Crowd control by multiple cameras

L.J.M. Rothkrantz
Man-Machine Interaction Group
Delft University of Technology
2628CD Delft, The Netherlands
L.J.M.Rothkrantz@tudelft.nl

SEWACO Faculty of Military Sciences
Netherlands Defence Academy
Den Helder, The Netherlands

Z.Yang
Man-Machine Interaction Group
Delft University of Technology
2628CD Delft, The Netherlands
Z.yang@tudelft.nl

ABSTRACT

One of the goals of the crowd control project at Delft University of Technology is to detect and track people during a crisis event, classify their behavior and assess what is happening. The assumption is that the crisis area is observed by multiple cameras (fixed or mobile). The cameras sense the environment and extract features such as the amount of motion. These features are the input to a Bayesian network with nodes corresponding to situations such as terroristic attack, fire, and explosion. Given the probabilities of the observed features, by reasoning, the likelihood of the possible situations can be computed. A prototype was tested in a train compartment and its environment. Forty scenarios, performed by actors, were recorded. From the recordings the conditional probabilities have been computed. The scenarios are designed as scripts which proved to be a good methodology. The models, experiments and results will be presented in the paper.

Keywords

Bayesian reasoning, computer vision, scenarios, scripts.

INTRODUCTION

Surveillance of public spaces is currently widely used to monitor areas and the behavior of people in these areas [1, 9]. Closed Circuit television (CCTV) systems around the world are used to monitor the safety of people in public spaces 24 hours a day (Fig1). Since events like the terrorist attack in Madrid and London there has been a further increasing need for video network systems to guarantee the safety of people in public areas. Those systems can be used in case of crisis or disasters. However the greater the number of cameras, the greater the number of operators and supervisors that are needed to monitor the video streams. The solution of that problem could be an automated surveillance system. A fully automated surveillance system is currently not available commercially. Some software packages do exist, but they mostly record video streams and provide little further functionality. Behavior detection and motion detection, as well as human tracking methods are a widely researched topic. A combination of these methods has been used in our project to develop an automated system capable of classifying human behavior and compute situational awareness.

Today the results shown by object detection and motion tracking demonstrate many opportunities in current situations and future applications [1,4]. Both fields benefit from great attention in current society. The power to predict information about object identity, behavior and position can be used by monitoring services, security applications and information retrieval.

(a) **(b)**

Figure 1. (a) Typical outdoor surveillance camera (b) Working algorithm detecting and tracking multiple people using one single camera

Identification in this paper is not only focused on finding credentials of a humanoid object but in a broader sense also on identifying certain tasks this object is performing or behavior it is showing. Together with other information, detection and tracking data results in predicting behavior and processed tasks, and gives semantic meaning to sensor data. Central to this monitoring task is the translation of the 'where they are' information to predict the 'what they are doing' question. This research focuses primarily on the 'where they are' question in specific environments. Our developed prototype has been developed and tested in railway station environment. The results of this research can then be used to predict the 'what they are doing' question. Also, when tracking humanoid objects, these applications usually cope with crowded scenes and tracking occluded objects. Furthermore lighting conditions, colors of people or environments and quality of image feeds are very determining for system results. Some solutions to these difficulties have been presented, most successfully by using a setup with multiple cameras to aid the system with more spatial information and less uncertainty of other variables about the objects being tracked. Such a system is capable of detecting and tracking multiple people in cluttered scenes with various lighting conditions.

The outline of the paper is as follows. In the next section we describe related work. Then we present our model. Then we present our experiments and results. The last section contains our conclusions.

RELATED WORK

Behavior Recognition Systems

In any system that will have some form of a behavior recognition module there will be some form of tracking module too, to follow certain movements of an actor in the scene. Such a system has been proposed by Cupillard et.al [11] where a behavior recognition module relies on a vision module composed of three tasks: (a) motion detection and frame to frame tracking, (b) multiple cameras combination and (c) long term tracking of individuals, groups of people and crowd evolving in the scene. For each tracked actor, the behavior recognition module performs three levels of reasoning: states, events and scenarios. The vision module is composed of three tasks. First a motion detector and a frame to frame tracker generates a graph of mobile objects for each calibrated camera. Second, a combination mechanism is performed to combine the graphs computed for each camera into a global one. Third, this global graph is used for long term tracking of individuals, groups of people and crowd evolving in the scene (typically on hundreds of frames). The motion detector and frame to frame tracker have three sub-tasks: detection of mobile objects, extraction of features and classification of mobile objects. A list of mobile objects is obtained at each frame. Once all mobile objects are extracted for each camera they are added to a graph. All the graphs for each camera with all mobile objects are now combined into a combined graph. For a survey of similar systems we refer to Hu et. al. [12].

