Recognizing Interactions Between People from Video Sequences

Kyle Stephens and Adrian G. Bors

Dept. of Computer Science, University of York, York YO10 5GH, UK
E-mail: adrian.bors@york.ac.uk

Abstract. This research study proposes a new approach to group activity recognition which is fully automatic. The approach adopted is hierarchical, starting with tracking and modelling local movement leading to the segmentation of moving regions. Interactions between moving regions are modelled using Kullback-Leibler (KL) divergence. Then the statistics of such movement interactions or as relative positions of moving regions is represented using kernel density estimation (KDE). The dynamics of such movement interactions and relative locations is modelled as well in a development of the approach. Eventually, the KDE representations are subsampled and considered as inputs of a support vector machines (SVM) classifier. The proposed approach does not require any intervention by an operator.

Keywords: Group Activity Recognition, Streaklines, Moving Regions, Kullback-Leibler divergence, Kernel Density Estimation, SVM.

1 Introduction

Human activity recognition has received considerable attention, by modelling and identifying the movement of isolated individuals. Nevertheless, many human activities take place in a social context of interaction with other people. Most human activity recognition methods start with extracting local features from video sequences which are then modelled either syntactically or statistically and the resulting modelling data is fed into a machine learning classifier. More recently, this area evolved towards detecting anomalies in the videos representing human activity, such as by using dynamic texture models [1], and Markov random fields [2]. An observational approach, detecting new activities in the scene, by using the Kullback-Leibler (KL) divergence from a dictionary of pre-observed events was proposed in [3, 4].

Following behaviour studies resulting from the complexity of modern life lead to the requirement of contextual modelling of human activities instead of that of simple movements by individual persons. Group activity requires more complex descriptions of how people interact with each other and with their surroundings. In the study by Ni *et al.* [5] group activities are recognized using manually initialized tracklets while a heat-map based algorithm was used for modelling human trajectories when recognising group activities in videos in [6].

A statistical approach of modelling data acquired by a multi-camera system was used in [7] and a hierarchical semantic granularity approach was employed for group activity in [8]. Movement trajectories have been represented as either histograms of features extracted from tracklets [10] or as Gaussian processes modelling time-series of movement trajectories [11]. Such approaches rely on either the training of a pedestrian detector for each scene, or on the manual annotation of trajectories.

This research study describes an automatic method for group activity recognition by modelling the inter-dependant relationships between human activity characteristic features over time. Features representing medium-term tracking of moving regions are extracted using the method from [12]. leading to the segmentation of compactly moving regions. The interdependency between moving regions is represented by evaluating the relative movement and location between pairs of segmented moving regions. Kernel Density Estimation (KDE) is then used to model the statistics of the movement, location, as well as their evolution in time, representing the dynamics of such interactions between moving regions. The group interaction model keeps track of stationary pedestrians by automatically marking the locations where these stop and then when they start an activity again. Section 2 describes the features used for representing moving regions, while the statistical modelling is provided in Section 3. Section 4 describes the classification approach. Section 5 provides the experimental results on two group activity datasets while Section 6 draws the conclusions.

2 Modelling Human Interactions

The proposed methodology for group activity recognition has three main processing stages: estimating streaklines of movement, modelling moving regions and their dynamics and group activity recognition. Optical flow estimation leads to tracking of regions of movement in the image [13, 14]. Streaklines [12], similarly to the approach from [14], represent the smooth movement of particles of fluid. Modelling streaklines relies on the Lagrangian framework for fluid dynamics, ensuring the robustness and the continuity of movement estimation. Unlike in the approach from [12], where streaklines are computed for each pixel, in this research study each streakline is associated with a block of pixels of fixed size by computing the marginal median of all streakline vectors located in a specific region. A streakline consists of several vectors head-to-tail located along a localized trajectory of movement which is then fit by a first degree polynomial for smoothing.

