Violent social interaction recognition

Adwan Alanazi[1], Abdul Basit[2]

[1]College of Computer Science and Engineering, University of Ha'il, Saudi Arabia
[2]University of Engineering and Technology, Peshawar, Pakistan

---

Abstract - *Modeling a method for the automatic analysis of surveillance videos in order to detect the presence of violent social interaction is of broad interest. In this work, we consider an effective and adaptive appearance model for the purpose of detecting unwanted social interaction in term of violence. We also consider a low-rank and structured sparse matrix decomposition (LSMD) model to better highlight the presence of violence. Our model is capable of capturing localized spatio-temporal features which enables the analysis of local motion taking place in the video. Potentially our method use adjacent frame differences as the input to the model thereby forcing it to encode the changes occurring in the video. The performance of our method is evaluated on a standard benchmark dataset in terms of detection accuracy. Results obtained with our technique revealed the promising capability of our method in detecting violent social interaction.*

Key Words: Social interaction, motion analysis, violence, features.

1. INTRODUCTION

There is no doubt that video surveillance equipment can be easily installed to monitor practically any environment and situation. The value of doing so, however, is practically questioned. Software and hardware systems for surveillance are often ineffective due to small number of trained supervisors watching the videos on screens and the natural limits of human attention capabilities. This is understandable, when thinking of the large numbers of cameras that require supervision, the monotonic nature of the surveillance videos, and the alertness required to detect events and provide quick response. In fact, even the seemingly easy task of searching recorded videos, off-line, for events that are known to have happened, requires the aid of automated tools for video retrieval and summarization. Another fact is greater ratio of the people in the world live in urban areas [1][2][3]. Hence, automated detection of unwanted generated by self-organization phenomena resulting from the interactions of many individuals, can cause significant hindrance in the social interactions [4]. Therefore, it is important to provide modern surveillance, possibly in lieu of, or as an assistance to, human operators. However, there is a lack of empirical studies of unwanted events where besides basic motion segmentation, also the analysis of unstructured behaviors, such as violence, or unacceptable social interaction, is decisive for the safety of people during, for example, gatherings of people in different public places, or in situations of social excitement (winning of a sport). These conditions can possibly trigger violence arising from the maximum density and irregular flow of people [5][6]. Moreover, the behavior of the people may transition from one state of collective behavior to a qualitatively different behavior depending on the current situation. Such transitions typically occur when individuals in the public accumulate, propagate, or non-uniformly move with the flow [7][8]. Activity analysis and scene understanding entail object detection, tracking and activity recognition [9]. These approaches, requiring low-level motion features, appearance features, or object trajectories, render good performance in low level gathering of people, but fail in real-world high level gathering of people at public places. Some recent works [10][11][12] presents an offline data-driven approach for crowd videos to learn unwanted patterns by performing long-term analysis. The approach tracks individuals in these gatherings, showing typical and rare behaviors. Other related works [13][14][15] proposed an interdisciplinary framework for the analysis of the social interactions, which integrates benefits of simulation techniques, pedestrian detection and tracking, dense crowd detection and event detection [16][17].

Considering the difficulty in performing detection and tracking in social interactions, the research has focused on gathering the motion information at a higher scale, thus not associating it to single objects, but considering the interaction as a single entity [18][19][20][21]. These approaches often require low-level features such as multi-resolution histograms [22], spatio-temporal volumes [23][24], and appearance or motion descriptors [25][26][46]. In the work [27], the optical flow constraint is exploited to estimate a conditional probability of the spatio-temporal intensity change. Furthermore, motion estimation and segmentation are integrated into a functional minimization strategy based on a Bayesian framework. In [28][29], authors used a mixture of dynamic textures to fit a video sequence and then assigned homogeneous motion regions to the mixture components. However, the methods presented in [30][31][33] are only targeted at addressing the cases of simple motion patterns. In [35][36], motion segmentation is performed without relying on the optical flow. In [37][38][39][40], a dynamic texture model is used to measure the similarity between neighboring spatio-temporal patches. These patches are grouped by connected component analysis, resulting into over segmentation in presence of low level social gatherings. In [41][42][43][44][45], the authors proposed a method to perform multi-target tracking in crowd using time integration of the dynamical system defined by the optical flow.