Tracking of people occlusion detection

Before tracking can begin, the tracking algorithm needs to know which objects to track. Some kind of motion segmentation needs to be done. One way to do this is by using a technique called optical flow. With this technique, every pixel has a corresponding motion vector that is updated every step. However, this is a computationally very expensive method, and so not so useful for real-time applications.

Another method of separating the background (static parts) from the foreground (moving parts), is to use background subtraction. A model of the background is kept in memory, and when there appears a moving object, that is not consistent with the background model, the object is seen as foreground. Because changes in lighting or weather can influence the background, most often an adaptive background model is used, that is, one that is updated every step. The method described by Grimson and Stauffer [10] uses a mixture of Gaussians for modeling the adaptive background and has proven to be a powerful and reliable background subtraction method.

State-of-the-art tracking methods can be divided in several ways. An important first distinction is the way in which a tracked person is represented. When there is no predefined explicit shape model, some possibilities are a box, an ellipse (generally a more appropriate shape for tracking of humans), the contours of a blob [4], or the blob itself. If there is an explicit shape model, a stick figure can be used, or every body part can have its own box. Then the tracking algorithm itself must be chosen. Condensation is a very well-known algorithm in the Computer Vision research with a lot of applications based on it that are already in use. It is actually an example of a particle filter, also known as a Sequential Monte Carlo method. The idea of a particle filter is that random sampling is used to estimate a Bayesian model. Large advantages of Condensation is that it is possible to use it in a very broad range of applications, and that it is usable more or less independent of the object representation. A problem is that particle filters are not very good for use in multimodal applications (tracking of more than one object at the same time), so Condensation does suffer the same problem. Improvements that have been suggested are to mitigate this problem, for example to use a MCMC (Markov Chain Monte Carlo) based particle filter, or to use a mixture of particle filters to create one multimodal particle filter [10]. A less known alternative for Condensation is tracking based on the Mean Shift algorithm [10]. Mean shift is an older theoretical method, but can be applied to tracking in a very promising way. The 'mean shift' is the estimated direction and distance in which the target moves, and this is computed by comparing an already defined model target with the current candidate target. The targets are defined by their color distribution (histogram). The advantage of this method is that no dynamic model is needed in advance. It is a fast and reliable procedure that has started to be used in many tracking systems.

Given that there is a reliable multiple-person tracking system in place, the occlusion detection can start to find the depth ordering of occluding objects. Several methods have been proposed to solve this problem. One solution, described in [10], adds occlusion handling to Condensation. A virtual object representing the occlusion relation between two objects is subjected to the same sampling process as all normal Condensation objects. Another method is used when objects are represented as blobs [10]. In these systems, when moving blobs merge into one and then split again, it is seen as the assembling and disassembling of groups of people. If the group consists of more people, it is segmented into its constituent people. Also the detection of body parts, like the head, can be used to count the number of people in the group. This method does not give real depth information, but it still can quickly find people in a group.

Face detection

Face detection is defined as a computer technology that determines the locations and sizes of human faces in arbitrary (digital) images. It is designed to detect facial features and ignore everything else. Face detection does not recognize identity nor does it describe anything about facial expressions or emotions seen on the face. Currently three notable face detection algorithms are possibly useful in solving our problem. The face detector that is most widely used and implemented in the OpenCV library is the Viola and Jones algorithm [3]. The University of Beijing is working on a successor based on algorithms proposed by Viola and Jones, created by Huang and Haizhou [10]. Finally Idiap is a possible face detector for the system.

SCRIPTS

Context awareness will be computed by using a Bayesian networks. A disadvantage of Bayesian networks is the great number of probabilities, which have to be defined in the conditional probabilities tables. If enough data is available the probabilities can be computed using special training algorithms. In the case of crisis situations and disasters it can be expected that only sparse data are available. In that case experts must usually provide the necessary data or probabilities. In our project we use an alternative: we designed scripts for possible disaster scenarios.

Scripts provide a way for understanding and enacting behavioral patterns and routines. A classic example being Schank and Abelson's [8] example of the restaurant scripts that includes a structure of elements for entering a restaurant, sitting down, ordering food, eating, conversing, paying the bill and leaving. Scripts used for both comprehension and behavior generation, represent a structure of cognitive functions that may include cognitive resources, perceptual interpretations and preconditions, decision processes, attention management and responsive motor actions. Story scripts are patterns representing a structure of understandable elements that must occur to make stories comprehensible. The presence of story scripts in the cognitive systems of

storytellers, listeners, readers or viewers of stories allow stories to be told and to be comprehended, including the inference of missing information. If a story deviates too far from known story scripts, it will not be perceived as a coherent story. Apart from the structure in time, scripts are also concerned with the structure in space. Spatial structures, such as the layout of a railway station or a train are commonly known as scenes.

