This is an author produced version of Human Group Activity Recognition based on Modelling Moving Regions Interdependencies.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/127769/

Proceedings Paper:

https://doi.org/10.1109/ICPR.2016.7899948
Human Group Activity Recognition based on Modelling Moving Regions Interdependencies

Kyle Stephens and Adrian G. Bors
Department of Computer Science, University of York, York YO10 5GH, UK
E-mail: adrian.bors@york.ac.uk

Abstract—In this research study, we model the interdependency of actions performed by people in a group in order to identify their activity. Unlike single human activity recognition, in interacting groups the local movement activity is usually influenced by the other persons in the group. We propose a model to describe the discriminative characteristics of group activity by considering the relations between motion flows and the locations of moving regions. The inputs of the proposed model are jointly represented in time-space and time-movement spaces. These spaces are modelled using Kernel Density Estimation (KDE) which is then fed into a machine learning classifier. Unlike in other group-based human activity recognition algorithms, the proposed methodology is automatic and does not rely on any pedestrian detection or on the manual annotation of tracks.

Index Terms—Group Activity Identification, Motion Segmentation, Streaklines.

I. INTRODUCTION

Several algorithms have been proposed for human activity recognition by considering individual actions. This research area has a significant importance for video surveillance, human-computer interaction, semantic annotations of multimedia, retrieval of video data, among many other applications. Meanwhile, group activity classification has attracted interest only very recently, despite being essential in defining the real intention and the context of human activities. Most of the human activity recognition methods begin by modelling low level local features from video sequences, for example using the Dollar gradient cuboids [1] or histograms of gradients (HOG) [2]. In other approaches, Baktashmotlagh et al. [3] applied non-linear stationary subspace analysis to activity recognition while Ryoo and Aggarwal [4] introduced a method named spatio-temporal relationship match.

More recently, the main focus of human activity has moved on from simple human activities to those that are more complex, where the main objective is scene analysis rather than determining the activities of a single individual. One group of approaches is to detect abnormalities or uncommon activity events. The method from [5] modelled the motion patterns using Gaussian Mixture Models (GMMs) of 3D distributions of local space-time gradients. Similarly, GMMs of Markov random fields (GMM-MRF) was used in [6] for abnormal activity detection. Dynamic texture models [7], which considers both appearance and dynamics, have also been considered for abnormal activity detection. An observational system, in which new activities are identified in the scene, based on a significant Kullback-Leibler divergence from a dictionary of activities pre-learnt during the training stage, was proposed in [8], [9]. In comparison to human activity recognition, group activity recognition requires more complex descriptions of the people’s interaction in the group. Ni et al. [10] recognizes group activities using manually initialized tracklets. Lin et al. [11] used a heat-map based algorithm for modelling human trajectories when recognising group activities in videos. Chang et al. [12] used a probabilistic approach to group human activity by forming various probabilities depending on the tracks between individuals using a multi-camera system. Choi et al. [13] proposed a framework for analysing collective group activities based on different levels of semantic granularity. Zhang et al. [14] addressed the problem of group event recognition by computing histograms of different features extracted from the tracklets, representing localized movement in the video. Similarly, Cheng et al. [15] modelled group activity as a framework composed of multiple layers and Gaussian defined processes were used for representing motion trajectories. One common issue with all these methods is that they rely on either the training of a pedestrian detector for each scene, or on the manual annotation of tracklets.

In this research study we propose an automatic method for group activity recognition by modelling the inter-dependant relationships between features over time. Unlike other methods, we do not rely on any manual initialisation of tracklets and instead make use of medium term tracking as provided by streaklines [16]. Compact moving regions are then segmented. The interdependency between moving regions is represented by evaluating the relative movement and location of each moving region with respect to all the others. Kernel Density Estimation (KDE) is used to model both time-location and time-motion spaces, resulting in representing the dynamics of such interactions. Moreover, the model keeps track of stationary pedestrians by marking the locations where they stop moving and considers these locations in modelling their following movements. We also propose a scaling procedure in order to compensate for the effect of perspective projection in video sequences acquired by lowly located cameras of wide view and compensate in the group activity model for such effects. Section II describes the features used for representing moving regions, while how their inter-dependencies are modelled in the context of group activity is explained in Section III. Section IV describes the classification of group activities. Section V shows the experimental results and Section VI draws the conclusions of this research study.
II. GROUP ACTIVITY MODELLING

The proposed methodology for group activity recognition has several stages, including extracting streaklines, representing medium-time trajectories of movement, identifying moving regions and their dynamics, using these for modelling group interactions, and then finally classifying the sequences into group activities using Support Vector Machines (SVM). A block diagram of the proposed method for recognising group activities is shown in Figure 1.

