Graph-based Crowds Modeling for Video Surveillance

Duan-Yu Chen and Po-Chung Huang
Department of Electrical Engineering, Yuan-Ze University
dychen@saturn.yzu.edu.tw, *pochunghung26@hotmail.com*

***Abstract*— Modeling human crowds is an important issue for video surveillance and is a challenging task due to their nature of non-rigid shapes. In this paper, for real time constraint, Haar-like features are first employed to approximately locate the position of an isolated region that comprise an individual person or a set of occluded persons. Each isolated region is considered a vertex and a human crowd is thus modeled by a graph. To regularly construct a graph, Delaunay triangulation is used to systematically connect vertices and therefore the problem of event detection of human crowds is formulated as measuring the topology variation of consecutive graphs in temporal order. To effectively model the topology variation, local characteristics such as triangle deformations and subgraph analysis, and global features such as moments are all computed and finally combined as an indicator to detect if any anomalies of human crowd(s) present in the scene. Experimental results obtained by using extensive dataset show that our system is effective in detecting anomalous events for uncontrolled environment of surveillance videos.**

I. INTRODUCTION

Nowadays to prevent criminal behaviors or traffic accidents, video surveillance systems have become more and more popular in many public places, such as airports, train stations, critical intersections, etc. People are usually the main objects of interest in surveillance tasks. In public places, pedestrians usually move towards some orientation(s) and thus form some crowd(s). In the literature, Andrade and Fisher [1] used optical flows to describe crowds and extract features from optical flows in an unsupervised manner. Ihaddadene and Djeraba [2] proposed a motion-based approach that can estimate the crowd density and detect abnormal events based on that without training. It also considers simultaneously density, direction and velocity and focuses analysis on specific regions where the density of motions is high. Cheriyadat and Radke [5] present a survey of computer vision algorithms that deal with crowds of people and review model-based crowd analysis algorithms, in which some type of human model is applied to segmentation or tracking. To detect typical motion patterns in crowded scenarios, Hu et al. [6] which utilizes the instantaneous motions of a video. The motion flow field is a union of independent flow vectors computed in different frames. Detecting motion patterns in this flow field can therefore be formulated as a clustering problem of the motion flow fields, where each motion pattern consists of a group of flow vectors participating in the same process or motion. Ryan et al. [3] proposed a scene independent approach that can count the number of people in crowded scenarios without training. A "global scaling factor" is used to compensate for camera angle and distance, based on a reference person in each scene. Chan et al. [4] developed a privacy-preserving system to estimate the attributes of inhomogeneous crowds. The approach does not depend on object detection or feature tracking. Jiang et al. [7] employed the concept of contextual anomaly for crowd analysis. Their system follows an unsupervised approach. It automatically discovers important contextual information from the crowd video and detects the blobs corresponding to contextually anomalous behaviors. Wang et al. [8] detect some activities and interactions between individuals in a crowded scene using their proposed unsupervised learning framework. Under their framework, hierarchical Bayesian models are used to connect three elements in visual surveillance: low-level visual features, simple "atomic" activities, and interactions. Mehran et al. [9] used social force model to detect and localize unusual events in a crowd of people. Ma and Cisar [10] detect crowd events using dynamic texture descriptor. The dynamic texture descriptor is an extension of the local binary patterns. The image sequences are divided into regions. A flow is formed based on the similarity of the dynamic texture descriptors on the regions. Wu et al. [11] proposed a learning-based approach that can analyze textures in a crowd and then detect abnormal crowd density. By using the perspective projection model, a series of multi-resolution image cells are generated to make better density estimation in the crowded scene.

Most related works focus on the analysis of activities induced by some individuals or interactions between them. However, the events induced by the interactions within crowds are also extremely important since the abnormal behaviors of crowds would usually result in or represent some massive damages. To solve this problem, Haar-like features are first employed to approximately locate the position of an isolated region that comprise an individual person or a set of occluded persons. Each isolated region is considered a vertex and a human crowd is thus modeled by a graph. To regularly construct a graph, Delaunay triangulation is used to systematically connect vertices and therefore the problem of event detection of human crowds is formulated as measuring the topology variation of consecutive graphs in temporal order. To effectively model the topology variation, local characteristics such as triangle deformations, and global features such as moments are all computed and finally combined as an indicator to detect if any anomalies of human crowd(s) present in the scene.