We designed scripts for 40 global scenarios; some of them are related to crisis others describe normal situations. Examples are fire, bomb explosion, terroristic attack, hooligans, fighting people, but also neutral and friendly situations such as waiting people for the train, movement of travelers around arriving and leaving trains. For every scenario we sketched a storybook (see Fig 2). To improve the consistency of both scripts and storybooks and the interaction between both we did the following experiments. Respondents were requested to tell the full story based reading the storybook or the other way around sketching the storybook after reading the scripts. Only the scripts that were verified to yield consistent results were used. Next the scripts were used to instruct actors to play the scenes (see Fig 3). Scenes were recorded and analyzed for discriminative and predictive features. Our goal was to design an automated system, so we also had to focus on those features that can be extracted in an automated way.

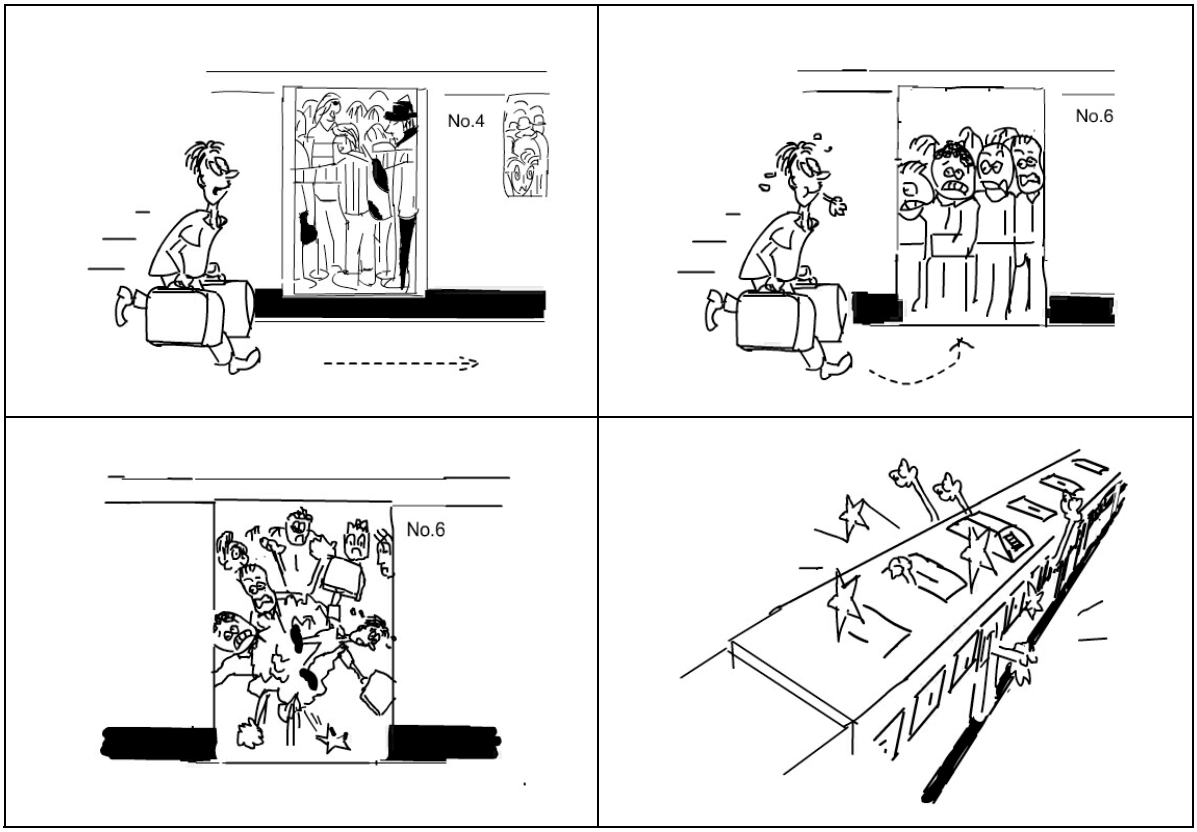


Figure 2. Intrusion in full train (scenes from the storybook)

Figure 3. Screenshots of intrusion in full train (video recordings from the story book)

MODEL

Architecture of surveillance system

Automated video surveillance is a task that includes many subtasks. For this reason we suggest a modular approach for such a system. A diagram of our prototype is shown in figure 4. The video input stage takes the raw input from the video cameras in the train. In the next step this data is processed so that it is fit for the next two stages where the image sequences will be analyzed. This involves tasks such as resizing the video, and applying corrections for orientation. The system then splits into two parallel tasks, motion based, and human detection based analysis. The motion based analysis will perform motion detection and some motion recognition. The human detection side is aimed to detect the humans in a scene and their locations, and track them accordingly. The results of these two modules are then analyzed in the behavior analysis module, which will try to infer behavior, and discern between aggressive and non-aggressive behavior. The final stage of the system will take this behavior analysis, and produce an alert when aggression is detected. This could then be used to alert the operator to evaluate the situation on the video stream, or even automatically dispatch the security personnel required for handling the situation.