The first processing stage consists of movement estimation. One issue that arises from using traditional optical flow is the difficulty in capturing unsteady movement in scenes with multiple pedestrians interacting and crossing each other. To alleviate this problem, we propose the use of a medium-time movement tracking method such as the streaklines proposed in [16] which was used in [8], [9] as well. Streaklines correspond to tracking fluid like flow in a scene, enabling the filling of spatial gaps. Unlike in [16], where streaklines are computed for each pixel, we associate each streakline with blocks of pixels of a fixed size by computing the marginal median as the streakline estimate for each block of pixels. Following this, we fit a first degree polynomial to each streakline in order to obtain a smoother representation. This differs from [8], where the principal direction of movement was obtained from applying PCA on the vectors forming each streakline. One issue with the approach from [8] is that it does not consider the motion consistency over several frames. In the approach from this study the consistency of the streaklines is enforced over several frames.

We make the assumption that each compact region of streakflows may contain several individual movements, which can be represented by clusters. Firstly, we begin by segmenting the streakflow field into distinct moving regions. The Expectation-Maximization (EM) algorithm, under the Gaussian Mixture Model (GMM) modelling assumption, is used for segmenting and modelling each inter-connected region. The number of clusters and the centres of the Gaussian functions in the EM Maximization (EM) algorithm, under the Gaussian Mixture Model (GMM) modelling assumption, is used for segmenting the streakflow field into distinct moving regions. The Expectation-Maximization (EM) algorithm, under the Gaussian Mixture Model (GMM) modelling assumption, is used for segmenting the streakflow field into distinct moving regions. The Expectation-Maximization (EM) algorithm, under the Gaussian Mixture Model (GMM) modelling assumption, is used for segmenting the streakflow field into distinct moving regions.

The first processing stage consists of movement estimation. One issue that arises from using traditional optical flow is the difficulty in capturing unsteady movement in scenes with multiple pedestrians interacting and crossing each other. To alleviate this problem, we propose the use of a medium-time movement tracking method such as the streaklines proposed in [16] which was used in [8], [9] as well. Streaklines correspond to tracking fluid like flow in a scene, enabling the filling of spatial gaps. Unlike in [16], where streaklines are computed for each pixel, we associate each streakline with blocks of pixels of a fixed size by computing the marginal median as the streakline estimate for each block of pixels. Following this, we fit a first degree polynomial to each streakline in order to obtain a smoother representation. This differs from [8], where the principal direction of movement was obtained from applying PCA on the vectors forming each streakline. One issue with the approach from [8] is that it does not consider the motion consistency over several frames. In the approach from this study the consistency of the streaklines is enforced over several frames.

We make the assumption that each compact region of streakflows may contain several individual movements, which can be represented by clusters. Firstly, we begin by segmenting the streakflow field into distinct moving regions. The Expectation-Maximization (EM) algorithm, under the Gaussian Mixture Model (GMM) modelling assumption, is used for segmenting and modelling each inter-connected region. The number of clusters and the centres of the Gaussian functions in the EM algorithm are initialised using the modes of the histogram of streakline flow in order to improve the convergence. Moreover, in this study we also address the perspective distortion effects by using a two-step approach to movement segmentation. Such effects are evident in the case of video sequences acquired with wide-angle lens cameras which are located at low heights. 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 considering this scaling factor, applied appropriately to the estimated movement, according to the location of its corresponding moving region in the scene. A moving region $i$ is scaled as follows:

$$s_i = \frac{1}{\sum_{j=1}^{n} h_j} (h_i + \frac{\sum_{j=1}^{n} h_j}{n})$$

Where $h_i$ is the height identified for each moving region in the first step, $j = 1, \ldots, n$ are the segmented moving regions, $h_m$ is the predetermined overall mean height of all moving regions and $s_i$ is the scaling factor for moving region $i$. This is repeated for all compact moving regions which are identified in the scene. The motion $M_i$ of region $i$ is then scaled by a factor $s_i$:

$$M'_i = s_i M_i.$$  

Each moving region is therefore represented by a GMM defined by its characteristic parameters representing its movement and location in the scene. Another issue that is addressed in this research study is the modelling of people who become stationary after they have moved through the scene. Under the optical flow detection and motion model such people would not be accounted for. To overcome this situation, we propose to identify when and where people stop moving in the scene. If no movement is present in a particular region where motion was previously detected, during $p$ consecutive frames, this indicates a stationary region that has previously moved. Such stationary regions are characterised by their location and by zero motion. Any movements of a person present near the edge of the scene that subsequently moves out of the scene is identified and the respective moving region is no longer considered. Finally, when movement occurs within a bounding box of the stopped pedestrian, the region is deemed to be no longer stationary and the new emerging moving region in the area is activated in the existing group activity model.