2. VIOLENT SOCIAL INTERACTION

We consider a robust violence social interaction algorithm inspired from [46] with an effective and adaptive appearance model. Within the proposed algorithm, the collaboration of the generative model and the discriminative classifier leads to a more flexible and robust likelihood function to verify the state predictions. Our method is adaptively updated with consideration of occlusions to account for appearance variations and alleviate drifts.

Our method is formulated within the Bayesian filtering framework (proposed by [46]) in which the goal is to determine a posteriori probability, of the target state by

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})d\mathbf{x}_{t-1},$$
$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1}),$$

In the equation, the object state, and the observation at time t has been mentioned, where lx , ly ,θ, s, α, φ denote x, y translations, rotation angle, scale, aspect ratio, and skew respectively. The assumption has been made that the affine parameters are independent and modeled by six scalar Gaussian distributions. The motion model narrates that the state at t based on the immediate previous state, and the observation model demonstrates the likelihood of observing zt at state xt . The particle filter is an effective way of Bayesian filtering, which finds the state regardless of the underlying distribution. The optimal state is determined by the maximum a posteriori estimation over a set of N samples as formulated in the equation,

$$\hat{\mathbf{x}}_t = \arg_{\mathbf{x}_t^i} \max p(\mathbf{z}_t|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}),$$

Using a particle filter, the samples at frame t can be inferred by a Gaussian function with mean xt−1 and variance σ2 as given in the equation.

$$p(\mathbf{x}_t^i|\mathbf{x}_{t-1}) = G(\mathbf{x}_{t-1}, \sigma^2)$$

The use of huge number of samples would possible improve the violence social interaction at the expense of increasing computational overhead. We investigate the detection result is the MAP estimation over the samples which can be modeled well with mode seeking, and propose a motion model. At time t, the sample set is achieved by the Gaussian function. The template set is composed of different tracking results in the latest set of frames and the template in the first frame. Given the sample set, the sparse coefficients of each template set are calculated according to the equation.

$$\min_{\gamma_j} \|t^j - X\gamma_j\|_2^2 + \lambda_1\|\gamma_j\|_1, \quad \text{s.t. } \gamma_j \geq 0, \quad j = 1, \ldots, m,$$

In the equation, each column of X is a sample at time t and lambda is a weight parameter. The sample set X models an over-complete dictionary, and the sparsity constraints force to identify the samples that are highly correlated to the templates. In fact, the samples that do not formulate the templates well are not considered as good candidates for the detection of social interaction. We investigate that this formulation is different from other methods which need solving 1-minimization problems. On the contrary, our method needs solving 1-minimization problems, therefore reducing the computational complexity to greater extent.

Most methods use rectangular image regions to represent detection, and thus background pixels are inevitably included as part of the foreground elements. Therefore, classifiers based on local representations may be significantly affected when background blocks are considered as positive ones for update. On the other hand, the holistic appearance encoded by a target template is more unique than the local appearance of local blocks. Thus, holistic templates are more effective for discriminative models to isolate foreground elements from the background. In addition, local representations are more useful for generative models because of flexibility. In our method, we present a collaborative observation model that fuses a discriminative classifier based on holistic templates and a generative model using local representations.

Additionally, we used the model of Peng et al. [47] who proposed a low-rank and structured sparse matrix decomposition (LSMD) model. For this purpose, a tree-structured sparsity-inducing norm regularization is firstly investigated to render a hierarchical description of the image structure to ensure the completeness of the extracted features. The similarity of feature values within the violence social interaction is then provided by the `1-norm. High-level priors are combined to guide the matrix decomposition and enhance the detection of violence.

Tree structure is widely existed and explored in natural image processing, e.g., tree-structured wavelet transforms, tree-based image segmentation. Recent advances in the sparse representation research also exploit tree structure to pursuit the structured sparsity in terms of relationships between different patterns. The work considers a tree structured sparsity-inducing norm, which essentially is a hierarchical group sparsity, to represent the underlying structure of violence in feature space.

The step is to build an index tree to represent the scene structure via divisive hierarchical k-means clustering. During the tree build up, for each image patch, we get its position coordinate and feature representation. All patches from an image composite a set of different data points. The divisive hierarchical clustering begins with all the points in a single cluster, and then recursively divides each cluster into k child clusters using k-means algorithm. The recursion ends when all the clusters contain less than k data points. We exploit a quad-tree structure where k = 4.