The general assumption is that movement in the scene corresponds to moving people, but interactions with other moving objects such as vehicles is accounted for in this model as well. Firstly, we begin by segmenting the streakflow field into distinctly moving regions. The Expectation-Maximization (EM) algorithm, assuming Gaussian Mixture Models (GMM) is used for segmenting and modelling each inter-connected region. The number of clusters and the centers of the Gaussian functions are initialized using the modes of the streakline flow histograms.

A two-step approach is adopted for movement segmentation in order to address the effects of perspective projection, which are mostly observed in the case of video sequences acquired with wide-angle lens cameras located at low heights. The assumption is that in the upper part of the video frames, objects and their motion is smaller than in the lower part, due to the perspective of the scene. In the first step, the segmentation is performed in order to estimate the height of the moving objects, which is used to derive a scaling factor. In the second step, the segmentation is repeated by considering this scaling factor, applied to the movements estimated from the video sequence, according to the location of its corresponding moving region in the scene. The motion $\mathbf{M}_i$ of region i is then scaled by a factor s_i:

$$\mathbf{M}_i^{'} = s_i \mathbf{M}_i, \tag{1}$$

where s represents the perspective projection scaling factor estimated for the given scene from the video sequence. Each moving region is therefore represented by a GMM, defined by its mean and variance.

3 Modelling interactions between moving regions

The key characteristics of group activities are often present in the interdependent relationship between the people present in the scene as well as between them and the surroundings. The general assumption is that moving regions correspond to human activities and in the following we model the relationship between such regions. In the first instance, we compute statistical differences between streakflow distributions $\mathcal{A}_{I(t)}$ and $\mathcal{A}_{J(t)}$, corresponding to two moving regions $I(t)$ and $J(t)$ at time t by

$$M(I(t), J(t)) = \mathrm{e}^{-\frac{D_{SKL}(\mathcal{A}_{I(t)}||\mathcal{A}_{J(t)})}{\sigma_m}} \tag{2}$$

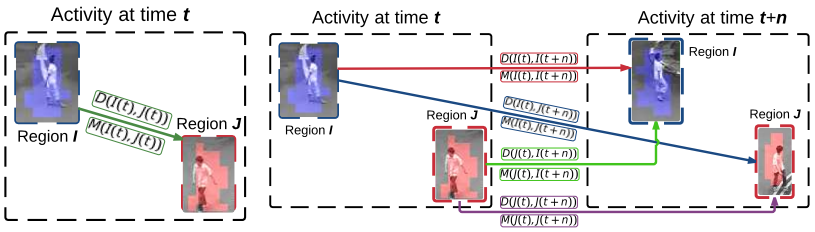
where $D_{SKL}(\mathcal{A}_{I(t)}||\mathcal{A}_{J(t)})$ is the symmetrized KL divergence between the local statistics of streaklines corresponding to the moving regions $I(t)$ and $J(t)$ at time t, [3] and σ_m is a scaling factor for movement differences. The background is considered as one of the regions as well. The calculation of equation (2) results in a value within the range $[0, 1]$ which models the inter-dependacy between regions $I(t)$ and $J(t)$. For example, individuals moving in completely opposite directions will have $M(I(t), J(t)) = 0$, whilst individuals moving in the same direction and at the same speed will have $M(I(t), J(t)) = 1$. These are then concatenated to form a vector representing the inter-dependant group relationships of the streakflows at a particular time t.

A similar approach is adopted for the locations of the moving regions by forming distributions of location coordinates corresponding to each moving region, including the background. The distributions of relative locations for the people from the scene, both moving or stationary, is modelled as well. The characteristic parameters of GMMs in this case correspond to the location, size and approximative size and shape of each moving region. Similarly to equation (2),

we model the interaction between two GMMs $\mathbf{C}_{I(t)}$ and $\mathbf{C}_{J(t)}$ representing the moving regions $I(t)$ and $J(t)$ at time t, as:

$$D(I(t), J(t)) = \mathrm{e}^{-\frac{D_{SKL}(\mathcal{C}_{I(t)}||\mathcal{C}_{J(t)})}{\sigma_l}} \tag{3}$$

where σ_l represents the characteristic scale parameter for locations. Similarly to the streakflow model, this provides a value in the range [0,1] representing the spatial relationship between the two moving regions. For example, individuals characterised by moving regions $I(t)$ and $J(t)$ at time t, located far apart, will have $D(I(t), J(t)) = 0$, whilst individuals located closer together will have $D(I(t), J(t)) = 1$. A vector, representing all the inter-relationships of locations for the group activity at time t, is then formed as shown in Fig. 1(a).



