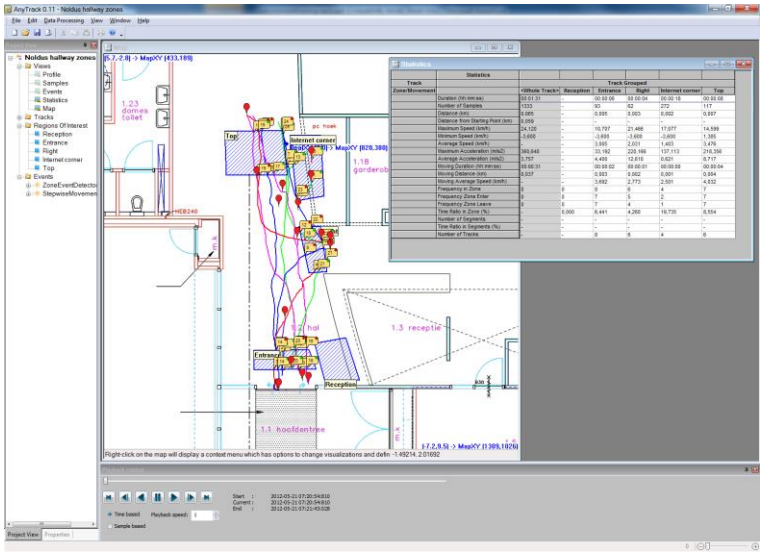


Figure 10: Track analysis and spatial event detection with TrackLab based on location tracking of EagleEye stereo cameras at the office reception desk.

The objective for behavior analysis is to detect abnormal or suspicious behavior at the ATM in the shopping mall. This suspicious behavior was defined as: if a person enters the ATM area more than two times within 10 minutes. This was implemented by defining a zone around the ATM in TrackLab for the spatial event detection by TrackLab and the same person and time condition as a set of rules in the RT-MMC. Another objective is to detect group behavior at a certain spot in the shopping mall: if more than 2 persons are located in the defined zone at the same time. This was implemented by defining a zone at a particular place on the floor plan in TrackLab for the spatial event detection by TrackLab and a set of rules to count the number of persons in the RT-MMC.

Figure 11: Monoscopic camera setup in the shopping mall.

4. RESULTS

4.1 Flexible distribution of video and metadata streams

The servers used for VMS duties were specified to support thousands of cameras and thus overbuilt for the test environments with only 20-30 cameras. For this reason and to test the newly built metadata support, all metadata services were also run on the video servers. The metadata support proved itself to be highly efficient by handling all metadata generated by the video analytics components with only low server resource usage. The maximum number of metadata events received in the largest test environment was 200 events per second, which the server handled without problems. To further test the performance of the metadata services, they were fed with simulated data, showing a maximum of 6000 events per second on a commodity level server. Video was also handled without problems by the same server, transporting it to 20 clients without impact to the cameras.

4.2 Tracking people with stereo cameras

The Eagle Eyes used in the WPSS project (Figure 12) were directed straight down. One advantage of this situation is that occlusion rarely occurs from this bird's-eye view. Wide angle lenses are used to increase the field of view. A full-length detection of a person is possible in an area of approximately $1.0h \times 1.5h$ where h is the height of the Eagle Eye. At the shopping mall h is 3.20 or 4.3 meters at two different locations.

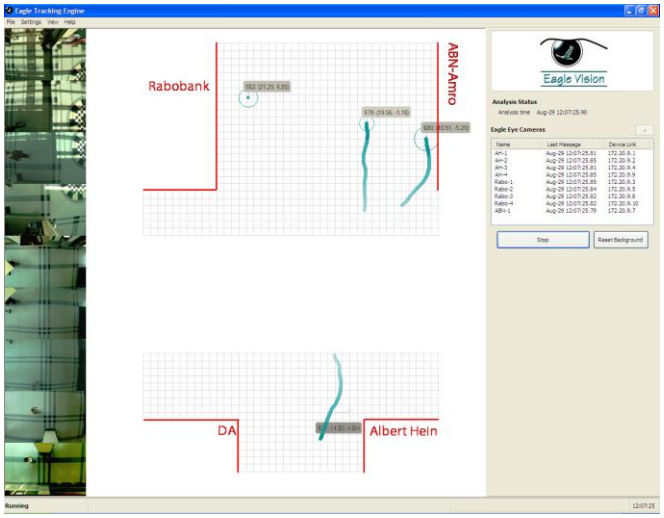


Figure 12: Screenshot of the Eagle Grid, combining the information of all Eagle Eyes.

