PEOPLE COUNTING SYSTEM IN CROWDED SCENES BASED ON FEATURE REGRESSION

Hajer Fradi, Jean-Luc Dugelay

EURECOM
Sophia Antipolis, France

ABSTRACT

While people counting has been improved significantly over the recent years, crowd scenes and perspective distortions remain particularly challenging and could deeply affect the count. To handle such problems, we propose a counting system based on measurements of interest points, where a perspective normalization and a crowd measure-informed density estimation are introduced into a single feature. Then, the correspondence between this feature and the number of persons is learned by Gaussian Process regression. Our approach has been experimentally validated showing more accurate results compared to other features-based methods.

Index Terms— People counting, SIFT interest points, crowd analysis, perspective, density, Gaussian Process.

1. INTRODUCTION

People counting is a crucial component in many domains, from marketing to video surveillance. It is useful information for safety, security and economic reasons. For instance, the automatic monitoring of the number of persons flooding public areas is extremely important for safety control mainly when this number exceeds a certain level of crowd. Also, the estimation of number of passengers is relevant to economic applications such as optimizing the schedule of public transportation systems.

In recent years, significant progress has been made in the field of people counting. But, crowd scenes still remain challenging mainly by applying detection-based methods which rely on detecting and separating individuals. Therefore, recent works typically bypass the challenge of detecting people and instead focus on learning a mapping between the number of persons and a set of features. In this context, two trends have been followed, either by augmenting the number of features (it reaches 28 features in Chan's method [1]) or by applying different regression functions to be able to select the one that fits better the selected features (e.g. linear [2], ϵ-SVR and ANFIS in [3], Bayesian Poisson [4] and Gaussian Process regression [1]).

This extensive study varying the features or the trainable function is caused by the fact that the features deviate from the perfect case where the number of persons is simply proportional to the features. Therefore, instead of training more features and testing different regression functions, we are interested in revealing the factors that affect the relationship between the features and the number of persons. In particular, we intend to explore distance and density cues. The first cue is employed to handle the problem of perspective distortions, whereas, the second cue is density, it is used as crowd feature to detect and to measure the overlap between individuals. To achieve this goal, we apply perspective map normalization and to weight the feature by a crowd measure in order to compensate the variations in distance and in density. Our intuition behind this is to make our selected feature invariant to the aforementioned factors, which could guarantee the linearity of the trainable function in challenging situations.
The remainder of the paper is organized as follows: In Section 2, we briefly review relevant works to people counting. Then, in Section 3, we introduce our counting system based on a single feature regression. The proposed approach is evaluated using PETS dataset and the experimental results are summarized in Section 4. Finally, we conclude in Section 5.

2. RELATED WORKS

The taxonomy of people counting methods embodies two paradigms: detection and features based methods. By using detection-based methods, the number of persons and their locations are provided simultaneously and the count is not affected as long as people are correctly segmented. But, the difficulty is that detecting people is a complex task by itself, mainly in presence of crowds and occlusions. This problem has been addressed by adopting part-based detectors, or by detecting only heads or the Ω shape formed by heads and shoulders. These attempts to mitigate occlusions are not applicable in very crowded scenes which are of primary interest for people counting.

The second paradigm consists of estimating the number of persons from various features. These methods are more efficient since it is easier to detect features than to detect persons. For this purpose, many features of foreground pixels have been used such as: total area, textures and edge count [1]. Other features based on interest points measurements such as SURF features [5] and corners points [2] are also introduced into counting methods. To perform the counting, a regression function has to be applied. It is required to learn the relationship between the features and the number of persons.

Among features-based methods, Albiol's method [2] has shown good performance using PETS dataset. It uses Harris corner points as features. Then, the count is performed by assuming a direct proportional relation between the number of moving corner points and the number of persons. Actually, the application of this method is limited by the fact that it does not consider the difference between the perceived size of persons at different distances from the camera and with different densities as well. These limitations were not revealed in the PETS contest since only videos characterized by short depth range and trivial occlusions were required for the tests. Therefore, more tests under situations with serious perspective distortions and occlusions are needed to evaluate this method.