(a) Frame model (b) Dynamic model

Fig. 1. Modelling the inter-dependencies of moving regions in both space and time.

We also model the dynamic changes of relative differences between moving regions over subsequent frames by computing the differences between all streakflow models $M(I(t), J(t)) = 1$ at time t and those identified at other times $t + n$. These are computed as in equation (2), except that the models are now calculated across subsequent sets of frames. A vector of streakflow differences representing all the inter-dependant relationships of streakflow models between the time instances t and $t + n$ is then formed. The same modelling of dynamic changes is applied for changes in the relative distances between the locations, sizes and shapes of the moving regions by using inter-distances $D(I(t), J(t)) = 1$ from (3) at times t and $t + n$. Another issue addressed in this research study is the modelling of people who become stationary after they have moved through the scene. If there is no movement detected in a particular area and its neighbouring surroundings of the scene where motion was previously detected, during p consecutive frames, this indicates that a previously moving region ceased to move. Such stationary regions are characterised by their location and by zero motion. Finally, when movement occurs again in the region of a stationary person, then such regions are considered to be moving again as components of the group activity model. The dynamic model is illustrated in Fig. 1(b).

4 Classifying types of interactions between people

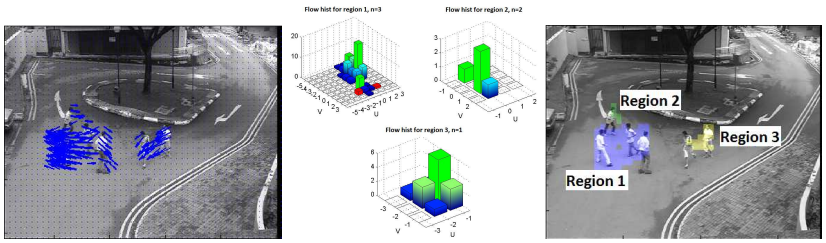
Kernel Density Estimation (KDE) is a non-parametric representation which provides a good model for complex data such as those defining human interactions. On the other hand KDE smoothes the data representation reducing the uncertainty when compared to assuming a certain parametric statistical model. The bandwidth parameters of the bi-variate Gaussian kernel are used to help control the smoothing effects of the kernel density estimator. In this study, we use the bivariate KDE method employing diffusions on data representations, proposed in [16], which considers a Gaussian kernel, and uses an automatic bandwidth selection method.

A discrete representation of the resulting KDE's for each set of features is represented on a grid of fixed size $K \times K$. By using a fixed grid size for representing the movement in the scene, the locations of the regions of movement, dynamics of movement and their region locations, we implicitly apply a data normalization, because such data representations do not depend on the frame size or on the actual number of frames. Such KDE's are then sampled and used as a feature vector representing the characteristics of the group activity taking place in the given video sequence. The feature vectors are then used to train a Support Vector Machine (SVM) algorithm, having K^2 inputs, while the outputs separate each group activity.

5 Experimental results

In the following we provide the experiments when considering two databases containing group interaction videos: NUS-HGA [5], and Colective [9] datasets. This first data set consists of six different group activities collected in five different sessions containing 476 video sequences, each session representing staged actions. Initially, streaklines are extracted for blocks of size 14×14 over 10 consecutive frames. The motion is segmented and each moving region is represented by the Gaussian Mixture Model (GMM) of streakflows vectors and their locations GMM. Fig. 2 shows an example of the estimated streakflows, motion histograms, and the moving region segmentation for the fight activity from the NUS-HGA dataset. In this particular activity, movement is intense and chaotic. In Fig. 2b the solid green bars correspond to peaks of the histogram, while the solid red bars are entries with the height below 15% of the maximum bar height which are eventually removed for not being significant enough in the context of the scene' movement. The moving regions are well segmented and the small regions obtained in region 1 of Fig. 2c help characterize the smaller atomic events performed in the group, for example pushing or kicking which usually happens during the fighting activity.

