Challenges of Deep Learning for Crowd Analytics

Muhammad Siraj
University of Bahrain, College of Information Technology;
University of Trento, Trento, Italy;
muhammad.siraj@alumni.unitn.it

Abstract: In high population cities, the gatherings of large crowds in public places and public areas accelerate or jeopardize people safety and transportation, which is a key challenge to the researchers. Although much research has been carried out on crowd analytics, many of existing methods are problem-specific, i.e., methods learned from a specific scene cannot be properly adopted to other videos. Therefore, this presents weakness and the discovery of these researches, since additional training samples have to be found from diverse videos. This paper will investigate diverse scene crowd analytics with traditional and deep learning models. We will also consider pros and cons of these approaches. However, once general deep methods are investigated from large datasets, they can be consider to investigate different crowd videos and images. Therefore, it would be able to cope with the problem including to not limited to crowd density estimation, crowd people counting, and crowd event recognition. Deep learning models and approaches are required to have large datasets for training and testing. Many datasets are collected taking into account many different and various problems related to building crowd datasets, including manual annotations and increasing diversity of videos and images. In this paper, we will also propose many models of deep neural networks and training approaches to learn the feature modeling for crowd analytics.

Keyword: anomaly; crowd analytics; congestion; crowd counting

 doi:10.20944/preprints201908.0291.v1

1. Introduction

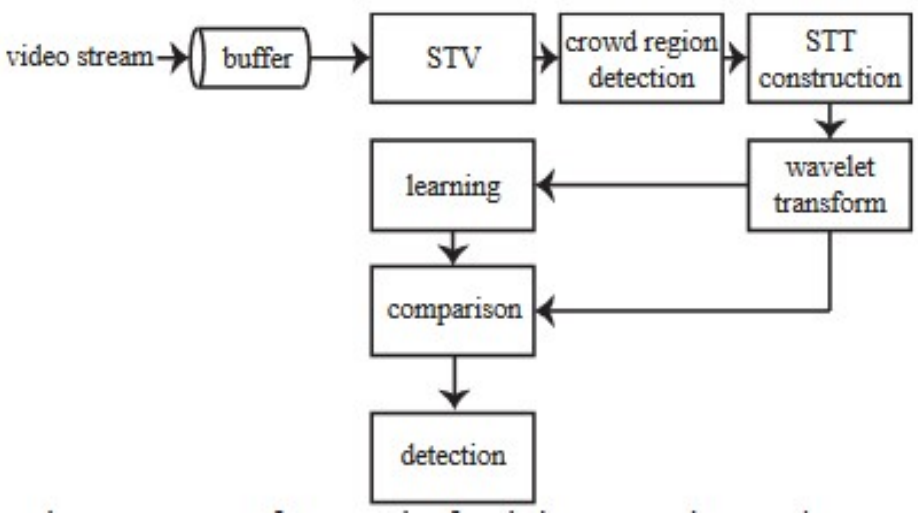
The important increase in CCTV videos in the lastdecade has led to more video data being collected than can be investigated by a computer operator [1][2]. Indeed, vision based and real-time computations of these significantly increasing databases has become important issue for the machine learning and computer vision researchers [3][4][5]. The performance considering real-time computational overhead is very significant in large-data driven scenes where scalability and a rapid response time are required [6][7].

Montgomery presented a report where he mentioned that, more than half of the people of the world reside in dense cities [8][9][10]. In fact, automated crowd investigation plays an crucial role in crowd analysis and visual surveillance videos considering these CCTV cameras and other installation systems [11][12]. Therefore, in terms of designing public spaces, visual surveillance systems, and intelligent controlled physical situations [13]. Therese kind of approaches will have various important applications such as the monitoring of crowd flows, taking care of accidents, and managing evacuation designs required in the bad event of a sudden and uncontrolled fire or in presence of riots in cities zones especially [14][15][16]. In the research documentation, the researchers have investigated the situation of gathering the motion information at a higher level [17][18]. It means that the motion information does not take into account individual moving or static objects [19][20][21]. These methods therefore, often need various features including multi-resolution histograms [22][23], spatio-temporal cuboids [24][25], appearance or motion descriptors [26][27] and spatio-temporal cubes [28][29].

