Article:
Stephens, Kyle and Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 (2018) Modelling of interactions for the recognition of activities in groups of people. Digital Signal Processing. 34–46. ISSN 1051-2004

https://doi.org/10.1016/j.dsp.2018.03.021

Modelling of Interactions for the Recognition of Activities in Groups of People

Kyle Stephens and Adrian G. Bors*

Department of Computer Science, University of York, York YO10 5GH, United Kingdom

Abstract

In this research study we adopt a probabilistic modelling of interactions in groups of people, using video sequences, leading to the recognition of their activities. Firstly, we model short smooth streams of localised movement. Afterwards, we partition the scene in regions of distinct movement, by using maximum *a posteriori* estimation, by fitting Gaussian Mixture Models (GMM) to the movement statistics. Interactions between moving regions are modelled using the Kulback-Leibler (KL) divergence between pairs of statistical representations of moving regions. Such interactions are considered with respect to the relative movement, moving region location and relative size, as well as to the dynamics of the movement and location inter-dependencies, respectively. The proposed methodology is assessed on two different data sets showing different categories of human interactions and group activities.

Keywords: Human Interactions, Human Group Activity, Kullback-Leilbler divergence, Kernel Density Estimation, Gaussian Mixture Models

1. Introduction

Modelling physical interaction between people and the recognition of group activities are important computational tasks in many applications including: security, human safety, human-computer interaction, video retrieval, designing better social spaces, personalised analytics, among others. Human activities are recognised based on recording the movement of people from a certain space followed by machine learning based training and decisions. While wearable devices can be used for the acquisition of precise, localised body movements [23, 32], video recordings of human activities provide the contextual information of the human activity under observation. In this research study we consider video recordings of a scene where a group of persons is involved in various activities. Research on human activity recognition (HAR) focused mostly on analysing video sequences showing single individuals. However, many human activities take place in a social context, where people interact with each other and with the rest of the scene. We address the challenges related to how the movements are related to each other and to the surroundings. Human activities can vary considerably from simple movements such as gestures, simple actions, human to human interactions, human interactions with the surroundings, to more complex group activities. Two types of interactions can be identified in groups of people: those involving physical contact and by imitation. Examples of first type include shaking hands or fighting, while for the second type we can consider walking or running within a group. In HAR there are several challenges, including movement occlusion due to other persons interposing with the field of view of the camera, non-uniform changes in illumination, involving shadows of moving persons and unexpected reflections of lighting in the scene, camera movement, noise and compression artefacts among many others.

Many of the existing group activity recognition (GAR) algorithms require manually placed markers in order to identify the persons and their movements in the scene, In this paper we propose an automatic method

*Corresponding author, E-mail: adrian.bors@york.ac.uk

for group activity recognition by modelling the inter-dependent relationships between features characteristic to human movements and interactions. Moreover, the proposed methodology extends the modelling of interactions to their dynamics in time and space. In order to ensure the robustness of localised movement modelling, we employ streaklines [27], which addresses the challenges posed by noise or illumination change in the scene. Compactly moving regions, are represented statistically as Gaussian Mixture Models (GMM), in both movement and location spaces, similarly to the approach from [6]. We address the challenges of modelling complex interactions under occlusions between multiple moving persons, by modelling the inter-dependency between moving regions, using the Kullback-Leibler (KL) divergence between their relative movement or their location in the scene. The dynamics of such models of movement interaction and relative inter-location dependencies is also considered in order to model the changes emerging in movement. The interactions with the surroundings are considered in the model by embedding the background as one of the moving regions. The proposed group interaction model keeps track of stationary pedestrians by automatically marking the locations where they stop and by identifying when they would start moving again. Eventually, sampled Kernel Density Estimation (KDE) of the feature vectors are used to represent normalized inputs to a machine learning classifier. Section 2 provides an overview of previous works in the human and group activity recognition literature. Section 3 describes the probabilistic framework of the approach, while Section 4 describes the moving regions segmentation. Section 5 describes the modelling of the inter-dependencies between moving regions. Section 6 presents the classification approach for group activities. Section 7 provides the experimental results on two different datasets showing group activities, and Section 8 draws the conclusions of this research study.

2. Related Works

Initial approaches for human activity recognition (HAR) relied upon extracting sparse spatio-temporal features [16], and then modelling them statistically or syntactically for recognizing activities from the video sequence. Appearance based features, representing solutions of Poisson equations, have been used by Gorelick *et. al.* in [19]. A generative method using the Probabilistic Latent Semantic Analysis algorithm was proposed by Niebels *et. al.* in [29]. In this approach, activities are represented as temporal successions of movements, which are modelled using the volumetric feature detector from [16]. Gaidon *et al.* [18] proposed to model activities as a sequence of actoms, which represent semantically meaningful parts of an activity, while Histograms of Gradients have been used in [3].

Another category of approaches consists of extracting and matching body postures between frames [31, 35, 40]. Simple image template matching was considered in [4], which was extended to 3-D spatial-temporal patches in [22, 33]. Graph-based modelling and matching of shape models, was proposed in [39]. The disadvantage of silhouette-based methods, aiming to model body postures, rests in the difficulty of the automatic extraction of precise and robust shapes representing moving bodies, particularly when other moving objects are around in the scene.

Trajectory-based approaches model the movement as a set of trajectories over groups of frames. Wang *et al.* [42] proposed a trajectory based method by tracking patches extracted at multiple scales for HAR. Probabilistic methods such as Hidden Markov Modells (HMM) have been used for representing interaction gestures in [30], for modelling activities in the office [44], and for modelling trajectories for HAR in [14]. The disadvantage of state-based sequential modelling approaches is their limited generalization ability. Li *et al.* [24] used dynamic textures for detecting anomalies in video sequences. An observational system, which after recording a dictionary of specific activities for a scene during a training stage, can identify new activities by using a statistical test, was proposed in [36, 38]. Neural networks and fuzzy systems have been used for identifying and combining a set of micro-behaviours in [1]. Long Short Term Memory (LSTM) networks, which are a variant of recurrent Neural Networks (RNN) using deep learning, have been used for extracting patterns of observations of human activities from image sequences in [26]. A two-stream LSTM architecture which incorporates spatial and temporal networks for detecting specific still frames and movement, respectively, was proposed in [34]. A deep network integrating LSTM with saliency-aware deep 3-D convolutional neural networks (CNN) features from video shots, was proposed in [43]. Using deep learning for video processing applications, such as HAR, is still in its infancy, and existing approaches consider the

information from individual frames or identify the changes from within short sequences of images. CNNs and especially RNNs require significant computation power and huge data sets for efficient training.

Table 1: Categories of algorithms used for human and group activity recognition.

Application	localised Spatio-temporal Features	Matching Body Postures	Trajectory based Approaches	Convolution Neural Networks (CNN)
Individual HAR	Sparse spatio-temporal features [16]	Matching shape sequences [40]	Motion boundary descriptors [42]	Two-Stream CNN [34]
	Non-Linear Stationary Subspace Analysis [3]	Viewpoint manifold [35]	Hierarchical Bayesian [30]	Long Short Term Memory model [26]
	Actions as space-time shapes [19]	Poselet Key-Framing [31]	Layered HMM [44]	Saliency-Aware 3-D CNN [43]
	Probabilistic Latent Semantic Analysis [29]	Image template matching [4]	Event probability Sequences [14]	
	Temporal Localization with Actoms [18]	Space-time Correlation [33]	Dynamic Textures [24]	
		Shape and Flow Correlation [22]	Hierarchical neuro-fuzzy [1]	
		Graph-based Matching [39]		
		Observational Dictionary [36]		
Group Activity Recognition	localised Causalities [28]	Interactive pose-based [21]	Probabilistic interactions [15]	
	Monte Carlo Tree Search [2]		Multiple-layered model [11]	
			Group interaction zone [12]	
			Probabilistic group-level [9]	
			Heat-Maps [25]	
			Collective activities [13]	

Algorithms used for individual person activity recognition can not always be extended in order to be used for group activity recognition (GAR). Group activities in video sequences involve multiple participants performing a wide range of movements, interacting with each other and with their surroundings. Through movement multiple persons would overlap each other from the field of view of the camera, raising challenges for GAR. Probabilistic analysis of group interactions in the dynamic context was proposed in [15]. A multi-camera system was used in [9] for tracking multiple people and their movements, while a hierarchical semantic granularity approach was employed for GAR in [13]. Interactive activity recognition using pose-based spatio-temporal relation features was used in [21]. In the study by Ni *et al.* [28], group activities are recognised using manually initialised tracklets, while Monte Carlo tree search in the context of bag of words mixtures was employed in [2]. A heat-map based algorithm was used for modelling human trajectories when recognizing group activities in videos, [25]. Gaussian processes modelling time-series of movement trajectories was employed in [11]. GAR by defining group interaction zones based on the relative distance between the humans in the scene was proposed in [12]. Most of these algorithms rely on either the manual annotation of trajectories, or by marking the people taking part in the activities. Modelling the inter-relationships between the moving regions, using an automatic approach based on the segmentation of moving regions [6, 7], was used in [37]. An outline of the main categories of approaches for HAR and GAR is provided in

Table 1.

