Proceedings of the 2004 IEEE
Conference on Cybernetics and Intelligent Systems
Singapore, 1-3 December, 2004

ON PIXEL COUNT BASED CROWD DENSITY ESTIMATION FOR VISUAL SURVEILLANCE

Ruihua MA, Liyuan LI, Weimin HUANG, Qi TIAN
Institute for Infocomm Research (I2R), Singapore
{ruihua,lyli,wmhuang,tian}@i2r.a-star.edu.sg

ABSTRACT

Surveillance systems for public security are going beyond the conventional CCTV. A new generation of systems rely on image processing and computer vision techniques and deliver more ready-to-use information and provide assistance for early detection of unusual events. Crowd density is a useful source of information because unusual crowdedness is often related to unusual events. Previous works on crowd density estimation either ignore perspective distortion or perform the correction based on incorrect formulation. Also there is no investigation on whether the geometric correction derived for the ground plane can be applied to human objects standing upright to the plane. This paper derives the relation for geometric correction for the ground plane and proves formally that it can be directly applied to all the foreground pixels. We also propose a very efficient implementation because it is important for a real-time application. Finally a time-adaptive criterion for unusual crowdedness detection is described.

1. INTRODUCTION

Surveillance systems are going beyond conventional CCTVs which are limited to record video data and/or display continuous video to human operators. The new generation of surveillance systems, relying on image processing and computer vision techniques, can deliver more ready-to-use information and provide assistance for early detection of unusual events [1]. Crowd density is a useful source of information because unusual crowdedness is often related to unusual events.

In the past, there have been a few works on crowd density estimation, designed for subway station crowd flow estimation and management. The techniques used can be classified into two categories: (1) (foreground) pixel counting based and (2) feature-based. In pixel counting based methods [2, 6], background segmentation is first performed to extract the foreground, mainly made of moving people. Then crowd density is computed as a function of the number of foreground pixels; the function itself is obtained by learning. In feature based methods [3, 4, 5], features are computed for the *whole* image, and segmentation may or may not be necessary. In [3], texture features are used based on the observation that crowded areas exhibit textures and the higher the crowd density the stronger the texture features (contrast, homogeneity, energy, entropy). The same authors proposed the use of the Minkowski fractal dimension [4]. However, their experimental results show slightly inferior performance compared to texture features. In both [3] and [4], crowd density is divided into 5 levels. The three classifiers tested, statistical, neural network and curve fitting, give similar performance. We note that both methods assume implicitly that the background is untextured and the entirety of the scene is of interest. Lo *et al.* [5] provides a more detailed study. Foreground pixel count and texture features are found to correlate the best with the crowd density and used as input to a 4-layer feed-forward neural network.

No matter the technique adopted, the *perspective distortion* becomes a factor that must be properly taken into account. Perspective distortion here refers to the simple fact that far objects in the scene appear smaller than near ones in the image. Pixel counting based methods are substantially affected by this perspective distortion. Texture or fractal feature measures in feature-based methods are also directly influenced. In the works mentioned above, only [5] and [6] take into account perspective distortion explicitly and have performed what is called "geometric correction (GC)". Unfortunately, the formulation in [5] is incorrect. As to [6], although no full detail is given, one still can affirm it is incorrect. First, the relation is not linear; and second, the relation is not the same for both vertical and horizontal directions in the image. These are different from the simple linear relation which is the same for both image axes that we derived in this paper. One more problem with [5] and [6] is, the equations for GC – suppose they were correct – are derived and valid only for the ground plane. But they are directly applied to human beings standing upright to the ground without providing further explanation.

In this paper, we first conduct an analysis of geometric correction for the ground plane in the context of pixel counting. Then we show formally that the geometric correction can be directly applied to all the foreground pixels with scales derived for the ground plane. We also

0-7803-8643-4/04/$20.00 © 2004 IEEE

show how to implement pixel counting with geometric correction efficiently. The paper is organized as follows. In Section 2, the pixel counting method is briefly described. Section 3 is concerned with geometric correction. An outline of our system is given in Section 4. Also we will discuss shortly how to speed up pixel counting as well as how we define a time-adaptive criterion for unusual crowdedness detection. Finally we conclude in Section 5 with perspectives.

2. PIXEL COUNTING FOR CROWD DENSITY ESTIMATION

The pixel counting method relies on the foreground segmentation result. It relates the total foreground pixel count to the number of persons. With the ideal setting (i.e. top-view, orthographic projection), perfect segmentation, further with the assumption of equal size of persons, the relationship is strictly proportionality: $N_{persons} = a * N_{pixels}$.