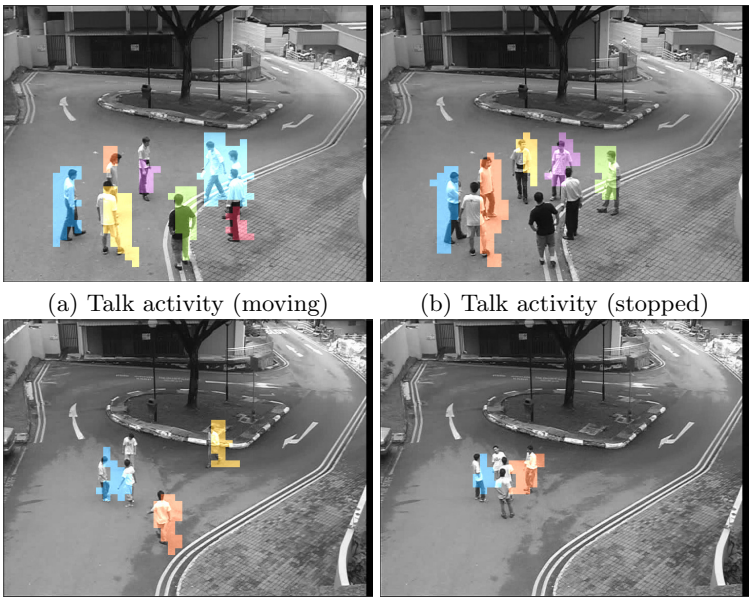
We account for the perspective projection effects, where smaller movements in smaller segmented regions would correspond to movements detected from farther away in the scene. The segmentation is done in two stages, where during the first segmentation stage a scaling factor is calculated and then the motion is scaled



(a) Streakflows (b) Histograms of flow (c) Moving regions

Fig. 2. Example of streakflows, histograms of flow and the moving regions before and after segmentation on a fight sequence from the NUS-HGA dataset. In b) "n" refers to the number of histogram peaks.

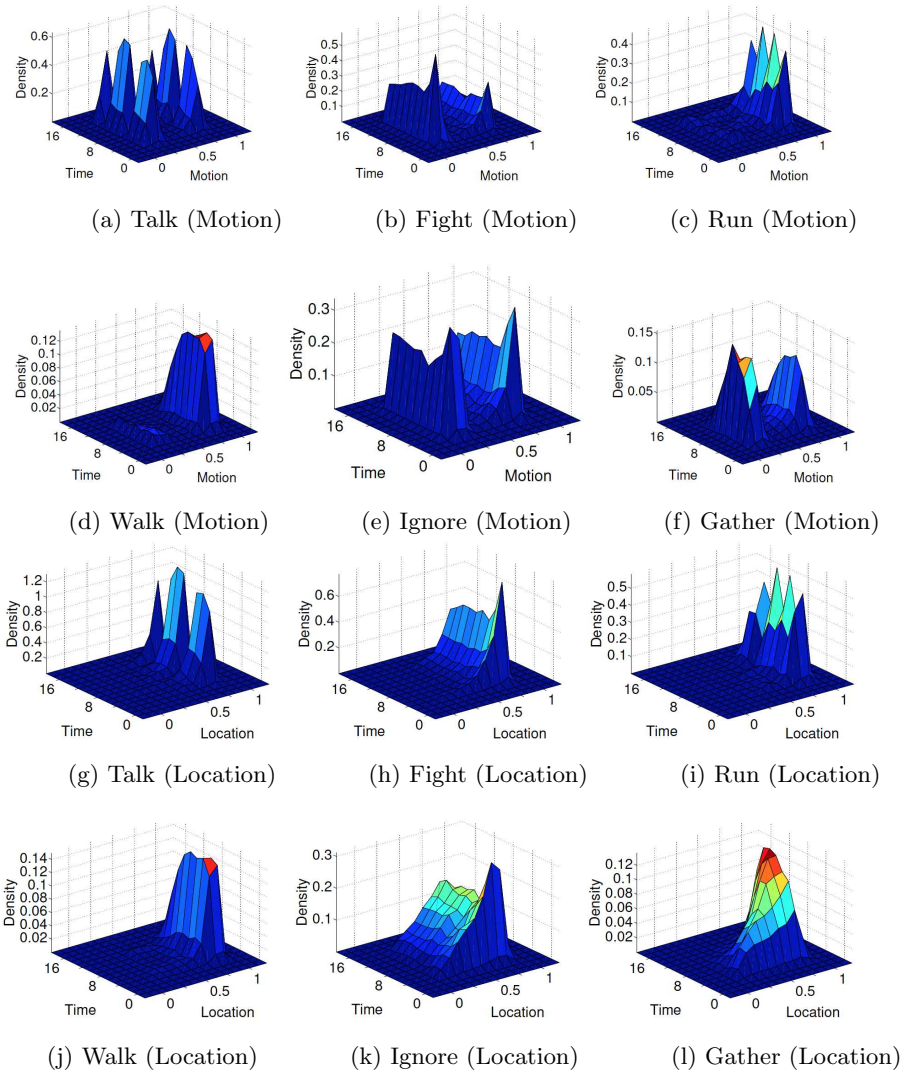
accordingly and the scene resegmented. The detection of the stationary regions detector is applied considering the number consecutive frames for estimating the streaklines as $p = 25$. Two examples of detecting stationary pedestrians are shown in Fig. 3 for the Talking and Gathering activities. In Figs. 3a and 3c the pedestrians are still moving and therefore their corresponding moving regions are properly detected. In Figs. 3b and 3d the individuals have stopped and their stationary regions are properly detected.



