Crowd Density Estimation Based on Voxel Model in Multi-View Surveillance Systems

Paweł Gardziński, Krzysztof Kowalak, Sławomir Maćkowiak, Łukasz Kamiński

Poznań University of Technology, Polanka 3, 60-965 Poznań, Poland
pgardzinski@multimedia.edu.pl

Abstract - **In this paper, a novel crowd density estimation method based on voxel modeling in multi-view surveillance systems is presented. The approach proposed in this paper is based on human silhouette modeling with an anthropometric cylinder. The performance of crowd density estimation was analyzed on two multi-view sequences datasets. For this propose PETS 2006 and PETS 2009 were used. Performance of the proposed approach has been evaluated for two metrics: people counting and crowd classification.**

Keywords **- Crowd density estimation; Multi-view; Voxel modeling**

I. INTRODUCTION

One of the most important problems in intelligent video surveillance systems, that recognize danger, is to observe "the crowd" and not to focus on "the unit". In sociology, the crowd is referred to as a disorganized human collectivity formed during a period of time. The people who are in the crowd feel stronger and lose the ability to assess the situation objectively. This often leads to spontaneous actions. People in the crowd tend to imitate other participants in the crowd and cannot evaluate the situation objectively. Which in turn leads to dangerous behaviors.

There are two different approaches used in order to estimate the crowd density – count-based approach and classification-based. In the count-based approach the number of people in the image are counted with the use of silhouette shape-based [1], tracking-based [2], and feature-based techniques. In the classification-based approach the crowd density is estimated based on local features of objects and their classification. To distinguish the texture features, a change of intensity gradients, edges or Chebyshev moments are used. Using different algorithms, the system learns the relation between the feature vector and the density. This approach is more effective in crowded scenes. In this type of approach features such as a change of intensity gradients, edges or Chebyshev moments are used to distinguish the texture.

In this paper, a technique of crowd detection and crowd density estimation based on volumetric modeling is proposed by the authors. The proposed technique uses the properties of the created models which are calculated based on observations from multiple cameras. The experiments were conducted with PETS 2006 and PETS 2009 test sequences. The results were compared to the commonly used methods of crowd density estimation based on local features (eg. Local Binary Pattern with a reduction in the size) [6] which are characterized by a better performance than the techniques based on the grey level co-occurrence matrix (GLCM) [3,5], Translation Invariant Orthonormal Chebyshev Moments (TIOCM)[4], and multi-channel Gabor filter [6]. The experiment results confirm the high effectiveness of the proposed techniques of crowd density estimation.

The paper is organized as follows. Section II presents the idea of volumetric modeling for crowd detection. In particular, Section II.4 describes the proposed approach to estimate the density of the crowd. Section III presents the experimental results. Section IV provides a summary and a proposal for further work.

II. GENERAL APPROACH

A method presented in [7] utilizes a multi-camera system in order to estimate a voxel model and analyze humans' behavior. An approach presented in this paper is based on results presented in [7] and is used to estimate crowd density. All consecutive steps required to achieve the goal are shown in Fig. 1.

- *Multi-View System Calibration*

Method presented in this paper utilizes information from at least three cameras. Moreover, the cameras should be placed in such a relative position that they oversee a mutual area from different angles as presented in Fig. 2. To calibrate such a system one of the known methods may be utilized. In this particular paper a method presented in [10] has been used. It calculates both intrinsic and extrinsic camera parameters that are necessary in the next stages of the proposed approach.

- *Background Subtraction*

Background subtraction is a technique that estimates changes (moving objects) against a static background. This operation results in a set of moving object silhouettes that are used in the voxel reconstruction process. Background subtraction is carried out for all views independently using the method described in [9].

- *Voxel Reconstruction*

Voxel reconstruction in one of the most computation efficient methods to recreate a 3D scene estimated from multiple cameras. It allows to create and analyze a 3D scene in

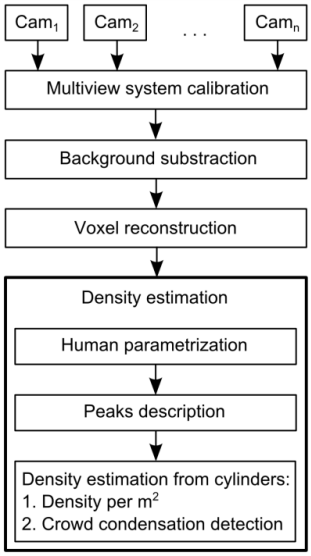

Figure 1: Block diagram of proposed approach.

real time, which is the most wanted characteristic in surveillance systems.