In real applications, the camera is often placed sideway, i.e. at some height forming an angle with the ground plane. With moderate crowd density whereby occlusion is not important, the linear relationship still holds, but with a constant term accounting for various other factors:

$$N_{persons} = a * N_{pixels} + b \qquad (1)$$

It must be noted, however, that for the above equation to be valid, perspective distortion must be corrected such that all the foreground pixels are brought to the same scale. This is done through "geometric correction".

3. GEOMETRIC CORRECTION

Geometric correction (GC) is necessary to bring all the objects at different distances in a scene to the same scale. Suppose the ground is a plane, a simple relation for the ground plane can be derived. In previous works, such a relation is directly applied to human objects (not lying on the ground!) without giving proof about the feasibility. In this section, we first compute the GC scale for the ground plane and then discuss its validity for crowd density estimation.

3.1 Computing the Scale for the Ground Plane

In Figure 1, projective imaging is shown with a road scene as example. Here we assume that the vanishing line (horizon) is parallel to the image horizontal scan lines. The four points $\{P_i,\ i=1,2,3,4\}$ form a quadrilateral that corresponds to a rectangle (*eg*., a portion of a road). $\overline{P_1P_2} = \Delta x_1$ is chosen as the reference scale. Note that the actual location of P_1 and P_2 in the image are not important; only important is that all the objects are brought to a same scale. The extension of the parallel road borders lines intersects at the vanishing point P_V which itself lie on the vanishing line. From the triangular relation

$$\frac{\Delta x_1}{\Delta x_2} = \frac{\overline{P_V P_1}}{\overline{P_V P_3}} = \frac{\overline{P_V P_1'}}{\overline{P_V P_3'}} = \frac{y_1 - y_V}{y_3 - y_V}$$

we get

$$y_V = \frac{y_1 \Delta x_2 - y_3 \Delta x_1}{\Delta x_2 - \Delta x_1}$$

Thus the scale for any horizontal line Δx on the road with respect to Δx_1 can be computed by

$$s_{Horiz} = \frac{\Delta x_1}{\Delta x} = \frac{y_1 - y_V}{y - y_V} \qquad (2)$$

From the same triangular relation, the same vertical scale is derived:

$$s_{Verti} = \frac{\Delta x_1}{\Delta x} = \frac{y_1 - y_V}{y - y_V} \qquad (3)$$

The scale is only a function of the vertical location of the line y and we write $s = s_{Horiz} = s_{Verti}$.

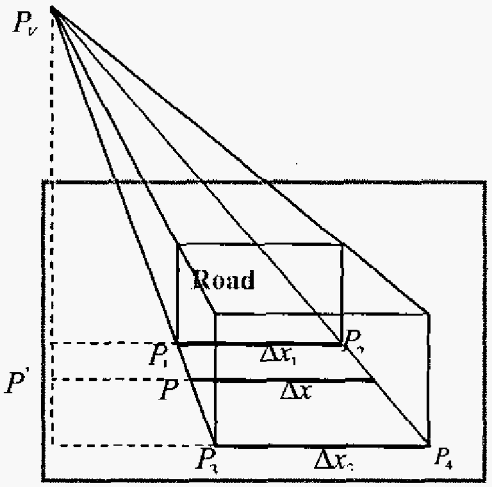

Figure 1 Geometric correction: compute the scale with respect to the reference line P_1P_2.

The ratios in (2) and (3) are applied in pixel counting, i.e., N_{pixels}. That is, a foreground pixel is counted as $s_{Total} = s_{Horiz} * s_{Verti} = s^2$, instead of 1. Thus Eq. (1) is now rewritten as

$$N_{persons} = a * \sum_{i \in I_{FG}} I_i(x,y) * s_{Total}(y) + b \qquad (4)$$

where I_{FG} represents the foreground image.

3.2 Applying the Scale

The scales just derived are valid to pixels on the ground plane. Our objects of interest are humans, walking or standing, which are vertical to the ground plane. In other words, the majority of the foreground pixels from a same person are not on the ground. Direct GC by Eqs. (2) and (3) is incorrect.

Under the condition that the ground plane and the image plane form roughly π/2 angle, the scales to apply for the body pixels of a person should be the same as that of his feet on the ground. Thus the right GC should find the scale of each person by his/her feet and apply the scale to the whole body. But this would require segmentation which is not a trivial task by itself. In fact if we could do that in an easy way we would be able to count persons directly! Previous works apply the affine transform for the ground plane directly, without considering the error induced or its viability. We will fill this gap.