Differently from the previous work, some methods take into account the effects of perspective distortions. To handle this problem, different techniques have been investigated. In [6], a geometric correction (GC) is proposed to bring all the objects at different distances to the same scale. Also, a perspective map normalization is proposed in [1] to weight the pixels according to their distance from the camera. Another technique has been investigated in [5] to account the effect of perspective. It is based on applying an Inverse Perspective Mapping (IPM) to compute the distance of each group of persons from the camera. In [5], the problem of density has also been addressed following these steps: first, a clustering algorithm is applied to partition different groups of persons, then, the density of each cluster is obtained as the ratio between the number of the detected points and the area of the bounding box. Although this method [5] proposes to deal with two major problems that usually affect the results of feature-based methods, it still suffers from many limitations and leaves rooms for improvements. One of the drawbacks is that it requires three parameters (number of detected points, distance, and density) for each cluster separately, which is a heavy annotation task. More details about the limitations of this method will be discussed through the development of our proposed approach.

3. PROPOSED APPROACH

In this section, our proposed approach for people counting is presented. It is basically inspired from [1] and [5] with substantial improvements.
To perform people counting, we follow the line of methods based on measurements of interest points [5, 2]. One major advantage using these methods is that they bypass intermediate steps like the segmentation of foreground pixels as used in [1]. Then, to filter out the static detected points, motion information has to be estimated. For this purpose, a block matching is applied in [5, 2]. Given the difficulty of this technique to deal with problems of occlusions and discontinuities at boundaries, we propose in this paper a more efficient solution based on computing the optical flow with reduced weights near the borders since the expansion coefficients are less reliable there, see Section 3.1. Moreover, in this study we explore distance and density cues in order to compensate the effects of perspective distortions and partial occlusions due to the crowd. On the one side, these two factors were not taken into account in [2]. Also in [1], the effects of people density were not considered, however, 28 features from foreground pixels were devoted to infer the contents of each frame. On the other side, Conte's method [5] is the only work that dealt with the two aforementioned factors, but the proposed approach is still problematic. More details about its limitations are highlighted in next sections.

Compared to [5], we propose to process the perspective normalization at pixel level which is more accurate than assigning one distance value to each group of persons, see Section 3.2. In addition, for density estimation, we apply density-based clustering which is better adapted for separating different groups of persons than the graph-based clustering proposed in [5]. Another problem is addressed in our study; it is the calculation of the area of each cluster. We apply α-shape technique which is more powerful than the bounding box proposed in [5]. This latter fails to define boundaries of a set of points by leaving large gaps which could amply deteriorate the estimated density. Added to that, one major contribution of our counting system consists of formulating a new weight function based on density estimation for crowd normalization, see Section 3.3. An overview of our counting system modules and their interactions is shown in Figure 1. The remainder of this section describes each of these system components.

3.1. Detection of moving interest points

To infer the contents of each frame under analysis, only interest points have to be detected. In this context, we propose to use SIFT (scale-invariant descriptor) [7] where interest point locations are defined as maxima/minima of the difference of Gaussians in scale-space. Under this respect our method is nearby similar to [5, 2] which applied Harris corner detector and SURF, respectively. Using corner detectors is not accurate since they are somehow dependent on the perceived scale of the considered object. Also for SURF detector, it is an extension of SIFT and according to [8], SIFT is more invariant to scale, rotation, and affine transformations.

After that, motion information has to be associated to the detected interest points to be able to distinguish between moving and static ones. By considering the same assumptions as in [5], the detected interest points with non-null motion vector typically belong to persons. To handle this problem, we compute the optical flow by the method proposed in [9]. It uses quadratic polynomial model to approximate each neigh-