The remainder of this paper is organized as follows. Section II illustrates the modeling and analysis of human crowds based on variation of graph topology. Section III shows the experimental results and Section IV gives some concluding remarks.

II. CROWDS MODELING AND MATCHING

Haar-like features [12] are first employed to approximately locate the position of an isolated region that comprise an individual person or a set of occluded persons. Each isolated region is considered a vertex and a human crowd is thus modeled by a graph. To regularly construct a graph, Delaunay triangulation is used to systematically connect vertices. For detecting abnormal events caused by the variation of graph topology, the constructed graph for a human crowd is first described by using adjacency matrix [13-14]. Since the adjacency matrix $\boldsymbol{A_G}$ is a real symmetric matrix, the principal axis theorem and Prerron-Frobenius theorem [15-16] are used for characterizing the graph. The adjacency matrix of a graph is indecomposable precisely when the graph is connected. Graph $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$ is connected since it is constructed by Delaunay triangulation that none of $\boldsymbol{V}$ is disconnected.

Principal axis theorem: *If* **A** *is a real symmetric matrix of n, then* **A** *has n real eigenvalues and a corresponding orthonormal set of eigenvectors.*

Prerron-Frobenius theorem: *If* **A** *is a non-negative matrix with eigenvalues* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ *, then* $|\lambda_1| \geq |\lambda_k|$ *, for k=1,2,…,n, and the eigenvalue* λ_1 *has an eigenvector with all entries non-negative. If* **A** *is indecomposable, then the eigenvalue* λ_1 *is simple* $(\lambda_1 \geq \lambda_2)$*, and the eigenvector has all entries posititve.*

Without loss of generality, let $\boldsymbol{G}$ and $\boldsymbol{H}$ be two constructed graphs with the relations $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. $\boldsymbol{H}$ is then a subgraph of $\boldsymbol{G}$. If $\boldsymbol{G}$ and $\boldsymbol{H}$ satisfy the conditions $V(H) = V(G)$ and $E(H) \subseteq E(G)$, $\boldsymbol{H}$ is named a spanning subgraph of $\boldsymbol{G}$. The subgraph induced by non-empty set $\boldsymbol{S}$ of vertices in $\boldsymbol{G}$ is that subgraph H with vertex-set $\boldsymbol{S}$ whose edge-set consist of those edges of $\boldsymbol{G}$ that join two vertices in *S*; it is denoted by $\boldsymbol{G}(\boldsymbol{S})$. A subgraph $\boldsymbol{H}$ of $\boldsymbol{G}$ is induced if $H = \langle V(H) \rangle$. In order to effectively recognize the subgraph relationship between graph $\boldsymbol{G}$ and $\boldsymbol{H}$, the Interlacing Theorem is adopted to achieve this aim. This theorem is effective since it analyzed a graph based on the principal components by graph eigenvalues. A principal sub-matrix corresponds to an indced subgraph with one fewer vertices.

Interlacing theorem: *Let* **A** *be a real symmetric matrix with eigenvalues* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$*, and let* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n-1}$ *be the eigenvalues of a principal submatrix of* **A**. *Then* $\lambda_i \geq \mu_i \geq \lambda_{i+1}$*, for i*=1,2,…,n-1.

Based on the corollary of induced subgraph in [14], we can recognize if the subgraph relationship exists between two graphs, i.e., human crowds present in consecutive frames. Subgraph analysis is important since the movement of human crowds would usually be smooth under normal cases. That means that a graph at time *t* with slight changes of one or few vertices present additionally or absent would still be a subgraph of graph at time *t*-1.

