Probabilistic Detection of Crowd Events on Riemannian Manifolds

Aravinda S. Rao, Jayavardhana Gubbi, Slaven Marusic, and Marimuthu Palaniswami
ISSNIP, Department of Electrical and Electronic Engineering, The University of Melbourne,
Parkville Campus, VIC - 3010, Australia.
Email: aravinda@student.unimelb.edu.au, {jgl, slaven, palani}@unimelb.edu.au

***Abstract*—Event detection in crowded scenarios becomes complex due to articulated human movements, occlusions and complexities involved in tracking individual humans. In this work, we focus on crowd event (activity) detection and classification. We focus on active crowd (continuously moving crowd) events. First, event primitives such as motion, action, activity and behaviour are defined. Furthermore, a distinction is made among event detection, action recognition and abnormal event detection. Further, event detection and classification are defined on Riemannian Manifolds that yields six different probabilities of the event occurring. Using a new probabilistic approach, an automated event detection algorithm is proposed that temporally segments the event using a novel framework. The results indicate that the proposed approach delivers superior performance in selected cases and similar results in other cases, whilst the detection model delay allows operation in near real-time. The Performance Evaluation of Tracking and Surveillance (PETS) 2009 dataset was used for evaluation. Existing crowd event detection approaches used supervised approach, whereas we eschew semi-supervised approach.**

I. Introduction

Analyzing events in videos is highly informative in learning behavioral characteristics of the objects. However, detecting and predicting events in videos is both exacting and challenging. Individual object detection and tracking in itself is a challenging task in multi-object scenarios and the difficulty further escalates during crowded scenarios. In particular, event detection in crowded scenarios becomes complex when faced with articulated human movements and occlusions [1]. The primary objective of the event analysis is to localize the events in space and time. Event detection can thus be summarized to recognizing and detecting those patterns from the video data in conjunction with object detection. This work is aimed at detecting and classifying the crowd events (as shown in Figure 1).

Human activity recognition has gained much importance in recent times especially at locations where people go about their daily activities (e.g. shopping malls), transits (e.g. airports), public gatherings (e.g. sports or music events) and so on. There has been a surge in video event detection applications in response to voluminous usage, distribution and sharing of videos (e.g. YouTube, Vimeo, VEVO and so on). Context-based search, content-based video retrieval, video event labeling and video surveillance are some of these associated applications of event detection.

There has been ongoing research on abnormality/anomaly detection, where the system endeavors to classify the event as normal or abnormal. Some of the examples of automated detection of abnormal/anomalous/unusual events can be found in [3], [4], [5], [6]. In [7], one can find a comprehensive study of vision-based anomaly detection methods. Anomaly detection, in general, operates on temporal domain data to identify the events.

(a) Walking (b) Running (c) Crowd formation (merging) (d) Splitting (e) Local dispersion (f) Rapid dispersion (evacuation)

Fig. 1. Examples of crowd events - walking, running, crowd formation (merging), splitting, local dispersion and rapid dispersion (evacuation) from PETS 2009 [2] dataset

On the other hand, human action recognition require understanding of both spatial and temporal characteristics. Initial reviews of human action recognition can be found in [8], [9]. A three-level hierarchical taxonomy based on object detection, tracking and activity recognition was proposed in [10]. A survey by Moeslund *et al.* [11] further elaborated on human detection, tracking and activity recognition. In [12], the focus of the study was more on human motion representation. In

literature, the terms “actions” and “activities” are used interchangeably. Turaga *et al.* [13] defined the contexts in which these terms are used. In [14], Aggarwal and Ryoo categorized the human activities as gestures, actions, interactions and group activities.

In this work, we focus on crowd event (activity) detection and classification. The crowd events targeted are: running, walking, crowd formation (merging), splitting, local dispersion and rapid dispersion (evacuation)—(as shown in Figure 1). Our approach to crowd event detection is different compared to previous works (for which a detailed study is provided in Section II) in that we do not track individual people, instead, we use motion patterns to recognize, detect and provide a probabilistic detection of crowd activities on Riemannian manifolds. Our work is further focussed on the *active crowd* (where the crowd is in motion) as opposed to the *static crowd* (where the crowd movement comes to a halt). Existing crowd event detection approaches used supervised approach, whereas we eschew a semi-supervised approach. Directional derivatives and geodesic distances on Riemannian manifolds are used for probabilistic detection of crowd events. Furthermore, we provide delay in detection of events for selected cases. The main contributions of this work are: (1) use of optical flow features for event detection on Riemannian manifolds, (2) semi-supervised approach to classification of crowd events and (3) probabilistic detection of crowd events without tracking.

Section II first provides the definition of motion, action, activity and behavior, followed by in-depth review of crowd events. Section III provides a brief introduction to Riemannian manifold and the problem formulation for crowd event detection. Section IV develops the methods to probabilistically detection of the crowd events. The information about the dataset and results are provided in Section V followed by discussion and conclusion.

