Recognizing Human Group Activities with Localized Causalities

Bingbing Ni, Shuicheng Yan, Ashraf Kassim
Electrical and Computer Engineering
National University of Singapore
Singapore, 117576
{g0501096, eleyans, eleashra}@nus.edu.sg

Abstract

The aim of this paper is to address the problem of recognizing human group activities in surveillance videos. This task has great potentials in practice, however was rarely studied due to the lack of benchmark database and the difficulties caused by large intra-class variations. Our contributions are two-fold. Firstly, we propose to encode the group-activities with three types of localized causalities, namely self-causality, pair-causality, and group-causality, which characterize the local interaction/reasoning relations within, between, and among motion trajectories of different humans respectively. Each type of causality is expressed as a specific digital filter, whose frequency responses then constitute the feature representation space. Finally, each video clip of certain group activity is encoded as a bag of localized causalities/filters. We also collect a human group-activity video database, which involves six popular group activity categories with about 80 *video clips for each in average, captured in five different sessions with varying numbers of participants. Extensive experiments on this database based on our proposed features and different classifiers show the promising results on this challenging task.*

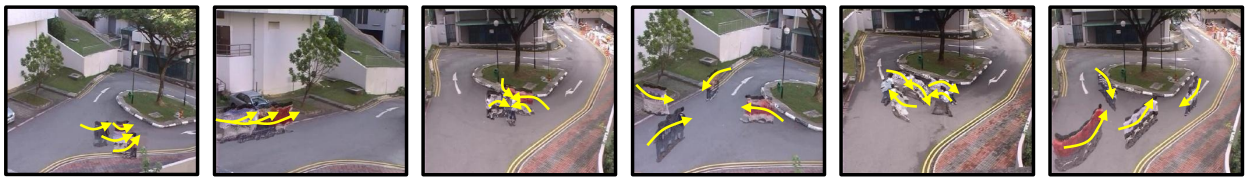
1. Introduction

Recently research on vision based human action and activity analysis has attracted much attention in computer vision literature. Previous works on this topic mainly focused on the relatively simple activities of single person [4, 12], the interactions between a person and the surrounding objects [11, 20], or the pair-activities between two persons [13, 22]. However, the interactions among a group of persons, namely human group activities, occur much more often in real scenarios, and the study of these human group activities based on visual cues has great potential for many applications such as smart video surveillance and human computer interfaces.

There has not been much research devoted to the problem of recognizing human group activities due to two factors, *i.e.*, the difficulties caused by the varying number of participants as well as mutual occlusions and the lack of usable benchmark databases. In literature, different visual features, *e.g.*, trajectories of the human body/body-parts [10, 12, 14, 15, 17], optical flows [3] and detected moving image regions [4, 5, 11, 13, 18, 20, 21], have been proposed for human activity representation. However, these features are limited in their use for recognizing human group activities, since 1) these features do not explicitly encode the information on the *interactions* among a group of persons, 2) it is difficult or even infeasible to robustly track a long trajectory of a person in crowded scenarios with occlusions, and 3) the segmentation of human body/body-parts in crowded scenes itself is a very challenging problem. Corresponding to the state-of-the-art of research on this topic, most currently available databases mainly include the activities of a single person (possibly including interactions with surrounding objects) or between two persons, *e.g.*, the CAVIAR database [2] and recent UIUC pair-activity database [22], or collected for abnormal event detection in crowded scenes [3]. The BEHAVE [1] database was recently released for human group activity analysis, but the very limited number of annotated video samples makes it unsuitable for statistically sufficient studies of this problem.

In this paper, we first investigate how to effectively represent human group activities. The number and identities of persons involved in a group activity video clip are changeable, and hence the high level visual information is generally of little value for characterizing human group activities. Low level visual information, *e.g.* color patches or optical flow patterns, are however not enough to describe the interactions among a group of persons. Thus the middle level visual cues, *e.g.* the motion trajectories of the participants and their interactions, are critical for representing human group activities. We propose to represent the human group activities with three types of localized causality

978-1-4244-3991-1/09/$25.00 ©2009 IEEE



walk-in-group run-in-group stand-and-talk gathering fighting ignoring

Figure 1. An illustration of the image frames (stacked together) for six types of human group activities. Note that the camera view angle may be different in different sessions, which brings greater difficulties in recognizing human group activities.

features, *i.e.*, self-causalities, pair-causalities, and group-causalities, which encode the causal properties of individual trajectories, trajectory pairs, and trajectory groups respectively. Each type of causality is expressed as a digital filter, whose frequency responses then constitute the feature space for human group activity representation. Due to the mutual occlusions as well as the inherent limitations of the tracking algorithms, the trajectory of a person within a video clip is often broken into several segments, and thus the causality features are extracted locally based on the short segments instead of the entire motion trajectory. These localized causalities describe the spatially local and temporally short-term behaviors of a human group, and the human group activities are then decomposed into a set of such local behaviors. Finally each group activity sample is represented as a bag of orderless localized causalities/filters.