For this purpose, we need to understand the combine distribution of the image pixels [30]. Furthermore, to take into account temporal along with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0291.v1

the spatial information of the data, the combine characteristics of the pixels across multiple adjacent video frames must be investigated [31][32]. For understanding, analyzing and learning, we make the general postulate that the distribution does not matter, it could be stationary over the learning interval or it could be mobile [33]. To consider the validity of this approach, it may be required to limit the temporal and spatial length of the learning window or time interval and therefore the number of videos in the training samples [34][35]. Our method is to understand, absorb and learn the distribution for a definite frame. Once this concept is understood and learned, it can be effectively prolonged to larger frame chunk sizes depending either an AR (Markov), MA, or ARMA process model [37]. Decreasing or imposing conditions on the training session reduce the number of learned parameters and therefore, the order of the process, hence reducing the learning variance. For this purpose, the flow diagram is presented as



2. Proposed Method

We compute the estimation of the spatio-temporal pixel distribution, the traditional machine learning method [6] is to estimate whether or not a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0291.v1

scene in the image or video is anomalous by computing its likelihood where we consider the learned distribution. This method is tending toward the fact that the anomalous situation is unknown, so the probability ratios cannot be properly computed. As an approach of how proper the method demonstrates the data, the Bayesian model is a significant feature. Thresholding the likelihood probability is encouraged by theoretical background considerations [6].

$$\ell(x) = (x - \mu_x)^T \Sigma^{-1} (x - \mu_x)$$

For each parameter, the partial derivative of the crowd model is computed for single training sequences of the videos under observations, i.e., one weight for each feature function Fj is considered. The partial derivative with respect to the parameter of the learning model corresponds to important value of the feature function for its true parameter, minus the averaged values of the feature function for all possible cases. Therefore, Eq. can be formulated as:

$$\frac{\partial}{\partial w_j} \log p\,(y/\mathbf{x};\mathbf{w}) = F_j(\mathbf{x}, y) - \sum_{y'} p\,(y'/\mathbf{x};\mathbf{w})\,[F_j(\mathbf{x}, y')]$$

During the model development, it has been found that with fixed values, the odel approximately consider Gaussian distribution. This experimental estimation works properly on many testing videos for anomaly detection, and has been exploited for developing the normal crowd activities in this research. For each learning video, the formulation is defined as

$$\mu_{jk} = \frac{1}{L}\sum_i f_{ijk},$$

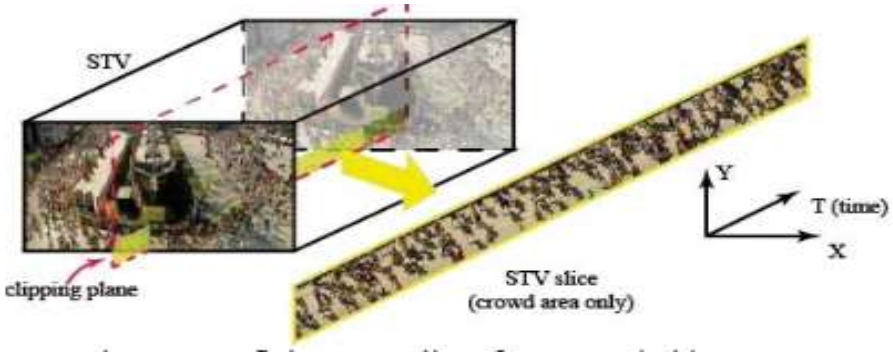
$$\sigma_{jk} = \sqrt{\frac{1}{L}\sum_i \left(f_{ijk} - \mu_{jk}\right)^2}.$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0291.v1

Due to the variety of moving individuals in a crowd video, well organized tracking individual objects is challenging. We demonstrate the crowd motion by the patch-based local motion formulation. Similarly to [6], the non-moving elements in the video are formulated as a collection of spatio-temporal cubes of dimension equal to p×p×q, where p(spatial size) and q(temporal size) must be big enough to encode the important characteristics of the different elements of the local motion flow. Every block is analyzed by a dynamic texture model [10], which is in fact a linear dynamic system of different parameters as formulated in the eq.

$$\begin{aligned} x_{t+1} &= Ax_t + Bv_t \\ y_t &= Cx_t + w_t, \end{aligned}$$

The same concept has been predicted in the Figure as shown below.