II. Related work

Much work has been conducted in event detection applications, yet only, limited focus has been devoted towards crowd event detection. It is worthwhile to note that event detection in crowded scenarios differs from the crowd event detection where the former refers to detection of events pertaining to individual subjects (such as running, walking etc) in the presence of multiple objects and possibly partially occluded, whereas the latter refers to events associated with crowd (group of people) movements. Additionally, there is lack of clarity among the terms activity, event and behavior—more often than not, these terms are used interchangeably. In this work, we follow the taxonomy laid down by Chaaraoui *et al.* [15], in which motion, action, activity and behavior have been organized in a hierarchy. We consider “event” to be synonymous to “activity” in this taxonomy.

- **Motion** - is considered to be the movement of the actor/objects in the scene. This includes movement of body parts with respect to a spatial location together with displacement of actor from one spatial location to another against time.
- **Action** - is considered as the interaction of the actor with surrounding actors/objects. This can be regarded as extraction of the frame-level semantics (*e.g.* handshaking between two actors at a particular frame).
- **Activity/Event** - is the resulting consequence of spatio-temporal interactions among actors. This can be regarded as extraction of the windowed, temporal-level semantics (*e.g.* detection of start and end timings of handshaking between two actors).
- **Behaviour** - is a high-level representation of activities of actors. This can be accounted as the unconstrained, temporal-level semantics of the actors and its scene (*e.g.* changes in actors’ behaviour before and after handshaking).

To handle emergency events in crowded scenarios, Andrade *et al.* [16], [17], [18] proposed spectral clustering of optical flow as features. An automatic model was extracted by fitting an Hidden Markov Model (HMM) for each of the video segments. Zhang *et al.* [3] used similar approach, but the normal events were first learned and later, using Bayesian framework, the abnormal events were detected. These approaches, although applied to crowd, are indeed determining whether the event is normal or abnormal, but not on classification of crowd events.

Recently, localized spatio-temporal based approaches are seen to be promising in event detection. Using flow matching technique depicted in [19], [20] and combining shapes in a volumetric setting, Ke *et al.* [21], [22] prosed event detection methods in crowded scenarios. Tran *et al.*[23] prosed spatio-temporal search paths in volumetric setting applied to crowded scenarios. However, these approaches are only applicable to action recognition of an individual rather than event detection of the crowd.

Agent-based modeling has been applied in many instances to study the behavior of the interaction of people in behaviourial analysis. Chen *et al.* [24] applied agent-based technique to detect queuing, gathering and dispersion events with the aid of tracking. It incorporates head features, template matching, Kalman filtering and a Support Vector Machine (SVM) for object agent analysis.

Garate *et al.* [25] used reference frame to extract motion information and 2-D Histogram of Gradients (HOG) descriptors as features and are tracked to categorize the crowd events applied on PETS 2009 dataset. Utasi *et al.* [26] combined optical flow with statistical filtering to separate background and also as low-level features for probabilistic event recognition. The event is classified based on evaluating the group membership of the mean probability of low-level features with that of training set. Chan *et al.* [27] utilized the dynamic texture model (generative probabilistic model) was used to detect events considering the sample output from a linear dynamic system as a video.

In a study by Benabbas *et al.* [28], optical flow was used to extract motion patterns and build a direction and magnitude model for crowd event detection. Li *et al.* [29] performed the crowd event detection using the intersection of motion vectors derived from Harris corner point and Kanade-Lucas-Tomasi feature tracking. They classify the events based on the motion vector patterns at local intersection points in the space and membership event voting. It is worth nothing that most of the

methods are supervised and use training data to classify the events.

III. Problem formulation on Riemannian manifold

A topological space $\mathcal{M}$ is regarded as a manifold $\mathcal{M}$ of dimension n or n-Manifold if it is a Hausdorff space, Second Countable, locally Euclidean ($\mathbb{R}^n$) and smooth (differentiable) [30]. The tangent space $T_p\mathcal{M}$ for a point $p \in \mathcal{M}$, can be considered as equivalence class of curves through p. The partial derivatives provide the local coordinates at point p. The $\mathrm{D}_p\gamma(t) : T_p\mathcal{M} \rightarrow T_{\gamma(p)N}$ is a linear map. On a finite-dimensional vector space T_pM, directional derivative on smooth function $\gamma(t)$ such that $\gamma(0) = p$ and $\dot{\gamma}_p(t)$ is the initial velocity vector at $p \in \mathcal{M}$ in direction of unit vector. The union (disjoint) of tangent spaces for all $p \in \mathcal{M}$ forms the tangent bundle. A Riemannian metric g on a smooth manifold $\mathcal{M}$ is a smoothly varying inner-product on each of the tangent space and is given by:

$$g_p : T_p\mathcal{M} \times T_p\mathcal{M} \rightarrow \mathbb{R} \tag{1}$$

A Riemannian manifold $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ together with Riemannian metric g. Video data can be considered as a 6-dimensional manifold with $p = f(r, g, b, x, y, t) \subseteq R^6$ and is assumed to be smooth (differentiable). Then, the video event lies on an embedded 5-d submanifold, $f(r, g, b, x, y)$, parameterized by time.