After Delaunay triangulation process, a triangle set Δ_t of $G_t(V,E)$ defined as $\Delta_t = \{\delta_1^t, \delta_2^t, ...\delta_k^t \mid \delta_i^t = (v_{i1}, v_{i2}, v_{i3}), i = 1 \sim k\}$ can be obtained, where *k* is the number of triangles of $G_t(V,E)$ and $v_{ij} \in V(G_t), j = 1 \sim 3$ are the three vertices of the triangle. Let Δ_{t-1} and Δ_t be the triangle set of graph $G_{t-1}(V,E)$ and $G_t(V,E)$, respectively. The centroid of the triangle Δ_{t-1} and Δ_t is defined as M_i^{t-1} and M_j^t respectively. Let $\mathrm{d}(M_i^{t-1}, M_j^t)$ be the distance between M_i^{t-1} and M_j^t. The triangle-pair $\delta_{\mathrm{p}}^{t-1}$ and δ_{q}^t that is of the shortest distance can thus be defined as

$$(\delta_{\mathrm{p}}^{t-1}, \delta_{\mathrm{q}}^t) = \arg\min(\mathrm{d}(M_i^{t-1}, M_j^t)). \quad (1)$$

Therefore, the similarity between $\delta_{\mathrm{p}}^{t-1}$ and δ_{q}^t is:

$$S_{p,q}^t = \left(1 - \frac{\left|e_{\mathrm{p12}}^{\mathrm{t\text{-}1}} - e_{\mathrm{q12}}^{\mathrm{t}}\right|}{\min(e_{\mathrm{p12}}^{\mathrm{t\text{-}1}}, e_{\mathrm{q12}}^{\mathrm{t}})}\right)\left(1 - \frac{\left|e_{\mathrm{p23}}^{\mathrm{t\text{-}1}} - e_{\mathrm{q23}}^{\mathrm{t}}\right|}{\min(e_{\mathrm{p23}}^{\mathrm{t\text{-}1}}, e_{\mathrm{q23}}^{\mathrm{t}})}\right)\left(1 - \frac{\left|e_{\mathrm{p31}}^{\mathrm{t\text{-}1}} - e_{\mathrm{q31}}^{\mathrm{t}}\right|}{\min(e_{\mathrm{p31}}^{\mathrm{t\text{-}1}}, e_{\mathrm{q31}}^{\mathrm{t}})}\right). \quad (2)$$

where $(e_{\mathrm{p12}}^{t-1}, e_{\mathrm{p23}}^{t-1}, e_{\mathrm{p31}}^{t-1})$ and $(e_{\mathrm{q12}}^{t}, e_{\mathrm{q23}}^{t}, e_{\mathrm{q31}}^{t})$ are the set of edges for the triangles $\delta_{\mathrm{p}}^{t-1}$ and δ_{q}^t, respectively.

The cost of the alignment of miss-matched triangles should be minimized to make the alignment meaningful. Let the set of miss-matched triangles at time *t*-1 be $\Psi_{t\text{-}1}$, where $\Psi_{t\text{-}1} = \{\phi_1^{t\text{-}1}, \phi_2^{t\text{-}1}, \phi_3^{t\text{-}1}, ..., \phi_\alpha^{t\text{-}1} \mid \phi_i^{t\text{-}1} = (\theta_{i1}, \theta_{i2}, \theta_{i3})\}$, $\Psi_{t\text{-}1} \subseteq \Delta_{t\text{-}1}$, α is the number of miss-matched triangles, and $(\theta_{i1}, \theta_{i2}, \theta_{i3})$ is the set of internal angles of the triangle $\phi_i^{t\text{-}1}$. Similarly, the set of miss-matched triangles at time t is Ψ_t, where $\Psi_t = \{\phi_1^t, \phi_2^t, \phi_3^t, ..., \phi_\beta^t \mid \phi_i^t = (\theta_{i1}, \theta_{i2}, \theta_{i3})\}$, $\Psi_t \subseteq \Delta_t$ and β is the number of triangles. Without loss of generality, the optimal matching for cost $\boldsymbol{C}$ is defined as the matching between subsets ζ_{t-1} and ζ_t :

$$(\zeta_{t-1}, \zeta_t) = \arg\min(C), \quad (3)$$

where the number of triangles in ζ_{t-1} and ζ_t is ω and τ respectively, and $\omega \leq \alpha \leq k_{t-1}$, $\tau \leq \beta \leq k_t$, $\zeta_{t-1} \subseteq \Psi_{t\text{-}1} \subseteq \Delta_{t\text{-}1}$, and $\zeta_t \subseteq \Psi_t \subseteq \Delta_t$. The convex hulls constructed from ζ_{t-1} and ζ_t are denoted by Γ_{t-1} and Γ_t :