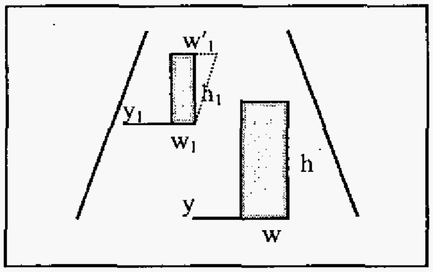

Figure 2 Geometric correction: error computation.

Refer to Figure 2 and Figure 1. We represent the human body using a rectangle. At the reference line y_l we have an average human body size $w_l \times h_l$. At the line y, the body size is $w \times h$, which is related to $w_l \times h_l$ by scaling defined in (2) and (3):

$$\begin{cases} w = w_1 / s \\ h = h_1 / s \end{cases}$$

On the other hand, we consider applying scaling to all foreground pixels only according to their vertical position in the image as done in previous works. While the feet level would be scaled to w_l, the height h scaled to h_1, the head level w_l would be scaled to w'_l, with the scale $s' = s\,(y\text{-}h)$:

$$w'_1 = s' * w = w_1 * \frac{s'}{s}$$

Referring to (2), we have

$$\begin{aligned} \frac{s'}{s} &= \frac{y - y_V}{y - h - y_V} \\ &= 1 + \frac{h}{y - h - y_V} \\ &= 1 + \frac{\frac{y - y_V}{y_1 - y_V} * h_1}{y - \frac{y - y_V}{y_1 - y_V} * h_1 - y_V} \\ &= 1 + \frac{h_1}{y_1 - h_1 - y_V} \end{aligned}$$

Thus the difference between w'_1 and w_1 is

$$w'_1 - w_1 = \frac{w_1 h_1}{y_1 - h_1 - y_V}$$

and the area of the extra triangle is

$$\Delta = \frac{1}{2}(w'_1 - w_1) * h_1 = \frac{w_1 w_1 h_1}{2(y_1 - h_1 - y_V)} \qquad (5)$$

which is a constant. Eq. (5) shows clearly that if the scale found for the ground planc is applied blindly to foreground pixels from human bodies, the difference between this and the correct size is a constant, regardless of the location of the human bodies in the image. This is to say that the linear relationship in Eq. (1) holds up to a scalar when blind scaling is applied to all the foreground pixels.

4. THE SYSTEM

We developed a system for crowd density estimation shown in Figure 3.

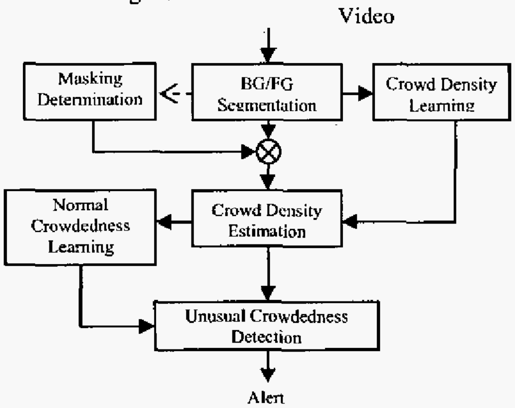

Figure 3 System diagram

Essentially, first, video images are segmented into foreground and background. Then pixel counting based crowd density estimation is performed, with scaling for geometric correction. Note that a mask representing region of interest is applied to foreground image to reduce effect in the segmentation. Finally the estimated crowd density is sent to unusual crowdedness detection, the output of which could be alert signal. In the following, we will discuss specific considerations in the system.

Foreground Segmentation This step is crucial to the performance of the system. The system has been developed with applications in outdoor public spaces in mind, whereby light condition varies a lot and there may be moving non-human objects such as trees. Thus we adopt a robust segmentation algorithm developed in house [7]. Figure 4 shows an example of the result.