After getting the feature matrix representation and the corresponding structured index tree of the input scene, we use the proposed LSMD model to divide it into a low-rank component and a structured sparse component. Considering the tree-structured sparsity regularization into the LSMD model, we can combine perceptually similar patches of the foreground elements and throwing away the irrelevant information.

3. EXPERIMENTS

Our method runs on a Windows system. Eight different people were tested performing various kinds of violent and non-violent activities in a set of videos [48]. Dataset was collected with a Sony stationary camera. The results obtained with our method are listed in the table below. Our method successfully detected most of the violent activities. It shows the strong capability of our method to detect violent social interaction.

Name	Total Frames	No. of violent activities	Correct detection
Omar Yun1	510	6	4
Omar Yun2	279	1	1
Yun Omar1	400	2	1
Yun Omar2	239	1	0
Zeeshan Cen1	474	4	3
Zeeshan Cen2	333	4	3
Jaime Jigna1	329	2	2
Jaime Xuan2	408	3	2
Joanna Xu1	281	4	3
Joey Dave1	310	7	6
Joey Dave2	249	2	1
Dave Will 1	130	1	1
Joey Joanna	210	2	1
Kris Rusty1	465	6	3
Kris Rusty2	304	2	1

4. CONCLUSION

We presented a novel method to detect the presence of violent social interaction in videos. For this purpose, we used a robust appearance model with the addition of a low-rank and structured sparse matrix decomposition (LSMD) model to better highlight the presence of violence. Our model is capable of capturing spatio-temporal features which enables the analysis of local motion taking place in the video. The performance of our method is tested on a standard dataset. The evaluation of our method present results revealing the promising capability of our method in detecting violent social interaction.

REFERENCES

[1] Ullah, Habib, Ahmed B. Altamimi, Muhammad Uzair, and Mohib Ullah. "Anomalous entities detection and localization in pedestrian flows." Neurocomputing 290 (2018): 74-86.

[2] Khan, Wilayat, Habib Ullah, Aakash Ahmad, Khalid Sultan, Abdullah J. Alzahrani, Sultan Daud Khan, Mohammad Alhumaid, and Sultan Abdulaziz. "CrashSafe: a formal model for proving crash-safety of Android applications." Human-centric Computing and Information Sciences 8, no. 1 (2018): 21.

[3] Ullah, Habib, Mohib Ullah, and Muhammad Uzair. "A hybrid social influence model for pedestrian motion segmentation." Neural Computing and Applications (2018): 1-17.

[4] Ahmad, Fawad, Asif Khan, Ihtesham Ul Islam, Muhammad Uzair, and Habib Ullah. "Illumination normalization using independent component analysis and filtering." The Imaging Science Journal 65, no. 5 (2017): 308-313.

[5] Ullah, Habib, Muhammad Uzair, Mohib Ullah, Asif Khan, Ayaz Ahmad, and Wilayat Khan. "Density independent hydrodynamics model for crowd coherency detection." Neurocomputing 242 (2017): 28-39.

[6] Khan, Sultan Daud, Muhammad Tayyab, Muhammad Khurram Amin, Akram Nour, Anas Basalamah, Saleh Basalamah, and Sohaib Ahmad Khan. "Towards a Crowd Analytic Framework For Crowd Management in Majid-al-Haram." arXiv preprint arXiv:1709.05952 (2017).

[7] Saqib, Muhammad, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. "Extracting descriptive motion information from crowd scenes." In 2017 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1-6. IEEE, 2017.

[8] Ullah, Mohib, Habib Ullah, Nicola Conci, and Francesco GB De Natale. "Crowd behavior identification." In Image Processing (ICIP), 2016 IEEE International Conference on, pp. 1195-1199. IEEE, 2016.

[9] Khan, S. "Automatic Detection and Computer Vision Analysis of Flow Dynamics and Social Groups in Pedestrian Crowds." (2016).

[10] Arif, Muhammad, Sultan Daud, and Saleh Basalamah. "Counting of people in the extremely dense crowd using genetic algorithm and blobs counting." IAES International Journal of Artificial Intelligence 2, no. 2 (2013): 51.