3. The Framework for Group Activity Modelling

The proposed methodology is characterized by a hierarchical modelling structure as shown in the block diagram from Figure 1. In the following we consider that the activity taking place in the scene is made up of all the inter-dependencies between any two moving regions found in the scene. The recognition of a group activity $\mathcal{G}_j$ is achieved for:

$$p(\mathcal{G}_j|\mathbf{I}(t)) > p(\mathcal{G}_i|\mathbf{I}(t)) \tag{1}$$

where we consider that we identify N regions of movement, characterized by consistent movement, and $\mathcal{G}_i$, $i = 1, \ldots, N^2 - 1$, $i \neq j$ represent all movement inter-dependencies, by pairing the given N regions from the video sequence $\mathbf{I}(t)$. A group activity is given by:

$$p(\mathcal{G}_i|\mathbf{I}(t)) = \prod_{i=1}^{N^2} p_i(\mathcal{A}_k, \mathcal{A}_l|\mathbf{I}(t)) \tag{2}$$

where $k, l = 1, \ldots, N$, and N is the number of moving regions identified in the scene and $p_i(\mathcal{A}_k, \mathcal{A}_l|\mathbf{I}(t))$ represents the probability of ith inter-dependence between two regions of movement $\mathcal{A}_k$ and $\mathcal{A}_l$, [37]. This model incorporates the interactions between the people and their surroundings, given that moving objects, such as cars for example, would constrain the movement of people and may interact with them as well.

The proposed system starts with identifying and estimating localised movement in the scene. Using the local consistency of local movement, we segment the moving regions, as in [6]. The moving regions, depending on the context, can represent the entire movement of a person or that of a specific body part of an individual. In recordings with strong perspective projection effects, the persons located far away would look small in the frame and may be identified as a single moving region. The interaction with other moving regions, representing vehicles for example, can be included in the model as well. We can identify two types of interactions in groups of people: by direct contact, as in fighting or shaking hands, and by synchronisation or imitation, as in walking or running in group. The interacting moving regions $\mathcal{A}_k$ and $\mathcal{A}_l$ can be bordering each other or they may be located in different regions of the same scene., even though the activities from the first category would rather be bordering each other. In this approach, we identify and segment the moving regions $\mathcal{A}_j$ $j = 1, \ldots, N$ found in the scene, and then consider the interactions within pairs of such regions in order to model the equations (1) and (2). The interactions are modelled as the relative motion between the moving regions, as well as the inter-dependencies between their locations and shape representations. Eventually, a feature space of such interactions, or of their dynamics in time, is extracted and fed into a machine learning classifier for identifying each group activity. In the following sections we describe each component of the proposed group activity recognition framework in detail.

4. Identifying the Moving Regions

We consider that each region k, representing a compact area of consistent movement in the scene, is defined by its movement statistics, as well as its location and shape characteristics. Each moving region k is characterized by a Gaussian probability density function (pdf) $G_{k,M}(\mu_{k,M}, \Sigma_{k,M}|\mathbf{I}(t))$, where $\mu_{k,M}$ represents the mean of the local movement vector and $\Sigma_{k,M}$ is the covariance matrix. Meanwhile, the area where the activity takes place is also defined by a Gaussian $G_{k,D}(\mu_{k,D}, \Sigma_{k,D}|\mathbf{I}(t))$, with the center in $\mu_{k,D}$ and modelling its shape by an ellipse characterized by the covariance matrix $\Sigma_{k,D}$. In this way we associate a specific area of the scene with the human activities taking place there, which is an appropriate assumption when modelling the activities for a particular scene, while it can be considered for the recognition of group activities in a similar context as well. The whole movement in the scene is modelled by a mixture of multidimensional Gaussian distributions, each representing a moving region:

$$\prod_{k=1}^{N} p(\mathcal{A}_k|\mathbf{I}(t)) = \sum_{k=1}^{N} \lambda_k G_{k,M}(\mu_{k,M}, \Sigma_{k,M}|\mathbf{I}(t)) G_{k,D}(\mu_{k,D}, \Sigma_{k,D}|\mathbf{I}(t)) \tag{3}$$

where N represents the number of Gaussian components, which make up N moving regions, defining their movement and location in the scene. The movement of a pedestrian located far away from the camera, may be entirely represented by a single Gaussian, while the more complex movement of a group activity, close to the camera, would be represented by several Gaussian functions. In the spatial domain a set of Gaussians would represent the shape of a moving region as a combination of ellipses.

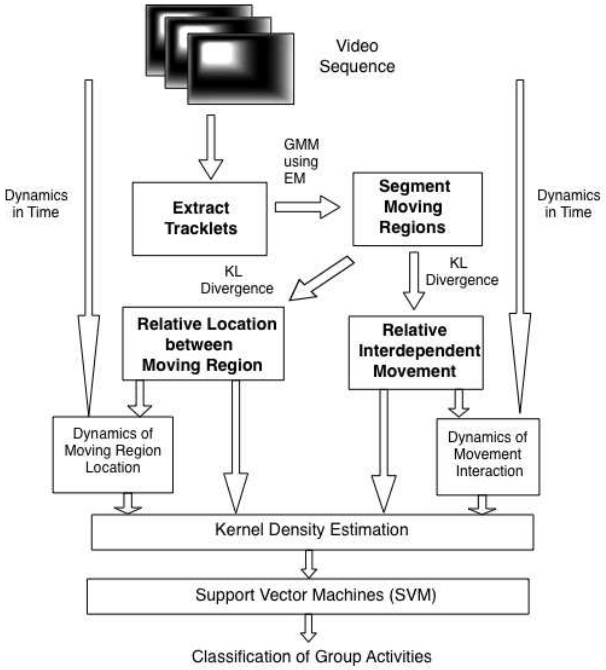


Figure 1: Overview of the proposed group activity recognition approach.

Optical flow vectors have been statistically modelled in [6] using the Median Radial Basis Functions (RBF) network and then used to segment the moving regions from video sequences. Such moving regions are then tracked in [7] using piecewise linear modelling. In this study, we model the probability density functions $G_{k,M}(\mu_{k,M}, \sigma_{k,M}|\mathbf{I}(t))$ of the local movement, by using streakflows proposed in [27]. Streakflows, called also tracklets, are able to represent complex localised movement using a model based on diffusions [17, 20]. The modelling of streaklines is based on the Lagrangian framework for fluid dynamics, ensuring the robustness and the continuity of movement estimation when considering particles of fluid with well defined trajectories. Median estimation of optical flow was shown in [6] to provide robust movement estimation for moving regions, while in [17, 20] it was shown to smooth the complex flows using fluid dynamics, while removing the outliers, instead of diffusing and mixing them with the flow's information. Unlike in the approach from [27], where each streakline corresponds to the movement of a single pixel, in this research study we associate each streakline with blocks of pixels of a fixed size by computing the marginal median of streakline vectors for each block of pixels as it was used for smoothing optical flow associated with moving regions in [6, 17]. Moreover, the computational complexity is reduced by a significant factor, by replacing the calculation of the motion estimation vectors for every pixel with the robust estimate for the movement of an entire block of pixels. However, when considering large blocks of pixels for movement estimation, the local precision of changes in movement will be rather coarsely estimated. In order to improve the resolution of the movement estimation we can consider overlapping blocks leading to a trade-off between the accuracy of the boundaries for moving regions with the required computational complexity. A streakline consists of several vectors head-to-tail located along a localised trajectory of movement to which a first degree polynomial is fitted. This provides a smooth model for the trajectory of local data. In this study, instead of using a single representative vector as in [36], we consider all the vectors, making up each streakline. In this way,

each moving region is characterized by several vectors of movement defined smoothly both spatially and temporally within their streakline.

The general assumption is that the movement in the scene corresponds to moving people, but interactions with other moving objects from the surroundings, such as vehicles for example is accounted for in this model. Firstly, we begin by segmenting the streakflow field into distinctly and compactly moving regions by assuming a probabilistic estimation model, as in [6]. The Expectation-Maximization (EM) algorithm, considering the Gaussian Mixture Model (GMM) modelling, of components $G_{k,M}(\mu_{k,M}, \sigma_{k,M}|\mathbf{I}(t))$ and $G_{k,D}(\mu_{k,D}, \sigma_{k,D}|\mathbf{I}(t)))$ from (3) are used for segmenting and modelling each inter-connected region. The EM algorithm estimates the relevant parameters, eventually leading to the partition of the scene using the maximum *a posteriori* estimation according to the local similarity in the local movement, while enforcing the compactness of the resulting moving regions. The number of clusters and the centers of the Gaussian functions in the EM algorithm are initialized using the modes of the histograms representing the streakline flows.

The perspective projection effect in the recording cameras can cause distortions which may significantly influence the modelling of the movement in the scene. Such perspective distortions are stronger in the case of video sequences acquired by wide-angle video surveillance cameras located at low to medium heights. The assumption here is that in the upper part of the video frames, objects and their motion are smaller than in the lower part of the scene, due to the perspective projection. In our approach we consider an automatic approach in two steps for correcting the effects caused by the perspective projection. During the first step, segmented regions of movement are used in order to estimate a scaling parameter which is associated with approximating the perspective effect in the scene as it is recorded. The movement vectors from the scene are then appropriately scaled according to their location in the scene and the scaling parameter. The scene is then segmented into moving regions, by using the rescaled movements.