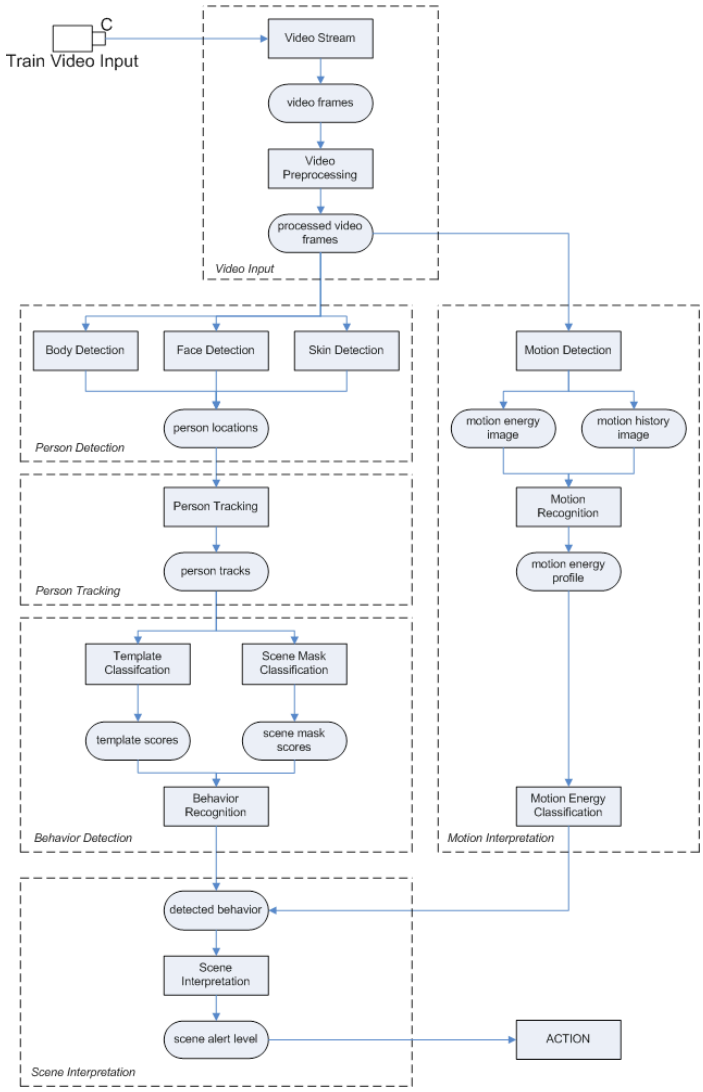


Figure 4. Framework for an automated surveillance system

Feature extracting methods

The main goal of the project is to develop an automated video surveillance system hat can detect and analyze types of human behavior. The environment where the prototype has been tested is a railway station and its

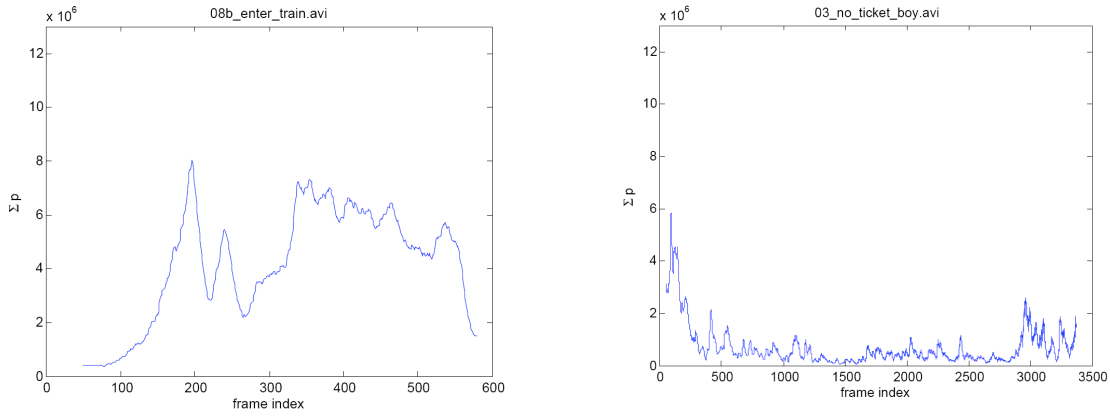
environment. The cameras at the entrance of the railway station records incoming and outgoing travelers. As soon some travelers start running, running forth and back our system should be alarmed. Also some behavior such as pushing other people, falling or lying on the ground is unusual behavior.

Being able to detect motion in a scene is the first and most basic step towards understanding behavior. Motion detection aims at segmenting regions of an image corresponding to moving objects from the rest of the image. A first and low level approach to scene understanding is to detect motion. Motion detection algorithms, including the ones implemented in our prototype will perform a comparison with a previous frame or background frame. The result of the motion detection then yields a so-called foreground image, which contains the moving pixels. With his information we determine several features about the detection motion which are nodes in our Bayesian network.

