4D Light-field Sensing System for People Counting

Guangqi Hou, Chi Zhang, Yunlong Wang, and Zhenan Sun

Center for Research on Intelligent Perception and Computing,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Institute of Automation, University Of Science and Technology Of China

ABSTRACT

Counting the number of people is still an important task in social security applications, and a few methods based on video surveillance have been proposed in recent years. In this paper, we design a novel optical sensing system to directly acquire the depth map of the scene from one light-field camera. The light-field sensing system can count the number of people crossing the passageway, and record the direction and intensity of rays at a snapshot without any assistant light devices. Depth maps are extracted from the raw light-ray sensing data. Our smart sensing system is equipped with a passive imaging sensor, which is able to naturally discern the depth difference between the head and shoulders for each person. Then a human model is built. Through detecting the human model from light-field images, the number of people passing the scene can be counted rapidly. We verify the feasibility of the sensing system as well as the accuracy by capturing real-world scenes passing single and multiple people under natural illumination.

Keywords: Light field photography, RGB-D sensor, people counting

1. INTRODUCTION

The people counting system based on vision sensors becomes a practical but challenging task in intelligent surveillance. Typically, such a system obtains the data of the people flow by zenithal (overhead) mounted cameras at the entrances or exits of public areas. The people flow statistical data are extremely useful for public security as well as business intelligence.

A people counting system is often implemented by tracking the heads passing a virtual line set in the field of view of a conventional camera. The major limitations in such a scheme are changes of illumination, shadows, compound objects for crowded scenes. To mitigate the impact of these problems, recent literature approaches resort to novel RGB-D sensors, e.g. the Microsoft Kinect sensor.

Since the Kinect sensor can output regular RGB images as well as depth maps simultaneously in real time, it has become the most popular RGB-D sensor applied in compute vision tasks. Such a multi-function vision sensor shows a

![Figure 1. Demonstration of the proposed people counting system. (a) is the demonstration of overall system. (b) is the raw data captured by a light field camera.](image_url)
promising prospect for objects detection, tracking, segmentation etc. The Kinect is equipped with an active depth sensor, which detects the range map via emitting active radiations and reconstructs high fidelity depth maps based on triangulation measurement in Kinect I and Time of Flight (TOF) in Kinect II. However, such active sensors are vulnerable to coherent radiations possibly existing in the scene, e.g. solar radiations of outdoor environments or mutual interference between the collaborative Kinect sensors. Furthermore, there are some disparities between the intensity images and the depth maps offered by Kinect sensors. Such disparities impede the application of Kinect to a few specific scenarios, i.e. detection, segmentation, 3D reconstruction, where RGB images and depth maps must be registered exactly.

Light field cameras are able to render exactly registered RGB images and depth maps without emitting any radiation, which is considered as a type of ideal vision sensor. In this paper, a people counting system based on a zenithal mounted light field camera is introduced, as shown in Figure[1]. Then we test the proposed system in practical surveillance scenes to quantitatively evaluate its performance.

The contributions of this paper are as follows:

(1) introducing a novel people counting system based on light field photography;

(2) building a novel light field camera oriented to video surveillance tasks;

(3) proposing a people counting scheme combining both the intensity image and the estimated depth map from a light field camera.

2. BACKGROUND

A variety of people counting schemes using zenithal mounted cameras have been introduced. Most of these schemes are designed based on a similar architecture that involves the following steps: (1) separating the foreground pixels from the static part of the image; (2) detecting every person in the scene with the prior separation of the foreground mask; (3) counting the detected people according to the requests of specific tasks. Both traditional cameras and novel RGB-D sensors have been explored for solving the problems in all of the three mentioned steps.

At the beginning, a number of approaches for people counting are designed to process the video streams recoded by traditional cameras.[1–5] These approaches have major drawbacks in crowded scenes due to variations in lighting, shadows, compound objects, etc. To weaken the impact of these problems, new approaches resort to 3D analysis of the scene.[6–8] These approaches began to use depth sensors, and thus are possible to solve people counting issues in crowded scenes. However, these approaches make use of stereo cameras and the ill-posed stereo matching to obtain the depth map, which leads to impractical solutions.