This study also considers the situation when people stop moving. The decision that a moving region is stopping, is taken when no movement is detected for p frames in region corresponding to the moving region and its surroundings. The surroundings are defined by a boundary around the region, representing a percentage w of the size of the moving region. Such stationary regions are characterized by their location and by zero motion. Eventually, in most cases, movement would be detected again in the region marked as corresponding to the stationary person, by considering its surrounding areas as well. Persons leaving the scene are identified and removed from the context of the group activity modelling.

5. Inter-dependent movement modelling

The relative movement of each moving region with respect to all the others is crucial for defining interactions. The inter-dependent relationships are calculated as the relative differences in the statistics of movement or in the relative locations and shape of the moving regions. At this processing stage we aim to identify whether moving regions are similar or antagonistic in their direction and intensity of movement. For example, if individuals in a group exhibit similar streakflows during a specific activity such as when for example running in a group, their corresponding probability density functions modelling their movement would be similar as well, and consequently statistical differences in their streakfow estimates would be negligible. On the other hand, if individuals in a group exhibit very different streakflows, for example as in the case when people are ignoring each other, or when fighting, the inter-dependent movement modelling would indicate significant differences, depending on the intensity of the physical actions involved. In the following, Kulback-Leibler (KL) divergence is used to estimate statistical inter-dependecies within pairs of moving regions. KL was used in [36] in order to identify new activities in a scene, by comparing their statistics with those representing a dictionary of activities identified during a training stage. The symmetric KL divergence between the streakflow models is used for characterizing the inter-dependence between two regions of movement $\mathcal{A}_k$ and $\mathcal{A}_l$, at time t:

$$D_{SKL}(\mathcal{A}_k, \mathcal{A}_l) = \int_{-\infty}^{\infty} \left(p(\mathcal{A}_k) \ln \left(\frac{p(\mathcal{A}_k)}{p(\mathcal{A}_l)} \right) + p(\mathcal{A}_l) \ln \left(\frac{p(\mathcal{A}_l)}{p(\mathcal{A}_k)} \right) \right) dx \tag{4}$$

The KL divergence is considered as the energy measuring the interaction between two regions of movement, represented as the probability density function of inter-dependence and interaction between the regions $\mathcal{A}_k$ and $\mathcal{A}_l$, given by:

$$p(\mathcal{A}_k, \mathcal{A}_l|\mathbf{I}(t)) = \exp\left(-\frac{D_{SKL}(\mathcal{A}_k, \mathcal{A}_l)}{S_M}\right), \tag{5}$$

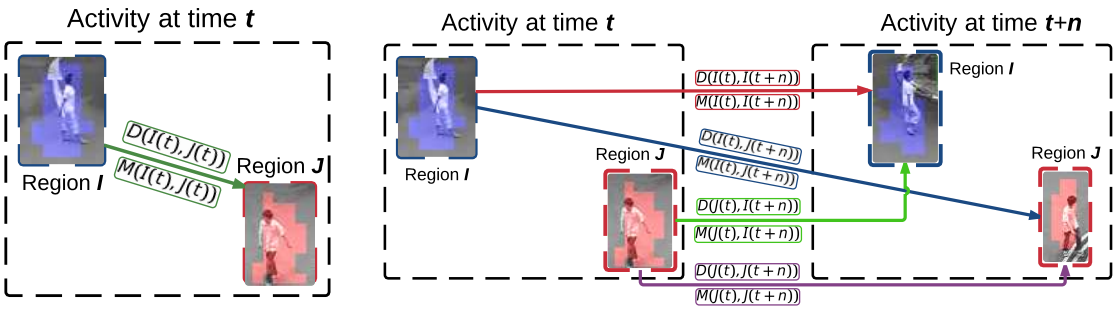
where S_M represents a scaling factor for the interactive energy of movement in between the given regions k and l. These pdfs become $p(\mathcal{A}_{k,M}, \mathcal{A}_{l,M}|\mathbf{I}(t))$ when representing the inter-dependence between two regions $\mathcal{A}_{k,M}$ and $\mathcal{A}_{l,M}$ in the motion domain and $p(\mathcal{A}_{k,D}, \mathcal{A}_{l,D}|\mathbf{I}(t))$, when representing the location, size and shape of the moving regions. By considering the scaling for compensating for the perspective projection effects, as described in the previous section, we will appropriately model the statistical distributions of the moving regions of the scene characterized by similar types of movement, regardless where they are located in the scene.

A similar approach is adopted for the relative localization and shape correspondence for pairs of moving regions. This model starts with forming distributions of location coordinates associated with each moving region and by comparing these statistically with the others. The distributions of relative locations for the moving regions characterizing people present in the scene, either moving or stationary, is modelled by considering differences between their corresponding GMMs. Each GMM $G_{k,D}(\mu_{k,D}, \sigma_{k,D}|\mathbf{I}(t)))$ models the spatial location, given by $\mu_{k,S}$, the size and the approximate shape of each moving region, represented by $\sigma_{k,S}$. The distributions for two regions k and l are compared statistically using the KL divergence, similar to the equation (4). An illustration for the evaluation of the inter-movement and that of relative locations of moving regions and shape correspondences is provided in Figure 2a.

Group activities are defined not only by instantaneous interactions, but also by the way how the interactions evolve in time. In the following we also consider the dynamic changes of relative differences between moving regions over subsequent frames by computing the variation of the inter-dependencies between segmented moving regions at different time intervals, as illustrated in Figure 2b. The dynamics of movement and location inter-dependences dynamics over time, is modelled using the KL divergence, similar to equation (4), except that the models are now located in a 3D spatio-temporal space, representing the dynamics across subsequent sets of frames. The scaling parameters for the energy functions, representing the dynamics of movement, similar to equation (5) are considered as S_{dM} and S_{dD}, for the dynamics of movement inter-relationships and for the differences in the localizations of the moving regions, respectively. The output of equation (5) provides a value in the range [0,1] representing the relationship between the two moving regions, either in the dynamics of their movement interaction or in their relative localization. For example, individuals characterized by the moving regions k and l at time t, located far apart, will have $p(\mathcal{A}_{k,D}, \mathcal{A}_{l,D}|\mathbf{I}(t)) \to 0$, whilst individuals located closer together will have $p(\mathcal{A}_{k,D}, \mathcal{A}_{l,D}) \to 1$. By accounting for the times when the moving regions would stop, as mentioned at the end of the previous section, the characteristics of the movement dynamics are better modelled for certain human activities. 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, indicating how such inter-dependences changes during n frames. The same modelling of dynamic changes is applied for changes in the relative distances between the localizations of the moving regions over n frames. In this way, we model the dynamics of the position of each moving region with respect to the others in the scene as well as their relative changes in size. This model assumes that a single group activity takes place, but it can easily be extended to consider several group activities in the same scene.

6. Feature representation by kernel density estimation

In this research study we propose four different feature representations for the inter-dependent region movement modelling, the relationships between the locations of the moving regions, and the dynamics of the inter-dependent movements and the corresponding dynamics of the changes in the relative location of the moving regions. In the case of the first and second data representations, the inter-dependence is modelled spatially, while in the case of the last two representations, the interactions are modelled in both



(a) Static model (b) Dynamic model

Figure 2: Modelling the inter-dependencies of moving regions, in the movement space $M(I(t), J(t))$, in the relative locations and shape correspondences space $D(I(t), J(t))$, as well as in their corresponding dynamic spaces.

time and space. In order to have a robust and normalized feature vector representation of the data, we use a non-parametric statistical data representation, such as the Kernel Density Estimation (KDE) [5, 8]. Then, the resulting feature vectors are fed into a machine learning classifier in order to decide the group activity, following a training stage.

In order to represent the dynamic data variation, we form two column matrices, where the motion or location inter-dependences for each pair of moving regions are represented in the first column, while their corresponding time instances are in the second column, denoting time reference markers. Such matrix representations are used for the dynamics of the moving region inter-dependence as well as for that of their relative locations and shape correspondences. The proposed representation, considering the localization of the moving regions in both space and time, their inter-dependence, as well as the dynamics of changes, ensures a good tracking of the interactions which take place in the scene. In this study, we use the bivariate KDEs employing diffusions on data representations, proposed in [8]:

$$\mathcal{F} = \sum_{i=1}^{L} \mathcal{K}_h(\mathbf{X}, \mathbf{X}_i) \tag{6}$$

where $\mathcal{F}$ is the pdf of the data, L is the number of kernels, each associated with a data sample, and $\mathcal{K}_h(\mathbf{X}, \mathbf{X}_i)$ is the kernel function calculated in the data space $\mathbf{X}$, which is considered as Gaussian:

$$\mathcal{K}_h(\mathbf{X}, \mathbf{X}_i) = \frac{1}{(2\pi)^{d/2} h^d} \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{2h^2}\right) \tag{7}$$

where h represents the bandwidth, d is the dimension of the data, $\mathbf{X}_i$, $i = 1, \ldots, L$ are the data used in the KDE estimation. For selecting the bandwidth h, which controls the smoothing, we use the method proposed in [8]. Some models of the movement interactions may exhibit very small changes in feature differences over time while others may have large, significant changes in the statistics of the interactions. KDE, provides good data smoothing and avoids overfitting to the training data, thus improving the generalization ability of the classifier when compared to a histogram-based data representation which may lead to training data overfitting. The smoothing provided by KDE, not only removes the noise but also provides smooth transitions between the models of the group activity features in time. Moreover, this allows for modelling the diverse granularity of different features to be represented in a consistent way.