Feature 1 - Location in the image The location in the image will be helpful to determine what action is being observed. We made scene models in different areas, to interpret motion.
Feature 2 - Direction The direction of movement tells us whether a person is entering or leaving, and where a moving object is moving to. Movement opposite to the main movements gives a high probability to suspicious persons.
Feature 3 - Speed We expect passengers to move at a certain walking speed. Objects moving at a higher or lower speed can be a sign of fleeing or victims respectively.
Feature 4 - Brownian motion We compute the average movement of a cloud of people as an entropy measurement of the human system

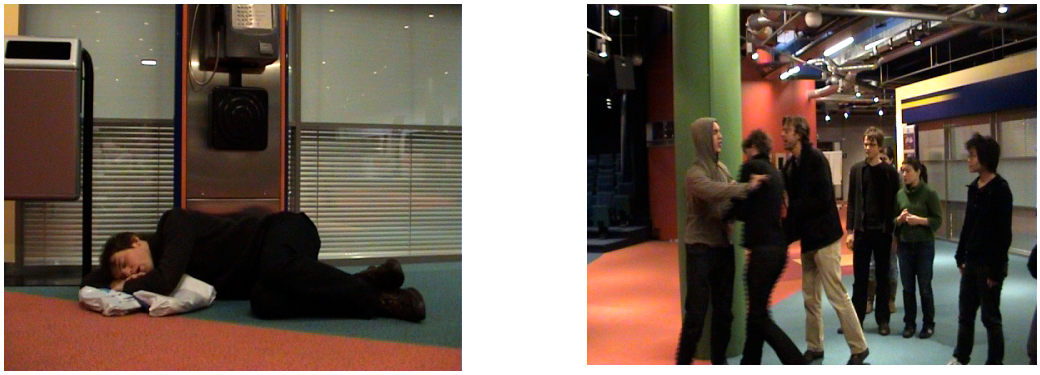
Motion detection algorithms can give us data on the direction and speed of moving objects in a scene, as well as the total amount of motion observed. We can compare this data to previously recorded scenarios to see if they are similar and recognize behavior in this manner. In our prototype we implemented the Motion Energy Image (MEI) as an indicator of motion of the last viewed frames, opposed to the Motion History Image (MHI) which describes the recency of motion. We expect that different scenarios have different amounts of motion (see figures 5 to 7). We designed models of running passengers, crawling passengers and simple walking passengers.

Feature 5 - Running
Feature 6 - Crawling
Feature 7 - Fighting



(a) (b)

Figure 5. Energy graph fighting people (a, figure 6b), victim laying on the ground (b, figure 6a)



(a) (b)

Figure 6. Low energy (a) vs. high energy (b)

Human detection

The main method we used for the purpose of human detection is face detection. When a face is detected in an image, this obviously means we have detected a human as well. We used the implementation provided with the openCV library, which is based on a method proposed by Viola & Jones [3]. The software was extended so that we can look up the face in a database to find out if that person is wanted, suspicious, a rescue worker etc.

Feature 8 - (Un-) Friendly persons

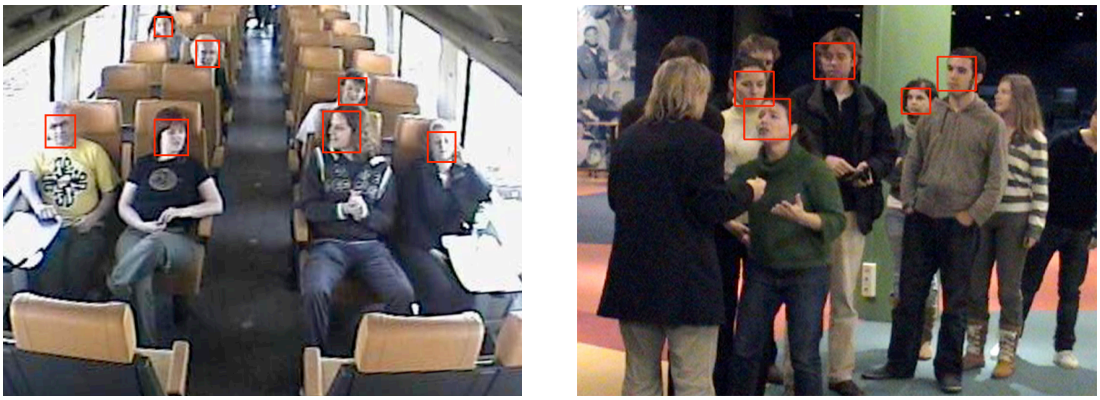


Fig 8. Face localization

Human tracking

Face detection is also used as a human tracker tool. As a second method we implemented Kalman filters. Most people enter a railway station/train or leave it in a straight way. Other people hanging around or move without any plan. Some path crosses paths of the people in the main stream

Feature 9 - Tracked path goal directed
Feature 10 - Tracked path without goal
Feature 11 - Exceptional path

BAYESIAN REASONING

Bayesian Networks are appropriate tools able to represent various sources of uncertain information and join them into an inference system. In probability reasoning, random variables are used to represent events and objects in the world. By assigning values to these random variables, we can model the current state of the world (in our case a railway station environment) and weight the states according to the joint probabilities.