$$\Gamma_{t-1} = \{\gamma_1^{t-1}, \gamma_2^{t-1}, \gamma_3^{t-1}, ..., \gamma_\rho^{t-1} \mid \gamma_i^{t-1} = (X_i, Y_i, \theta_i), \gamma_i^{t-1} \in V(G_{t-1})\},$$

$$\Gamma_t = \{\gamma_1^t, \gamma_2^t, \gamma_3^t, ..., \gamma_\varepsilon^{t-1} \mid \gamma_i^t = (X_i, Y_i, \theta_i), \gamma_i^t \in V(G_t)\}$$

where ρ and ε are the number vertices of Γ_{t-1} and Γ_t, X_i, Y_i is the X and Y coordinate respectively , θ_i is the internal angle of vertex γ_i , $\Gamma_{t-1} \subseteq V(G_{t-1})$, and $\Gamma_t \subseteq V(G_t)$. The difference between Γ_{t-1} and Γ_t is first computed based on their internal angles as defined by

$$diff = \sum_{i=1}^{\min(\rho,\varepsilon)} (|\theta_i^{t-1} - \theta_i^t|) + \begin{cases} \sum_{j=\min(\rho,\varepsilon)+1}^{\varepsilon} \theta_j^t & \text{if } \rho < \varepsilon \\ \sum_{j=\min(\rho,\varepsilon)+1}^{\rho} \theta_j^{t-1} & \text{if } \varepsilon < \rho \end{cases}. \quad (4)$$

However, without the alignment of the convex hulls, the exact vertex alignment is missing. Therefore, the order of vertices can be determined by the x and y coordinates, and thus the respective cost C_X and C_Y can then be obtained by

$$C_X = \frac{diff_x + |\rho - \varepsilon| * \Theta + (\alpha - \omega) * \Theta + (\beta - \tau) * \Theta}{\min(\rho, \varepsilon)}, \quad (5)$$

$$C_Y = \frac{diff_y + |\rho - \varepsilon| * \Theta + (\alpha - \omega) * \Theta + (\beta - \tau) * \Theta}{\min(\rho, \varepsilon)}, \quad (6)$$

where $diff_x$ and $diff_y$ denote the angle difference between Γ_{t-1} and Γ_t according to the order of *x*-coordinate and y-coordinate, respectively. Θ is the penalty for miss-alignment between ζ_{t-1} and ζ_t. Finally, the matching cost $\boldsymbol{C}$ based on triangles is the combination of C_X and C_Y are defined by

$$C = 0.5 * C_X + 0.5 * C_Y. \quad (7)$$

Considering global features, moment invariants are effective features for shape analysis [17]. For contour comparison, we can integrate over all of the pixels of the contour. In general, we define the (p, q) moment of a contour as

$$m_{p,q} = \sum_x \sum_y I(x,y) x^p y^q, \quad (8)$$

where $I(x, y)$ is intensity of pixel in coordinate (x, y). A central moment is basically the same as the moments just described except that the values of x and y used in the formulas are displaced by the mean values:

$$\mu_{p,q} = \sum_x \sum_y I(x,y)(x - x_{\text{avg}})^p (y - y_{\text{avg}})^q, \quad (9)$$

where $x_{\text{avg}} = m_{10}/m_{00}$ and $y_{\text{avg}} = m_{01}/m_{00}$. The normalized moments are the same as the central moments except that they are all divided by an appropriate power of m_{00}:

$$\eta_{p,q} = \frac{\mu_{p,q}}{m_{00}^{(p+q)/2+1}}. \quad (10)$$

We then combine Hu's invariant moments by the linear combinations of the central moments. Naturally, with Hu moments we'd like to compare two areas enclosing by convex hulls and determine whether they are similar as follows:

$$M(H_{t-1}, H_t) = \frac{\sum_{i=1}^{7} \left| \frac{m_i^{H_{t-1}} - m_i^{H_t}}{m_i^{H_{t-1}}} \right|}{7}. \quad (11)$$