After the advent of the Microsoft Kinect that provides low-cost range data through an active depth sensor, researchers introduce new solutions of exploring range data offered by the Kinect. Specifically, Zhang et al. proposed to detect heads by searching for local minima within the depth map using watershed.[9] Vera et al. proposed a scheme that trains a SVM classifier to detect heads of people by applying Histograms of Oriented Gradients (HOG) descriptor.[10]

Because it is hard to register the disparities between the RGB image and the depth map in Kinect sensor, current methods tend to use single depth map to solve people counting problems.[9,10] However, those solutions ignore the texture information of the scene and are easily misled by human-like objects. Moreover, the Kinect sensor is apt to be interfered by unexpected radiations in the scene as discussed above. Thus the implementation of these solutions is limited.

Light field cameras were firstly proposed by Gabriel Lippmann in 1908. In 1992, Adelson and Wang produced the first light field camera by inserting an array of microlenses between the main lens and the photo sensor.[11] This camera can significantly simplify corresponding problems in stereo matching. Ng introduced the first hand-held light field camera via directly installing a microlens array in front of the digital image sensor, which eventually make the lenselet-based light field camera to be the most practical light field sensor applied in computer vision tasks.[12]

Light field cameras can provide the densely accurate depth map without executing ill-posed stereo matching. Adelson and Wang analyzed the light distribution of objects placed at different depth on a light field sensor, and propose a depth estimation method.[13] Wanner et al. computed local slope estimates on EPI for each pixel in each view using the structure tensor, and integrated these local estimates into a high quality depth map based on total variation regularization.[14] Tao et al. proposed a fusion framework with defocus and correspondence cues using the full 4D EPI, and the globally smooth
Figure 2. The proposed framework for counting the number of people.

depth map is refined through MRFs with local estimation cues. Then they refine the depth map by exploring lighting and shading cues with the reference of the raw depth map.

Combination of the intensity images and the depth maps rendered by a light field camera shows a promising prospect for boosting the performance of our people counting system, which encourages us to apply the light field camera to solve relevant issues.

3. METHODOLOGY

In this section we explicitly introduce the proposed framework for people counting from the zenithal mounted light field camera. Firstly, images captured by the light field camera are decoded into a set of sub-aperture images. Then depth estimation is invoked to get the high quality depth map. The central-view all-in-focus sub-aperture image and its corresponding depth map are coupled as the input data.

Secondly, we leverage the sequence of depth maps to perform foreground segmentation via Ostu’s threshold. With the foreground binary segmentation, we mask the pairwise sample separately to extract ROIs of head candidates. In ROIs of the disparity map, local maxima are selected and grouped which imply potential head candidates.

Finally, the potential head candidates can be used as prior for searching the head position in the intensity images. A Mean Shift algorithm is proposed to iteratively search the heads positions by setting the results acquired from the depth map as the initial heads positions. The predicted head positions are reinforced by merging the intensity information, whose accuracy is improved dramatically.

The overview of the proposed framework is shown in Figure 2.

3.1 Preprocessing and depth estimation

The lenselet light field camera is able to record 4D light fields via inserting a microlens array in front of its image sensor. The lights crossing the main aperture deposit on pixels according to their spatial positions and propagation directions. The raw light field image comprises a large number of lenselet sub-images, as shown in Figure 1(b). A decoding process should be operated before image analysis, since a 4D function tends to represent the light field,

\[ L(u, v, x, y) \]  

where \( u, v \) are indexes of angular dimensions, and \( x, y \) are indexes of spatial dimensions. We implement a decoding algorithm that builds a map from the 2D raw lenselet image to the 4D light field representation under the rules of geometrical optics.
Intuitively, the 4D light field can be viewed as an array of sub-aperture images that are formed by gathering the pixels of the same position in the coordinates of each microlens, as shown in Figure 2. The sub-aperture images are equivalently captured by a pinhole camera array settled at the aperture plane. There are a variety of disparities along the sub-aperture images, which show the cues of scene of depth.