A Bayesian Network is a graphical model for probabilistic relationships among a set of variables. Figure 8 shows an example of a Bayesian network which models aggression in railway stations and trains. At the lowest level we have our input features, extracted from the footage, next we want to compute the probability of the 40 different scenarios and finally we want to know if the situation is dangerous or not and if help or assistance is needed. There is no direct relation between the appearance of some feature and the amount of danger. If a person is detected lying on the ground it can be a victim, a sleeping person, a person with some illness etc. The

observation of other features and the relation between them as modeled by our BN let us conclude about the outcome. The frequencies of objects and relations between nodes provide us basic information to compute the conditional probability tables. Policemen, firemen and first responders in crisis situations were interviewed about the probabilities between nodes.

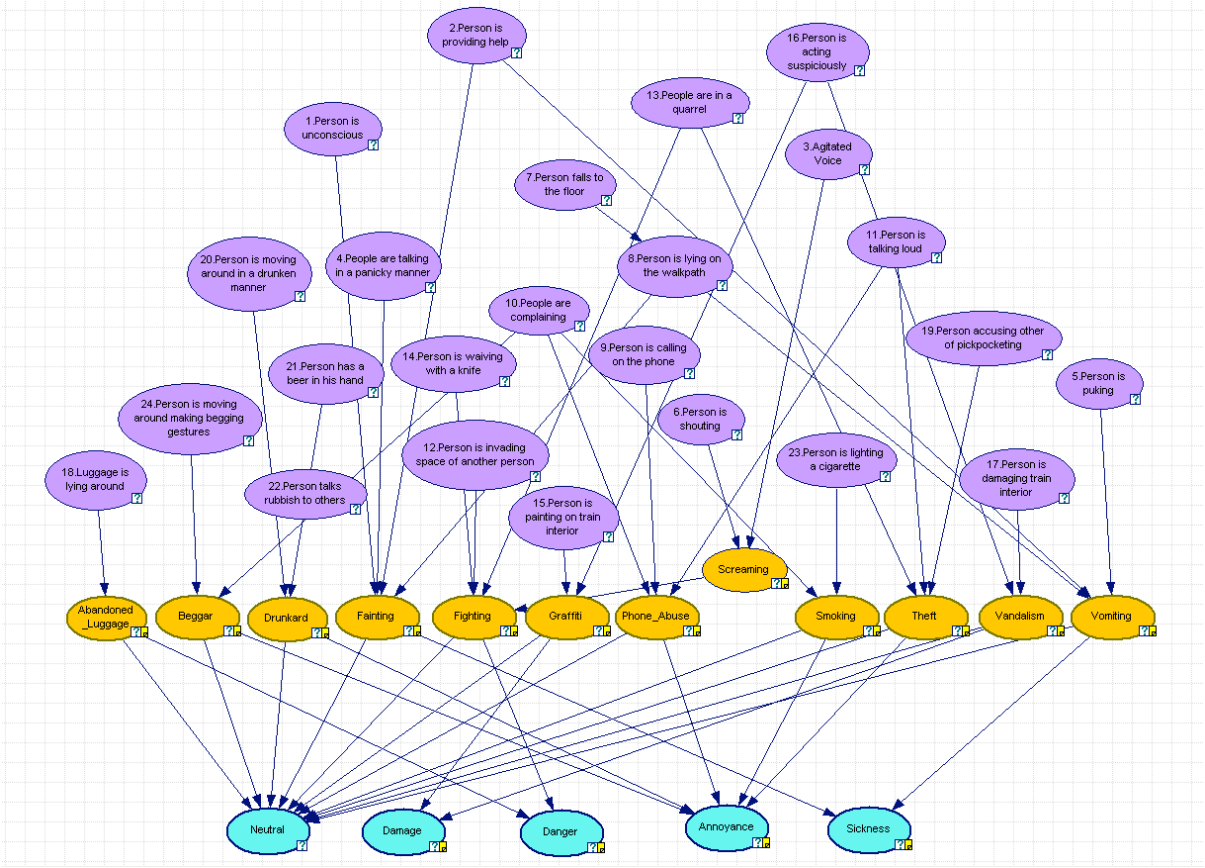


Figure 8: Bayesian net

IMPLEMENTATION AND RESULTS

To detect the behavior previously described, there are currently a number of computer vision techniques available. Tracking multiple persons in a group can be done by using multiple single-person trackers at the same time. We adapted available single person tracking methods (Condensation and Mean Shift). To solve the problem of occlusion detection we implemented two different methods. A detailed presentation of these algorithms is beyond the scope of this paper. The algorithms and test results are described in [10]. On average the off-line result of the recognition score in our experiments of acted scenes is 90% (given sufficient video resolution and stable lighting conditions).