IV. Methodology

Before we proceed further, the crowd events as defined in PETS 2009 dataset are first defined below:

- **walking** $(\mathcal{W})$- is the *event* where objects move at a particular velocity collectively, which is less than the velocity of the events defined in running. Further, *subevents* are defined such as *standing* (W_s), *slow walking* (W_{sw}) and *fast walking* (W_{fw}) for efficient recognition and detection of events. Therefore, $\mathcal{W} = \{W_s, W_{sw}, W_{fw}\}$.
- **running** $(\mathcal{R})$ - is the *event* where objects take spatial-temporal paths that are faster than those described in walking. Furthermore, *slow running* (R_{sr}) and *fast running* (R_{fr}) are defined as *subevents* of running. Hence, $\mathcal{R} = \{R_{sr}, R_{fr}\}$.
- **crowd formation (merging)** $(\mathcal{F} = \{F_f\})$ - is the event where the spatio-temporal analysis reveals that objects are converging to a single point or multiple points. Additionally, the tendency of objects portraying this phenomenon is categorized under this event.
- **crowd splitting** $(\mathcal{S} = \{S_s\})$ - is the opposite of crowd formation. The objects in the scene would diverge from a single point or from multiple points.
- **local dispersion** $(\mathcal{D} = \{d\})$ - is a conditional event where walking event is recorded in association with crowd splitting.
- **rapid dispersion (evacuation)** $(\mathcal{E} = \{E_e\})$ - a conditional event where the running event is observed in conjunction with crowd splitting.

The crowd events — walking, running, crowd formation, crowd splitting, local dispersion and evacuation — are subdivided into three subsets as $\mathcal{A} = \{\mathcal{W}, \mathcal{R}\}$, $B = \{\mathcal{F}, \mathcal{S}\}$, and $\mathcal{C} = \{\mathcal{D}, \mathcal{E}\}$ and $C = \{(W_s, W_{sw}, W_{fw}, R_{sr}, R_{fr}), (f, S_s), (d, e)\}$ and are brought under a single set $\mathcal{C}$ as:

$$\mathcal{C} = \{\mathcal{A}, \mathcal{B}, \mathcal{C}\} \tag{2}$$
$$= \{(\mathcal{W}, \mathcal{R}), (\mathcal{F}, \mathcal{S}), (\mathcal{D}, \mathcal{E})\} \tag{3}$$
$$= \{(W_s, W_{sw}, W_{fw}, R_{sr}, R_{fr}), (f, S_s), (d, e)\} \tag{4}$$

A. Walking and running events

Intuitively, one of the key distinguishing characteristic of pairwise walking–running events is the velocity vectors. For instance, we can safely assume that when we are walking, we have a gait pattern that generates motion vectors that has a Gaussian distribution $\mathcal{N}(\mu_W, \sigma_W)$; likewise, for running event, the distribution will be $\mathcal{N}(\mu_R, \sigma_R)$, where μ_W and μ_R are the mean length of the optical flow vectors for walking and running events respectively, and σ_W and σ_R are the standard deviations of length of walking and running events accordingly. To support this claim, Fig. 2 shows the variations of the optical flow vectors' length. The standard deviation and the mean of the motion vector lengths' during the walking and running events from the PETS 2009 [2] dataset were computed. It is evident that the mean of the length of optical flow vectors increase for running events. Further statistics such as standard deviation for each of the waveforms indicate that the walking and running events can be distinguished using the length of the optical flow vectors. This forms the fundamental basis for other events as well.

Let $\mathcal{M}$ be a manifold of 5 dimensions. At first, the directional derivatives in the directions of x and y cartesian planes are computed using the standard bases on the 5-D manifold. This is tantamount to computing tangent vectors $T_p\mathcal{M}$ at different points $p \in \mathcal{M}$ on the 5-D manifold. Here, the functional derivatives are assumed to be smooth function and that the optical flow vectors are assigned for 2-D manifold of the 5-D manifolds. Tangent bundle $T\mathcal{M} = \sqcup_{p\in\mathcal{M}} T_p\mathcal{M}$. Using the vectors over the bundle, the temporal variations are used to recognize the walking and running events.

The distribution of the vectors are updated over time with a window measure (history) to compute the probability of the walking and running events. Initially, the distribution is assumed to be uniform and later the system is allowed to evolve as per the new samples. The event is classified as walking or running based on the sorted order of probability distributions. The distribution with highest probability is regarded the current event in the video scene. The model has the constraint of minimum number of samples required to compute the probability to avoid prejudiced output at the initialization stage. The probability of walking and running events are denoted by $\Pr(\mathcal{W})$ and $\Pr(\mathcal{R})$.