(a) Talk activity (moving) (b) Talk activity (stopped)

(c) Gather activity (moving) (d) Gather activity (stopped)

Fig. 3. Identifying when pedestrians stop during the video frames showing gathering and talking activities from the NUS-HGA dataset.



(a) Talk (Motion) (b) Fight (Motion) (c) Run (Motion)

(d) Walk (Motion) (e) Ignore (Motion) (f) Gather (Motion)

(g) Talk (Location) (h) Fight (Location) (i) Run (Location)

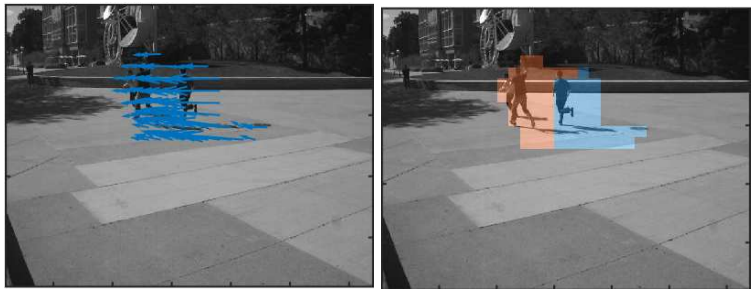
(j) Walk (Location) (k) Ignore (Location) (l) Gather (Location)

Fig. 4. KDEs and histograms representing motion dynamics in (a)-(f) and for location in (g)-(l) for the NUS-HGA dataset.

