PEOPLE COUNTING WITH STEREO CAMERAS
Two Template-based Solutions

Tim van Oosterhout[1], Ben Kröse[1,2] and Gwenn Englebienne[2]
[1]*University of Applied Sciences of Amsterdam, Amsterdam, The Netherlands*
[2]*University of Amsterdam, Amsterdam, The Netherlands*

Keywords: Stereo Camera, People Counting, Template Projection.

Abstract: People counting is a challenging task with many applications. We propose a method with a fixed stereo camera that is based on projecting a template onto the depth image. The method was tested on a challenging outdoor dataset with good results and runs in real time.

1 INTRODUCTION

For events such as large-scale concerts, fairs and outdoor art festivals, it is important to know the number of people attending. This number can be inferred if people are counted at all entrances to the area. Manual counting is very accurate but is costly since humans can only perform the counting task for a limited time. Mechanical solutions such as shaft devices restrict the throughput and placement may be impractical. Infrared sensing devices are only applicable with one person at a time in the passage. Therefore, cameras have been suggested for automatic counting of people.

Figure 1: Example frame from the (stereo) camera.

This paper focuses on counting of people for an outdoor event. We developed and compared two template-based methods for people detection that are used for counting using stereo cameras. Our method has the following advantages: (1) It is robust to changes in lighting. (2) It incorporates prior knowledge about locations where people may walk.

2 RELATED WORK

When using static cameras usually a model of the background is learned to classify pixels as foreground or background. A noise-cleaning step is performed to eliminate too small or short-lived regions. Connected components of pixels can then be found, resulting in foreground "blobs". However, a single person may give rise to multiple blobs, and parts of multiple people can be combined into a single blob because of visual overlap. In (Bahadori et al., 2007) an approach is described where the foreground pixels are mapped to the floor plane where the segmentation is done.

Template-based methods typically model the foreground but not the background. Examples are the Viola-Jones face detector (Viola and Jones, 2001), or the use of edge templates for pedestrian detection (Gavrila, 2000). They have been applied successfully to a variety of situations, including the tracking of rigid objects (Nguyen et al., 2001) and segmenting and tracking humans in crowded scenes (Zhao et al., 2007). However, such methods typically require very large training sets (Gavrila, 2000), adapting the templates to the test data over time (Nguyen et al., 2001) or extra parameters which need to be optimized.

Our methods takes the best of both worlds. As we will discuss in section 3, we model the background

van Oosterhout T., Kröse B. and Englebienne G..
PEOPLE COUNTING WITH STEREO CAMERAS - Two Template-based Solutions.
DOI: 10.5220/0003865304040408
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 404-408
ISBN: 978-989-8565-04-4

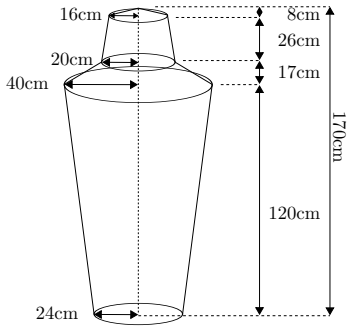


Figure 2: Illustration and dimensions of the template used. Note that the total width of 80cm might seem large but is needed to accommodate foreground fattening.

and additionally use the shape of the foreground.

3 METHODS

Our approach uses a stereo camera that looks straight down and has the volume delimited by the ground area and about 2 meters above it in view, so that people moving through the passage will be fully visible while they do so. Figure 1 gives a typical frame.

We define a 2.5D range image $\mathbf{r}$ for which the pixel values $\mathbf{r}_i$ are zero on the floor and positive towards the camera. In the first method a 2D foreground region is extracted from the range image onto which 3D shapes are projected. This method is based on (Englebienne and Kröse, 2010), but now using range data instead of color images. In the second method 3D shapes are used to reconstruct the 2.5D range image, such that a similarity measure is maximized.