[11] Ullah, Habib, Mohib Ullah, Hina Afridi, Nicola Conci, and Francesco GB De Natale. "Traffic accident detection through a hydrodynamic lens." In Image Processing (ICIP), 2015 IEEE International Conference on, pp. 2470-2474. IEEE, 2015.

[12] Ullah, Habib. "Crowd Motion Analysis: Segmentation, Anomaly Detection, and Behavior Classification." PhD diss., University of Trento, 2015.

[13] Khan, Sultan D., Stefania Bandini, Saleh Basalamah, and Giuseppe Vizzari. "Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows." Neurocomputing 177 (2016): 543-563.

[14] Shimura, Kenichiro, Sultan Daud Khan, Stefania Bandini, and Katsuhiro Nishinari. "Simulation and Evaluation of Spiral Movement of Pedestrians: Towards the Tawaf Simulator." Journal of Cellular Automata 11, no. 4 (2016).

[15] Khan, Sultan Daud, Giuseppe Vizzari, and Stefania Bandini. "A Computer Vision Tool Set for Innovative Elder Pedestrians Aware Crowd Management Support Systems." In AI* AAL@ AI* IA, pp. 75-91. 2016.

[16] Saqib, Muhammad, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. "A study on detecting drones using deep convolutional neural networks." In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-5. IEEE, 2017.

[17] Khan, Sultan Daud, Giuseppe Vizzari, Stefania Bandini, and Saleh Basalamah. "Detection of social groups in pedestrian crowds using computer vision." In International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 249-260. Springer, Cham, 2015.

[18] Khan, Sultan Daud, Fabio Porta, Giuseppe Vizzari, and Stefania Bandini. "Estimating Speeds of Pedestrians in Real-World Using Computer Vision." In International Conference on Cellular Automata, pp. 526-535. Springer, Cham, 2014.

[19] Khan, Sultan D., Luca Crociani, and Giuseppe Vizzari. "Integrated Analysis and Synthesis of Pedestrian Dynamics: First Results in a Real World Case Study." From Objects to Agents (2013).

[20] Khan, Sultan D., Luca Crociani, and Giuseppe Vizzari. "PEDESTRIAN AND CROWD STUDIES: TOWARDS THE INTEGRATION OF AUTOMATED ANALYSIS AND SYNTHESIS."

[21] Ullah, Habib, Mohib Ullah, and Nicola Conci. "Dominant motion analysis in regular and irregular crowd scenes." In International Workshop on Human Behavior Understanding, pp. 62-72. Springer, Cham, 2014.

[22] Saqib, Muhammad, Sultan Daud Khan, and Michael Blumenstein. "Detecting dominant motion patterns in crowds of pedestrians." In Eighth International Conference on Graphic and Image Processing (ICGIP 2016), vol. 10225, p. 102251L. International Society for Optics and Photonics, 2017.

[23] Ullah, Habib, Mohib Ullah, and Nicola Conci. "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, 2014.

[24] Ullah, Habib, Lorenza Tenuti, and Nicola Conci. "Gaussian mixtures for anomaly detection in crowded scenes." In Video Surveillance and Transportation Imaging Applications, vol. 8663, p. 866303. International Society for Optics and Photonics, 2013.

[25] Rota, Paolo, Habib Ullah, Nicola Conci, Nicu Sebe, and Francesco GB De Natale. "Particles cross-influence for entity grouping." In Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European, pp. 1-5. IEEE, 2013.

[26] Ullah, Habib, and Nicola Conci. "Structured learning for crowd motion segmentation." In Image Processing (ICIP), 2013 20th IEEE International Conference on, pp. 824-828. IEEE, 2013.

[27] Ullah, Habib, and Nicola Conci. "Crowd motion segmentation and anomaly detection via multi-label optimization." In ICPR workshop on Pattern Recognition and Crowd Analysis. 2012.

[28] Khan, Wilayat, and Habib Ullah. "Authentication and Secure Communication in GSM, GPRS, and UMTS Using Asymmetric Cryptography." International Journal of Computer Science Issues (IJCSI) 7, no. 3 (2010): 10.

[29] Ullah, Habib, Mohib Ullah, Muhammad Uzair, and F. Rehman. "Comparative study: The evaluation of shadow detection methods." International Journal Of Video & Image Processing And Network Security (IJVIPNS) 10, no. 2 (2010): 1-7.