One main method we used for the purpose of human detection is face detection. When a face is detected in an image, this obviously means we have detected a human as well. We used the implementation of the OpenCV library, which is based on the method proposed by Viola & Jones in [3]. When applied to a single frame, the algorithm will return the regions in the image in which a face is believed to be. This gives us the location and the size of face for each frame, as well as the total number of faces present. This information can be stored for each frame in the video stream. The data is then used to construct paths of detected faces over consecutive frames, so we can track movement of passengers. Having detection data over a longer period of time will allow us to use filtering method to smooth out the paths of passengers, as well as fill in the gaps if detection should fail. Using these filtering techniques, we were able to perform person tracking with only several percentage true positive measurements per seconds.

The Viola & Jones face detector needs features to detect a face. Thus, the smaller a face, the lower the detection rate will be. The faces we wish to detect in the video taken from cameras already installed in trains are usually small in size, due to the limited resolution of the cameras. We determined that a person standing in front of a camera in our train setup will produce an image in which the size of the face is roughly about 1024 pixels. We found a steep threshold at an average face size of about 1200 pixels, after which detection rates fall

dramatically. The results from this experiment therefore suggest that we need to have image in which the faces are at least this resolution, and preferably larger.

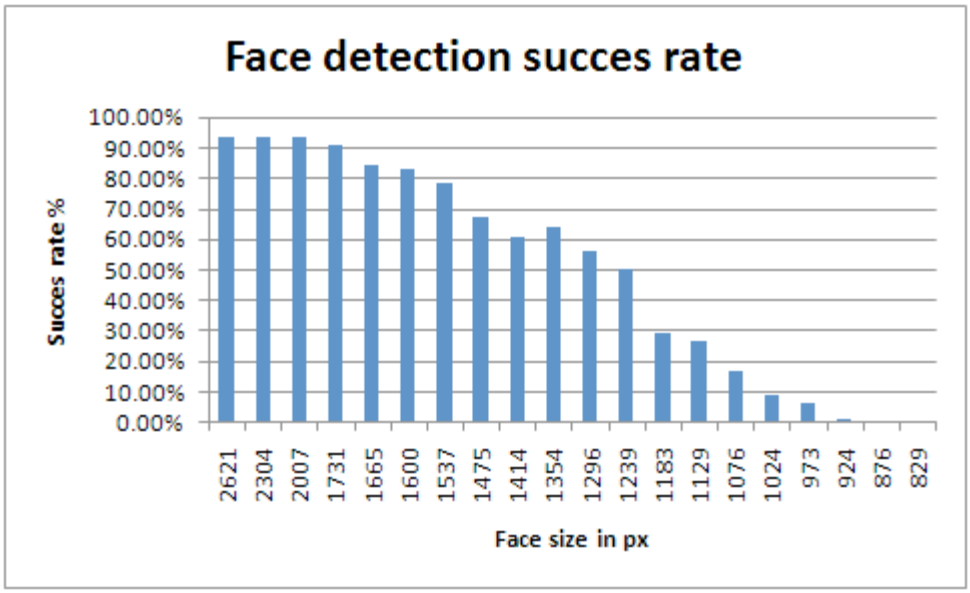


Figure 9: Face detection success rate for various face sizes

Our Bayesian network is implemented under SMILE and GeNIe. The Structural Modeling, Inference, and Learning Engine (SMILE) is a fully platform independent library that implements graphical probabilistic and decision-theoretic models and is suitable for direct inclusion in intelligent systems. SMILE was first released in 1997 and has been thoroughly used in the field since. The classes defined in SMILE enables us to create, edit, save and load graphical models, and use them for probabilistic reasoning and decision making under uncertainty. The Graphical Network Interface (GeNIe) is the graphical user interface to the SMILE library.

EXPERIMENTAL RESULTS

The goal of our project was to develop a surveillance system of a network of CCTV cameras for crowd control and situational awareness. We developed a running prototype for a railway system. The input to the system consists of features automatically extracted from video recordings. Most features are based on the movement of (groups of) people. In the past years we developed some computer vision modules to localize faces and to track faces based on Viola Jones algorithms [3] and on the color of the face. A high resolution camera is required to get video frames with faces more then minimal size. At this moment we process the data offline, because the algorithm is time consuming. Another problem is that people are recorded from above, from an angle of about 25 degree. Because our system is trained on frontal faces we rotate the recorded pictures (figure 10). In the future we plan to train our system on recognition of faces under varying angles, but a huge amount of data and computer power is needed.

