[DOAS] Tracking of Multiple Occluded People using Combined Camera Feeds

Dimitris Karapapas, Arjan Nusselder, Jeroen Vlek
{dkarapap,anussel,jvlek}@science.uva.nl

January 29, 2009

Abstract

Tracking objects, and specifically people, in an image sequence is often complicated by the problem of occlusion. In order to solve this problem, a method using multiple cameras is presented and implemented. By combining multiple images, an object representation in a three dimensional space is created. These objects are labeled and tracked throughout a sequence, using a Kalman filter and object comparison based on colour histograms. It is shown that a basic setup is able to correctly track people in real-world data sets. Possible improvements are given, that should allow a more robust tracking under more difficult circumstances.

1 Introduction

An increasing amount of video information is becoming available to monitor and hopefully prevent misdeeds in public areas. This increase effects the usefulness of an automated analysis and interpretation of video, that can signal possible disturbances. Many of such systems are developed, but there is not yet an encompassing solution that works in the general case. Looking at one such system, we propose an addition to the complete system where we keep track of all people in the ground plane projection of a three dimensional space. Using a data set with three simultaneous videostreams per scene, the addition is tested and improved.

The existing system detects aggression by combining sound and video. The sound can give a general indication of excitement. To identify whether this corresponds to aggressive behaviour –as opposed to other forms of exaltation or background noise– it is important to know which people are in the video feed. Questions like 'how many people are there' and 'what is their movement' must be answered to attribute behaviour to a specific person present at a scene. To have a better visual view of the people, multiple cameras are used. In this project we focus on the problem of differentiating between people and tracking them in a cluttered scene.

Keeping track of people can be a difficult task for a number of reasons. First the quality of the images or generic clothing may make it hard to distinguish between two people. Second, naturally or artificially occuring illumination changes complicate both properly detecting foreground objects and matching people in sequential frames. Third the people will, depending on the view, at some time occlude each other. Last it is very well possible that at times the background moves as much as the people in front of it. Either the objects that should be considered background still move –like a train passing by– or people stand still in one location long enough to appear as background.

To robustly track people in a changing environment, a model of each person is kept and is continually updated. When a sufficiently clear patch of foreground is detected in all three movies, this model will decide to either update an existing person or introduce a new person to the scene. This person is then placed in a three dimensional scene representation. Using a tracking Kalman filter, the estimated new position is then corrected with previous infor-

mation about the persons position.

To give a broader view of the problems the existing system tries to solve, relevant related work is mentioned in Section 2. A description of the data follows in Section 3. Then the main approach is detailed in Section 4. First the foreground segmentation (Section 4.1) and the three dimensional space (Section 4.2) are discussed. Next the person model –that will describe and decide whether some visual information is in fact a person– is described in Section 4.3. The tracker that incorporates motion estimation is described in Section 4.4. An evaluation of different extensions is tested on the data and interpreted in Sections 5 and 6 respectively. We conclude on the current status in Section 7 and point to possible enhancements in Section 8.

2 Related work

An approach similar to the one currently presented was done by Khan and Shah, who also track people in a ground plane representation of a scene.[1] For that, they use the information from multiple cameras and combine it into synergy maps [sic]. Our approach differs mainly in that we use the camera calibration to explicitly create a three dimensional voxel space, whereas Khan and Shah project rays from the uncalibrated camera images through the ground plane.

Foreground segmentation of objects moving in the frames is done in accordance with the *Gaussian mixture model* described by Zivkovic and Van der Heijden in [5].

An addition to the person model could be an enhanced histogram representation using different body parts, as done by Quinn et. al. in [2].

To facilitate our approach, a specifically created data set was used that is described in Section 3.[4]

3 Data

To have clear video footage in multiple cameras of –in the end– aggressive behaviour, several actors were hired to play out scenes that might occur in real life (see [4] for a description of the full dataset and its use). These scenes were filmed at the *Amstel Station* in Amsterdam. Available for this project are three scenes, each consisting of three movies (one for each of the three views). The scenes impose increasing difficulties for the tracker. The first scene only has two people that hardly occlude each other. The second scene has more people walking around, with plenty of occlusion happening. In the third scene, a train arrives in the scene, complicating the correct detection of people because there is no clear foreground. An example frame of the first scene can be seen in Figure 1.

