Contents lists available at ScienceDirect
Digital Signal Processing
www.elsevier.com/locate/dsp

# Modelling of interactions for the recognition of activities in groups of people

Kyle Stephens, Adrian G. Bors *
*Department of Computer Science, University of York, York YO10 5GH, United Kingdom*

ARTICLE INFO
*Article history:*
Available online xxxx
*Keywords:*
Human Interactions
Human Group Activity
Kullback–Leibler divergence
Kernel Density Estimation
Gaussian Mixture Models

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 Kullback–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.


## 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 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 interdependency 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

* Corresponding author.
*E-mail address:* adrian.bors@york.ac.uk (A.G. Bors).

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

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 Models (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.

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 Fig. 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 $i$th 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

Please cite this article in press as: K. Stephens, A.G. Bors, Modelling of interactions for the recognition of activities in groups of people, Digit. Signal Process. (2018), https://doi.org/10.1016/j.dsp.2018.03.021