The resulting KDE's are eventually sampled on a grid of fixed size $\mathbf{X} \in K \times K$. By using a fixed grid size for KDE, when representing (6), representing either the statistics of movement, or that of the location of moving regions, or that of their dynamics, respectively, we implicitly enforce that all feature sets are of the same size. By using a fixed grid size, we will be able to extract features of identical size from video sequences of different lengths, normalizing the feature space in both space and time. The grid size is an important parameter for representing the KDE. If K is small, the data would be coarsely represented and

it would result in over-smoothing, missing important characteristics. On the other hand if the grid size is too large, then the representation may lead to overfitting the training set, similar to when using histograms. The choice for K is analyzed experimentally in Section 7.

The sampled values of the KDE, using equations (6) and (7), on the $K \times K$ grid are then used as a feature vector for a classifier. In this study we consider the Support Vector Machine (SVM) algorithm, with the RBF kernel, having K^2 inputs, while the output separates each group activity class from all the others in a winner takes all architecture, [10]. The SVM with RBF kernels depends on two parameters C and γ. C is the parameter for the soft margin cost function, which controls the influence of each isupport vector. This process involves trading error penalty for stability. The γ parameter defines the influence of each data sample, similarly to the function of h in equation (7). If γ is large, then the support vector does not have wide-spread influence.

7. Experimental results

The proposed approach has been evaluated on two group activity datasets: the NUS-HGA [28] and the New Collective [41]. Both datasets contain several short video sequences, each showing only a single group activity at any one time. In some frames there are certain persons, who are not part of the group activity on display, who are crossing through the scene. In the NUS-HGA dataset the activities are pre-segmented temporally into separate video sequences, while the New Collective dataset contains video sequences where the activities flow from one activity to the next. Examples of Fighting and Gather activities from NUS-HGA dataset are shown in Figures 3a and 3b, while examples of the Queuing and Chasing activities from Colective dataset are shown in Figures 3c and 3d. The perspective projection effect is strong in the New Collective videos, as it can be observed in Figures 3c and 3d, because these video sequences are recorded by a hand-held camera of wide angle from a low vantage point.

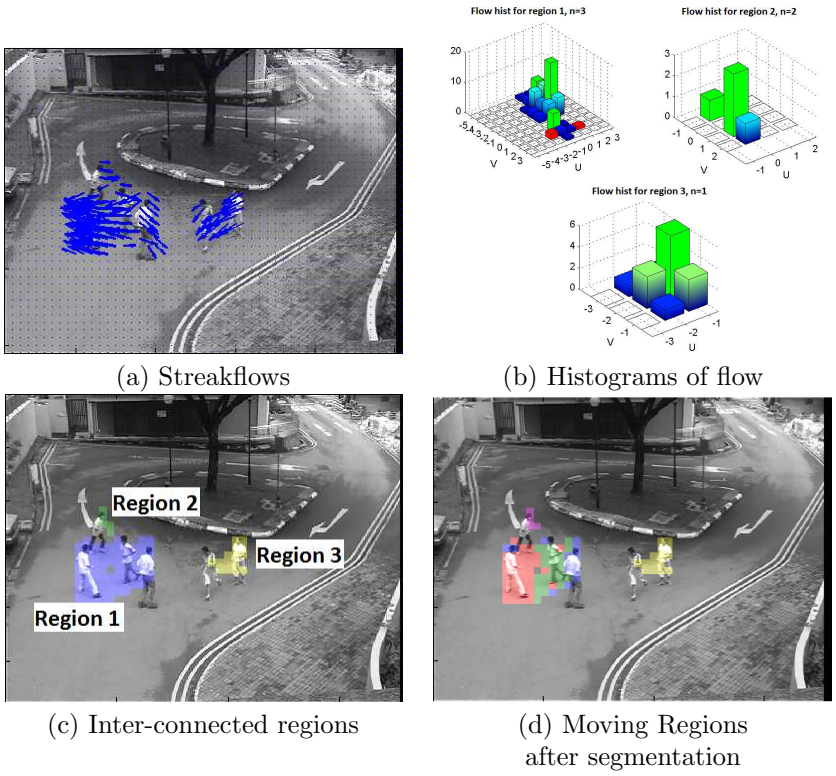
(a) Fighting (b) Gather (c) Queing (d) Chasing

Figure 3: Examples of images from NUS-HGA dataset in (a) and (b) and New Collective dataset in (c) and (d).

The NUS-HGA dataset consists of six different group activities: $\mathcal{G} = \{$Walk in Group, Ignore, Gather, Stand and Talk, Fight, Run in Group$\}$. In the following we aim to identify each of them, according to the framework from Section 3. The videos represent staged actions, which had been collected in five different sessions, each containing 476 video streams. The resolution of the video is 720×576 at 25 frames per second (fps), and each activity sequence is approximately 4 to 10 seconds long. The videos of the scene vary from one to another due to changes in the camera angle and lighting, while the tree visible in the center of the scene often casts shadows of various orientations, as it can be observed in Figures 3a and 3b. The streaklines are estimated from overlapping blocks of size 14×14 pixels over 10 frames, ensuring the motion consistency while contributing to the removal of camera movement or that of frivolous human hand movements during recording. Figure 4a shows an example of the estimated streakflows for the Fight activity from the NUS-HGA dataset. The movement in frames displaying the Fighting activity is intense and chaotic. Motion histograms are computed and any entry in the histogram with a height below 15% of the maximum bar height is considered to be noise and is subsequently removed. This procedure ensures the removal of movement vectors, which are not significant enough, from further processing. Histograms for the Fight activity are shown in Figure 4b. The moving regions are identified as corresponding to the peaks in these histograms. The moving regions for the Fight activity scene, corresponding to the histograms from

Figure 4b, are identified in the scene as shown in Figure 4c. The motion is segmented and each moving region is represented by a Gaussian Mixture Model (GMM) of statistics of streakflows vectors and their locations, respectively. The modes of the histograms are then used as inputs to the EM algorithm and the moving regions are segmented accordingly. In the example from Figure 4 it can be observed that small moving regions from the larger Region 1 of Figure 4d help characterize the smaller actions performed in the group such as pushing or kicking, which usually happen during the fighting activity.

We account for the perspective projection effects, where movements from smaller segmented regions correspond to regions located far away. The segmentation is done in two stages, where during the first segmentation stage, a scaling factor is calculated according to the location of the moving region along the y axis. Then the motion is scaled according to the perspective effect of the scene, assuming that moving regions located in the lower part of the scene are closer to the camera, while those from the upper part are farther away. The detection of the stationary regions is applied as explained in the last paragraph from Section 4, by estimating and modelling the movement during a set of $p = 25$ frames, preceding the identified stillness. We define a boundary parameter for identifying when the stationary region is moving again as being $w = 10\%$ of the region size. Two examples of detecting stationary pedestrians are shown in Figure 5 for the Talking and Gathering activities, respectively. In Figures 5a and 5c the pedestrians are still moving and therefore their corresponding moving regions are detected. In Figures 5b and 5d the individuals have stopped and their stationary regions are properly identified by the stationary pedestrian detector procedure.



(a) Streakflows

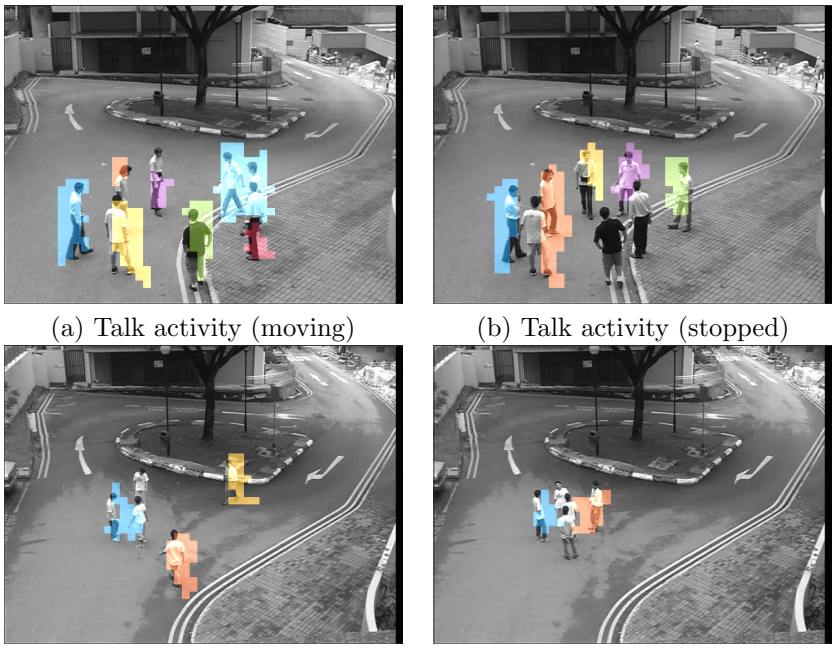
(b) Histograms of flow

(c) Inter-connected regions

(d) Moving Regions after segmentation

Figure 4: Data representations for the moving regions in one of the Fighting activity sequences from the NUS-HGA dataset. In b) "n" refers to the number of histogram peaks.