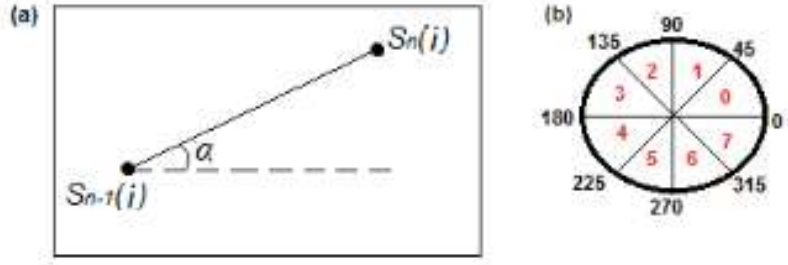
The important parameters for a motion pattern to be considered and learned are mean vector and covariance matrix specific to crowd scene under observation. Assume we have a motion pattern computed by m samples, if we get a new motion pattern whose training number is n, we estimate the new mean vector and covariance matrix of the new motion pattern as formulated:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0291.v1

$$\boldsymbol{\mu}_c = \frac{m\boldsymbol{\mu}_a}{m+n} + \frac{n\boldsymbol{\mu}_b}{m+n}$$

$$\boldsymbol{\Sigma}_c \approx \boldsymbol{\Sigma}_c' = \frac{(m-1)\boldsymbol{\Sigma}_a + n\boldsymbol{\Sigma}_b}{m+n-1}$$

The same concept of the formulation of the above equation is highlighted in the figure below.



3. Experimental Analysis and Evaluations

In this section we discuss about the experimental analysis and evaluations, results, evaluation and performance of our proposed method. Our proposed method is implemented using MATLAB by modeling user interface. This method is implemented considering image processing libraries and tested on the dataset from University of Minnesota. For this purpose, we first convert videos into frames. These videos have normal and abnormal situations occurrences at different parts of the videos. In our experiment we have used the GROUND sequence for performance analysis. In the figure given below we show the results of our method obtained with the marathon video sequence. In the beginning, user interface is launched which consists of the input option for the video of the crowd. Secondly, each video frame is taken as input in a given window interval. As can be seen in the figure, important tracklets are highlighted in different colors to show these individuals in the crowd.

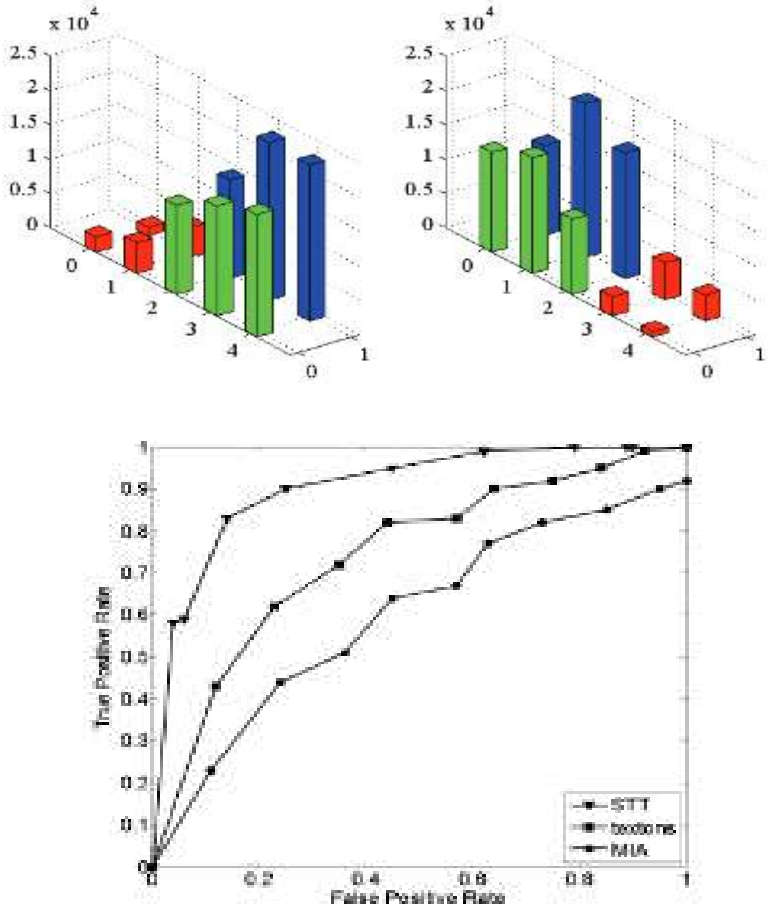
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0291.v1

The Table shows the experimental analysis results in term of accuracy all video sequences. The results demonstrate that most of video sequences are accurately understood and learned by the algorithm. On average our proposed method achieved 88.83% accuracy when applied on the videos from the same dataset. Hence, it shows the effectiveness of our method.