III. MODELLING INTERDEPENDENT RELATIONSHIPS OF MOVING REGIONS

The key characteristics of group activities are often present in the interdependent relationship between the pedestrians and moving objects. In this research study we propose to model the interdependent relationship between the features of each pair of moving regions detected in the scene. In this section, we describe how we model four distinct features for representing group activities: streakflows, streakflow dynamics, locations and location dynamics.

To begin, we model the interdependent relationship by evaluating the differences between streakflow models in the scene for each pair of moving regions. This models the interdependent relationship of the movement of the group at a particular time instance. We compute the differences between streakflows, $A_{I(t)}$ and $A_{J(t)}$ for two moving regions $I(t)$ and $J(t)$ at time $t$ by:

$$M(I(t), J(t)) = e^{-\frac{D_{SKL}(A_{I(t)} || A_{J(t)})}{\sigma_m}}$$

where $\sigma_m$ is a scaling factor for movement differences and $D_{SKL}(A_{I(t)} || A_{J(t)})$ is the symmetrised KL divergence between the streakline distribution of moving regions $I(t)$ and $J(t)$ at time $t$. This results in a value within the range $[0,1]$ which models the difference between two streakflow models, each characterising the movement of one region in the scene, associated to a moving person. For example, individuals moving in completely different directions will have
The moving regions location differences, representing all the individuals located close together will have $M(I(t), J(t)) = 1$. The differences are computed by considering all pairs of moving regions in the scene at a particular time $t$ by using equation (3). These are then concatenated to form a vector representing the inter-dependant group relationship of the streakflows at a particular time $t$.

We also model the dynamic changes of differences between moving regions over subsequent frames by computing the differences between all streakflow models at time $t$ and those identified at time $t + n$. These are computed as in equation (3), except that the models are now across subsequent sets of frames instead of at the same time instance. 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 distributions of relative locations for the people from the scene, both moving or stationary, is modelled similarly by considering differences between the GMM representing the spatial-location of their corresponding moving region. The means will approximate the centres of moving regions, whilst the variance will provide some characteristics of the size and shape of the region. Similarly to the streakflows, the differences between such location GMMs are then computed. Given two location GMMs $C_{I(t)}$ and $C_{J(t)}$ for moving regions $I(t)$ and $J(t)$ at time $t$, the differences between their locations can be computed by:

$$D(I(t), J(t)) = e^{-\frac{D_{SKL}(C_{I(t)}||C_{J(t)})}{\sigma_1}}$$

where $\sigma_1$ 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 close 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.

Similarly to the streakflow model, the dynamics of the locations over time is computed as well. The dynamic changes of differences over subsequent frames are computed by the differences between all location points at time $t$ and all location points at time $t + n$ using equation (4). A vector representing the moving regions location differences, representing all the inter-dependant relationships of location points between time $t$ and $t + n$, is then obtained. These movement models are illustrated in Figure 2.

One further issue that arises when computing such differences is that the rate of movement change and the rate of location change are not clearly characterised. For example, when using the dynamics in both movement and locations alone, the dynamics between walking and running activities may appear quite similar. In order to avoid this situation we consider the background as an additional region for both the streakflow and the location models. In the former case, the background object is defined as the GMM model comprising of all the motion in the scene that does not belong to a moving region (often zero motion if the camera is stationary). In the latter case, the location object is defined as the GMM representing the centre of the scene. By adding the background model, the change in both motion and location relative to the background represents the absolute movement in the scene. In the case of camera movement, such a model would account for this. Given a streakflow background model $A_{B(t)}$, at time $t$ the difference between the streakflow model $A_{I(t)}$, for moving region $I(t)$, at time $t$, and the background $B(t)$ is computed as:

$$M(I(t), B(t)) = e^{-\frac{D_{SKL}(A_{I(t)}||A_{B(t)})}{\sigma_m}}$$

Similarly, given the centre point $C_{B(t)}$ defined as the location of background model $B(t)$ (the centre of the scene) at time $t$ and the location model $C_{I(t)}$ for moving region $I(t)$ at time $t$, the difference is computed as:

$$D(I(t), B(t)) = e^{-\frac{D_{SKL}(C_{I(t)}||C_{B(t)})}{\sigma_2}}$$

Such differences are then computed between every region in the scene and the background model $B(t)$. Finally, the vector of differences in both cases are concatenated with the vector representing the other pairwise movement and location differences, corresponding to the pairs of moving regions.


IV. GROUP ACTIVITY CLASSIFICATION

To model the change in feature relationship over the whole sequence, we propose to use bi-variate Kernel Density Estimation (KDE). KDE would provide smoothing on the dynamics of feature changes over time increasing the robustness of the group activity model. We form two column matrices where the motion and location interdependences for each pair of moving regions are represented along the first column and their corresponding time instances are located in the second column. This matrix representation is used for each feature representing streakflow, streakflow dynamics, locations and location dynamics, separately. The bi-variate kernel density estimation is applied over a fixed grid size of $K \times K$, given the normalized matrix data.