B. Merging and splitting events

The key distinguishing character of merging and splitting events is the temporal evolution of the distances between groups of people. As the distance between groups increase, there is a very high likelihood that the people in the scene

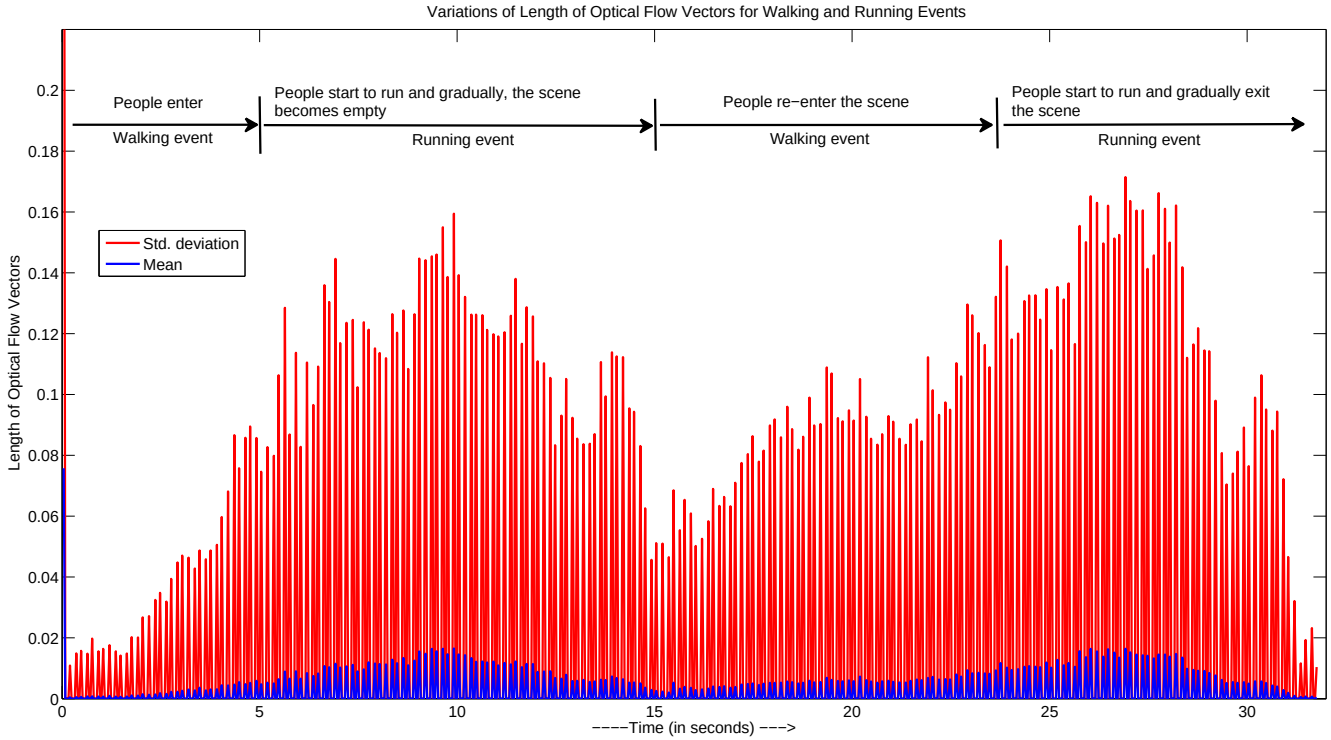


Fig. 2. The standard deviation and mean of the length of the optical flow vectors of the walking and running events for 14-16, View-001 of the PETS 2009 [2] dataset. Statistics for Mean (Std:0.0480, mean:0.0022, maximum:0.0758); statistics for Std. deviation (Std:0.0495, mean:0.0272, maximum:0.4076). The waveforms clearly show that the length of the optical flow vectors can be used as key distinguishing feature.

are splitting and merging if the distance between them is decreasing. This has been the general approach so far in the literature. In this work, an additional constraint has been employed to decide with higher degree of certainty. In addition to distances between groups, the directions of the vectors are also significantly important.

To start with, groups of people require to be first identified and clearly demarcated. This is accomplished by identifying the points on the manifold $\mathcal{M}$ that result in nonzero tangent vectors. Except for noise generated due to image characteristics (quantization) and noise due to other isolated movements (such as tree leaves, shadows), the majority of the tangent vectors are due to movements of the people in the scene. The boundary of the movements on the manifold $\mathcal{M}$ is found by using boundary conditions on the $\mathcal{M}$. Isolated tangent planes are eliminated to remove unwanted noise. This establishes the boundaries such that people in the video scene are identified. Secondly, to find the distances between groups, the geodesic distance between groups are computed using tangent planes. A center of mass for a particular group of people is calculated to identify the tangent planes. This is achieved computing the spherical mean given by:

$$\frac{1}{w_{n-1}} \int\limits_{\partial B(p,r)} f \, dS(f), \tag{5}$$

where $p \in U \subset \mathcal{M} \in \mathbb{R}^n$, ∂B is the boundary with radius r, f is the function is 5D parameter space, dS denotes the surface integral and w_{n-1} is the surface area.