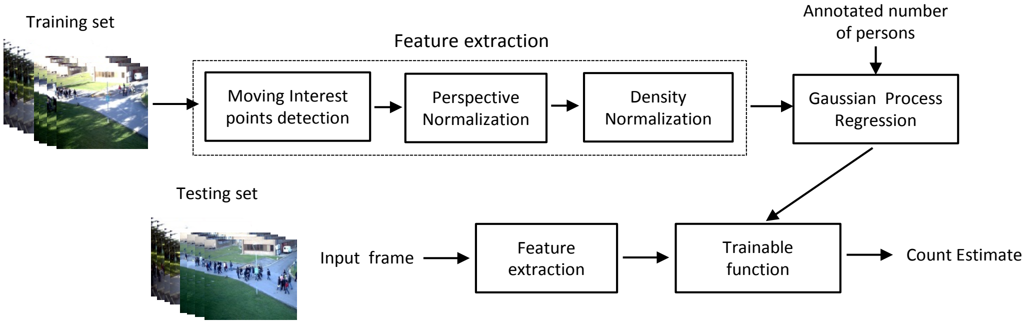

Fig. 1. *Flowchart of the proposed counting system*

borhood of two consecutive frames. Then, the displacement fields are estimated from the polynomial expansion coefficients. This method has the advantage of reducing the errors near the borders by computing the polynomial expansions with certainty set to zero off the border and with a reduced weight for pixels close to the borders.

3.2. Perspective normalization

The objective here is to compensate for changes in number of interest points because of perspective distortions. The effects of perspective can be simply explained by the fact that objects far from the camera appear smaller than the closest ones. This makes any extracted feature for farther away persons account for a smaller portion compared to closer persons. This problem could be addressed by weighting each interest point according to a perspective map with assigning larger weights for farther points in the scene.

Similar to [1], we estimate the perspective map by linearly interpolating between the two extremes of the scene. First, the ground plane is marked. Then, the distance d_1 and d_2 of the two extreme lines are measured. After that, the difference between the perceived height of persons in these two lines can be derived by manually calibrating two frames, where the center of a reference person belongs to the first line in the first frame while belonging to the second extreme line in the second frame. A weight of 1 is assigned to pixels on the first line, and the pixels on the second line are weighted by $\frac{h_1 * d_1}{h_2 * d_2}$, where h_1 and h_2 denote the two heights of the reference person in the two frames. A linear interpolation is applied to compute the remaining weights between the two extreme lines. Finally, the weights W_p are assigned according to the y-coordinate of each interest point.

After perspective normalization, the number of moving SIFT points in each frame i under analysis is updated as follows:

$$FeatN_i = \sum_{y=1}^{Y} W_p(y) * N_T(y) \tag{1}$$

Where $N_T(y)$ is the number of moving points in the y^{th} row.

3.3. Density estimation for Crowd measurement

In addition to perspective distortions, the density of people could also affect the number of detected points. When persons are closer to each other, more partial occlusions occur. Thus, we aim at estimating the density of people by measuring how close the detected points are, in order to handle the underestimation of the number of persons in high crowd situations. Therefore, a clustering algorithm has to be applied in order to distinguish the different groups of persons. The most appropriate solution for this problem is density-based clustering, where clusters are identified according to the spatial density of the points. It has also the advantage of being flexible enough to discover clusters of arbitrary shape.

3.3.1. Density-based clustering

For density-based clustering, we apply DBSCAN (Density Based Spatial Clustering of Applications with Noise) [10]. This algorithm does not require any prior knowledge about the number and the shape of the clusters. Added to that, it fits well our requirements by adopting the concept of density-reachable to form the clusters with respect to $MinPts$ and Eps input parameters which denote a threshold of points needed in a neighborhood and a neighborhood radius. Moreover, points which are not density-connected are labeled as noise.

3.3.2. Density estimation

The density is measured by computing the ratio between the number of moving interest points and the area covered by the clusters. For the area computation, we propose to delineate the boundaries of each cluster by α-shape [11] which is an accurate technique to extract the shape of a set of points. α-shape has not only the advantage of closely following variations in the outer-edge but it reveals also the inner gaps.

This technique is reliable to accurately estimate the density of clusters mainly with the association of the selected density-based clustering algorithm that picks out the clusters using the density relevance and filters out the noise.