In the WPSS project people are tracked using the Eagle Eyes at two of the entrances of the shopping mall. The Eagle Grid creates an appearance-based representation of the tracked person in order to recognize this person in one of the other video systems. The setup showed that people can effectively be tracked within the covered region, meaning that each person received a single id upon entering the view and kept that same id during the entire stay in the covered area. This also includes situations where people moved from one Eagle Eye to the other. The Eagle Grid proved to be a very precise people detection and tracking mechanism with a high accuracy in estimating people's position (±1 centimeter).

Furthermore, the tracks can efficiently be distributed to multiple other computer vision systems, using different protocols at the same time; TCP, UDP and HTTP (POST) at 10 to 20 Hz. In the WPSS project, the Eagle Grid was integrated with various systems provided by Camera Manager, Noldus, Vicar Vision and TNO.

The RGB histograms did not capture enough information about a person's appearance to recognize a person when he or she reappeared in the Eagle Grid. An explanation is that many people wear similarly colored clothes, or at least clothes that look the same from the sensor's birds-eye view. Furthermore, direct sunlight or shadows greatly affect the perception of colors in the RGB space. The snapshots in the middle of Figure 2 show the impact of these issues.

4.3 Tracking people with mono cameras

Tracklet generation consists of pedestrian detection and within-camera tracking. An example of generated tracklets in a camera view is shown in Figure 13. The user is able to interact with the system to show more or less candidates with the tracking and re-identification GUI (see Figure 14), which can be used for a forensic search or live tracking. By selecting the correct candidates, a complete track of an individual can quickly be generated.

Figure 13: The system generates tracks in all cameras.

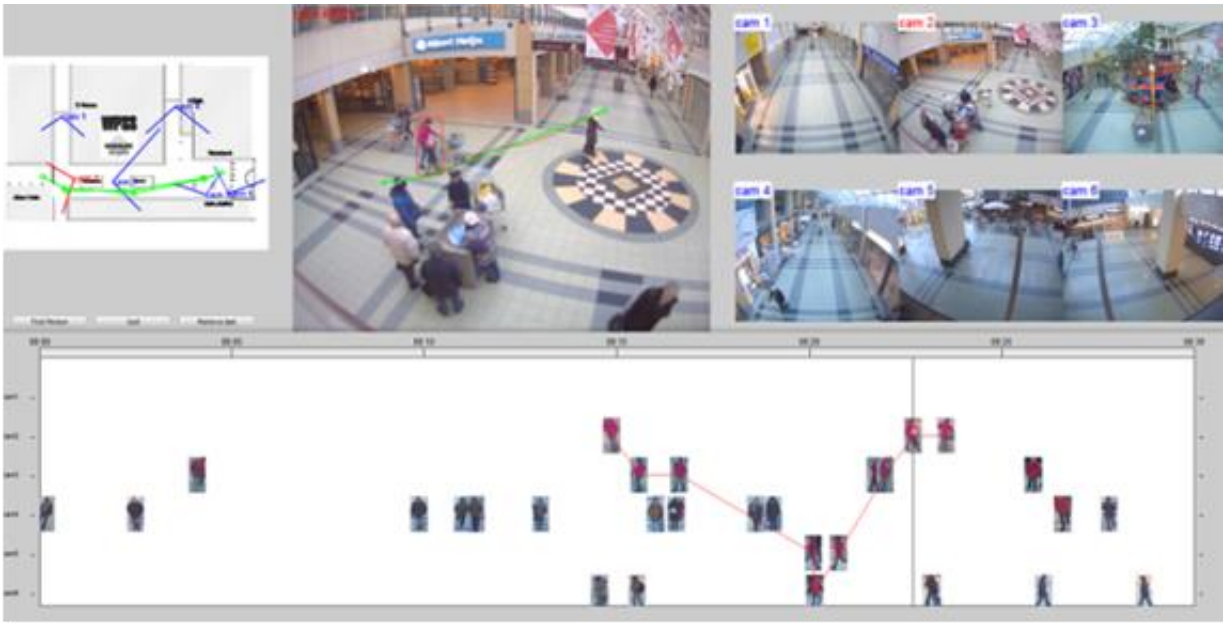


Figure 14: The graphical user interface of the tracking and re-identification system.