The geodesic distance between the groups by utilizing the Riemannian metric tensor is given by (assuming the curves are admissible):

$$\mathcal{L}_a^b(\gamma)_{ij} \stackrel{\text{def}}{=\!=} \int_a^b g_{\gamma(t)} \langle \dot{\gamma}_i(t), \dot{\gamma}_j(t) \rangle \tag{6}$$

where γ is the geodesic curve, a and b are the point on the tangent planes where spherical mean for each group has been identified such that $\gamma : [a, b] \to \mathbb{R}$, i and j subscripts indicating different crowd groups.

The distance between different crowd groups are computed for each frame using the temporal variations between the frames. The direction constraint is added by computing the divergence considering radius r such that r is within ∂B at points on the manifold. The boundary points on the manifold determine the extent of the group from the spherical mean. This is equivalent to finding the boundaries of the crowd groups from the centroid in 2D cartesian plane. Depending on whether the crowd motion is acting as source or sink, and the geodesic distance, the crowd events (merging–splitting) are determined. Fig. 3 shows the merging and splitting patterns. If the distance between the groups is increasing and the source patterns are identified using the temporal evolution, then the crowd event is regarded as splitting. If on the other hand, the geodesic distance is decreasing and the motion from the divergence is forming a sink at points on the manifold, then the event is considered to be merging. In all other cases, the crowd event is termed as a *group movement*. The probability of merging and splitting events are denoted by $\Pr(\mathcal{F})$ and $\Pr(\mathcal{S})$.

Fig. 3. Merging and splitting of crowd groups. In the scene, there is white crisscrossed point in the middle of the scene. (a), (b) – the motion pattern indicates that the crowd is merging, where the crowd group is acting as a source. (c), (d) – the pattern of crowd splitting, where the crowd groups are in effect sinking. The arrows indicate the direction of movement. The arrows are the output from the proposed approach. In the case of merging [(a) and (b)], the crowds gather at that point. On the other hand, the crowd starts to split from the same point in case of splitting [(c) and (d)].

C. Local dispersion and evacuation events

The local dispersion event occurs when the crowd groups move from a point in the scene towards outwards in all directions while they are walking. This can be represented as the conditional event given the people are walking and also they are splitting and can be denoted as $\Pr(\mathcal{D}|\mathcal{W},\mathcal{S})$. On other hand, the evacuation event is the conditional event given the people are running and are splitting and can be denoted as $\Pr(\mathcal{D}|\mathcal{R},\mathcal{S})$.

Thus far, in this section, the crowd events were defined on Riemannian manifolds. Optical flow features were used to detect the crowd events. The detection of six crowd events are based on the probabilistic framework without tracking of the individuals. The classification of the events are based on the distribution of the probabilities for each events.

V. Results and Discussion

In this work, PETS 2009 [2] has been used for identifying six crowd events. This is the only dataset where the events are clearly manifested into six crowd events. The dataset was manually annotated to find the events as the ground truth. Later, this was used to compare with the results available from the proposed method. The proposed method was implemented in OpenCV 2.3 on a Virtual Box Linux machine (32-bit Ubuntu 12.04 LTS) equipped with 1.5GB RAM and Intel® i7 – 2600 CPU running at 3.4 GHz

A. Performance

The performance results have been presented in four layers of assessment. At the first layer, the results have been presented based on the classification results. Table I provides the confusion matrix for all the six crowd events. From Table I–(a), the walking events were correctly identified 76% of the times, with an error of 24% as running. On the other hand, 63% of the running events were correct and 37% of the events were identified as walking. From Table I–(b), we see that merging events were matching with the expected results up to 88%. Nevertheless, 60% of the splitting events matched with the true crowd splitting events. From Table I–(c), dispersion events were 94% of the times matched with ground truth; evacuation events matched up to 65%.

A second layer of comparison in terms of the delay in the detection is provided in the Table II for View-001. The results in Table II provides a comparison of detection of start and end timings of the crowd events from the selected video sequences. Figure 4 shows the corresponding temporal output. For the walking events a minimum delay of one second and a maximum of 4 seconds were observed. Likewise, for the running events, a minimum of 2 seconds and a maximum of 4 seconds delay were recorded. In the case of merging and splitting events for View-001 (timestamp : 14-33), a two seconds delay in detecting merging events and a second delay in splitting events were registered. Both, dispersion and evacuation events were reported by one second delay in detecting the events.

TABLE I. Confusion matrices for crowd events tested on the PETS 2009 dataset [2]. (a)- the confusion matrix for walking and running events; (b) - the confusion matrix for merging and splitting events; and (c) - the confusion matrix for dispersion and evacuation events.

	$\mathcal{W}$	$\mathcal{R}$
$\mathcal{W}$	0.76	0.24
$\mathcal{R}$	0.37	0.63

(a) Confusion matrix for walking and running events.