The modelling of the inter-dependency movement using the streakflows, the dynamics of such inter-dependencies, the relative location of moving regions and their dynamics are computed as explained in Section 5. The number of frames, considered for the dynamic window, when calculating the dynamic inter-dependencies in both movement and locations, is set to $n = 13$. The scaling parameters when calculating the inter-dependences between moving regions is set as $S_M = 15$ in equation (5), while the scaling parameter



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

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

Figure 5: Identifying when pedestrians stop during the video frames showing Gathering and Talking activities from the NUS-HGA dataset.

for representing the relationships between locations is $S_D = 550$. The corresponding scaling factors for the dynamics of motion and location inter-dependences are $S_{dM} = 17.5$, $S_{dD} = 650$, respectively. The scaling factors for the energy functions representing the dynamics of the inter-dependencies are higher than when modelling the static interdependencies, because the variation is higher in the KL divergencies calculated for the dynamics. Then, the bivariate kernel density estimation from [8] is computed over a fixed grid size of 16×16 for the estimates of the moving regions $p(\mathcal{A}_k, \mathcal{A}_l|\mathbf{I}(t))$. Representations of the KDEs, using the equations (6) and (7) considering Gaussian kernels, as described in Section 6, for various human interaction activities are shown in Figures 8a-8f for the inter-dependencies of the moving regions, and in Figures 8g-8l when modelling the differences in the locations of the moving regions. It can be observed from Figure 8e that the KDE's for location differences for the Ignoring activity are grouped into two large groups representing large and small differences, because individuals are well spread and constantly moving around. The KDE's of the dynamics of movement and location for the Walking activity, shown in Figures 8d and 8j are largely located in the same data range as those for Running, which are shown in Figures 8c and 8i. This is because the two activities are rather similar, but the dispersion of variations in the movement and locations is higher for the latter, resulting in more peaks in their KDE representations. The density of movement in the Fighting activity is higher than in the Ignoring activity, as it can be observed from Figures 8b and 8e. The movement in the Gathering activity, shown in Figure 8f displays a wide variety of difference values, which is expected because individuals are coming from different directions during the Gathering activity. The Walking activity location differences shown in Figure 8e are all close to 1. This implies that the individuals are tightly grouped, which is expected for the Walk in Group activity. The distributions for the location differences for Gathering activity, shown in Figure 8l, clearly display transitions between locations situated far apart varying towards becoming closer-together, as it would be expected to happen during this activity.

When considering the background as one of the regions of relative movement, the results have improved significantly as it can be observed from Figure 6. Actually, the background characterized by null movement represents the dominant region in the video sequences from NUS-HGA database. By using the background as one of the regions of movement, we implicitly consider the relative movement of all moving regions with

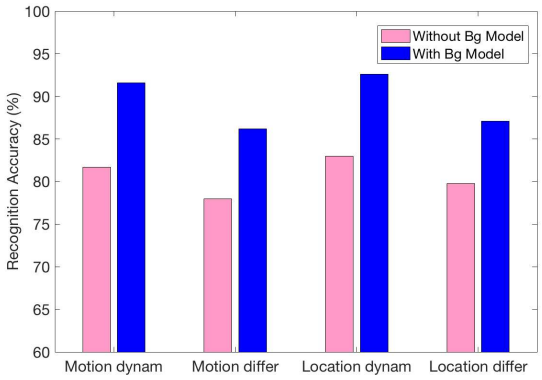


Figure 6: Results when adding the background as one of the moving regions in the NUS-HGA database.

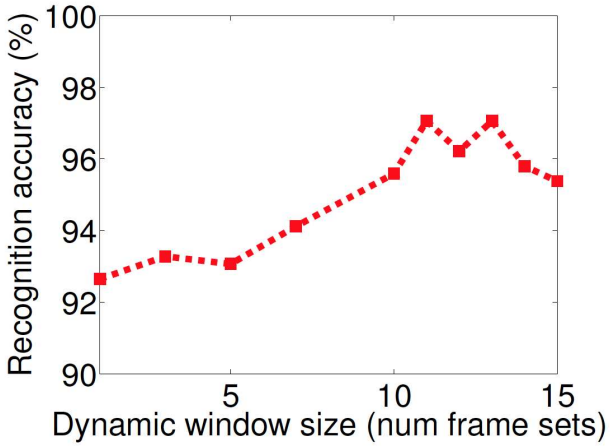
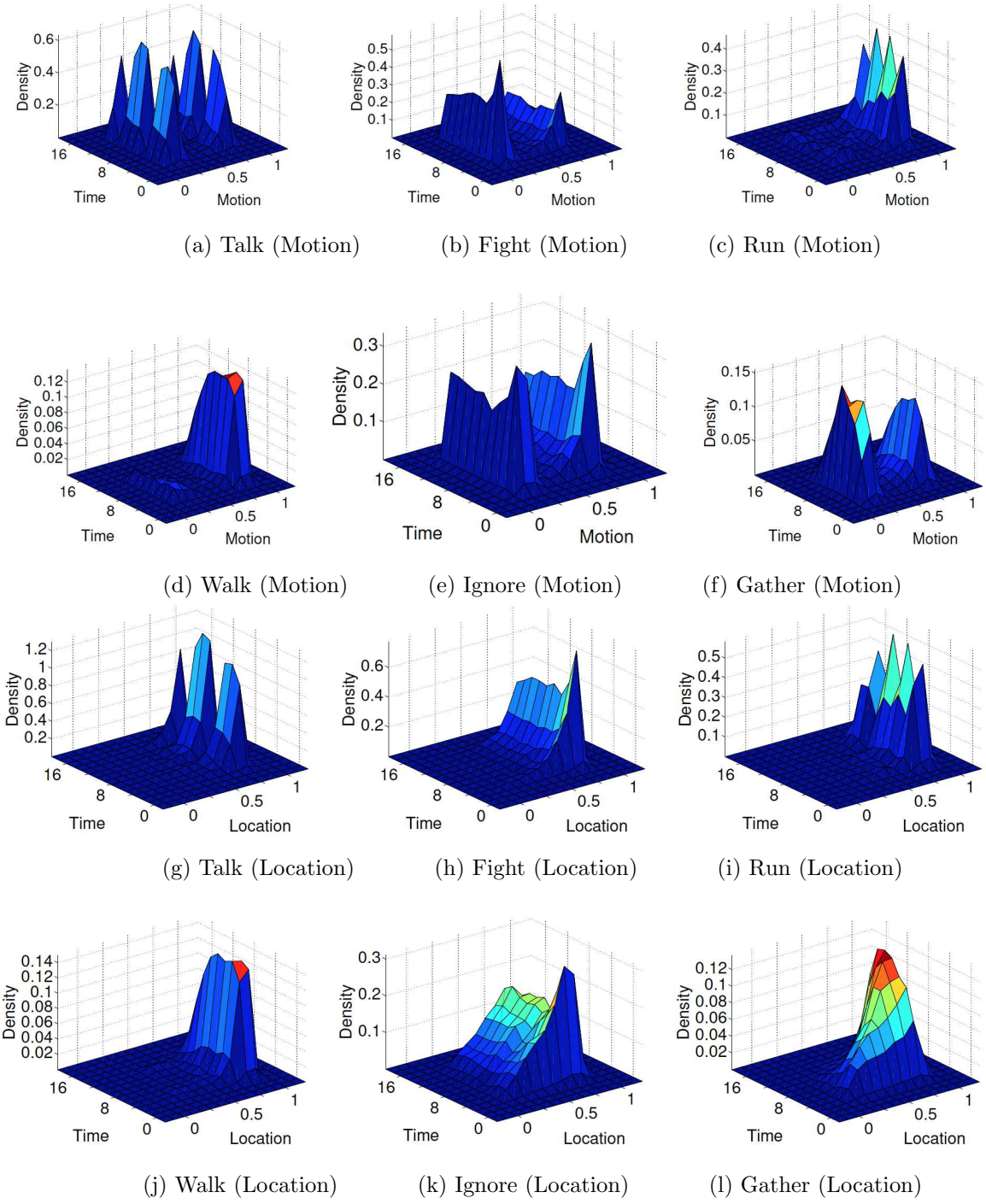


Figure 7: Results when varying the number of frames used for estimating the streakflows.

respect to their surroundings and thus implicitly model the interactions of the persons in the scene with the objects from the background. We consider varying the number of frames used for estimating the streakflows and the results are shown in Figure 7. From this figure we can observe that best results are achieved for either $n = 11$ or $n = 13$ frames, for estimating a streakline, confirming the efficiency of medium-term tracking measure for group activity classification in this database. If we would use fewer frames, we would only capture a very localised motion, while by considering longer streams of frames we would miss the interactive movements between people.