The system is benchmarked on international datasets [16] and in a realistic crowded environment (the demonstration environment in the shopping mall) with multiple surveillance cameras. The results showed that our search engine allows five times faster retrieval in a database than a manual search [7] (Figure 15). Furthermore, an operator can track a person more efficiently, with 37% less misses, which is a significant improvement [8].

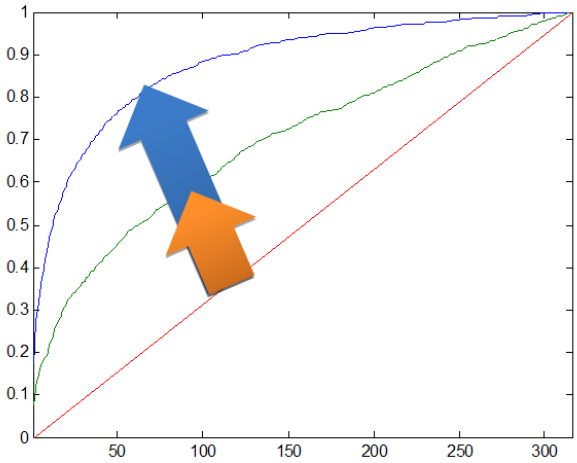


Figure 15: The TNO system (blue) is much better than a color histogram (green) and five times faster than a manual search in a database (red).

4.4 Measurement of behavioral features

The face analysis module is able to detect and analyze faces of a resolution larger than 100 by 100 pixels from a video stream of a network camera. Figure 16 shows a screenshot of the output of our module.

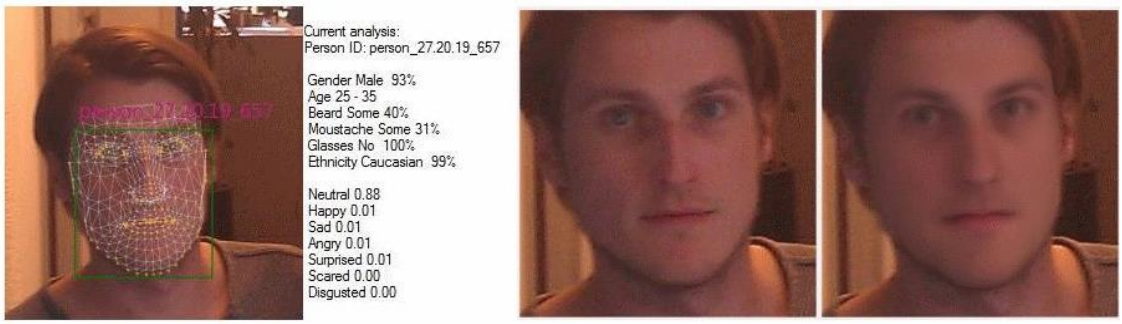


Figure 16: A screenshot of the output of the face analysis module. Left: The mesh of the 3D face model and the estimated face characteristics. Right: the original face and a reconstruction of the face using the Active Appearance Model.

A screenshot of the output of the rest of our system is shown in Figure 17. The figure displays the results of the body part detection module, height estimation module, and attention tracking module of two persons in the shopping mall. An example of the output of the clothing color estimation module is given in Figure 18.