Figure 10: Comparison of an original image (left) from the train camera and the same image after scaling and adjustment (right)

The tracking of faces is based on Kalman filtering. It proves that we are able to localize and track facial expressions given good quality of cameras and stable lighting conditions. We have to realize that there is a lot of redundancy in the recordings. And to record the movement of a crowd, only recordings of some individual are needed. The video recordings of our experiments are of high quality so our system was able to extract the features with high probability. Up to now we didn't test our system on real life recordings.

Our methodology to generate crisis models based on scripts and storybooks enables us to define features and to make assessments or good guesses of the corresponding probabilities. We defined 40 scripts but our results are probably dependent on these scripts. An open question is in how far our prototype is able to handle new situations. To split our data in an experimental and test set makes no sense because we have sparse data. Figure 8 shows that we have many variables and need much more data to compute the conditional probabilities.

Currently the probabilities are mostly defined by experts and our recordings. As a consequence our Bayesian network gives almost perfect results with perfect input. But thanks to our architecture of three layers, the lowest layer for the sensor features, the middle-layer for the scenarios and the upper layer for awareness in terms of (un-)friendly, needs assistance, we hope that our BN generalize to new scenarios. We assume that our scenarios cover the whole space of possible disasters around railway stations. In real life we have incomplete and noisy data but still we are able to compute the most probable scenarios. During a crisis we can expect that in the course of time more complete and less noisy data becomes available, so the results will improve. In the future we will extend to Dynamic Bayesian networks to take time into account.

SUMMARY AND CONCLUSION

The goal of our project was to design an automated surveillance system. The system can be used in crisis situations, aggressive situations or to observe and control crowd of people. As location a railway station has been chosen. The system is composed of a network of cameras. Special software from the area of computer vision enables the cameras to extract features from the video recordings. These basic features are related to assessment of different forms of movement. Next the features are fed into a Bayesian network to reason about the probabilities of (not-) dangerous situations. To compute the values in the conditional probability tables we designed 40 scenarios and scripts about disaster and aggressive scenes. These scenes were acted by players. From the video recordings we defined salient and discriminative features for the different scenes and computed basic probabilities.

We implemented a running prototype and tested it on recordings of 40 acted scripts/storyboards. The recordings are of good quality with respect to lighting and camera position. In real life less ideal situations can be expected. The test results are quite good. But we realize that our model, probabilities are based on the same recordings. But we expect that our 40 scripts are good representatives of all possible scripts, so our results will generalize to unexpected not modeled situations.

The main innovative aspect of this paper is the combination of techniques to achieve the goal. The overall performance of the system is as usual in pipeline systems dependent of the weakest link. The weaknesses of the system are the low level detection algorithms that depend on good quality input and stable conditions. The strength of the overall system lies in the way the components work together e.g. the tracker that is able to smooth tracks and fill in gaps left by the face detector and the Bayesian network that is able to generalizes and reason with uncertain data.

In the future we hope to get recordings of real life situation to test our system. But for example privacy aspects have to be solved before data is freely available for research.

REFERENCES

1. Weiming Hu, Tieniu Tan, Liang Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transaction on,* 34(3):334-352, Aug.2004

2. Sethi, I.K. and Jain,R. Finding trajectories of feature points in a monocular image sequence. *IEEE Trans. Pattern Anal. Mach. Intell.,* 9(1):56-73, 1987

3. Viola, P. and Jones, M. Robuust real-time face detection. *Int. J. Comput. Vision*, 57(2):137-154, 2004

4. Liang Wang, Weimang Hu, and Tieniu Tan, Recent developments in human motion analysis. *Pattern recognition*, 36(3):585-601, 2003

5. Bo Wu, Haizou Ai, Chang Huang, and Shihong Lao. Fast rotation invariant multiview face detection based on real adaboost. *Automatic face and gesture Recognition, 2004. Proceedings Sixth IEEE International Conference on*, pages 79-84, May 2004

7. Yang. Z, Keur, A., Rothkrantz, L.J.M. Behavior detection in Dutch train compartments. *In Proceedings of Euromedia 2008*, pages 52-57. April 2008

8. Zhao, W., Chellappa, R, Philips, P.J., and Rosenfeld, A., Face recognition: A literature survey. *ACM Comput. Surv.,* 35(4):399-458, 2003

9. Schank, R., and Abelson, R., (1977) Scripts , plans, Goals and Understanding, Hillsdale, NJ: Erlbaum

10. Datcu, D., Yang, Z., and Rothkrantz, L.J.M., Multimodal workbench for automatic surveillance applications. Computer Vision and Pattern recognition, 2007. CVPR'07, IEEE Conference on, pages 1-2, June 2007

11. Veen van der, C. Crowd surveillance by video camera, Technical Report TUDelft, MMI-2006-7

12. Hu, W., Tan, T., Wang, L., Maybank, S., A Survey on Visual Surveillance of Object Motion and Behaviors, *IEEE Transactions on Systems, Man, And Cybernetics.,* 34(3):334-352, 2004