3.3.3. *Crowd measurement*

At this stage, we aim at formulating a weighting function by using the density as a crowd measure. In particular, our goal is to weight the proposed feature defined in (1) by inflating its value in high crowd frames, while reducing it in low crowd frames. Therefore, we use the estimated density values d_i, $i = 1...M$, where M is the total number of frames for the video sequences. Then, the weight function is defined as:

$$W_d(i) = \frac{d_i - \mu}{\sigma_{max}} + 1 \qquad (2)$$

Where $\mu = \frac{1}{M}\sum_{i=1}^{M} d_i$ and σ_{max} is the maximum of standard deviation values σ_i.

This weight function ensures somehow kind of crowd normalization. It is achieved by setting $W_d = 1$ if the crowd is medium $(d_i = \mu)$, $1 < W_d \leq 2$ if the crowd is high, and $0 \leq W_d < 1$ otherwise.

Consequently, to take into account the effects of the crowd on the detected interest points, our proposed feature defined in (1) is updated again:

$$FeatN_i = W_d(i) * \sum_{y=1}^{Y} W_p(y) * N_T(y) \qquad (3)$$

3.4. Gaussian Process regression

Our proposed feature defined in (3) has been formulated to be invariant to perspective and to crowd variations. This could involve the linearity of the trainable function mapping the feature to the number of persons. But for more flexibility, we suggest to take into account the possible errors that could occur in the motion estimation or in any other step of our counting system. Therefore, we propose to learn the trainable function from a set of labeled examples by using Gaussian Process (GP) regression which is well adopted for linear features with local non-linearities (more details about GP are available in [12]). Once the function is estimated, the number of the persons could be predicted from the value of the proposed feature for each frame under analysis.

4. EXPERIMENTAL RESULTS

In this section, we present the experimental results on the PETS 2009 dataset [1] to evaluate our counting system described in Section 3. From this dataset, we are interested in the section S1 used to assess *Person count and Density estimation* algorithms. In our experiments, we employ not only the 4 videos tested in people counting contest held in PETS 2009, but also other 4 videos from the second view. In fact, videos referring to the second view are more challenging: they are characterized by a large depth range with significant variations in density compared to the first view. The main characteristics of these videos are summarized in Table 1.

			Number of people	
Video Sequence	View	Length	Min	Max
S1.L1.13-57	1	221	5	34
S1.L1.13-59	1	241	3	26
S1.L2.14-06	1	201	0	43
S1.L3.14-17	1	91	6	41
S1.L1.13-57	2	221	8	46
S1.L2.14-06	2	201	3	46
S1.L2.14-31	2	131	10	43
S3.MF.12-43	2	108	1	7

Table 1. Characteristics of 8 sequences from the PETS 2009 dataset used for the counting experiments.

Apart from the testing set, the counting regression function is learned from a training set built by 2 other videos from S1 Section. The training frames have to guarantee different cases in terms of number of persons, distance and density.

For the ground-truth of the count, it is obtained by annotating the number of persons by hand in every 5th frame. The count for the remaining frames is obtained using linear regression.

To compare the estimated number of persons to the ground truth, we calculate the Mean Absolute Error (MAE) and the Mean Relative Error (MRE) which are defined as:

$$MAE = \frac{1}{M} . \sum_{i=1}^{M} |E(i) - G(i)| \qquad (4)$$

$$MRE = \frac{1}{M} . \sum_{i=1}^{M} \frac{|E(i) - G(i)|}{G(i)} \qquad (5)$$

Where M is the total number of frames in a video sequence. E(i) and G(i) denote, respectively, the estimated and the ground-truth number of persons in the $i-$th frame. The MAE metric was used to compare the performance of the algorithms submitted to the PETS contest. But, the same error could be negligible if the number of persons is high. Therefore, in [5], the authors propose to use also the MRE metric, which relates the error to the number of the persons.