After a calibration process, which is performed in order to estimate the camera parameters, the reconstruction space is initialized where the objects visible from all views will be reconstructed. The light re-projection algorithm with regard to a pinhole camera model is used. For each view a foreground mask is projected into voxel space, where all voxels along the projection are marked as visible. The final model is a result of all voxels visible from all views. An example of this method has already been described in [7].

Another advantage of the approach using voxel reconstruction is that a high fidelity model is not necessary in surveillance systems. A coarse model that enables an efficient object distinction in a given scene is more than enough. The time complexity is influenced by different factors such as the number of input views or reconstruction quality (number of voxels used in reconstruction process). Therefore, for such system that is supposed to work in real time (or at least near real time) it is necessary to determine a sweet spot between computational time and reconstruction quality.

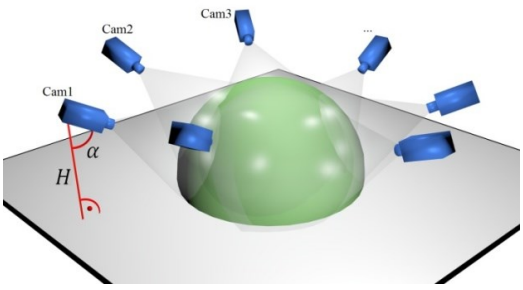

Figure 2: An example of cameras deployment that oversee mutual area (green). H – camera altitude over ground level, α – camera angle of inclination.

- *Density Estimation*

The idea of crowd density estimation is based on peak detection in voxel model that are used to describe people with parameterized cylinders. Crowd density is calculated as a number of people per square meter of the mutual area observed by cameras. Additionally, crowds density is categorized as dense or sparse. Algorithm's block diagram has been shown in Fig. 1 in the density estimation block.

A. Human parametrization

Despite differences between people in their stature, shape or generally silhouette it is possible to create a unified model that can be used to describe a majority of population. The measures of people used in the approach presented in this paper were taken from [10]. The proportion of an average human height to their width can be used as a scaling value to estimate one's measurements from their height. Utilizing that information it's possible to describe human silhouette volume with an anthropometric cylinder that describes human body ratios. The cylinder's radius is based on a value of a width to height ratio of an average human. This ratio is as follows:

$$r_{ave} = \frac{Human_w}{2 \cdot Human_h} = \frac{464}{2 \cdot 1755} = 0.1322 \quad (1)$$

where $Human_w$ and $Human_h$ are width and height of an average human respectively. Knowing ones height and using (1), calculating cylinder's radius that will be used to describe that person is trivial:

$$C_{r_1} = r_{ave} \cdot C_h \quad (2)$$

where C_h is a cylinder height.

B. Peaks Description

In this step voxel model peaks are detected in flood-like manner. A peak is a local maximum in voxel model and is detected by "flooding" the model from bottom to top. In each step a layer of voxels is marked as "flooded" and checked if any of them do not have neighboring voxel in the next (higher)

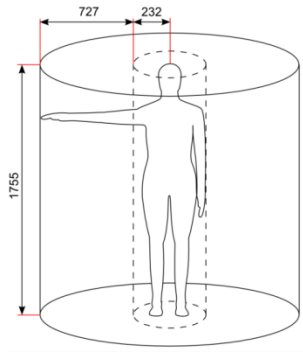

Figure 3: Selected measures of an average human (in mm) and a silhouette description with anthropometric cylinder.

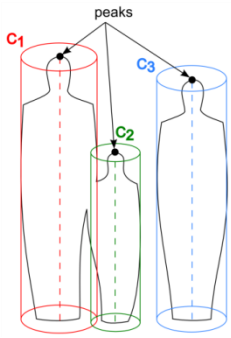

Figure 4: Human silhouettes describes by cylinders. Black points represent peaks of each object.

layer. A detected voxel is taken to measure local objects height and then modeled with an anthropometric cylinder. Its height and radius are calculated according to previous paragraph. An example of description of a cylinder with regard to their corresponding peaks is shown in Fig. 4.