For both methods we use a simplified 3D model of a human being of average size. The model consists of several stacked cone segments, which is illustrated in figure 2. The advantage of not including limbs in the model is that it makes the model rotationally invariant in the horizontal plane and forgoes the multitude of poses people could possibly assume at the cost of leaving small foreground sections unaccounted for.

Using the camera's intrinsic and extrinsic parameters we can project the template at any position. For both methods we generate hypotheses about the number and locations of people compare the fit under projection. By scanning the passage at discrete intervals we can find and localize any number of people.

3.1 Method 1: Generative Model

We consider the observation vector $\mathbf{r}$, containing the range image data. Each pixel $\mathbf{r}_i$ is assigned a probability that it is foreground $p_f(\mathbf{r}_i)$ using a dynamic statistical background model. We denote the vector of pixelwise foreground probabilities as $\mathbf{p}_f(\mathbf{r})$. Similarly, we use $p_b(\mathbf{r}_i)$ and $\mathbf{p}_b(\mathbf{r})$ to denote the background probabilities. We further introduce $\mathbf{l}$ to indicate the location of a person and $\mathcal{L} = \{\mathbf{l}_1 \dots \mathbf{l}_n\}$ to indicate a set of locations corresponding to multiple people.

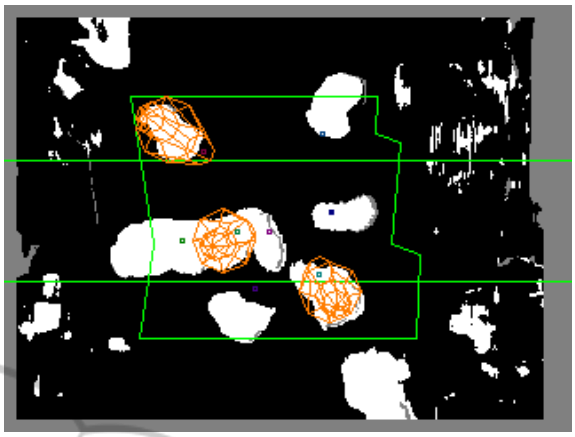


Figure 3: Three templates have been placed manually for illustration. The middle template is in between two people but seems to match well because only the silhouette is preserved. The range image (figure 4(a)) provides more detail.

The quantity we are interested in is the posterior probability of the people given an observed image, $p(\mathcal{L}|\mathbf{r})$. We compute this using Bayes' theorem,

$$p(\mathcal{L}|\mathbf{r}) = \frac{p(\mathbf{r},\mathcal{L})}{\sum_{\mathcal{L}} p(\mathbf{r},\mathcal{L})}\,, \qquad (1)$$

where $p(\mathbf{r},\mathcal{L})$ is the joint probability of the observed image and the locations of the visible people. We detail how this quantity is computed below.

3.1.1 Modelling People's Locations

In (Englebienne and Kröse, 2010) a "mask" vector $\mathbf{m}$ is defined, with the same dimensionality D as the observed frame, which is a vector whose elements are one for the components of foreground pixels and zero otherwise. This allows us to express the probability of an image, given a mask, as

$$p(\mathbf{r}|\mathbf{m}) = \mathbf{m}\cdot\mathbf{p}_f(\mathbf{r}) + (\mathbf{1}-\mathbf{m})\cdot\mathbf{p}_b(\mathbf{r}) \qquad (2)$$

where $\cdot$ indicates the elementwise product and $\mathbf{1}$ is a D-dimensional vector containing all ones.

We can project the 3D template of a person in any location on the ground plane. This is illustrated in figure 3. We can compute which pixels would belong to the object resulting in a mask vector $\mathbf{m}$. Because of our approximations, the mask $\mathbf{m}_\mathbf{l}$ depends only on the position, $\mathbf{l} = (x,y)$, of the person on the ground plane. When more than one person is considered, we take

the union of the masks created for each person. The resulting mask contains all pixels that fall inside the contours of templates located at the set of locations $\mathcal{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_n\}$, and is denoted as $\mathbf{m}_\mathcal{L}$.