The disparities can be estimated by analyzing the slope of stripes in epipolar plain image (EPI). The EPI can be formed by slicing along $u, x$ with constant $v_c, y_c$.

$$S(u, x) = S_{v_c, y_c}(u, x) = L(u, v_c, x, y_c)$$  \hspace{1cm} (2)

We adopt Wanner’s algorithm for estimating the depth map from 4D light fields, which uses the structured tensor $J$ for describing the stripes,

$$J = \begin{bmatrix} G_{\sigma} \ast (S_xS_x) & G_{\sigma} \ast (S_xS_u) \\ G_{\sigma} \ast (S_xS_u) & G_{\sigma} \ast (S_uS_u) \end{bmatrix} = \begin{bmatrix} J_{xx} & J_{xu} \\ J_{xu} & J_{uu} \end{bmatrix}$$  \hspace{1cm} (3)

Where $S_x$ is the derivative of $S$ on $x$ direction. Thus, the depth $Z$ can be calculated by

$$Z = \frac{f}{\tan \varphi}$$  \hspace{1cm} (4)

where

$$\varphi = \frac{1}{2} \arctan \left( \frac{J_{uu} - J_{xx}}{2J_{xu}} \right)$$  \hspace{1cm} (5)

The raw depth map is noisy with some holes on textureless surfaces. The edge-preserved filter is applied to smooth the raw depth map according to its intensity image. A variety of edge-preserved filters are tested for smoothing the depth map, including the bilateral filter, the guided filter, the weighted least squared (WLS) filter, etc. And the WLS filter gets the best performance in the experiments. We also render an all-in-focus image by refining the center-view sub-aperture image on the basis of other sub-aperture images. Finally, an all-in-focus intensity image and a smoothing depth map are acquired for heads detection.

3.2 Heads positions from depth maps

It is hard to robustly locate heads positions by simply setting a fixed threshold, e.g., the empirical stature, since it may be confused for discerning the real heads form objects that have similar height in the scene. In addition, the depth map has an uncertain bias introduced by smoothing operations in depth estimation, which also degrades the performance of policy of a fixed threshold.

We compute rough heads positions via Otsu’s method, which is used to automatically perform clustering-based image thresholding. The algorithm assumes that the image contains two classes of pixels following the bi-modal histogram (foreground pixels and background pixels). Then it calculates the optimum threshold separating the two classes so that their combined spread (intra-class variance) is minimal, or equivalently (because the sum of pairwise squared distances is constant) so that their inter-class variance is maximal.

The Otsu’s method outputs a background mask of depth map, as shown in Figure 2. Then the local maximum points of the masked depth map can be considered as the potential heads positions. To estimate these local maximum points, we establish a Gaussian pyramid of the masked depth map and merge the local maximum points of each layer. The merged outputs are able to robustly predict the heads positions, although these positions have various degrees of deviation from their ground truth points, as shown in Figure 2. Such deviation can be corrected by combining the intensity information embedded in the all-in-focus image, which will be discussed in the next section.

3.3 Heads positions from intensity images

The potential heads positions from the depth map can be easily interfered by non-human objects, since the position is acquired via detecting the local maximum of the depth map. To enhance the robustness and accuracy of the proposed people counting system, the intensity information is applied to correct the bias of the predicted heads positions from the corresponding depth map. Mean shift is a procedure for locating the maxima of a density function given discrete data sampled from that function. Thus, a Mean-Shift algorithm is proposed to refine the predicted heads positions:
The correlation coefficient between the template and the scanning window is computed via

\[
I(i, j) = \begin{cases} 
    d(i, j), & \forall d(i, j) \geq T_d \\
    0, & \forall d(i, j) < T_d
\end{cases}
\] (6)

where \(T_d\) is the threshold; \(d\) is the chi-square coefficient defined as

\[
d_{(h_{\text{win}}, h_{\text{tem}})}(i, j) = \sum_{k=1}^{m-1} \frac{(h_{\text{win}}(k) - h_{\text{tem}}(k))^2}{(h_{\text{win}}(k) + h_{\text{tem}}(k))}
\] (7)

where \(h_{\text{win}}, h_{\text{tem}}\) are the histograms of the scanning window and the template respectively.

Then the mean shift vector can be calculated as

\[
\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} M_x/M \\ M_y/M \end{bmatrix}
\] (8)

where

\[
M_x = \sum_{x=x_0}^{m_x} \sum_{y=y_0}^{m_y} x \cdot I(x, y),
M_y = \sum_{x=x_0}^{m_x} \sum_{y=y_0}^{m_y} y \cdot I(x, y)
\] (9)

where \((x_0, y_0)\) is the initial position predicted from the depth map.