	$\mathcal{F}$	$\mathcal{S}$
$\mathcal{F}$	0.88	0.12
$\mathcal{S}$	0.4	0.60

(b) Confusion matrix for merging and splitting events

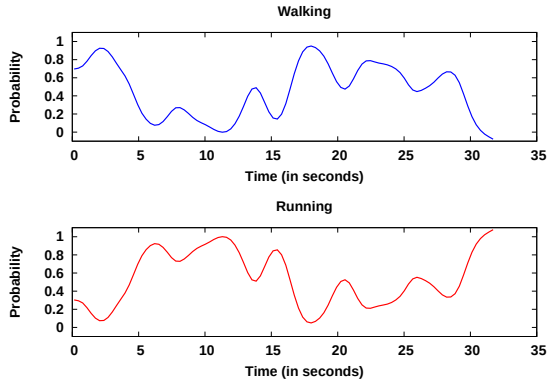
	$\mathcal{D}$	$\mathcal{E}$
$\mathcal{D}$	0.94	0.06
$\mathcal{E}$	0.35	0.65

(c) Confusion matrix for local dispersion and evacuation events

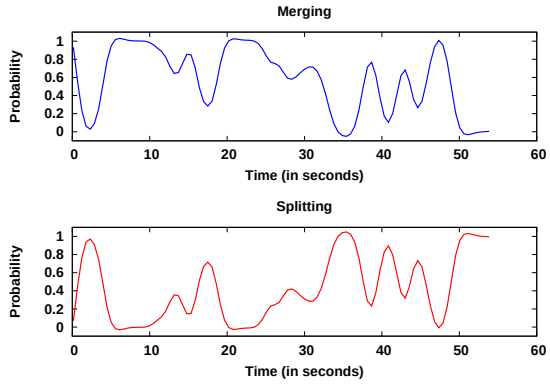
TABLE II. Comparison of start and end timings (in seconds, fps=7) of crowd events detection results with ground truth from selected video samples. The resulting temporal graphs have been provided in Figure 4.

Video sequence	**Ground Truth** [Start–End] (in seconds)		**Detected** [Start–End] (in seconds)	
14-16, View-001	Walking	Running	Walking	Running
	[0–6]	[6–15]	[0–7]	[7–17]
	[13–24]	[24–31]	[17–28]	[28–31]
14-33, View-001	Merging	Splitting	Merging	Splitting
	[0–29]	[48–53]	[0–27]	[49–53]
14-33, View-001	Dispersion	Evacuation	Dispersion	Evacuation
	[0–48]	[48–53]	[0–49]	[49–53]

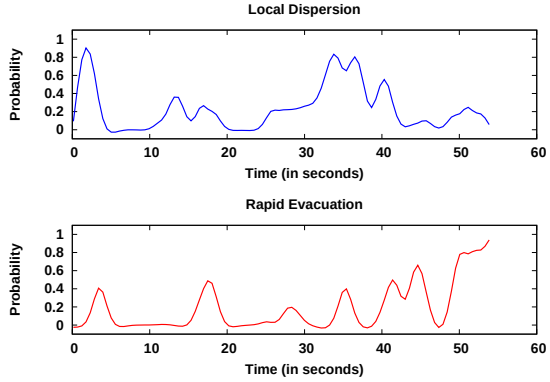
As a third layer of comparison, the events were compared in terms detection rates utilizing the precision and recall measures. Table III provides the comparison of different methods to detect crowd events. In this work, the output is shown for View-001 of the PETS 2009 dataset [2]. Leveraging upon the domain knowledge, pragmatic knowledge and the experience from other video surveillance projects, we conducted an experimental evaluation of events detection from different views and found that View-001 best captures the crowd events. The parameters used in the View-001 were applied to other



(a) Dataset: 14-16, View-001, number of frames=223, fps=7



(b) Dataset: 14-33, View-001, number of frames=377, fps=7



(c) Dataset: 14-33, View-001, number of frames=377, fps=7

Fig. 4. Demonstration of detection of crowd events. Refer to Table II for delay in detection of events in the above results.

three views. Thus the proposed method is termed as semi-supervised. The results in the Table III is the combined results of all the different views. The comparison is conducted with statistical filters [26] and motion pattern [28]. In [26], background modelling has been used followed by optical flow. In [28], motion pattern from optical flow is used for event detection. Since, our method uses optical flow, these two methods have been provided for comparison and analyses. From the Table III, it is evident that merging and dispersion events are best detected using the proposed approach with a precision of 0.85 and 0.9 respectively; likewise, the recalls were 0.88 and 0.94 accordingly. Other pairwise events such as walking–

TABLE III. COMPARISON OF CROWD EVENT DETECTION RESULTS.

Crowd Event	Measure	Statistical Filters [26]	Motion Pattern [28]	Our Approach
Walking	Precision	-	0.97	0.61
	Recall	-	0.96	0.75
Running	Precision	0.99	0.75	0.78
	Recall	0.99	0.81	0.63
Merging	Precision	-	0.59	**0.85**
	Recall	-	0.45	**0.88**
Splitting	Precision	0.65	0.47	**0.66**
	Recall	1	0.47	0.6
Dispersion	Precision	-	0.67	**0.9**
	Recall	-	0.45	**0.94**
Evacuation	Precision	-	0.69	**0.75**
	Recall	-	0.82	0.65

running events' performance is equally well compared to the other methods (precision of 0.61 and 0.78, recall of 0.75 and 0.63). Similarly, splitting and evacuation events maintained their performance (precision of 0.66 and 0.75, recall of 0.6 and 0.65).