Runs	**1**	**2**	**3**	**4**	**5**
Accuracy	91.67%	93.33%	93.33%	100%	93.33%
Runs	**6**	**7**	**8**	**9**	**10**
Accuracy	100%	100%	60%	63.33%	93.33%
Average	88.83%				

The same analysis are also performed in terms of different graphs. Both graphs below show that our method performs very correctly irrespective of the challenge of the crowd events.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0291.v1



References

[1] Khan, S. D., Tayyab, M., Amin, M. K., Nour, A., Basalamah, A., Basalamah, S., & Khan, S. A. (2017). Towards a Crowd Analytic Framework For Crowd Management in Majid-al-Haram. arXiv preprint arXiv:1709.05952.

[2] Ahmad, K., Conci, N., & De Natale, F. G. (2018). A saliency-based approach toevent recognition. Signal Processing: Image Communication, 60, 42-51.

[3] Ullah, H., Altamimi, A. B., Uzair, M., & Ullah, M. (2018). Anomalous entities detection and localization in pedestrian flows. Neurocomputing, 290, 74-86.

[4] Saqib, M., Khan, S. D., Sharma, N., & Blumenstein, M. (2017, December). Extracting descriptive motion information from crowd scenes. In 2017International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

[5] Basalamah, S., Khan, S. D., & Ullah, H. (2019). Scale Driven Convolutional Neural Network Model For People Counting and Localization in Crowd Scenes. IEEE Access.

[6] Saqib, M., Khan, S. D., & Blumenstein, M. (2016, November). Texture-based feature mining for crowd density estimation: A study. In Image and Vision Computing New Zealand (IVCNZ), 2016International Conference on (pp. 1-6). IEEE.
[7] Ullah, H., Ullah, M., & Uzair, M. (2018). A hybrid social influence model for pedestrian motion segmentation. Neural Computing and Applications, 1-17.
[8] Bisagno, N., Zhang, B., & Conci, N. (2018, September). Group LSTM: Group Trajectory Prediction in Crowded Scenarios. InEuropean Conference on Computer Vision(pp. 213-225). Springer, Cham.
[9] Ahmad, F., Khan, A., Islam, I. U., Uzair, M., & Ullah, H. (2017). Illumination normalization using independent component analysis and filtering. The Imaging Science Journal, 65(5), 308-313.
[10] Ullah, H.,Uzair, M.,Ullah, M., Khan, A., Ahmad, A., & Khan, W.(2017). Density independent hydrodynamics model for crowd coherency detection. Neurocomputing, 242, 28-39.
[11] Trabelsi, R., Jabri, I., Melgani, F., Smach, F., Conci, N., & Bouallegue, A. (2017, November). Complex-Valued Representation for RGB-D Object Recognition. In Pacific-Rim Symposium on Image and Video Technology (pp. 17-27). Springer, Cham.
[12] Ullah, M., Ullah, H., & Alseadonn, I. M. (2017). Human action recognition in videos using stable features.
[13] Xu, M., Ge, Z., Jiang, X., Cui, G., Zhou, B., & Xu, C. (2019). Depth Information Guided Crowd Counting for Complex Crowd Scenes. Pattern Recognition Letters.
[14] Alameda-Pineda, X., Ricci, E., & Sebe, N. (2019). Multimodal behavior analysis in the wild: An introduction. In Multimodal Behavior Analysis in the Wild (pp. 1-8). Academic Press.
[15] Ullah, M., Ullah, H., Conci, N., &De Natale, F. G. (2016, September). Crowd behavior identification. In Image Processing (ICIP), 2016IEEE International Conference on(pp. 1195-1199). IEEE.
[16] Kim, H., Han, J., & Han, S. (2019). Analysis of evacuation simulation considering crowd density and the effect of a fallen person. Journal of Ambient Intelligence and Humanized Computing, 1-11.
[17] Hao, Y., Xu, Z. J., Liu, Y., Wang, J., & Fan, J. L. (2019). Effective crowd anomaly detection through spatio-temporal texture analysis. International Journal of Automation and Computing, 16(1), 27-39.
[18] Ullah, H., Ullah, M., Afridi, H., Conci, N., & De Natale, F. G. (2015, September). Traffic accident detection through a hydrodynamic lens. In Image Processing (ICIP), 2015IEEE International Conference on (pp. 2470-2474). IEEE.
[19] Shimura, K., Khan, S. D., Bandini, S., & Nishinari, K. (2016). Simulation and Evaluation of Spiral Movement of Pedestrians: Towards the Tawaf Simulator. Journal of Cellular Automata, 11(4).