[30] Khan, Wilayat, and Habib Ullah. "Scientific Reasoning: A Solution to the Problem of Induction." International Journal of Basic & Applied Sciences 10, no. 3 (2010): 58-62.

[31] Uzair, Muhammad, Waqas Khan, Habib Ullah, and Fasih-ur Rehman. "Background modeling using corner features: An effective approach." In Multitopic Conference, 2009. INMIC 2009. IEEE 13th International, pp. 1-5. IEEE, 2009.

[32] Ullah, Mohib, Habib Ullah, and Ibrahim M. Alseadoon. "HUMAN ACTION RECOGNITION IN VIDEOS USING STABLE FEATURES."

[33] Khan, Wilayat, Habib Ullah, and Riaz Hussain. "Energy-Efficient Mutual Authentication Protocol for Handhled Devices Based on Public Key Cryptography." International Journal of Computer Theory and Engineering 5, no. 5 (2013): 754.

[34] Arif, Muhammad, Sultan Daud, and Saleh Basalamah. "People counting in extremely dense crowd using blob size optimization." Life Science Journal 9, no. 3 (2012): 1663-1673.

[35] Saqib, Muhammad, S. D. Khan, and S. M. Basalamah. "Vehicle Speed Estimation using Wireless Sensor Network." In INFOCOMP 2011 First International Conference on Advanced Communications and Computation, IARIA. 2011.

[36] Khan, Sultan Daud. "Estimating Speeds and Directions of Pedestrians in Real-Time Videos: A solution to Road-Safety Problem." In CEUR Workshop Proceedings, p. 1122. 2014.

[37] Khan, Sultan Daud, and Hyunchul Shin. "Effective memory access optimization by memory delay modeling, memory allocation, and buffer allocation." In SoC Design Conference (ISOCC), 2009 International, pp. 153-156. IEEE, 2009.

[38] Khan, Sultan Daud, Giuseppe Vizzari, and Stefania Bandini. "Facing Needs and Requirements of Crowd Modelling: Towards a Dedicated Computer Vision Toolset." In Traffic and Granular Flow'15, pp. 377-384. Springer, Cham, 2016.

[39] Saqib, Muhammad, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. "Person Head Detection in Multiple Scales Using Deep Convolutional Neural Networks." In 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-7. IEEE, 2018.

[40] Ullah, Mohib, and Faouzi Alaya Cheikh. "Deep Feature Based End-to-End Transportation Network for Multi-Target Tracking." In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3738-3742. IEEE, 2018.

[41] Ullah, Mohib, Mohammed Ahmed Kedir, and Faouzi Alaya Cheikh. "Hand-Crafted vs Deep Features: A Quantitative Study of Pedestrian Appearance Model." In 2018 Colour and Visual Computing Symposium (CVCS), pp. 1-6. IEEE, 2018.

[42] Ullah, Mohib, Ahmed Mohammed, and Faouzi Alaya Cheikh. "PedNet: A Spatio-Temporal Deep Convolutional Neural Network for Pedestrian Segmentation." Journal of Imaging 4, no. 9 (2018): 107.

[43] Ullah, Mohib, and Faouzi Alaya Cheikh. "A Directed Sparse Graphical Model for Multi-Target Tracking." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1816-1823. 2018.

[44] Ullah, Mohib, Ahmed Kedir Mohammed, Faouzi Alaya Cheikh, and Zhaohui Wang. "A hierarchical feature model for multi-target tracking." In Image Processing (ICIP), 2017 IEEE International Conference on, pp. 2612-2616. IEEE, 2017.

[45] Ullah, Mohib, Faouzi Alaya Cheikh, and Ali Shariq Imran. "Hog based real-time multi-target tracking in bayesian framework." In 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 416-422. IEEE, 2016.

[46] Zhong, W., Lu, H., & Yang, M. H. (2014). Robust object tracking via sparse collaborative appearance model. IEEE Transactions on Image Processing, 23(5), 2356-2368.

[47] Peng, H., Li, B., Ji, R., Hu, W., Xiong, W., & Lang, C. (2013, July). Salient Object Detection via Low-Rank and Structured Sparse Matrix Decomposition. In AAAI (pp. 796-802).

[48] Datta, A., Shah, M., & Lobo, N. D. V. (2002). Person-on-person violence detection in video data. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 1, pp. 433-438). IEEE.