For the comparisons, unfortunately, we are not able to compare our proposed method to Chan's method [1]. In fact, for their work [13] submitted to PETS 2009, only tests with videos from the first view were provided. Since, we are interested to test more challenging videos; our results are compared to those of Albiol and Conte methods [2, 5] which are reported in [5]. A summary of the counting results, with respect to our hand-annotated ground-truth, are given in Table 2. From these results, we show clearly a big difference between the performance of Albiol's method [2] between the first and the second views. That could justify the inability of this method to deal with challenging situations. For this reason, similarly to [5], we proposed to deal with the problems of

[1] `http://www.cvg.rdg.ac.uk/PETS2009/`

	Video Sequence	Albiol et al. [2]		Conte et al. [5]		Our approach	
		MAE	MRE	MAE	MRE	MAE	MRE
View1	S1.L1.13-57	2.80	12.6%	1.92	8.7 %	1.38	7.10 %
	S1.L1.13-59	3.86	24.9 %	2.24	17.3 %	2.25	15.02 %
	S1.L2.14-06	5.14	26.1 %	4.66	20.5 %	4.58	21.75 %
	S1.L3.14-17	2.64	14.0 %	1.75	9.2 %	1.54	8.99 %
View2	S1.L1.13-57	29.45	106.0 %	11.76	30.0 %	3.64	11.67 %
	S1.L2.14-06	32.24	122.5 %	18.03	43.0 %	6.87	18.30 %
	S1.L2.14-31	34.09	99.7 %	5.64	18.8 %	2.53	10.93 %
	S3.MF.12-43	12.34	311.9 %	0.63	18.8 %	2.20	40.31 %

Table 2. Quantitative evaluation of our proposed approach compared to other methods

perspective distortions and crowd density. That could justify as well the better results of our method and Conte's method compared to [2].

A comparison of our results with the results of [5] reveals the effectiveness of our proposed modifications. In particular, the tests with S1.L1.13-57(2) and S1.L2.14-06(2) show the effects of the proposed crowd measure to compensate the underestimation of number of persons because of dense crowd occurring at the last frames of S1.L1.13-57(2) and at the first frames S1.L2.14-06(2). We also note that Conte's method requires to compute the ground-truth inside each cluster separately, which is a burdensome task. It is also important to mention that we obtained linear trainable function estimated by Gaussian Process regression, which means that the non-local linearities were not significant. That could justify the success of our proposed feature to make the number of interest points independent from distance and density variations.

5. CONCLUSION

In this paper, we propose an approach for people counting based on regressing a single frame-wise feature independent from variations of perspective and crowd density. Our contribution regarding the related works in people counting is discussed through the details of the proposed approach. Experiments on PETS dataset demonstrate that our approach is able to maintain a linear relationship between the proposed feature and the number of persons under situations with heavy occlusions and serious perspective distortions. Also, by means of comparisons with other existing features-based methods in the literature, our approach has shown its ability to improve significantly the accuracy of the count.

6. REFERENCES

[1] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–7.

[2] A. Albiol, M. J. Silla, A. Albiol, and J. M. Mossi, "Video analysis using corner motion statistics," in *IEEE International Workshop on PETS*, 2009, pp. 31–37.

[3] G. Acampora, V. Loia, G. Percannella, and M. Vento, "Trainable estimators for indirect people counting: A comparative study," in *FUZZ-IEEE*, 2011, pp. 139–145.

[4] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 545–551.

[5] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, "A method for counting people in crowded scenes," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010.

[6] R. Ma, L. Li, W. Huang, and Qi Tian., "On pixel count based crowd density estimation for visual surveillance," in *IEEE Conference on Cybernetics and Intelligent Systems*, 2004, pp. 170–173.

[7] David G. Lowe, "Distinctive image features from scale-invariant keypoints," in *Int. J. Comput. Vision*, 2004, pp. 91–110.

[8] L. Juan and O. Gwun, "A Comparison of SIFT, PCA-SIFT and SURF," in *International Journal of Image Processing (IJIP)*, 2009, vol. 3.

[9] G. Farneback, "Two-frame motion estimation based on polynomial expansion," in *Proc. of 13th Scandinavian Conference on Image Analysis*, 2003, pp. 363–370.

[10] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 226–231.

[11] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," in *IEEE Transactions on Information Theory*, 1983, vol. 29, pp. 551–559.

[12] C. E. Rasmussen and C. K. I. Williams, "Gaussian Processes for Machine Learning," in *The MIT Press*, December 2006.

[13] A. B. Chan, M. Morrow, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," in *IEEE International Workshop on PETS*, 2009.