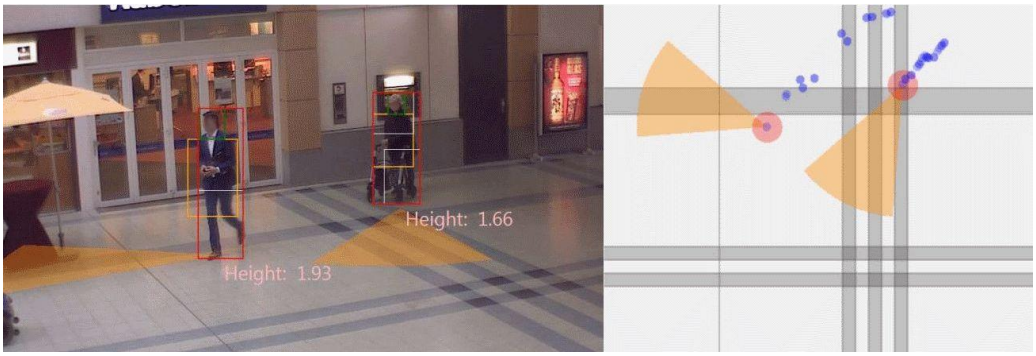
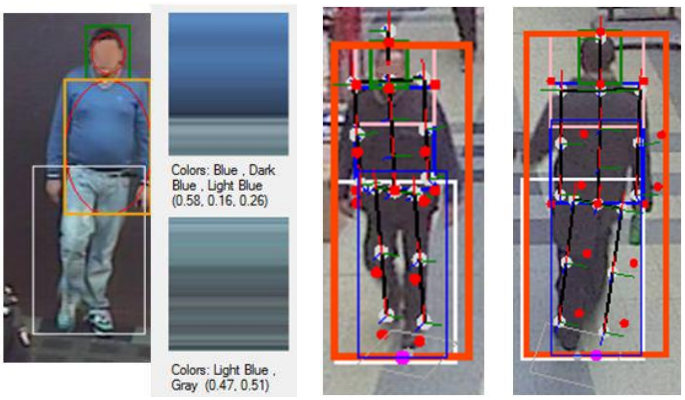


Figure 17: A screenshot of the attention tracking module. Left: Using the camera calibration data, the estimated gaze direction is projected on the image as an orange beam. Right: The location and gaze direction estimate of each person is projected onto a map. The blue dots represent the complete tracklets of both persons.



(a) (b) (c) (d)

Figure 18: (a+b) Estimation of clothing color: the upper body color is estimated as a combination of different shades of blue and the lower body consists of light blue and dark colors. (c+d) Example 3D pose estimation of head and torso superimposed on the images (black lines).

4.5 Track analysis and behavior analysis

The interactive system setup with the mono and stereo people tracking systems, the movement analysis system TrackLab and reasoning system RT-MMC showed that's able to detect abnormal or suspicious behavior and grouping behavior of people in the shopping mall. Figure 19 shows an example of TrackLab's visualization of spatial events for suspicious behavior detection. The resulting behavior events can be used to alert security services or to display the video stream and indicate the location and possibly the subjected person at the moment the event occurs. Besides spatial behavior analysis TrackLab also allows researchers to study visitor behavior by using the numerical analysis functions in combination with the defined zones of interest and visualizations of quantitative results in table format.

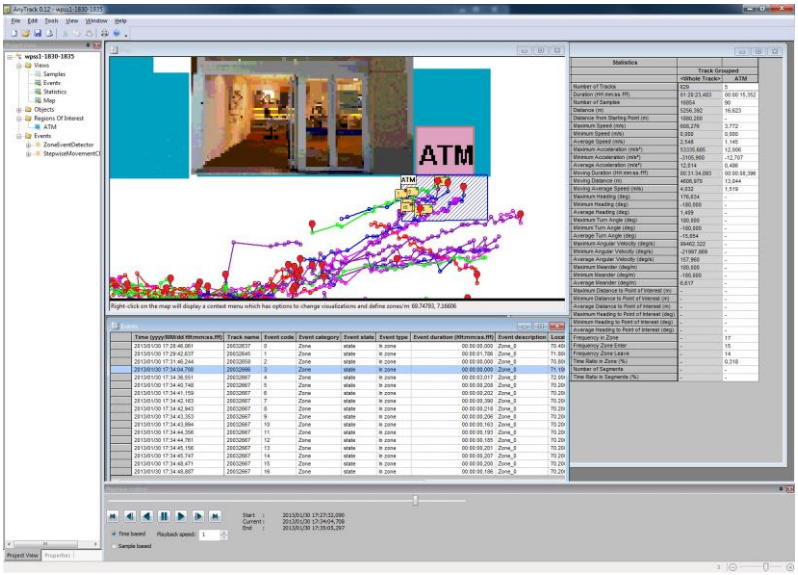


Figure 19: Track visualization in TrackLab. The colored lines are subject tracks (user-defined style). Position markers show the location of the subjects at the current time during play-back. The shaded area is an user-defined zone. The scenario is a test to determine suspicious behavior at the ATM, i.e. same person is entering the ATM zone more than 2 times within 10 minutes. A suspicious behavior event has been sent and is displayed in the user interface of CameraManager.