Firstly, from the Table III we observe that, the proposed approach is comparable to the existing approaches. Although, the detection of walking and running events is slightly low, splitting and evacuation moderately good, merging and dispersion are well captured compared to others. One of the possible reasons for low detection rates is that the estimation of velocity vector based on optical flow during crowded scenarios poses some limitations. This can be improved with use of group tracking techniques to estimate group velocity. Also, if the tracking algorithms are lightweight and sufficiently fast, then region-based optical flow can be implemented to improve the running and walking events.

Secondly, it is important to note that the proposed approach outperforms (in merging and dispersion events) where existing methods did not prove to be efficient. One of the reasons for this performance is the incorporation of temporal tangential gradients. Existing methods used the training set to achieve higher detection rates. Our final goal is to design automated event detection model by reducing human intervention in the detection of crowd events. From a video surveillance perspective, merging and dispersion are more important for behavioural analysis than merely walking and running events. For instance, in the event of crowd panic in response to possible injury or threat to human life at a stadium, then our probabilistic model indicates this trend immediately, which is an indispensable model compared to existing methods. Previous methods combined all events, except running and walking into a single class. We separated the merging/splitting events from local dispersion/evacuation events in order to facilitate the detection of exact events as in video surveillance applications. Further improvement was made by combining the regular event with local dispersion, since we found a significant overlap between them.

As a fourth layer of comparison, in the proposed approach, it was assumed inherently that the final decision on the events will be made by the end users depending on the events that they are interested by giving priority to those events among others. However, if the end users are not acquainted or unable to interpret the crowd events either because they do not have experience or due to complexities in deciding, then the system itself requires to provide a single output based on the crowd events. Moreover, this affects complete automation

TABLE IV. COMBINED CONFUSION MATRIX FOR FOUR CROWD EVENTS (MERGING, SPLITTING, LOCAL DISPERSION AND EVACUATION).

	Merging	Splitting	Dispersion	Evacuation
Merging	0.56	0.22	0.08	0.14
Splitting	0.35	0.48	0.03	0.14
Dispersion	0.03	0.03	0.66	0.28
Evacuation	0.06	0.16	0.00	0.78

that is envisaged in visual surveillance and is a critical step that has been addressed for the first time. Investigating on this, we modeled events such that the walking–running events were considered to be primitive events (as mentioned in Section IV-A) and the other four events to be derivative of the primitive events. In particular, the merging events and splitting were conditioned by walking events. Likewise, the dispersion events were conditioned by *slow walking* instead of walking and evacuation events by *speed running* instead of running. The result of this is tabulated in Table IV. It is evident from the Table IV that there is a slight performance deterioration in merging and splitting events because of change in probability conditioned compared to Table I–(b). The confusion between merging to dispersion (0.08) and evacuation (0.14) is less. Likewise, the detection from splitting to dispersion (0.03) and evacuation (0.14) follow the same trend. Although there is some confusion from dispersion to evacuation (0.66 to 0.28), there is a clear distinction of evacuation (0.78). The confusion between dispersion to merging (0.03) and to splitting (0.03) are nominal. Similar trend is observed between evacuation to merging (0.06) and splitting (0.16). The inefficiencies in the proposed model in detecting merging and splitting events are largely due to occlusions.

In the proposed method we considered $t = 5$ for all crowd event detection purposes, which was chosen empirically. This is the main contributor in detection of actual events as well as detection delay. From the Table II, we observe that there is a maximum delay of 4 seconds between the actual start of an event and the detection. The same was reported between event occurrence and detection in all cases across different camera views (View-001—View-004). The start of an event may be slightly delayed because of camera views and occlusion. The detection delay is the delay incurred by the model (time-window) and not the computation delay. Previous works did not mention the delay at the start and the end timings of events. Optical flow values vanish for a static crowd in the scene in which case we used Gaussian Mixture Model (GMM) [31] for background modeling followed by optical flow for crowd detection. Future work in this direction includes derivation of efficient velocity vectors in crowded scenes without tracking on manifolds. A further improvement in processing and feature space can be brought in with the help of manifold learning while detecting the events.

Riemannian metric provides a system independent structure tensor that aids in computing the geodesic lengths. All tensors for that matter are co-ordinate independent. Furthermore, calculus of variations has been used since many centuries as a tool to measure the nonlinear structure of the data. Application of nonlinear methods helps to uncover the geometric structures that otherwise would be compromised due to linear subspace projections. Hence, in this work, Riemannian metric has been used to preserve those nonlinear geometric objects and their shapes independent of coordinate system.