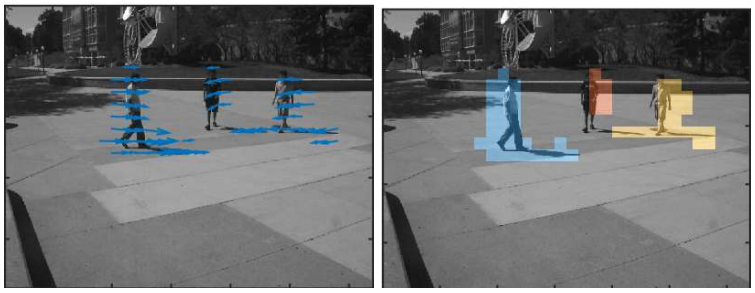
The streakflow movement model, streakflow dynamics, location and location dynamics relationship differences are computed as described in Section 3, considering the scaling parameters $\sigma_m = 15$, $\sigma_l = 550$ for motion and location differences respectively, and $\sigma_m = 17.5$, $\sigma_l = 650$ for the motion and location dynamics. The number of frames, considered for the dynamic window from Section 3, is set to $n = 13$. The bivariate kernel density estimation from [16] is computed over a fixed grid size of 16×16. Representations of the KDEs using Gaussian kernels for various human interaction activities are shown in Figs. 4a-4f when considering motion estimation and segmentation, and in Figs. 4g-4l when modelling the locations of moving regions. The gathering motion shown in Fig. 4f displays a diversity of differences in movement, which is expected as some individuals are gathering coming from different directions. The Walking activity location differences, shown in Fig. 4e, are all close to 1. This implies that the individuals are tightly grouped, which is expected in the Walk in Group activity. The Gather activity location differences shown in Fig. 4l display clear transitions between locations situated far apart leading to closer-together locations. This is expected, as the gathering activity involves individuals coming from far away towards gathering in a tight group at the end of the activity.

For classification purposes, the density estimations are subsampled and fed into the classifier. The motion and location features represent complimentary information and can be combined for the final activity classification. We use SVM with the RBF kernel as a classifier, considering the parameters $C = 2.83$ and $\gamma = 0.00195$ for the SVM margin and kernel bandwidth. For all experiments, we follow the evaluation protocol described in [5], where the NUS-HGA dataset is split into 5-fold training and testing.

The Collective dataset [9] consists of 6 different activities: Gathering, Talking, Dismissal, Walking Together, Chasing and Queueing. The dataset consists of 32 video sequences, where each video sequence contains multiple examples of each activity. The video sequences are recorded using a hand-held camera, and therefore the perspective distortion is quite strong in the scenes from this dataset. The spatio-temporal segmentation of these video sequences takes place into blocks of 20×20 pixels by 10 frames, where the streaklines are extracted for each block of 10 frames. Examples of the streakflows and movement segmentation are shown in Fig. 5 for the Chasing and Gather activities. In both cases, the moving regions are well segmented, particularly in the chasing example where the chaser and chasee are segmented separately despite forming one connected region moving in the same direction. The next step involves applying the stationary pedestrian detector as in Section 2, assuming the number of prior frames used as $p = 25$. The videos from the Collective dataset show different activities, displaying transitions from one activity to another, including times when people are stationary. Such situations are identified and an example of transitions through activities is shown in Fig. 6. Initially, as in Fig. 6a, the pedestrians are moving towards each other performing the gathering activity. People are eventually gathered together towards the end of this activity, and the transition to the talking activity is evident in Fig. 6b. The stationary people detection has



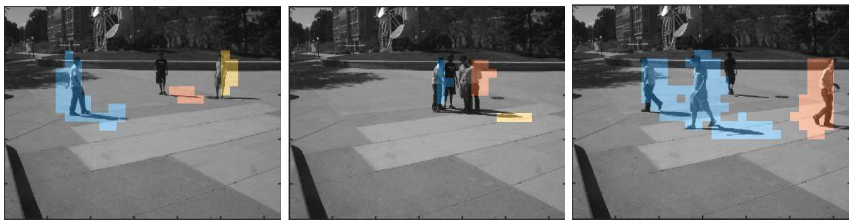
(a) Streakflow (Chasing) (b) Segmentation (Chasing)



(c) Streakflow (Gather) (d) Segmentation (Gather)

Fig. 5. Examples of streakflow and segmentation from the Collective dataset.