4.6 Visualization

When visualizing metadata in the live video view, the location of detected persons are shown only when the user moves his mouse over the image. This allows easy selection of interesting persons, but avoids clutter when the user is not interacting with the view.

The metadata is shown graphically with a video loop of the person flanked by a drawing of a figure, colored with the detected appearance of his clothes and extra icons and graphs indicating height, gender and other metadata. The data is shown graphically for two reasons: it is easier to understand and it is easier to translate to multiple languages to further improve understandability.

Detected events can be shown in two ways: as a popup giving a textual description and a screenshot, or by highlighting "interesting" cameras and showing a small clickable textual description of events that happened in the video. When the user then clicks the description, he gets a full description of the event, including a screenshot or small video.

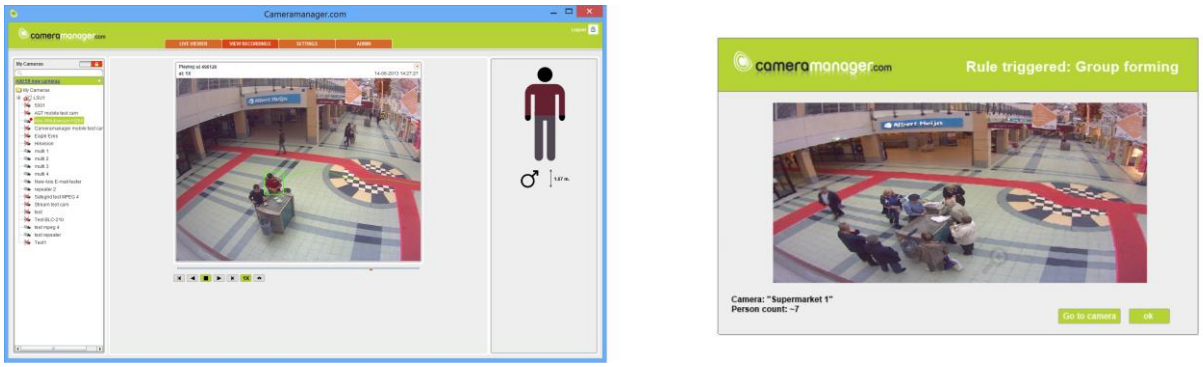


Figure 20: Visualization in the simplified GUI. Left: Video loop of a selected person and icon based appearance and details. Right: Event pop-up showing a grouping event including details such as the number of people forming the group.

5. CONCLUSIONS

In this paper, we presented a software platform that contributes to the retrieval, observation and analysis of human behavior. The platform consists of mono- and stereo-camera tracking, re-identification, behavioral feature computation, track analysis, behavior interpretation and visualization. This system is demonstrated in a crowded shopping mall with multiple cameras and different lighting conditions.

ACKNOWLEDGEMENT

The work for this paper was supported by the ‘Maatschappelijke Innovatie Agenda - Veiligheid’ in the project: “Watching People Security Services” (WPSS). This project is a collaboration between TNO, Eagle Vision, Vicar Vision, Noldus Information Technology, Cameramanager.com and Borking Consultancy. This consortium acknowledges the “Centrum voor Innovatie en Veiligheid” (CIV) and the “Diensten Centrum Beveiliging” (DCB) in Utrecht for providing the fieldlab facilities and support. The work related to 3D pose estimation is partially based on the research code from the GATE project of Utrecht University, funded by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

[1] Aa, N. van der, Noldus L., Veltkamp, R., “Video-based multi-person human motion capturing,” Proc. Measuring Behavior, 75-78 (2012).