The algorithm iteratively updates \(x_0 = x_t, y_0 = y_t\), and repeats the calculation until \(x_t, y_t\) converges.

4. EXPERIMENTS

In this section, we evaluate the performance of the proposed people counting system. Both the qualitative and quantitative assessments are performed on two representative datasets.

We build a novel light field camera oriented to the application in surveillance systems. In this camera, the micro-lens array is settled in front of the image sensor, and the distance between the micro-lens array and the image sensor is equal to the focal length of the micro-lens. To improve both the spatial resolution and the angular resolution of this camera, we select a 11 Megapixels image sensor and design a high-precision micro-lens array. The resolution of our sampled light field is \(398 \times 265 \times 9 \times 9\). The light field camera equipped with a 25mm Nikon lens is zenithally mounted on a gantry with optical axis approximately vertical to the ground plane. The camera is connected to a PC via CamLink and is set to record images at 6fps.

We capture two datasets for testing our system, a single-passenger dataset, as shown in Figure 3, and a multi-passengers dataset, as shown in Figure 4. In the single-passenger dataset, a person is requested to walk forth and back towards the specified direction. In the multi-passengers dataset, up to three persons walk around without any regulation.
The single-passenger dataset can simulate the scene that a person go through a constrained passage, as the ticket wicket, the narrow entrance, etc. In the experiment, the accuracy of the single-passenger dataset is approximately equal to 100%, since the constrained scene is quite easy for the proposed scheme. Thus, we concentrate on evaluating the proposed scheme on the multi-passengers dataset.

We use two indexes to quantitatively evaluate the accuracy of the proposed algorithm: Mean Absolute Error (MAE) and Mean Relative Error (MRE). The MAE and MRE are defined as

$$ MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{c}_i - c_i|, \quad MRE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\hat{c}_i - c_i}{c_i} \right| $$

where $\hat{c}_i$ is the estimated number of people in the No.$i$ frame, $c_i$ is the ground truth. In addition, the detection error (DE) is also provided for evaluating the proposed scheme.

The quantitative result is tabulated in Table 1. Moreover, the qualitative results of heads detection are shown in Figure 5. From the experiments, it is reasonable to draw some conclusions,

(1) The light field camera is applicable as a surveillance sensor in people counting tasks. It can output exactly aligned depth maps and all-in-focus intensity images, which combine the advantages of both the texture and the shape information for localizing tasks, e.g. the head detection in this paper.

(2) The proposed people counting scheme can accurately output the number of people appeared in the constraint as well as unconstraint scene. Such results verify the feasibility and superiority of the light field sensing system.

(3) In addition, the proposed sensing system can precisely locate the heads of people, as shown in Figure 5, which shows a possibility that such a system can be easily extended to solve more complicated tasks.

### Table 1. Accuracy of people counting.

<table>
<thead>
<tr>
<th>Index</th>
<th>Multi-person dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MAE$</td>
<td>0.15</td>
</tr>
<tr>
<td>$MRE$</td>
<td>0.08</td>
</tr>
<tr>
<td>$DE_x$</td>
<td>8.45 (pixels)</td>
</tr>
<tr>
<td>$DE_y$</td>
<td>12.87 (pixels)</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

Counting the number of people is still an important task in social security applications and many methods based on video surveillance have been proposed in recent years. In this paper, we design a novel optical sensing system to directly acquire the depth map of the scene from a light-field camera. The light-field sensing system can count the people crossing the...
passageway and record the direction and intensity of rays in a snapshot without any assistant light devices. Depth maps are extracted from the raw light-ray sensing data. Our smart sensing system can distinguish the depth difference between the head and shoulders of each person naturally. Through detecting the heads in the light-field image, the number of passing people can be counted rapidly.

The proposed system can capture the passengers from two passageways. After decoding, depth maps are extracted from the 4D light-field data through structure tensor method. In the depth maps, there are disparities from the head to feet. Furthermore, the disparity between the head and shoulders is very robust and accurate. Through the experimental results, the feasibility of the sensing system is verified, and the experiments containing more people will be investigated for enhancing our system in the future.

REFERENCES