Mask Determination In practice, all segmentation algorithms generate some spurious foreground regions. To reduce this effect in crowd density estimation, we apply a

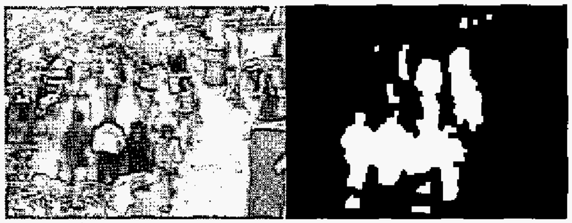

Figure 4 Example of segmentation result.

mask to the foreground image. The mask is determined once for all, either by designating a region of interest (ROI) manually or by the following foreground accumulation method we use:

$$I_{mask} = \sum^{N} I_{FG} > t, t = median(\sum^{N} I_{FG})$$

The result of the later approach, based on the same location as in Figure 4, is shown in Figure 5.

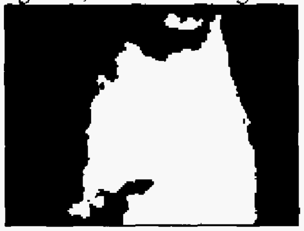

Figure 5 Mask by foreground accumulation.

Fast Implementation of Scaling There are two equivalent ways to perform pixel counting with GC. One is to apply the affine transform found previously and then call for Eq. (1). The other is to integrate GC in pixel counting, i.e., to scale each pixel in Eq. (1). We adopt the second method, which is computationally more efficient, as in Eq. (4).

Further, in Section 3 we have seen the scale is only a function of y, the vertical coordinate. We build a lookup table (LUT) is for $s(y)$ in Eq. (4). Thus pixel counting becomes again a simple addition even with scaling for GC. Thus this integrated GC and pixel counting approach suppresses the need of general image transform. Non-integer counting, aided by the LUT, adds only very slight computational cost.

Time Adaptive Criterion for Unusual Crowdedness Detection In practice, even for the same place, the criterion of unusual crowdedness may vary depending on the time. We group time slots into different time groups (TG); for each TG, we build crowd density histogram H_{TG} using recorded data over long periods. H_{TG} is then normalized, leading to H_{TG-N}. We call it normal crowdedness histogram. From this we derive the threshold for unusual crowdedness detection. We do this based on 95% trust level (i.e. number of persons covering 95% area under H_{TG-N}).

The system has given consistent results on various surveillance scenes. When the vanishing line is not horizontal in the image, it can be made so by rotating the image by the angle of the vanishing line. The vanishing line can be easily obtained using methods, say, as in [8].

5. CONCLUSION AND PERSPECTIVES

In this paper, we have discussed the pixel counting based method for crowd density estimation. In particular, we investigated the geometric correction to account for perspective distortion. Normally the derived transform for geometric correction is valid only for the ground plane, but not human bodies, an issue neglected in previous works. We proved formally that applying the geometric correction to human bodies also leads to a linear relationship between the number of pixels and the number of persons. Therefore, geometric correction can be carried out with pixels of human bodies regardless of their relative position in the scene, without more complex consideration of 3D information. We also described a crowd density estimation system. It is fast because no image level transform is needed and a LUT is used. For unusual crowdedness detection a time-adaptive criterion is proposed. In the future, we plan to perform automatic calibration for geometric correction, as well as better fitting technique (non-linear) for high crowd density estimation.

6. REFERENCES

[1] **Proceedings of** the **IEEE**, Special Issue on Visual Surveillance, October 2001, Vol. 89, No.10

[2] J.H. Yin, S.A. Velastin, A.C. Davies. *Image Processing Techniques for Crowd Density Estimation Using a Reference Image.* ACCV 1995: 489-498A.

[3] N. Marana, S. A. Velastin, L. da F. Costa and R. A. Lotufo. *Automatic Estimation of Crowd Density Using Texture*, Safety Science, 28(3), 165-175, 1998.

[4] N. Marana, L. da F. Costa, R. de A. Lotufo, and S. A. Velastin. *Estimating Crowd Density With Minkowski Fractal Dimension.* Proceedings 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 6, Phoenix, Arizona, USA, pp. 3521-3524.

[5] B. P. Lo and S. A. Velastin. *Automatic Congestion Detection System for Underground Platforms.* Proc of IEEE 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing (Hong Kong), p159-161, May 2001. (See also http://www.doc.ic.ac.uk/~benlo/)

[6] Paragios, V. Ramesh. *A MRF-based Approach for Real-Time Subway Monitoring.* IEEE ICPR, 2001.

[7] L. Li, W. Huang, Irene Y.H. Gu, and Q. Tian. Foreground Object Detection from Videos Containing Complex Background. ACM Multimedia, 2003.

[8] F. Lv, T. Zhao and R. Nevatia. Self-Calibration of a camera from video of a walking human. ICPR 2002.