In order to form feature vectors of identical size, the KDE representations are sampled and considered as inputs to the SVM classifier with RBF kernels. The parameter, which trades off the misclassification of the training examples against the simplicity of the decision surface, is chosen as $C = 2.83$, while the parameter which defines the influence of a single training example is chosen as $\gamma = 1.95 \cdot 10^{-3}$. The results are then combined in order to form discriminant boundaries as the motion and location features usually represent complementary information in the video sequence, [6]. For all experiments, we follow the evaluation protocol described in [28], where the NUS-HGA dataset is split into 5-fold training and testing sets, while the performance is evaluated by the average classification accuracy.

Confusion matrices when using the motion inter-dependence as well as the relative localizations and shape correspondences between the moving regions for all the categories of group activities from NUS-HGA database are shown in Figures 9a and 9b. The dynamics of movement inter-dependence and of the relative



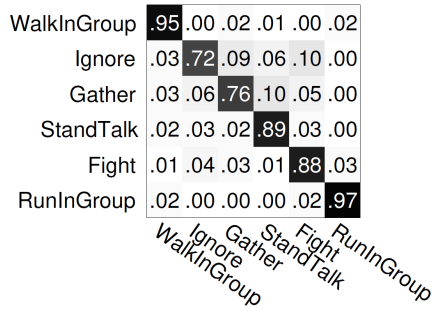
(a) Talk (Motion) (b) Fight (Motion) (c) Run (Motion)

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

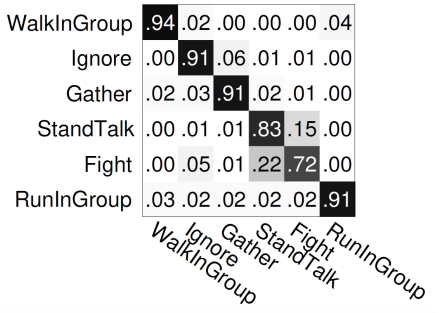
(g) Talk (Location) (h) Fight (Location) (i) Run (Location)

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

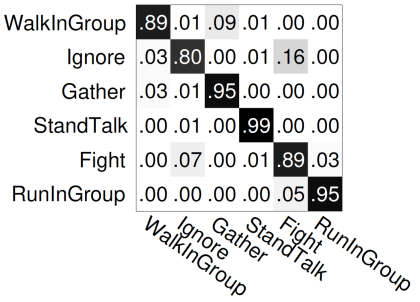
Figure 8: KDEs representing motion inter-dependencies in (a)-(f) and location differences in (g)-(l) for the NUS-HGA dataset.



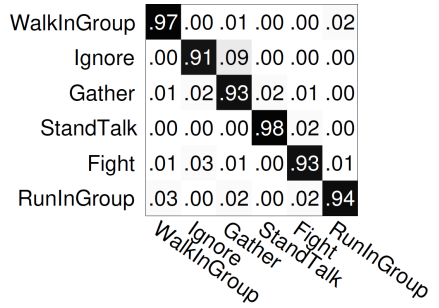
(a) Motion - 86.16%



(b) Location - 87.10 %)

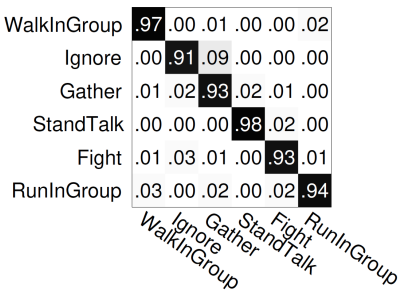


(c) Motion Dynamics - 91.59 %

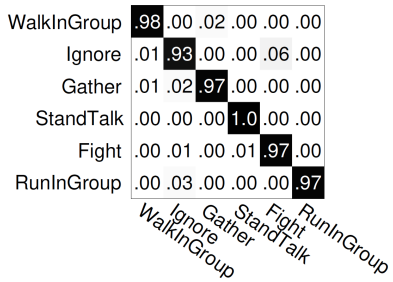


(d) Location Dynamics - 92.64 %

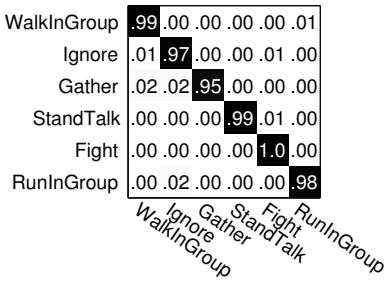
Figure 9: Confusion matrices for the group activities from NUS-HGA, when considering individual features.



(a) Combining Motion and Location - 94.50 %



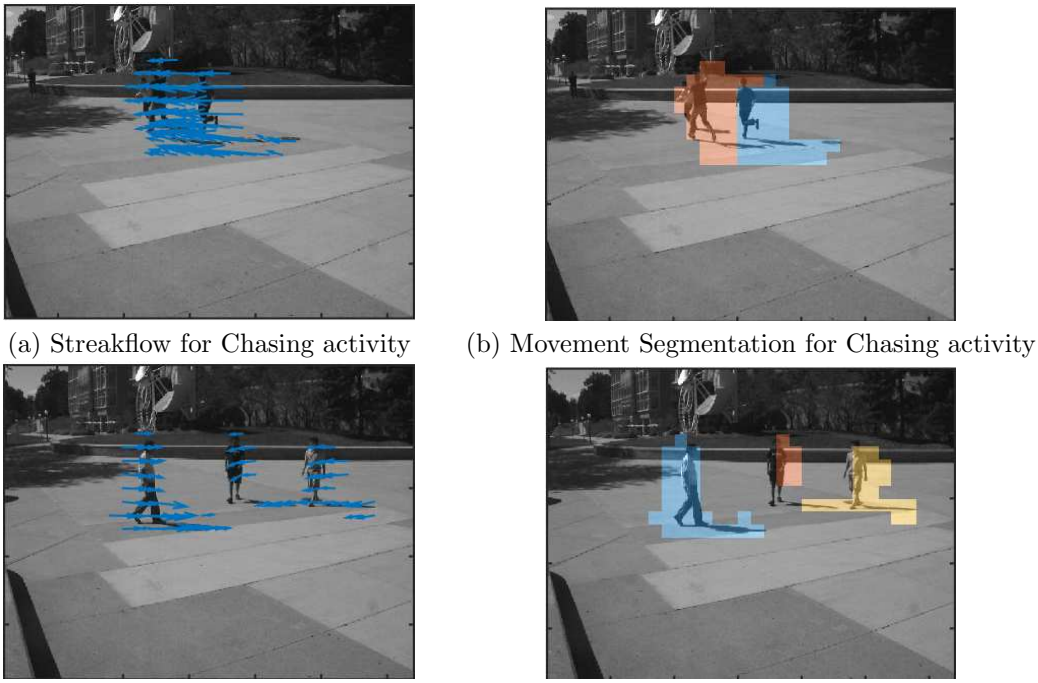
Combining the Dynamics of Motion and Locations - 97.07 %



(c) Combining all features - 98.00 %

Figure 10: Confusion matrices for the group activities from NUS-HGA, when combining features.

location inter-dependence are shown in Figures 9c and 9d. The location and shape correspondences features provides a better recognition result than the motion features, while the results improve significantly when considering the dynamics of either the movement or of changes in the relative locations of regions of movement in the scene. We also combine the movement and the location inter-dependence and the results are shown in Figure 10a. When combining the dynamics of motion and that of location inter-dependences, we obtain the results provided in the confusion matrix from Figure 10b. Eventually, we consider the combination of all four feature types, and the results achieved are provided in the confusion matrix from Figure 10c. The combination of all features results in 98 % classification accuracy, significantly improving the recognition results for the group activities from this database when using other methods.



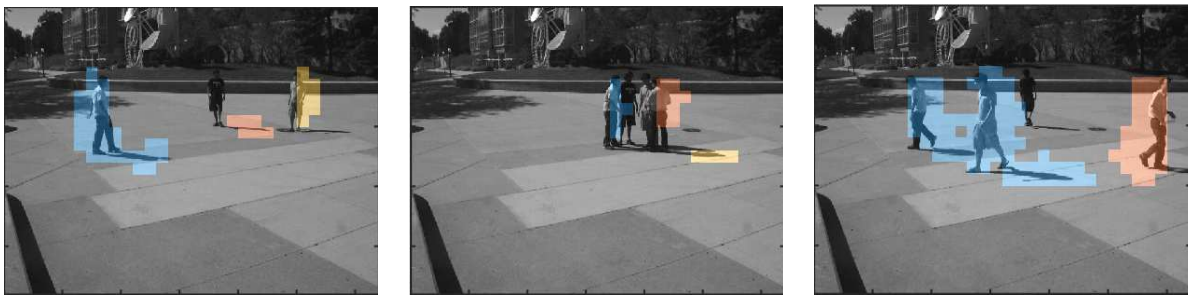
(a) Streakflow for Chasing activity (b) Movement Segmentation for Chasing activity

(c) Streakflow for Gathering activity (d) Movement Segmentation for Gathering activity

Figure 11: Examples of streakflows estimation and movement segmentation for two group activities from the New Collective dataset.