Use of texture properties in combination with motion features may prove to effective in some instances [27]; however, the interplay of texture and motion features at the base of the feature extraction and their dynamics in relation to crowd events is still unexplored and requires further research in this direction. Moreover, the handling of occlusions using the texture methods utilizing the regression approaches have been in the literature for sometime now, but unraveling the supervised concept from the texture regression and devising the unsupervised/semi-supervised interactions of texture models and motion features maybe of aiding in analysing the occlusion states. Occlusion handling is beyond the scope of this work and a thorough analysis in this is yet to be conducted. The future work will also include the use of texture features and the occlusion handling mechanism to ameliorate the existing performance.

VI. CONCLUSION

Crowd event detection and classification is key to understanding behavioural characteristics of a crowd. Of the many applications associated with the crowd events, automated video surveillance is an important aspect of computer vision. In this regard, we developed probabilistic detection of crowd events (running, walking, merging, splitting, local dispersion and evacuation) on Riemannian manifolds. Previous work used supervised approach to detect and classify events. However, our approach was semi-supervised and delivered superior performance in selected cases. Importantly, the method also enables identification of the specific timings associated with an event. From an automated surveillance perspective, a simple probabilistic approach is proposed in order to combine different event probabilities with new results. Behaviour analysis, which relies on merging/splitting and dispersion/evacuation events can be enabled by our proposed approach.

VII. ACKNOWLEDGEMENT

This work is partially supported by the ARC linkage project LP100200430, partnering the University of Melbourne, Melbourne Cricket Club and ARUP. Authors would like to thank representatives and staff of ARUP and MCG.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, Dec. 2006. [Online]. Available: http://doi.acm.org/10.1145/1177352.1177355

[2] J. Ferryman, "PETS 2009 benchmark data," 2009, http://www.cvg.rdg.ac.uk/PETS2009/a.html.

[3] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted hmms for unusual event detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 1. IEEE, 2005, pp. 611–618.

[4] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[5] F. Jiang, W. Ying, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 907–913, 2009.

[6] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*. IEEE, 2011, pp. 3313–3320.

[7] O. P. Popoola and W. Kejun, "Video-based abnormal human behavior recognition—a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 6, pp. 865–878, 2012.

[8] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review," in *Proceedings IEEE Nonrigid and Articulated Motion Workshop*. IEEE, 1997, pp. 90–102.

[9] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314298907445

[10] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, no. 3, pp. 585–601, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320302001000

[11] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 23, pp. 90–126, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314206001263

[12] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609002704

[13] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[14] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.

[15] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 873–10 888, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412004757

[16] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *18th International Conference on Pattern Recognition (ICPR 2006)*, vol. 1. IEEE, 2006, pp. 175–178.

[17] E. L. Andrade, R. B. Fisher, and S. Blunsden, "Detection of emergency events in crowded scenes," in *The Institution of Engineering and Technology Conference on Crime and Security*. IET, 2006, pp. 528–533.

[18] E. L. Andrade, O. J. Blunsden, and R. B. Fisher, "Performance analysis of event detection models in crowded scenes," in *IET International Conference on Visual Information Engineering (VIE 2006)*. IEEE, 2006, pp. 427–432.

[19] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR 2005)*, vol. 1. IEEE, 2005, pp. 405–412.

[20] E. Shechtman and M. Irani, "Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 2045–2056, 2007.

[21] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *IEEE 11th International Conference on Computer Vision (ICCV 2007)*. IEEE, 2007, pp. 1–8.

[22] Y. Ke, R. Sukthankar, and M. Hebert, "Volumetric features for video event detection," *Int. J. Comput. Vision*, vol. 88, no. 3, pp. 339–362, 2010.

[23] D. Tran, J. Yuan, and D. Forsyth, "Video event detection: From subvolume localization to spatio-temporal path search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2013.

[24] Y. Chen, Z. Zhong, L. Ka Keung, and X. Yangsheng, "Multi-agent based surveillance," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 2810–2815.

[25] C. Gárate, P. Bilinsky, and F. Bremond, "Crowd event recognition using hog tracker," in *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*. IEEE, 2009, pp. 1–6.

[26] Á. Utasi, Á. Kiss, and T. Szirányi, "Statistical filters for crowd image analysis," in *Proceedings of the 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (in conjunction with CVPR 2009)*. IEEE, 2009.

[27] A. B. Chan, M. Morrow, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," in *Performance Evaluation of Tracking and Surveillance workshop at CVPR*. IEEE, 2009, pp. 101–108.

[28] Y. Benabbas, N. Ihaddadene, and C. Djeraba, "Motion pattern extraction and event detection for automatic visual surveillance," *J. Image Video Process.*, vol. 2011, pp. 1–15, 2011.

[29] G. Li, J. Chen, B. Sun, and H. Liang, "Crowd event detection based on motion vector intersection points," in *2012 International Conference on Computer Science and Information Processing (CSIP)*. IEEE, 2012, pp. 411–415.

[30] J. M. Lee, *Riemannian Manifolds: An Introduction to Curvature*. Springer, 1997, vol. 176.

[31] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 1999, pp. 246–252.