4 Method

The basic input for the entire tracking system is a sequence of timesteps. A timestep is defined as a set of three frames, one from each camera, that show the same scene at the same time, but from different perspectives. Several subproblems can be defined that, when solved and used together, allow for automatic detection and tracking of people. After a global description of the system, each subproblem is discussed in more detail.

First a preliminary distinction between people and background must be made, by highlighting moving objects. This is described in Section 4.1.

The camera calibration gives a real-world coordinate for each pixel in the image, as if it were on the ground plane. Combining the highlighted parts of an image with the cameras calibration information, the location of the moving object can then be projected into a three dimensional voxelspace (see Section 4.2). Such a projection is not necessarily correct however, especially if multiple objects start to occlude each other. Because the information of three separate views are projected into one space however, some occlusion can already be (implicitly) detected. If we assume that people can be clearly distinguished at some point, for example when they enter a scene, it is possible to project a clear *voxel blob* into the three dimen-

Figure 1: Three views from the same frame of the first scene, showing occlusion in the left view, but a clearer distinction as seen from the other two cameras.

sional space. Such a blob is then labeled with an identifier, and added to a list of possible people present in the current scene. Once we have decided upon a labeled person, a model describing this person is constructed, consisting of a colour histogram and a Kalman filter.

A colour histogram can be extracted for each person in each view (see Section 4.3). Rather than using the original segmented foreground, the information in the voxel space can be projected back onto the original image. With this, the original segmentation can be enhanced to match the people more closely. This is done in the form of a mask. In a mask, each pixel coordinate is labeled as either belonging to a specific person or to the background.

Additionally, a Kalman filter is instantiated (see Section 4.4). The filter has a state representation including both the position and velocity of a person. It can account for probable future positions by applying the velocity and an estimated measurement- and model-noise to the last known location.

When an initial person model is created, it is compared to labeled objects in successive frames. If an object corresponds to a known person in both location and histogram description, the person model is updated with the new information. If there is no known person model available however, one of two things can happen. The first possibility is that an actual new person entered the scene, and concurrently a new person model is created.

A second possibility is that the voxel blob is actually a artifact resulting from the occlusion of two people (see Section 4.2), in which case the object should be discarded. Mostly the artifacts could be discarded using the projection in voxel space, where objects are clearer defined than in the original images.

In the ideal case then, people are properly detected. Their positions can be displayed in a visualisation of the voxel space ground plane (see Figure 4), as well as be projected back onto the original images for easy visual interpretation (see Figure 5.

4.1 Foreground Segmentation

Before any information about the targets can be processed, they must be detected in the frames. For this we used foreground segmentation based on a mixture of Gaussians model as described in [5]. The cameras that filmed the data sequences were stationary at all times. It follows then, that if a region in a frame does not change, it is probably part of the background, while a region of the image that does change is a foreground object.

This distinction between foreground and background is not definitive. Many parts of what is in fact background still change slightly and parts of a moving object might look the same for a few frames, for instance when the person is standing still. The result of the segmentation is therefore input for the trackers, but does not determine conclusively what is to be tracked. Once a target is being tracked, it does not need to be identified as foreground anymore.

The foreground segmentation delivers binary images, as can be seen in Figure 2. These three images, one per camera, are then enhanced with morphological opening and closing

to remove noise. The resulting blobs are used as masks to project the pixels of the original frames into a three dimensional voxel space.

Figure 2: Segmented foreground of an image, depicting the moving objects.