C. *Density Estimation From Cylinders*

The crowd density can estimated with a metric consisting of two values. The first one, describes the number of objects per square meter of a given scene. Based on the number of detected peaks, the rough area occupied by them can be estimated with a density value (3).

$$\rho = \frac{n}{area} \tag{3}$$

However, this value does not reflect the relative position of each object in a scene. As an example we can consider a scenario of two persons in one room. No matter where they stand, each time the density ρ is the same. A crowd appears when people concentrate near one location. Therefore a second value that would describe the crowd relative position is required. It must describe whether the crowd is dense – when people concentrate near a point, or when the crowd is sparse when they are relatively far from one another.

The second measure that would describe if the crowd is dense or sparse utilizes presented idea of anthropometric cylinders. A crowd is defined as group of people that violate their personal area as it was described in [11]. Therefore, the cylinder radius is increased by the value that defines personal area, which is calculated as an averaged arm length.

$$r'_{ave} = \frac{arm_{length}}{Human_h} = \frac{727}{1755} = 0.4143 \tag{4}$$

$$C_{r_2} = C_{r_1} + r'_{ave} \cdot C_h \tag{5}$$

where arm_{length} is a length of arm of an average human.

In order to estimate the intersecting part of cylinders, a $n \times n$ matrix is used, where n is the number of cylinders detected in scene (estimated number of people in the scene). It is calculated as follows:

$$SV_{i,j} = \int (V_{C_i} \cap V_{C_j})\, dV_C \tag{6}$$

where V_C is a volume of cylinder. Values greater than 0 in the matrix $\boldsymbol{SV}$ mean that *i-th* and *j-th* object are near each other and may be treated as a crowd.

This matrix is used to determine which objects belong to their respective groups of people. Visualization of proposed approach is shown in Fig. 5.

III. EXPERIMENTS

The experiments required multi-view video sequences. In this case PETS2006 and PETS2009 were used. The advantages of these evaluation sets are that they: are publicly available, contain calibration data and additionally all views of these sequences are synchronized. These databases allow to perform experiments in *indoor* and *outdoor* conditions. Crowd density in both datasets covers all density categories that were described in [12]. The following categories were distinguished: *free flow, restricted flow, dense flow, very dense flow, jammed flow*.

TABLE I. DEFINITION OF DIFFERENT CROWD LEVELS ACCORDING TO THE RANGE OF DENSITY [14].

Levels of crowd density	Range of density ($people/m^2$)
Free flow	< 0.5
Restricted flow	$0.5 - 0.8$
Dense flow	$0.81 - 1.26$
Very dense flow	$1.27 - 2.0$
Jammed flow	> 2.0

In order to perform experiments the number of people in test sequences was counted and taken as a *ground truth*. The efficiency of crowd density estimation method was defined as a number of detected people to *ground truth* ratio. The results are shown in Table 2 and were divided into different crowd density levels for both the *indoor* and *outdoor* conditions.

TABLE II. RESULTS OF OBJECT COUNTING EFFICIENCY.

Level of crowd density	Indoor	Outdoor
Free flow	97.34%	92.12%
Restricted flow	97.13%	91.74%
Dense flow	95.62%	88.37%
Very dense flow	90.28%	86.15%
Jammed flow	85.39%	80.41%
Mean	93.15%	87.75%

Results show that the proposed method of crowd density estimation achieves a mean efficiency of 93.15% for *indoor* and 87.75% for *outdoor* conditions. The efficiency of proposed method is slightly decreasing with the increase of crowd density in both scenarios. It is caused by difficulty to separate people in the crowd and limited resolution of voxel

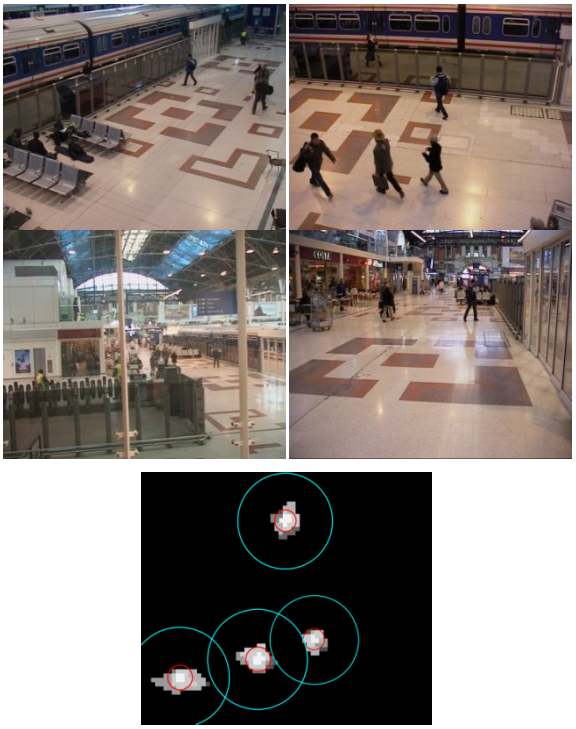

Figure 5: Example of 3D reconstruction of objects in a scene from top view (bottom image) created from 4 frames from PETS2006 sequence (four upper images).