We can now express the complete likelihood, ie. the joint probability of an observed image $\mathbf{r}$ *and* that a number of people should be present in locations $\mathcal{L}$:

$$\begin{aligned} p(\mathbf{r}, \mathcal{L}) &= p(\mathcal{L})\, p(\mathbf{r}|\mathcal{L}) \\ &= p(\mathcal{L})\, [\mathbf{m}_\mathcal{L} \cdot \mathbf{p}_f(\mathbf{r}) + (\mathbf{1} - \mathbf{m}_\mathcal{L}) \cdot \mathbf{p}_b(\mathbf{r})] \end{aligned} \quad (3)$$

Here, $p(\mathcal{L})$ indicates the prior probability that there should be n people in the set, and that these people are located in the locations $\mathcal{L}$. In (Englebienne and Kröse, 2010) it is explained that this probability can be factorized in components $p(|\mathcal{L}|)$, given the priors on the number of persons $p(\mathbf{l}_i)$, the location and the distance between persons$p(D(\mathbf{l}_i, \mathbf{l}_j))$. In our work the prior on location was set manually to represent the regions in which people can appear. In figure 3 this area S is delineated in green. The probability of locations outside S is 0 and uniform otherwise.

3.1.2 Inferring the Position of Multiple People

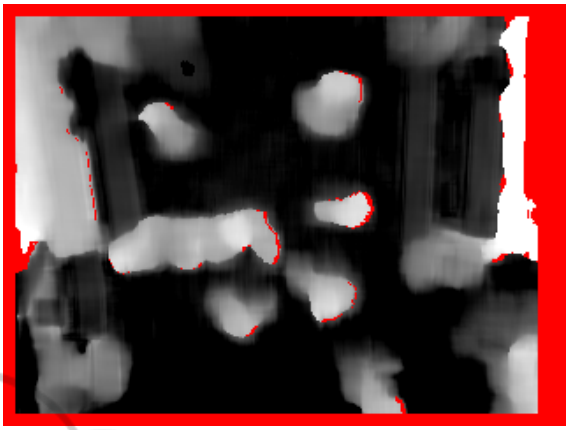
We can now infer the probability of a set of person locations $\mathcal{L}$ using Bayes' rule. However, even a low-resolution 320×240 pixel image yields 76800^m combinations of locations for m people. In (Englebienne and Kröse, 2010) it was shown that the search space can be reduced and that a fast greedy solution can be applied, similar to the one suggested by (Williams and Titsias, 2004).

1. Compute $p(\mathcal{L} = \emptyset)$, ie. no people.
2. Compute and store $p(\mathbf{l}_i)$ for all $\mathbf{l}_i \in S$.
3. Find the most likely position of the first person.
4. If this position improves the likelihood, add it to $\mathcal{L}$, otherwise exit the algorithm.
5. Find the next most likely position. Go to 4.

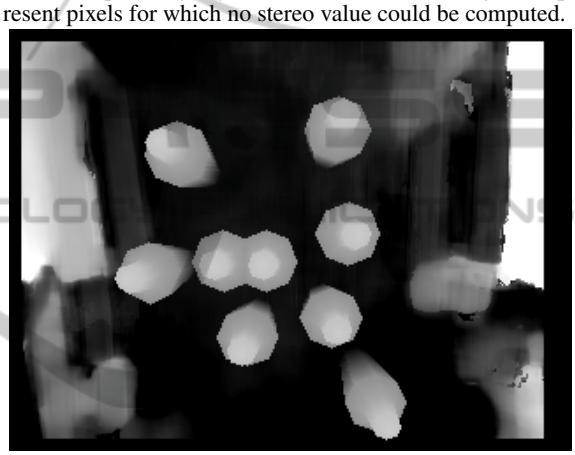
This algorithm results in the most likely number and locations of the detected persons.