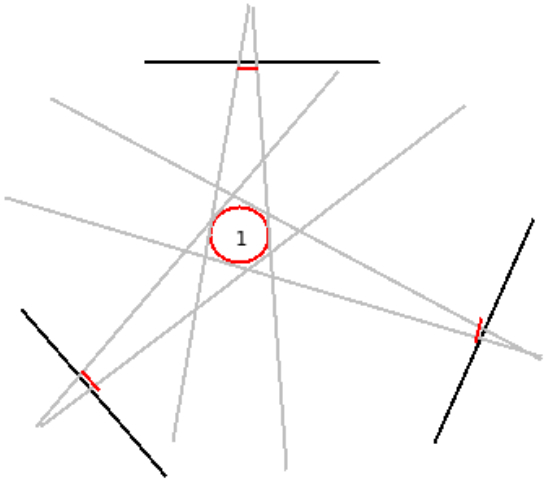
4.2 Voxel Space

To make optimal use of the three cameras a three dimensional voxel space is constructed for each timestep. The calibration information of the cameras is known. This information consists of the matrices that describe the transformation of camera coordinates to world coordinates.

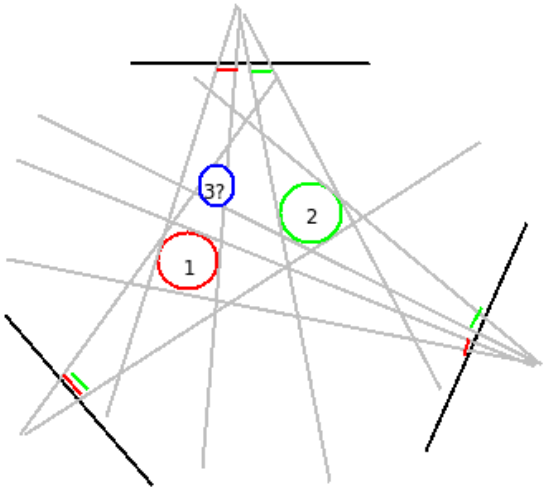
By masking the original frames with the binary blobs of the foreground segmentation the appropriate pixels are selected. These pixels are then transformed into voxel space using the above mentioned transformation matrices. This is done for all appropriate pixels for each of the three frames of the timestep and results in a three dimensional grid filled with voxels.

At first each view projects a cone through the voxel space, *carving* out part of this space. Only voxels that are carved out by –or originated from– all three views are kept in the voxelspace, as is illustrated in Figure 3(a). As a side effect, only objects that are within all three views can be tracked.

A disadvantage of using a three dimensional reconstruction is the appearance of artifacts. If two people occlude each other in one view, they create one large segmented foreground blob instead of two separate ones. When the other two views carve out their projection, it is possible that an additional part of the voxelspace remains visible, even though it does not correspond with an actual person. An example of how these artifacts can come into existence is depicted in Figure 3(b).



(a) Only the area projected by all three cameras is assigned a label.



(b) Because the red and green person occlude each other in one camera, a third blue spot is incorrectly carved as a person.

Figure 3: Carving voxel-blobs, shown in a top-down view of the groundplane. 3(a) shows a correct carving, while 3(b) shows the occurence of artifacts.

After carving out the three dimensional voxel blobs, they are projected down onto the ground plane by accumulating all voxel in the z-direction. In the ground plane (see Figure 4) the projected blobs are labeled as possible people. Since the blobs change position on the ground plane, the labeling order differs and cannot be trusted. This means that the labels can not be used to connected the projected blobs with their trackers. To disambiguate the projected blobs we developed a his-

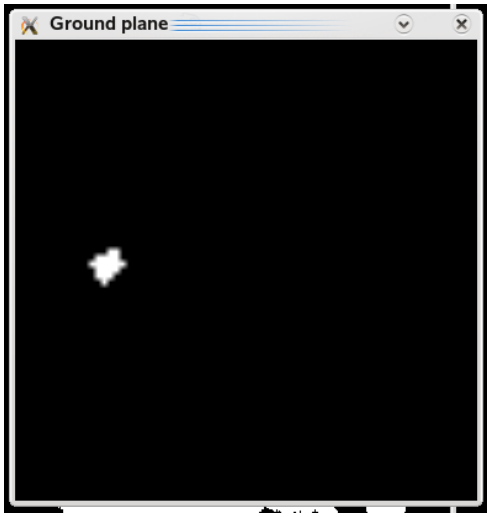


Figure 4: Top-down view of the voxel space ground plane, with one object.

togram based person model.