[2] Aa, N. van der, Luo, X., Giezeman, G. J., Tan, R. T., Veltkamp, R. C., "UMPM Benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction," IEEE ICCV workshops, (2011).
[3] Alavi, A., Yang, Y., Harandi, M., Sanderson, C., "Multi-shot person re-identification via relational stein divergence," IEEE Int. Conf. Image Processing ICIP, (2013).
[4] An, L., Kafai, M., Yang, S., Bhanu, B., "Reference-based person re-identification," IEEE Int. Conf. Adv. Video and Signal-based Surveillance AVSS, (2013).
[5] Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P., "A database for person re-identification in multi-camera surveillance networks," IEEE Int. Conf. DICTA, (2012).
[6] Benfold, B., Reid, I., "Guiding visual surveillance by tracking human attention," Proc. BMVC, (2009).
[7] Bouma, H., Borsboom, S., Hollander, R. den, Landsmeer, S., Worring, M., "Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination," Proc. SPIE 8359, (2012).
[8] Bouma, H., Baan, J., Landsmeer, S., Kruszynski, C., Antwerpen, G. van, Dijk, J., "Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall," Proc. SPIE 8756, (2013).
[9] Bouma, H., Burghouts, G., Penning, L., and others, "Recognition and localization of relevant human behavior in videos," Proc. SPIE 8711, (2013).
[10] Burghouts, G., Eendebak, P., Bouma, H., Hove, J.M., "Improved action recognition by combining multiple 2D views in the Bag-of-Words model," IEEE Int. Conf. Advanced Video and Signal-Based Surveillance, (2013).
[11] Burghouts, G.J., Schutte, K., Bouma, H., Hollander, R., "Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos," Machine Vision and Applications, (2013).
[12] Criminisi, A., Reid, I., Zisserman, A., "Single view metrology," Int. J. Comput. Vision 40(2), 123-148 (2000).
[13] Dalal, N., Triggs B., "Histograms of oriented gradients for human detection," Proc. CVPR, (2005).
[14] Friedman, J., Hastie, T., Tibshirani, R., "Additive logistic regression: a statistical view of boosting," Annals of Statistics, (2000).
[15] Gevers, T., "Color-based retrieval," Principals of Visual Information Retrieval, 11-49 (2001).
[16] Gray, D., Brennan, S., Tao, H., "Evaluating appearance models for recognition, reacquisition, and tracking," IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, (2007).
[17] Groot, K. de, Boughorbel, S., Ambroise, N., Buter M., Kang J., Loke B., Spreeuwers L., Vandenabeele J., "Multimodal sensing system to enhance the safety of infants in the home environment," Proc. Measuring Behavior, 491-494 (2012).
[18] Iodice, S., Petrosino, A., "Person re-identification based on enriched symmetry salient features and graph matching," Pattern Recognition LNCS 7914, 155-164 (2013).
[19] Kuilenburg, H. van, Wiering, M., Uyl, M.J. den, "A model based method for automatic facial expression recognition," Proc. ECML, (2005).
[20] Li, X., Tao, D., Jin, L., Wang, Y., Yuan, Y., "Person re-identification by regularized smoothing KISS metric learning," IEEE Trans. Circuits and Systems for Video Technology, (2013).
[21] O'Hara, J., "Toward a commodity enterprise middleware," ACM Queue 5(4), 48-55 (2007).
[22] Ramakrishna, V., Kanade, T., Sheikh, Y., "Reconstructing 3D human pose from 2D image landmarks," Proc. ECCV, (2012).
[23] Simo-Serra, E., Ramisa, A., Alenya, G., Torras, C., Moreno-Noguer, F., "Single image 3D human pose estimation from noisy observations," Proc. CVPR, (2012).
[24] Soori, U., Yuen, P., Han, J. W., and others, "Target recognitions in multiple-camera closed-circuit television using color constancy," Optical Engineering 52(4), (2013).
[25] Stoeller, B., "Recognising individuals by appearance across non-overlapping stereo cameras," M.Sc. thesis University of Amsterdam, (2012).
[26] TrackLab, URL: http://www.noldus.com/innovationworks/products/tracklab.
[27] Tuzel, O., Porikli, F., Meer, P., "Region Covariance: A Fast Descriptor for Detection And Classification," Proc. ECCV, (2006)
[28] Viola, P., Jones, M., "Robust real-time face detection," Int. J. Comput. Vision, 57(2), 137-154 (2004).
[29] Welman, C., "Inverse kinematics and geometric constraints for articulated figure manipulation," Master's thesis Simon Fraser University Canada, (1993).
[30] Wozniak, M., Odziemczyk, W., Nagorski, K., "Investigation of practical and theoretical accuracy of wireless indoor-positioning system UBISENSE," EGU General Assembly 15, 7845-7845 (2013).