[20] Ullah, H. (2015). Crowd Motion Analysis: Segmentation, Anomaly Detection, and Behavior Classification (Doctoral dissertation, University of Trento).
[21] Kang, D., Ma, Z., & Chan, A. B. (2018). Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. IEEE Transactionson Circuits and Systems for Video Technology.
[22] Rota, P., Ullah, H., Conci, N., Sebe, N., & De Natale, F. G. (2013, September). Particles cross-influence forentity grouping. In Signal Processing Conference (EUSIPCO), 2013Proceedings of the 21st European (pp. 1-5). IEEE.
[23] Ullah, H., Ullah, M., & Conci, N. (2014, March). Real-time anomaly detection in dense crowded scenes. In Video Surveillance and Transportation Imaging Applications 2014(Vol. 9026, p. 902608). International Society for Optics and Photonics.
[24] Arif, M., Daud, S., & Basalamah, S. (2013). Counting of people in the extremely dense crowd using genetic algorithm and blobs counting. IAES International Journal of Artificial Intelligence, 2(2), 51.
[25] Ullah, M., & Alaya Cheikh, F. (2018). A Directed SparseGraphical Model for Multi-Target Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1816-1823).
[26] Khan, S. D., & Ullah, H. (2019). A survey of advances in vision-based vehicle re-identification. Computer Vision and Image Understanding, 182, 50-63.
[27] Ullah, M., Mohammed, A., & Alaya Cheikh, F. (2018). PedNet: A Spatio-Temporal Deep Convolutional Neural Network for Pedestrian Segmentation. Journal of Imaging, 4(9), 107.
[28] Wang, LiMin, Yu Qiao, and Xiaoou Tang. "Mining motion atoms and phrases for complex action recognition." Proceedings of the IEEE international conference on computer vision. 2013.
[29] Wang, Heng, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. "Dense trajectories and motion boundary descriptors for action recognition." International journal of computer vision 103, no. 1(2013): 60-79.
[30] Wang, Xingxing, LiMin Wang, and Yu Qiao. "A comparative study of encoding, pooling and normalization methods for action recognition." In Asian Conference on Computer Vision, pp. 572-585. Springer, Berlin, Heidelberg, 2012.
[31] Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14, no. 3(2004): 199-222.
[32] Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: IEEEconference on computer vision and pattern recognition (CVPR), pp 1–8.
[33] Khan, Sultan Daud, and Habib Ullah. "A survey of advances in vision-based vehiclere-identification." Computer Vision and Image Understanding (2019).

 doi:10.20944/preprints201908.0291.v1

[34] Ullah, Habib, Muhammad Uzair, Arif Mahmood, Mohib Ullah, Sultan Daud Khan, and Faouzi Alaya Cheikh. "Internal Emotion Classification Using EEG Signal with Sparse Discriminative Ensemble." IEEE Access (2019).

[35] Saqib, M., Khan, S. D., Sharma, N., & Blumenstein, M. (2017, December). Extracting descriptive motion information from crowd scenes. In *2017 International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6). IEEE.

[36] Coluccia, A., Ghenescu, M., Piatrik, T., De Cubber, G., Schumann, A., Sommer, L., ... & Amandi, R. (2017, August). Drone-vs-bird detection challenge at IEEE AVSS2017. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.

[37] Ullah, M., Ullah, H., & Cheikh, F. A. (2019). SINGLE SHOT APPEARANCE MODEL (SSAM) FOR MULTI-TARGET TRACKING. *Electronic Imaging*, *2019*(7), 466-1.