$m_i^{H_{t-1}}$ and $m_i^{H_t}$ are defined as:

$$\begin{aligned} m_i^{H_{t-1}} &= \text{sign}(h_i^{H_{t-1}}) \cdot \log \left| h_i^{H_{t-1}} \right|, \\ m_i^{H_t} &= \text{sign}(h_i^{H_t}) \cdot \log \left| h_i^{H_t} \right| \end{aligned} \quad (12)$$

where $h_i^{c_{t-1}}$ and $h_i^{c_t}$ are the Hu moments of H_{t-1} and H_t, respectively. H_{t-1} and H_t are the contour of G_{t-1} and G_t. From Eq.(14), the smaller the value of $M(H_{t-1}, H_t)$, the more similar between G_{t-1} and G_t.

To detect if an abnormal event appearing in consecutive frames, a multi-cue measure, say P_E, by combining the properties mentioned above is defined as

$$P_E = \max(1 - P_s, P_n) \times [\xi \times P_m + (1 - \xi) P_t], \quad (13)$$

where ξ is a pre-defined weight for balancing the global cost of moment P_m and the local cost of triangle matching P_t. P_s is the probability that two graphs are of the subgraph relationship and is defined as

$$P_s = \frac{R}{\min(n_{t-1}, n_t)}, \quad (14)$$

where R is the number of eigenvalues of $G_t(V,E)$ and $G_{t-1}(V,E)$ that satisfies Corollary 3.2.2. When the value of P_s is closer to 1, that means graphs $G_t(V,E)$ and $G_{t-1}(V,E)$ are more similar. P_n is a probability to measure the number nodes that are miss-aligned and is defined by

$$P_n = 1 - \frac{Q}{\max(k_{t-1}, k_t)}, \quad (15)$$

where Q denotes the number of matched triangles for graphs $G_t(V,E)$ and $G_{t-1}(V,E)$. Fig.2 shows some examples of matched triangles that are of red or blue colors. In Eq.(15), we select the maximum value among $1 - P_s$ and P_n since graphs that are of the subgraph relationship cannot always be of high similarity even though the value P_s evaluated from them is 1. Finally, considering the temporal consistency, a temporal sliding window with width γ (γ is set to 5 empirically) is used for accumulating the value P_E . An abnormal event is detected if the accumulated value is larger than a pre-defined threshold.

III. Experimental Results

For performance comparison, we use the UMN dataset [18] to conduct the event detection process and compare to the social force model based approach (SFM) in [9] based on the detected time instant that the unusual event begins. The movement of human crowds in the dataset can be classified into two categories. The first is the gradual movement of the crowd initially in the center of the camera of view, abruptly moves to the right and then disappears. The second is the gradual movement of the crowd initially and then moves in an explosive manner. The video clip of scenario 1 consists of 658 frames and that of scenario 2 consists of 231 frames. Two clips are of the resolution 320×240. In Fig.1, Our system performance is compared with the social force model (SFM) based approach in [11]. The actual ground truth of the beginning of the unusual event is in the frame 570. Our approach outperforms SFM since the unusual event can be detected exactly in frame 564 that is earlier than that detected in frame 594 by using SFM.

Fig. 1. Our system performance is compared with the social force model (SFM) based approach in [9]. The actual ground truth of the beginning of the unusual event is in the frame 570. Our approach outperforms SFM since the unusual event can be detected exactly in frame 564 that is earlier than that detected in frame 594 by using SFM. (The green bar denotes usual status and the red one represents unusual status).

In Fig. 2, the actual ground truth of the beginning of the unusual event is in the frame 181. Our approach outperforms SFM since the unusual event can be detected exactly in frame186 that is earlier than that detected in frame 198 by using SFM. On average, we can achieve around 92% detection rate and thus it shows the efficacy of our proposed approach.