3.2 Method 2: Reconstruction of Range Image

Again we consider a vector $\mathbf{r}$ containing range data. We construct a range image $\mathbf{h}$ and maximize the similarity between $\mathbf{r}$ and $\mathbf{h}$. The reconstructed image $\mathbf{h}$ is a function of $\mathcal{L}$ and the objective is to find the $\mathcal{L}$ that minimizes the distance $D(\mathbf{r}, \mathbf{h}(\mathcal{L}))$.



(a) The depth image by the stereo camera. Red regions represent pixels for which no stereo value could be computed.



(b) Adding 2.5D templates to the background to minimise the difference between the range image and the reconstruction.

Figure 4: Range image reconstruction.

3.2.1 Modelling People's Locations

We construct an image $\mathbf{h}(\mathcal{L})$ with range values for each pixel. We use the same 3D model as in method 1, tessellate it into triangles and render it on a new image $\mathbf{h}$. Self-occlusion and other occlusion handling is easy since the calculated values represent distances to the camera. An example is shown in figure 4.

The method can search for any number of people by computing the union of templates. We can then compute the difference between $\mathbf{h}$ and the range image $\mathbf{r}$ as such:

$$D(\mathbf{r}, \mathbf{h}(\mathcal{L})) = \sum_{i=1,D} (r_i - h_i)^2 \quad (4)$$

3.2.2 Inferring the Position of Multiple People

We use the same greedy approach as we used in method 1. To account for the fact that only one person

can occupy a physical space, an interpersonal distance s is defined that specifies how close people can be to one another, and this is used in the selection of locations l.

1. Compute the $\mathbf{h}(\emptyset)$, ie., no people.
2. Compute and store $D(\mathbf{r},\mathbf{h}(\mathcal{L}+l))$ for each location $l \in S$.
3. Find the location m that reduces the distance most: $\forall l \in S : D(\mathbf{r},\mathbf{h}(\mathcal{L}+m)) \leq D(\mathbf{r},\mathbf{h}(\mathcal{L}+l))$
4. If $D(\mathbf{r},\mathbf{h}(\mathcal{L}+m)) < D(\mathbf{r},\mathbf{h}(\mathcal{L}))$, add m to $\mathcal{L}$ and remove locations within the interpersonal distance s from consideration, otherwise stop.
5. Find the next best position m. Go to 4.

This algorithm results in the number and locations of the persons such that the distance between the range image and the reconstructed image is minimal.

3.3 Tracking

The output for each method is the number and locations of people present in a single video frame. To establish whether people are entering or leaving the area of interest these individual observations must be linked over time. We use a basic tracker that predicts people's next locations and smooths the data using a Kalman filter per person.

4 EXPERIMENTS

The methods were tested on an especially challenging dataset. The data was recorded in cooperation with a day long outdoor music festival. Recording started when the area was opened to visitors and ended after the last visitor left. In total the video covered about 11 hours and contained over 498.000 frames. The proposed methods' counts are compared against a manually created ground truth and additionally against a simpler method. Using the manual figures, a busy (Fig. 1) and quiet sequence were selected for comparison, in addition to a sequence in bad lighting condition (Fig. 5). The details are described next.

4.1 Alternative Method

The described methods are compared against a naïve method that works on the same range image. This method creates a foreground mask that depends directly on the values in the depth map as such:

$$\mathbf{f}_i = \begin{cases} 1 & \text{, if } g - \mathbf{r}_i > c \\ 0 & \text{, otherwise} \end{cases} \quad (5)$$

Figure 5: At the end of the sequence the lighting conditions have changed dramatically. Despite this the stereo matching algorithm is still able to generate decent range images.

with g representing ground level and c a cut-off value chosen at waist height. In this intersection people will appear as islands. These are then detected using a connected component algorithm and tracked.