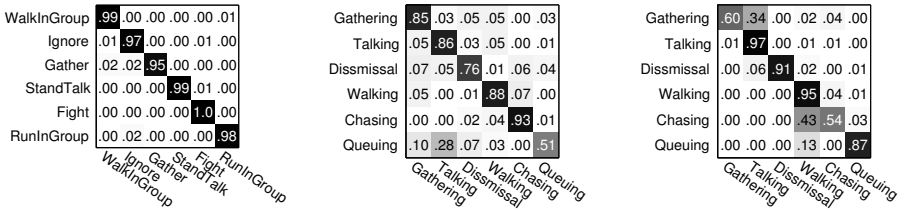
successfully recorded the locations of the individuals when stopping, as seen in Fig. 6b. Finally, after a period of time, the individuals begin to move again performing the dispersing activity shown in Fig. 6c. In Fig. 6c, the new moving regions are detected replacing the previously identified stopped regions which are no longer present.



a) Gathering b) Talking b) Dispersing

Fig. 6. Pedestrians transitioning through various activities in the Collective dataset.

In the following, the human activity features, representing the streakflow differences, streakflow dynamics, location differences and location dynamics are computed for each moving region as described in Section 3. The scaling parameters are $\sigma_m = 15$ and $\sigma_l = 450$ for motion features and location features,



(a) NUS-HGA database (b) Collective database (c) Collective using [11]

Fig. 7. Confusion matrices for the recognition results of the proposed method when combining all features modelling movement, location distribution and their dynamics as well on (a) NUS-HGA, resulting in 90 % classification accuracy and (b) Collective dataset, resulting in 79.7 % accuracy. (c) The confusion matrix for Colective dataset when using [11], resulting in 80.3% classification accuracy.

respectively, while the size of the dynamic window for the motion dynamics and location dynamics is $n = 5$. Then, the data is represented over time using KDE, as described in Section 4, over a grid size of 8×8, using the 2-column feature matrices as input data. The grid-based representation of the KDE is then used as input to the SVM classifier with the RBF kernel.

Table 1. Group activity recognition results on the NUS-HGA and Collective datasets

Method	NUS-HGA dataset (%)	Collective dataset (%)
Localized Causalities [5]	74.2	-
Group interaction zone [17]	96.0	-
Multiple-layered model [11]	96.2	80.3
Monte Carlo Tree Search [18]	-	77.7
Collective activities [8]	-	79.2
Motion	86.2	75.4
Location	87.1	64.3
Motion dynamics	91.6	76.8
Location dynamics	92.6	71.6
Motion+Location	94.5	76.5
Motion Dynamics+Location Dynamics	97.1	78.4
Motion+Location+Motion Dynamics+Location Dynamics	98.0	79.7

For the tests on the Collective dataset we divide the dataset into 3 subsets for 3-fold training and testing according to the tests in [9]. We split the sequences during training and testing into short sequences of 60 frames for the evaluation and then calculate the average recognition accuracy across all classes. Confusion matrices for all features combined are compared to the approach from [11] as shown in Fig. 7. The results for the Queuing activity are not that good because

that stationary pedestrians forming queues are not moving at all for the duration of the sequence, and therefore are not detected. However, it can be observed from Fig. 7 that the results of the proposed methods show a greater consistency across all the other activities then other approaches.

Comparative results are provided in Table 1 for NUS-HGA and Collective datasets. The location features provide a better recognition result than the motion features while the results for the dynamics models for motion and location emphasise their importance for the Group activity recognition. The combination of all features account for movement, location, as well as the dynamics of both movement and location, and gives the best result of 98% for the NUS-HGA dataset. The group interaction method from [17] does not evaluate the results using the 5-fold training and testing as suggested in [5] for the NUS-HGA dataset. The proposed methodology, which is fully automated, provides a clear improvement of about 2% over the best other approach for the NUS-HGA dataset. For the Collective dataset, the proposed method is comparative to the state-of-the-art and superior to the other methods when not considering the queuing activity. The motion and movement dynamics outperform the location inter-dependency features, while the dynamic features outperform their equivalent frame-by-frame features. Similarly to the results on the NUS-HGA database, the best results for the Collective dataset are achieved when combining all features. However, all the other comparative methods are not fully automatic and use some form of human intervention during the experiments. Meanwhile, the proposed methodology is completely automatic and does not require any human intervention.