model. In addition, the efficiency of the proposed approach is lower for *outdoor* conditions due to additional difficulties that appear during background subtraction like an illumination changes or reflections. The next step of experiments consisted of the comparison of achieved efficiency with another method. For this purpose LBP+DR method [13] was used as a reference. This method was chosen because the authors performed experiments on PETS2009 dataset and used the same crowd density levels. The additional experiments were performed to check performance of proposed approach in case of density level classification. The results are shown in Tab 3.

TABLE III. EVALUATION OF CROWD CLASSIFICATION RESULT.

Density level	LBP+DR	Ours method
Free flow	95.25%	96.8%
Restricted flow	91.50%	92.7%
Dense flow	87.00%	90.1%
Very dense flow	97.50%	89.2%

The proposed approach achieves higher classification efficiency than LBP+DR method in case of *free flow*, *restricted flow* and *dense flow*. It is caused by high efficiency of people counting in crowds. In case of *very dense flow* the efficiency of proposed approach is decreasing and LBP+DR achieves higher efficiency. The reason of this result is inaccurate reconstruction of voxel model for dense crowds. In such case the estimation of number of people achieves lower accuracy.

IV. CONCLUSIONS

In this paper a novel multi-view crowd density estimation method based on voxel model of scene was presented. Proposed approach describes three dimensional silhouette of human with cylinder. This is the base of crowd density estimation. The experiments were performed on PETS2006 and PETS2009. This approach achieves high efficiency of crowd density estimation of up to 93.15% and 87.75% for indoor and outdoor conditions respectively. The efficiency is slightly lower in case of very dense crowd and varied lighting in outdoor conditions. Therefore, some additional work is required in order to increase the efficiency in critical cases.

ACKNOWLEDGMENT

This work has been supported by the public funds as a DS research project.

REFERENCES

[1] D.L. Swets, B. Punch, J. Weng. Genetic algorithms for object recognition in a complex scene. International Conference onImage Processing, 1995,vol. 2, pp. 595-598.

[2] X. Yuan, Y. J. Lu, S. Sarraf. A computer vision system for measurement of pedestrian volume. In TENCON'93. Proceedings. Computer, Communication, Control and Power Engineering. 1993 IEEE Region 10 Conference on.

[3] A.N. Marana, S.A. Velastin, L.F. Costa, R.A. Lotufo. Estimation of crowd density using image processing. IEE Colloquium on Image Processing for Security Applications (Digest No.: 1997/074), 1997, pp. 11/1-11/8.

[4] H. Rahmalan, M.S. Nixon, J.N. Carter. On crowd density estimation for surveillance. The Institution of Engineering and Technology Conference on Crime and Security, 2006.

[5] R. M. Haralick. Statistical and structural approaches to texture. Proceedings of the IEEE, 67(5): 786-804, 1979.

[6] M. Jalali Moghaddam, E. Shaabani, R. Safabakhsh. Crowd Density Estimation for Outdoor Environments. 8th International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) BICT 2014, pp. 306-310.

[7] S. Maćkowiak, P. Gardziński, Ł. Kamiński, K. Kowalak. Human Activity Recognition in Multiview Video. 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, South Korea, 2014, pp. 148-153.

[8] Z. Zhang. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11): 1330-1334, 2000.

[9] Z. Zivkovic. Improved Adaptive Gaussian Mixture Model for Background Subtraction. Proceedings of ICPR, 2004.

[10] A. R. Tilley. The Measure of Man and Woman. John Willey & Sons, New York, 2002.

[11] F. Solera, S. Calderara, R. Cucchiara. Structured learning for detection of social groups in crowd. 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Kraków, Poland, 2013, pp. 7-12.

[12] A. Polus, J. Schofer, A. Ushpiz. Pedestrian Flow and Level of Service. Journal of Transporting Engineering, 109(1):46-56, 1983.

[13] H. Fradi, X. Zhao, J.-L. Dugelay. Crowd density analysis using subspace learning on local binary pattern. IEEE International Conference on Multimedia and Expo Workshops, San Jose, 2013