4.3 Person Model

A person model consists of a Kalman filter (discussed in Section 4.4) and three colour histograms, one for each view in each scene. The initial values of a person model are based on the first observation of that object in voxel space.

To have at least some colour constancy –to correct for the different lighting conditions in consecutive frames– the image is first converted to `normalised RGB`. Each `R` and `G` value is divided by the intensity (see Equation 1, with the added benefit that b is then always defined by $1-(r+g)$ and can be ignored. This decreases the dimensionality of the histograms by one. Next the histograms are normalised to correct for size differences in the different views.

$$r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B} \qquad (1)$$

To check if a detected object in the current scene is a person that we have already seen before, we compare the model of the detected object with the person models that already have been detected in the previous scenes. Two similarity measures are defined for the colour histograms and the location in the two dimensional ground plane respectively. Both similarity measures are in fact distance measures, ergo a smaller score equals a higher similarity. We use the Euclidean distance as measure to compare the positions of the detected object o and the existing person models m (see Equation 2.

$$e = \sqrt{(m_x - o_x)^2 + (m_y - o_y)^2} \qquad (2)$$

A Bhattacharyya distance is the basis for calculating the histogram similarity. This distance measures the similarity between two probability distributions p and q as in Equation 3. In our case, p and q are two normalised histograms –discrete probability distributions– with X the different bins.

$$b_v = -\ln\left(\sum_{x \in X} \sqrt{p(x) \cdot q(x)}\right) \qquad (3)$$

For each view $v \in V$, i.e. for each histogram pair, we calculate the Bhattacharyya distance, after which all three distances are summed to one final histrogram similarity score. Using Equation 2 and 3 we define a combined distance measure d as Equation 4.

$$d = e \cdot \sum_{v \in V} b_v \qquad (4)$$

By keeping the views separate when doing a histogram comparison –rather than combining the histograms and than doing one comparison– we acknowledge the fact that the histograms might differ under the lighting conditions of the different cameras. But more important, it also defines a person from different angles, as a collection of two dimensional viewpoints.[2] This knowledge can be used for improved artifact detection, as mentioned in Section 8.

4.4 Tracking

Apart from the general tracking problem discussed here, a Kalman filter is used to track the positions in the voxelspace ground plane. All of the elements within a Kalman filter are defined as matrices, which carry information about the tracked object and how to track it. In essence, it is able to project current location

information to a future timestep, and then correct it with an actual observation, resulting in a probabilistic estimate of the true location. The Kalman filter is instantiated with a movement model, an initial object position and velocity, and an estimate of the noise of both the observations and model.[3]

The movement model includes both the position and velocity of an object, that consist of two values each since we are tracking the two dimensional groundplane coordinates. This model can be interpreted as a constant velocity model, because acceleration is not explicitly represented meaning the velocity does not change.[1] For slowly moving objects this is not a real problem, because the velocity can adapt gradually if it changes at all. Sequences with people running and stopping abruptly might require explicit modelling of the acceleration.

Initialisation of the position can be done easily by taking the first observed location. The velocity is initialised as zero because there is no known direction and speed of movement known yet. It could be useful to instantiate a tracker after a few frames instead of immediately, creating a rough velocity estimate based on the initial positions; however, this was not evaluated.

The noise present in the model and observations are assumed to be normally distributed, given that the cameras are calibrated and of reasonable quality. The matrices representing the noise are therefore instantiated as the identity matrix.[2]

A specifically useful feature of the filter is its possibility to represent non-movement. For example, when a person stands still for the duration a few frames, the segmentation will start to regard it as background. When this person resumes movement, the detected blob can be matched to a person using its last known location. Presently the encompassing system does not remove a person from the list of detected people. A possible solution is to delete a person from the scene if there has not been any update to the person model for a specified period of time. This could be extended by defining 'exit points' in the voxel space –camera borders or stairs et cetera within an image– where this deletion can happen. As long as a person is still kept as present in the scene, but there is no accompanying voxel blob detected, its position according to the tracker can be supplied as a 'last known location'.