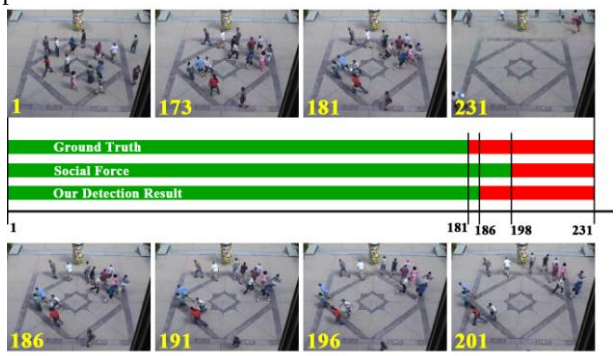

Fig. 2. The actual ground truth of the beginning of the unusual event is in the frame 181. Our approach outperforms SFM since the unusual event can be detected exactly in frame186 that is earlier than that detected in frame 198 by using SFM. (The green bar denotes usual status and the red one represents unusual status).

IV. Conclusions

In this paper, for real time constraint, Haar-like features are first employed to approximately locate the position of an isolated region that comprise an individual person or a set of occluded persons. Each isolated region is considered a vertex and a human crowd is thus modeled by a graph. To regularly construct a graph, Delaunay triangulation is used to systematically connect vertices and therefore the problem of event detection of human crowds is formulated as measuring the topology variation of consecutive graphs in temporal order. To effectively model the topology variation, local characteristics such as triangle deformations and eigenvalue-based subgraph analysis, and global features such as moments are all computed and finally combined as an indicator to detect if any anomalies of human crowd(s) present in the scene. Experimental results obtained by using extensive dataset have shown that the detection rate of abnormal events is about 91% and thus our system is effective in detecting anomalous events for uncontrolled environment of surveillance videos.

References

[1] E. Andrade, and R. Fisher, "Hidden Markov models for optical flow analysis in crowds," Proc. *International Conference on Pattern Recognition*, vol. 01, pp. 460–463, 2006.

[2] N. Ihaddadene and C. Djeraba. "Real-time Crowd Motion Analysis," *International Conference on Pattern Recognition*, Dec. 2008 pp.1 – 4.

[3] D. Ryan, S. Denman, and C. Fookes, S. Sridharan, "Scene Invariant Crowd Counting for Real-Time Surveillance," Proc. *IEEE ICSPCS*, 2008

[4] A. B. Chan, Z.-S. John Liang,and Nuno Vasconcelos," Privacy Preserving Crowd Monitoring: Counting People without People Models or Tracking" *Computer Vision and Pattern Recognition*, 2008.

[5] A.M. Cheriyadat and R.J. Radke, "Detecting dominant motions in dense crowds," *IEEE Journal of Selected Topics in Signal Process.*, vol. 2, no. 4, pp. 568–581,Aug. 2008.

[6] M. Hu, S. Ali, and M. Shah, "Learning motion patterns in crowded scenes using motion flow field," in Proc.*IEEE Int'l Conf. on Pattern Recognition*, Dec. 2008, pp.1–5.

[7] F. Jiang, Y. Wu, and A. K. Katsaggelos, "Detecting contextual anomalies of crowd motion in surveillance video" Proc. *IEEE ICIP*, 2009.

[8] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555,Mar. 2008.

[9] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," *IEEE International Conference on Computer Vision and Pattern Recongition* (CVPR), Miami, 2009.

[10] Y. Ma, and P. Cisar, "Event Detection Using Local Binary Pattern Based Dynamic Textures" cvprw, pp.38-44, 2009 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2009.

[11] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd Density Estimation Using Texture Analysis and Learning" robio, pp.214-219, 2006 *IEEE International Conference on Robotics and Biomimetics*, 2006

[12] P. Viola and M. Jones, "Robust Real-Time Face Detection", I*nternational Journal of Computer Vision*, 2004.

[13] J. L. Gross, and J. Yellen, "Graph Theory and Its Applications (2nd ed.), ", 2006.

[14] L. W. Beineke and R. J. Wilson, "Topics in Algebraic Graph Theory," 2004.

[15] F. R. Gantmacher, "The Theory of Matrices," Chelsea, 1959.

[16] H. Minc and M. Marcus, "A Survey of Matrix Theory and Matrix Inequalities," Prindle, Weber & Schmidt, 1964.

[17] M. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory 8* (1962), pp. 179-187.

[18] University of Minnesota - Crowd Activity Dataset, http://mha.cs.umn.edu/Movies/Crowd- Activity-All.avi.