6 Conclusion

A completely automatic approach for modelling interactions between people is proposed in this paper. Streakflows of localized movement along several frames are estimated from the video sequence. Statistical distributions of vectors forming streakflows, as well as their locations are represented using kernel density estimation (KDE) and are used in order to identify compactly moving regions. We also consider the dynamics of change in the streakflows and in the locations of the moving regions. The relative movement of each moving region with all the other moving regions, including the background, is then represented statistically. Scaling is used in order to mitigate the effects of perspective projection in the scene, while the dynamics of change in the moving regions considers the timing when people are stationary. Eventually, SVM with RBF kernels, considering sampled KDE representations of movement, location, and their dynamics, as inputs, is used as a classifier.

Acknowledgment

This research work was supported by DSTL grant DSTLX1000074616 "Human Activity Recognition."

References

1. W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):18–32, (2014).
2. H. Nallaivarothayan, C. Fookes, S. Denman, and S. Sridharan, "An MRF based abnormal event detection approach using motion and appearance features," *Proc. IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance*, pp. 343–348, (2014).
3. K. Stephens, A. G. Bors, "Observing human activities using movement modelling," *Proc. IEEE Int. Conf. on Advanced Video and Signal-based Surveillance,* paper # 44, pp. 1-6, (2015).
4. K. Stephens, A. G. Bors, "Grouping multi-vector streaklines for human activity identification," *Proc. IEEE Workshop on Image, Video and Multidimensional Signal Processing*, Bordeaux, France, (2016).
5. B. Ni, S. Yan, and A. Kassim, "Recognizing human group activities with localized causalities," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1470–1477, (2009).
6. W. Lin, H. Chu, J. Wu, B. Sheng, and Z. Chen, "A heat-map-based algorithm for recognizing group activities in videos," *IEEE Trans. on Circuits and Systems for Video Technology*, 23(11):1980–1992, (2013).
7. M. Chang and W. Ge, "Probabilistic group-level motion analysis and scenario recognition," *Proc. Int. Conf. on Computer Vision*, pp. 747–754, (2011).
8. W. Choi and S. Savarese, "Understanding collective activities of people from videos," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(6):1242–1257, (2014).
9. W. Choi and S. Savarese, "A unified framework for multitarget tracking and collective activity recognition," *In Proc. European Conf. on Computer Vision, vol. LNCS 7575*, pp. 215–230, (2012).
10. Y. Zhang, W. Ge, M. C. Chang, and X. Liu, "Group context learning for event recognition," *Proc. IEEE Work. on Applic. of Comp. Vision*, pp. 249–255, (2012).
11. Z. Cheng, L. Qin, Q. Huang, S. Yan, and Q. Tian, "Recognizing human group action by layered model with multiple cues," *Neurocomputing*, 136:124–135, (2014).
12. R. Mehran, B. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," *Proc. European Conference on Computer Vision, vol. LNCS 6313*, pp. 439–452, (2010).
13. A. G. Bors, I. Pitas, "Prediction and Tracking of Moving Objects in Image Sequences," *IEEE Trans. on Image Processing,* 9(8):1441-1445, (2000).
14. A. Doshi and A. G. Bors, "Robust processing of optical flow of fluids," *IEEE Trans. on Image Processing,* 19(9):2332-2344, (2010).
15. A. G. Bors, N. Nasios, "Kernel bandwidth estimation for nonparametric modelling," *IEEE Trans. on Systems, Man and Cybernetics, Part B: Cybernetics,* 39(6):1543-1555, (2009).
16. Z. Botev, J. Grotowski, and D. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, 38(5):2916–2957, (2010).
17. N.-G. Cho, Y.-J. Kim, U. Park, J.-S. Park, and S.-W. Lee, "Group activity recognition with group interaction zone based on relative distance between human objects," *Int. Jour. of Pattern Recog. and Artif. Intel.*, 29(5), #1555007:1–15, (2015).
18. M. R. Amer, S. Todorovic, A. Fern, and S. C. Zhu, "Monte Carlo tree search for scheduling activity recognition," 'em Proc. IEEE Int. Conf. on Computer Vision, pp 1353–1360, (2013).