The New Collective dataset [41] consists of video frames showing human-group activities which correspond to six different classes: $\mathcal{G} = \{$Gathering, Talking, Dispersing, Walking together, Chasing, Queueing$\}$. This dataset consists of 32 video sequences, each containing multiple successive examples of several activities. Because these video sequences are acquired by a hand-held camera kept at a low height, they are characterized by perspective projection distortions. It can be observed in the frames shown in Figures 3c and 3d that people, moving closer to the recording camera, appear larger in the scene. The streaklines are extracted from overlapping image blocks of size of 20×20 pixels for each set of 10 frames. Examples of the streakflows for Chasing and Gathering activities are shown in Figures 11a and 11c, respectively. The streakflows capture well the human movement, free of the camera shaking, which is often present in some of the video sequences. This is due to the robust adaptation by using blocks of pixels for the estimation of streaklines. The corresponding movement segmentations for Chasing and Gathering activities are shown in Figures 11b and 11d, respectively. The moving regions are well segmented, particularly for the Chasing activity where the chaser and chasee are identified and segmented separately. The videos from the New Collective dataset contain a wide variation of human movement, with transitions from one activity to another, including times when the people taking part in the activities become stationary. The stationary pedestrian detector is used as described in Section 4 by considering a number of frames equal to $p = 25$

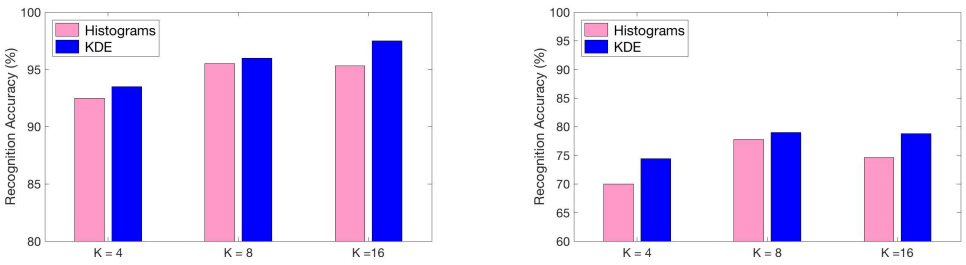
for detecting the stationarity of the people. An example of transition from a specific activity to a stillness status in pedestrians is illustrated in Figure 12. Initially, as shown in Figure 12a, the pedestrians are moving towards each other while performing the Gathering activity. Towards the end of this activity, people have gathered and the transition to the Talking activity takes place, as shown in Figure 12b. The pedestrians who are becoming stationary were successfully detected and their stopping locations recorded, as indicated in Figure 12b. Eventually, after some time, these pedestrians start moving again, performing the Dispersing activity as shown in Figure 12c. The newly emerging moving regions are then properly detected, replacing the previously stationary regions. The boundary parameter for identifying the activation of the movement is set as a bordering region, representing $w = 15\%$ of the moving region size. The human activity features, representing the streakflow differences, streakflow dynamics, relative localization and shape correspondences of moving regions, as well as their dynamics, are computed as described in Section 5. The scaling parameters for the inter-dependences between the moving regions for the New Collective database are the same as for the NUS-HGA database, except for that used for scaling the relationships between the moving region locations, which is set as $S_D = 450$.



a) Gathering b) Talking c) Dispersing

Figure 12: Pedestrians transitioning through various activities in the New Collective dataset.

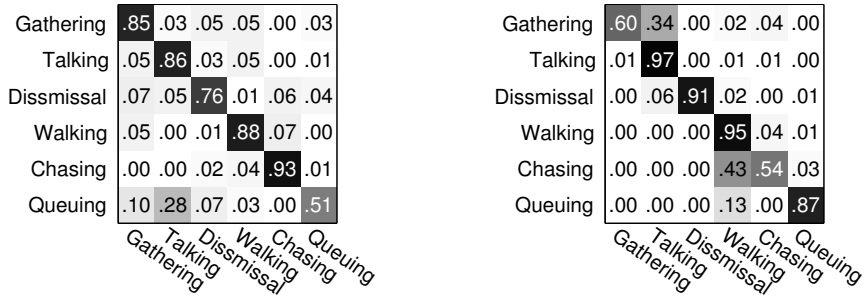
Unlike the NUS-HGA database, the activities from the New Collective dataset are unevenly distributed temporally and consequently the size of the dynamic window used when modelling the dynamics of motion inter-dependencies and the dynamics of the inter-region distances is set as $n = 5$ for this dataset. Then, the data is represented over time using KDE, as described in Section 6. Results when calculating KDE for various grid sizes, compared against those achieved when using histograms in the case of both NUS-HGA and New Collective datasets, are provided in Figures 13a and 13b. It can be observed that in all cases, the KDE representation provides better results than the histograms. Moreover, the computational complexity increases significantly when using histograms or when KDE's representations are sampled on a grid size of $K > 16$. In the following we consider a grid-based representation of size 8×8 for sampling the KDE. The sampling of KDE provides a vector of K^2 input entries for the SVM with the RBF kernel classifier.



a) NUS-HGA dataset b) New Collective dataset

Figure 13: Comparison between KDE with windows of various sizes and histograms for the NUS-HGA and New Collective datasets.

We divide the New Collective dataset into 3 subsets for 3-fold training and testing as in [41]. The sequences are split for the training by the start and end point of each activity as provided by the ground truth information. This approach is different from that from [41] where the video sequence is split into short sequences of a fixed length without taking into account the recorded activity. However, our approach for preparing the training data does not change the results with regard to the supervised learning approach, where the label of each activity at any given time is provided by the ground truth. Confusion matrices for all categories of movement activities, when considering the proposed set of features, are provided in Figure 14a. These results are slightly worse than those provided by the Multiple-layered model [11], shown in Figure 14b. However, it can be observed from Figure 14 that the results provided by the proposed methods show a greater consistency across all group activities, except for the Queuing, when compared to the Multiple-layered model [11].



(a) Combining all features - 79.7 % (b) Multiple-layered model, [11] - 80.3 %

Figure 14: Confusion matrices for GAR results when combining all features modelling movement, location distribution and their dynamics in the New Collective dataset, compared with the results from [11].

Table 2: Group activity recognition results for the NUS-HGA and New Collective datasets.

Method	NUS-HGA dataset (%)	New Collective dataset (%)
localised Causalities [28]	74.2	-
Group interaction zone [12]	96.0	-
Multiple-layered model [11]	96.2	80.3
Monte Carlo Tree Search [2]	-	77.7
Collective activities [13]	-	79.2
Motion inter-dependence	86.2	75.4
Moving regions location relationships	87.1	64.3
Motion inter-dependence dynamics	91.6	76.8
Dynamics of moving region location relationships	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

Comparative results for group activity classification are provided in Table 2 for NUS-HGA and New Collective datasets. The methods proposed in this study consider as discriminating features the modelling of the moving regions inter-dependencies, of the relative localization and shape correspondences, as well as the dynamics of inter-dependencies in movement among the moving regions and among their locations. We also consider the combinations of the two static features, as well as for their corresponding dynamical

feature sets, respectively. Moreover, we consider combining all four feature sets to be used as the input for the SVM classifier. These results are compared against those of other approaches: localised Causalities [28], Group interaction zone [12], Multiple-layered model [11], Monte Carlo Tree Search [2], and the New Collective activities [13]. The relative localization and shape correspondence features provide better results than the movement inter-dependence among the moving regions in the case of NUS-HGA, while these results are worse in the case of the New Collective database. Meanwhile, the models accounting for the dynamic inter-dependent movement and location of the moving regions, clearly provide better results that those when considering statically either the inter-dependence among moving regions or their relative localization. The results improve further when combining these features two by two, either for the stationary moving and locations features or for their corresponding dynamics, respectively. The best results are achieved when combining all four features sets discussed in this study, providing 98% recognition for the group activity results for NUS-HGA and 79.7% for those from the New Collective database. For the comparative algorithms, the group interaction zone method from [12] does not evaluate the results using the 5-fold training and testing as it was evaluated in [28] for the NUS-HGA dataset. The proposed methodology provides a clear improvement of about 2% over the next in line approach for the NUS-HGA dataset. For the New Collective dataset, the proposed method is comparable to the state-of-the-art and superior to the other methods for all group activities, except for the Queuing activity. However, while the proposed methodology is fully automatic, all the other comparative methods use some form of human intervention when identifying the human activities from video sequences.