By using a fixed grid size, video sequences of different lengths will be normalized in length. This helps to normalise the difference in speeds at which the activities are performed. The grid size is an important parameter in the density estimation as a too small grid would result in over-smoothed feature data and consequently important characteristics in the relationship features may be lost. If the grid size is too large, then the data will appear too sparse and would not model well the underlying pattern of the data. The kernel for density estimation is assumed to be Gaussian. The bandwidth parameters of the bi-variate Gaussian kernel are used to help control the smoothing effects of the kernel density estimator.

The densities computed over the fixed grid are used as the defining feature vector representation for the group activity. Such densities are computed independently for each dimension, representing the relationships of the moving regions in the movement, movement dynamics, location and location dynamics, respectively. Finally, the feature vectors representing each activities are used to train a Support Vector Machine (SVM).

V. EXPERIMENTAL RESULTS

For all experiments, we follow the same recognition routine. Firstly, the streakflows are extracted for each set of frames as in [16] and the moving regions are segmented based on the streakflows aiming to obtain compact inter-connected regions. Streakflows and their location are calculated for the moving regions in each set of frames. The features of the moving regions are then modelled by the differences between all pairs moving regions across the given set of frames. The dynamic changes of the features are modelled by the differences between all moving regions in one set of frames. The size of the number of frames, considered for the dynamic window from Section III, is set to $n = 13$. The data is represented by a 2-column matrix over time as described in Section IV. KDE is applied over a fixed grid size using the 2-column feature matrices as input data. In this study, we use the bivariate KDE method proposed in [17] which is based on using linear diffusion processes. The KDE methodology from [17] assumes the kernel to be Gaussian and uses a bandwidth selection method such that the bandwidth
In b) \(n\) refers to the number of histogram peaks.

Fig. 3. 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.

Fig. 4. Identifying when pedestrians stop during the video frames showing gathering and talking activities from the NUS-HGA dataset.

Fig. 5. Recognition results as \(K\) is varied when using KDE and histograms.

The bivariate kernel density estimation is computed over a fixed grid size of \(K \times K\). In our experiments, we examine the difference in recognition results as \(K\) is varied for KDE, when compared to histograms of the same size. Figure 5 shows the difference in recognition results between the histograms and KDE for grid sizes of 4, 8 and 16. In all three cases, a notable improvement can be seen when the KDE is used. We use the value \(K = 16\), because the results do not improve further when increasing \(K\), despite a higher computational complexity of the required processing. Representations of the PDFs are shown in Figure 6 for both motion and location. The walking motion shown in Figure 6a has a difference value close to 1 for the entire sequence, this implies that the motion is all quite similar, which is expected of the walking in group activity. The gathering motion shown in Figure 6b displays a variety of difference values, which is expected as some individuals are gathering coming from different direction. The walking activity location differences shown in Figure 6c are all close to 1. This implies that the individuals are tightly grouped, which is expected in the walk group activity. The gather activity location differences shown in Figure 6d display clear transitions between locations far apart to locations close together towards the end of this activity. This is expected, as the gathering activity involves individuals coming from a distance towards gathering in a small group at the end of the activity.

Fig. 6. KDEs for the motion and location differences of activities from the NUS-HGA dataset.

For classification purposes, the density estimations are sub-sampled and fed to the classifier independently. The results are then combined to form a discriminant model as the motion and location features are often complimentary. For the classifier we use SVM with the RBF kernel, considering the parameters \(C = 2.8284\) and \(\gamma = 0.0019531\). For all experiments, we follow the evaluation protocol described in [10], where the NUS-HGA dataset is split into 5-fold training and testing and the performance is evaluated by average classification accuracy.
A comparison of the results when compared to the state-of-the-art in group activity recognition is shown in Table I. 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 group activity recognition. The combination of all features provides the best overall result of 98%. We should remark that the group interaction zone method from [18] does not evaluate the results using the 5-fold training and testing as suggested in [10], therefore slightly different results are expected from their method. In comparison to the state-of-the-art methods, we achieve a clear improvement in results of about 2%, while using a fully automated method.

### VI. Conclusion

In this paper, we present an automatic approach for group activity recognition. We propose a model to describe the discriminative characteristics of group activity by considering the relations between motion flows and locations of moving regions in the scene as well as their dynamics in time. We also propose a scaling method to compensate for the effect of wide angles located at low height. Moreover, we propose a stationary pedestrian detector to keep track of stationary pedestrians by marking the locations where they stop moving. Kernel Density Estimation (KDE) is used to model both time-location and time-motion spaces for representing such interactions. Experimental results show the effectiveness of the approach, without relying on any manual annotation of tracks like in other approaches.

### ACKNOWLEDGMENT

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

### REFERENCES