[1] This means, the effect of the velocity as defined in the model does not change. The actual estimated velocity in the state vector of a tracked person does change however.

[2] For our purposes, the covariance matrix is presumed to be well represented by an identity matrix.

5 Evaluation

Because the cameras are calibrated, a real-world coordinate system is available from which the true positions of the people can be derived. To see whether the tracking system is able to correctly retrieve the positions of the people therein, a ground truth set of people was also created. In practice, this means defining the positions of the people by hand, for each view in each frame. Since all tracking is done on the groundplane, the location of a person is defined as the position between the feet of this person, projected into the voxel space. Manually picking such a position introduces some ambiguity and noise, but within reasonable limits.

The main evaluative measure is the Euclidian distance between a ground truth position and the position of the closest person. For each person in each scene, this distance is measured with two different settings: once using only the locational person matching e (Equation 2) and once extended with the histogram comparison e (Equation 4).

The mean error and standard deviation for the first person in scene one is shown in Table 1.

6 Results

We expected both measures to perform equal on the first sequence, because there is only one object to be tracked, with a difference between the two in the more difficult sequences with multiple people.

A visual comparison of back projected com-

\error scene\	distance measure	mean error person 1	standard deviation person 1
scene 1	e	1.3	0.4
	d	1.3	0.4

Table 1: Mean localisation error (over a sequence) for one person. Measured are the euclidian distance e and the combined distance d.

puted groundplane positions (see Figure 5) shows that an average error of around 1.5 falls within the limits of a useful tracker. In principle the system allows for displaying visual aids to assist human observers.

Figure 5: A person is tracked, with its location visualised by a cross marked at its feet.

The histograms did not improve the results enough to also have a good tracking in more occluded scenes, given a preliminary visual comparison. This means that the current histogram based person model does not suffice. Additional measures detecting blobs as artifacts should be implemented, either by using more elaborate histogram models, or by improving the Kalman filter.

7 Conclusion

We approached the problem of tracking people by creating an explicit three dimensional representation, based on multiple camera views. People are then identified by their location and colour histogram description. From the results it follows that this twofold approach is at least feasible for simple scenes where people are clearly distinguishable objects. By extrapolating to more difficult scenes it becomes clear that the approach should be refined and extended, if general robustness is to be achieved. There is however merit in continuing in this direction. To do so, several possible additions are given in Section 8.

8 Future Research

Several additions to the current system can be thought of. First, the person model is very generic. Currently the histograms depict an entire person. It is possible to segment such a histogram into several parts, such as the head, body, legs and shoes. If for example the lower part of a person were to be occluded in all views, the body could still be used to find a definitive match. Because we want to track only people, additional person descriptors that set people apart from other objects is also a viable direction of research, and is currently explored in related projects.

Second, the Kalman filter assumes a constant velocity model. For more hectic scenes this might not suffice, requiring the acceleration to be modelled as well. Additionally the noise is assumed to be normally distributed, which is not necessarily the case. Tuning both the movement model and the noise to match the data more closely would probably increase the performance. On the other hand this also makes the tracker less generic, suggesting evaluation on a different data set to compare the overall performance.

Third, the detection of artifacts is not solved in the general case. Within the approach described here, an additional layer in the person model comparison could introduce some logic, or probabilistic estimation, that classifies

a voxel blob as an artifact. This could trigger for instance if some criteria for artifact appearance are met, like heavy occlusion in the image views and close proximity in the voxel space.

References

[1] S. M. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *ECCV*, pages IV: 133–146, 2006.

[2] M. J. Quinn, T. Kuo, and B. S. Manjunath. A lightweight multiview tracked person descriptor for camera sensor networks. In *ICIP*, pages 1976–1979, 2008.

[3] G. Welch and G. Bishop. An introduction to the kalman filter. In *Technical Report*, 1995.

[4] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila. CASSANDRA: audio-video sensor fusion for aggression detection. In *i-LIDS: Bag and vehicle detection challenge*, pages 200–205, 2007.

[5] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, May 2006.