8. Conclusion

This paper proposes a method for recognizing complex human actiivities involving several people interacting with each other. The proposed approach, consists of a hierarchical modelling, which at the base level relies on the calculation of the medium-term trajectories of movement from the video sequence. Maximum *a posteori* estimation, implemented through the EM algorithm, is then used for partitioning the scene into moving regions. The relative movement of each moving region with respect to each of the others identified in the scene, including the background, is then represented statistically using the Kulback-Leibler (KL) divergence. KL is also used for evaluating the relative localization and shape correspondences between pairs of moving regions. We also consider modelling the temporal dynamics of changes in the inter-dependences between pairs of moving regions or between the locations of the moving regions. The dynamics of change in the moving regions takes into account the timeline of events, including when people become stationary, as well. Kernel density estimation (KDE) is then used to represent consistently the statistics for each of these features. Uniformly sampled KDEs are then used as inputs for an SVM classifier. The best results are achieved when combining all four feature sets, modelling the inter-dependencies in movements among the moving regions, their relative location, as well as the dynamics of such inter-dependencies in the movements and the locations. The parameters corresponding to activities learnt based on the training video sets from a scene can be easily adapted to be used for recognizing the same group activities from video recorded from a different scene. Human activities in real spaces are more complex and more diverse than those analysed in this study. Such activities would invariably lead to occlusions which would cover certain movements in the scene from the view angle of the recording camera. Recording from multiple view points by multiple video cameras would provide a better representation of the interaction movements in the 3D context of the scene. Meanwhile, better deep learning training algorithms can be used in order to improve the recognition in real-life group activities assuming that many video streams are available for training. The proposed methodology can be applied for video surveillance, human-computer interaction, semantic annotation of multimedia, retrieval of video streams displaying human interactions.

Acknowledgment

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

References

[1] Acampora, G., Foggia, P., Saggese, A., Vento, M.. A hierarchical neuro-fuzzy architecture for human behavior analysis. Information Sciences 2015;310:130–148.

[2] Amer, M.R., Todorovic, S., Fern, A., Zhu, S.C.. Monte carlo tree search for scheduling activity recognition. In: Proc. IEEE Int. Conf. on Computer Vision (ICCV). 2013. p. 1353–1360.

[3] Baktashmotlagh, M., Harandi, M., Bigdeli, A.. Non-linear stationary subspace analysis with application to video classification. In: Proc. Int. Conf. on Machine Learning. 2013. p. 450–458.

[4] Bobick, A.F., Davis, J.W.. The recognition of human movement using temporal templates. IEEE Trans Pattern Analysis and Machine Intelligence 2001;23(3):257–267.

[5] Bors, A.G., Nasios, N.. Kernel bandwidth estimation for nonparametric modelling. IEEE Trans on Systems, Man and Cybernetics, Part B: Cybernetics 2009;39(6):1543–1555.

[6] Bors, A.G., Pitas, I.. Optical flow estimation and moving object segmentation based on median radial basis function network. IEEE Trans on Image Processing 1998;7(5):693–702.

[7] Bors, A.G., Pitas, I.. Prediction and tracking of moving objects in image sequences. IEEE Trans on Image Processing 2000;9(8):1441–1445.

[8] Botev, Z., Grotowski, J., Kroese, D.. Kernel density estimation via diffusion. Annals of Statistics 2010;38(5):2916–2957.

[9] Chang, M., Ge, W.. Probabilistic group-level motion analysis and scenario recognition. In: Proc. IEEE Int. Conf. on Computer Vision (ICCV). 2011. p. 747–754.

[10] Chang, Y.W., Hsieh, C.J., Chang, K.W., Ringgaard, M., Lin, C.J.. Training and testing low-degree polynomial data mappings via linear svm. Jour Machine Learning Research 2010;11:1471–1490.

[11] Cheng, Z., Qin, L., Huang, Q., Yan, S., Tian, Q.. Recognizing human group action by layered model with multiple cues. Neurocomputing 2014;136:124–135.

[12] Cho, N.G., Kim, Y.J., Park, U., Park, J.S., Lee, S.W.. Group activity recognition with group interaction zone based on relative distance between human objects. Int Journal of Pattern Recognition and Artificial Intelligence 2015;29(5):1555007:1–15.

[13] Choi, W., Savarese, S.. Understanding new collective activities of people from videos. IEEE Trans Pattern Analysis and Machine Intelligence 2014;36(6):1242–1257.

[14] Cuntoor, N.P., Yegnanarayana, B., Chellappa, R.. Activity modeling using event probability sequences. IEEE Trans on Image Processing 2008;17(4):594–607.

[15] Dai, P., Di, H., Dong, L., Tao, L., Xu, G.. Group interaction analysis in dynamic context. IEEE Trans on Systems, Man and Cybernetics, Part B: Cybernetics 2009;39(1):34–42.

[16] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.. Behavior recognition via sparse spatio-temporal features. In: Proc. IEEE Int. Workshop on Visual Surveillance and Performance. 2005. p. 65–72.

[17] Doshi, A., Bors, A.G.. Robust processing of optical flow of fluids. IEEE Trans on Image Processing 2010;19(9):2332–2344.

[18] Gaidon, A., Harchaoui, Z., Schmid, C.. Temporal localization of actions with actoms. IEEE Trans Pattern Analysis and Machine Intelligence 2013;35(11):2782–2795.

[19] Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.. Actions as space-time shapes. IEEE Trans Pattern Analysis and Machine Intelligence 2007;29(12):2247–2253.

[20] Gudivada, S., Bors, A.G.. Robust processing of optical flow of fluids. Pattern Recognition 2015;48(12):4097–4115.

[21] Huynh-The, T., Le, B.V., Lee, S., Yoon, Y.. Interactive activity recognition using pose-based spatio?temporal relation features and four-level pachinko allocation models. Information Sciences 2016;369:317–333.

[22] Ke, Y., Sukthankar, R., Hebert, M.. Spatio-temporal shape and flow correlation for action recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2007. p. 1–8.

[23] Labrador, M.A., Lara Yejas, O.. Human Activity Recognition: Using Wearable Sensors and Smartphones“. Chapman and Hall/CRC, 2013.

[24] Li, W., Mahadevan, V., Vasconcelos, N.. Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Analysis and Machine Intelligence 2014;32(1):18–32.

[25] Lin, W., Chu, H., Wu, J., Sheng, B., Chen, Z.. A heat-map-based algorithm for recognizing group activities in videos. IEEE Trans on Circuits and Systems for Video Technology 2013;23(11):1980–1992.

[26] Ma, S., Sigal, L., Sclaroff, S.. Learning activity progression in LSTMs for activity detection and early detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 1942–1950.

[27] Mehran, R., Moore, B., Shah, M.. A streakline representation of flow in crowded scenes. In: Proc. European Conference on Computer Vision, vol. LNCS 6313. 2010. p. 439–452.

[28] Ni, B., Yan, S., Kassim, A.. Recognizing human group activities with localized causalities. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2009. p. 1470–1477.

[29] Niebles, J.C., Wang, H., Fei-Fei, L.. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 2008;79(3):299–318.

[30] Park, S., Aggarwal, J.K.. A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems 2004;10(2):164–179.

[31] Raptis, M., Sigal, L.. Poselet key-framing: A model for human activity recognition. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. 2013. p. 2650–2657.

[32] Ronao, C.A., Cho, S.B.. Human activity recognition with smartphone sensors using deep learning neural networks. Expert Systems with Applications 2016;59:235–244.

[33] Shechtman, E., Irani, M.. Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? IEEE Trans Pattern Analysis and Machine Intelligence 2007;29(11):2045–2056.
[34] Simonyan, K., Zisserman, A.. Two-stream convolutional networks for action recognition in videos. In: Proc. of 27th Int. Conf. on Neural Information Processing Systems (NIPS). 2014. p. 568–576.
[35] Souvenir, R., Babbs, J.. Learning the viewpoint manifold for action recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2008. p. 59–65.
[36] Stephens, K., Bors, A.G.. Observing human activities using movement modelling. In: Proc. IEEE Int. Conf. on Advanced Video and Signal-based Surveillance (AVSS). 2015. p. 44.1–44.6.
[37] Stephens, K., Bors, A.G.. Group activity recognition on outdoor scenes. In: Proc. IEEE Int. Conf. on Advanced Video and Signal-based Surveillance (AVSS). 2016. p. 59–65.
[38] Stephens, K., Bors, A.G.. Grouping multi-vector streaklines for human activity identification. In: Proc. IEEE Workshop on Image, Video and Multidimensional Signal Processing. 2016. p. 1–5.
[39] Tseng, C.C., Chen, J.C., Fang, C.H., James Lien, J.J.. Human action recognition based on graph-embedded spatio-temporal subspace. Pattern Recognition 2012;45(10):3611–3624.
[40] Veeraraghavan, A., Roy-Chowdhury, A.K., Chellappa, R.. Matching shape sequences in video with applications in human movement analysis. IEEE Trans on Pattern Analysis and Machine Intelligence 2005;27(12):1896–1909.
[41] W., C., Savarese, S.. A unified framework for multitarget tracking and new collective activity recognition. In: Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 7575. 2012. p. 215–230.
[42] Wang, H., Klaser, A., Schmid, C., Liu, C.. Dense trajectories and motion boundary descriptors for action recognition. Int Jour of Computer Vision 2013;103(1):60–79.
[43] Wang, X., Gao, L., Song, J.. Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition. IEEE Signal Processing Letters, 2017;24(4):510–514.
[44] Zhang, D., Gatica-Perez, D.. Modeling individual and group actions in meetings with layered HMMs. IEEE Trans on Multimedia 2006;8(3):509–520.