4.2 Performance

The naïve method ran while recording in real-time at the full capture rate which was set at 25 fps. The generative and reconstructive methods used the recording and did stereo processing as well as their own computation. They ran single threaded on an Intel Core 2 Duo 8400. The reconstruction method was slowest achieving frame rates varying from 20 fps to 6 fps depending on the amount of people visible.

4.3 Results

Results were collected for people entering and exiting separately. People turning around in view were discarded, as were tracks for which either the starting or ending point could not be established. Table 1 lists the number of people reported per method and per sequence. To better compare the individual methods table 2 lists the counting error relative to the ground truth.

5 DISCUSSION

A first striking result is that all methods seem to undercount. This can be easily explained since the method specifies a number of constraints on each track to be counted at all. Overcounting seems harder

Table 1: Numeric results per method.

	Manual		*Height Threshold*		*Generative model*		*Reconstruction model*	
	In	Out	In	Out	In	Out	In	Out
Entire video	14276	10934	4871	3688	10046	6741	11425	8789
Busy	779	1124	109	230	331	461	557	719
Quiet	417	43	273	23	435	56	446	46
Night	22	597	29	237	15	482	45	549

Table 2: Error percentages per method.

	Height Threshold		*Generative model*		*Reconstruction model*	
	In	Out	In	Out	In	Out
Entire video	-65.9%	-66.3%	-29.6%	-38.3%	-20.0%	-19.6%
Busy	-86.0%	-79.5%	-57.5%	-59.0%	-28.5%	-36.0%
Quiet	-34.5%	-46.5%	4.3%	30.2%	7.0%	7.0%
Night	31.8%	-60.3%	-31.8%	-19.3%	104.5%	-8.0%

to explain but is seen when people bring in baby carriages or when children carry helium filled balloons.

Undercounting is most prominent in the naïve method. The reason is that people do not produce separate blobs when the cut-off is applied. Raising the cut-off height will not solve this problem until it is raised to above shoulder height, but at then short people will be overlooked. Moreover, with foreground fattening (Scharstein and Szeliski, 2002) people's blobs may merge even at head height.

Another remarkable result is the over 100% error margin of people entering in the night sequence for the reconstruction method. The only people entering in that sequence are personnell bringing in objects such as trash bins as illustrated in figure 5 which match the template close enough.

6 CONCLUSIONS

We have shown two novel template based people counting and localisation methods that work with range images by stereo cameras. The methods were tested on a dataset that featured many people in view at once, a changing background and big changes in lighting conditions. We found that as the methods make more use of the available information from the range image the detection and tracking results improve. The methods run in real-time, making them suitable for live deployment.

ACKNOWLEDGEMENTS

The research reported in this paper was supported by the Foundation Innovation Alliance (SIA - Stichting Innovatie Alliantie) with funding from the Dutch Ministry of Education, Culture and Science (OCW), in the framework of the 'Mens voor de Lens' project.

REFERENCES

Bahadori, S., Iocchi, L., Leone, G., Nardi, D., and Scozzafava, L. (2007). Real-time people localization and tracking through fixed stereo vision. *Applied Intelligence*, 26(2):83–97.

Englebienne, G. and Kröse, B. (2010). Fast bayesian people detection. In *proceedings of the 22nd benelux AI conference (BNAIC 2010)*.

Gavrila, D. (2000). Pedestrian detection from a moving vehicle. *Computer VisionECCV 2000*, pages 37–49.

Nguyen, H., Worring, M., and Van Den Boomgaard, R. (2001). Occlusion robust adaptive template tracking. *Computer Vision, IEEE International Conference on*, 1:678.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:511.

Williams, C. and Titsias, M. (2004). Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Computation*, 16(5):10391062.

Zhao, T., Nevatia, R., and Wu, B. (2007). Segmentation and tracking of multiple humans in crowded environments. *IEEE transactions on pattern analysis and machine intelligence*, pages 1198–1211.