We have developed a human group activity video database, which includes six categories of popular group activities: *walk-in-group*, *run-in-group*, *stand-and-talk*, *gathering*, *fighting*, and *ignoring* (*i.e.*, the subjects walk independently). Some sample frames are shown in Figure 1. The database was captured in five different sessions, where the number of participants and the capture time are different, resulting in about 80 video clips for each category of group activity in average. This database is then used to evaluate the effectiveness of our proposed localized causality features for recognizing human group activities.

2. Representation by Localized Causalities

The problem of recognizing human group activities was rarely studied in computer vision literature. In this section, we first introduce the problem definition with the assumption that the sample of certain human group activity has been initially instantiated as a set of broken segments of the human motion trajectories, extracted by some object tracking algorithm, and then introduce how to design the causality features within, between, and among these segments.

2.1. Problem Definition

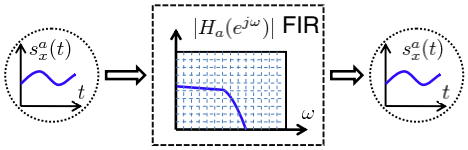
Denote the motion segments within an example video clip as $S = \{s^1, s^2, \cdots, \}$, where $s^k(t) = [s_x^k(t), s_y^k(t)]^T$ is the center position of the human body in 2D image plane. As a specific machine learning problem, a set of training samples are given as $\mathfrak{S} = \{S_1, S_2, \cdots, S_N\}$, where N is the number of training samples, and their human activity labels are denoted as $\{c_1, c_2, \cdots, c_N\}$, where $c_i \in \{1, 2, \cdots, N_c\}$ with N_c as the number of human group activity categories. The task of recognizing human group activities is to learn an activity category prediction function based on these labeled training samples, so as to predict the group activity categories for the new samples.

2.2. Motivations

Under the video surveillance scenarios, the size of human body is often very small and thus the tracking of the human body/body-parts for accurate human interaction modeling becomes infeasible, instead only segments of the motion trajectories are obtainable with certain popular tracking algorithm. In this subsection, we explain the motivations to utilize the localized causality features introduced afterwards for human group activity representation.

Why use causalities for group activity representation? The human group activities are mainly characterized by the dynamic interaction properties among a set of persons (generally >2), and conventional features for solo-activity analysis cannot fully convey the interaction information among a group of persons. These dynamic interactions are essentially embodied as the affection, or mathematically stated as causality and feedback, between two persons or among multiple persons. Thus the representation of group activities finally ends at how to model the causalities among a group of persons. These causalities can be further divided into three types: 1) *self-causality*, which describes the affection of the past status history to the current status of a person and mainly characterizes the behavior properties of a single person; 2) *pair-causality*, which measures the interactions between two persons, *e.g.*, meeting, chasing; and 3) *group-causality*, which shows the affect of other persons' behaviors to the behavior of the concerned person, and is unique for group activity modeling.

Why use localized causality features? Instead of calculating the causality features by fitting a global model using the complete motion trajectories throughout a video clip, we propose to conduct localized causality analysis, namely, the causality features are calculated based on the short segments of the human motion trajectories. On one hand, the dynamic interaction property of a human group activity sample may vary within the video clip, *e.g.*, for the activity *fighting*, the identity of the person to interact with a



$$H(z) = \sum_{n=1}^{k} h(n) z^{-(n+l)}$$

Figure 2. Digital filter representation for self-causality analysis of the human motion segment.

certain person may change frequently, and hence the localized interactions are more repeatable for different samples. On the other hand, the possible mutual occlusions make the extraction of the complete motion trajectories difficult, and only segments of these motion trajectories are achievable. Thus, it is more feasible and practical under real scenarios to calculate the causalities in a spatially and temporally local way.

In this work, we present three types of localized causality features, namely self-causalities, pair-causalities, and group-causalities introduced in the next three subsections, to characterize the dynamic interaction properties of different human group activities. More specifically, each style of causality is finally expressed by the frequency responses of a specific digital filter, with the trajectory segments as input and output signals respectively. Finally these frequency responses are further clustered with K-means approach to build the *visual word dictionary*, and the histogram vector based on these visual words is then used as final feature representation for recognizing human group activities.

2.3. Self-causality Analysis

The concept of self-causality refers to the affect of the past status history of a person to his/her current status, namely how the previous motion trajectory positions of a person affect her/his current position. For a human trajectory segment denoted as s which can be regarded as a digital signal, this self-causality is modeled as a digital filter as shown in Figure 2 with the input and output signals as the same digital signal s. Illuminated by the definition of self-causality, a natural form for this filter is the finite impulse response (FIR) digital filter denoted as,

$$s(t) = \sum_{n=1}^{k} h(n) s(t-n-l) + \epsilon(t), \tag{1}$$

where $h(n)$ is the impulse response of the FIR digital filter, k is the order of FIR filter, $\epsilon(t)$ is a stationary Gaussian noise with the variance as σ^2, and l is a manually set time lag which may avoid the overfitting issue in estimating the model parameters for the FIR filter. Note that in Eqn. (1), the values of the signal s at x and y coordinates are assumed to take the same model.

To obtain the values of the impulse response $h(n)$ and σ^2, we use all the frames in the motion trajectory segment s and calculate based on the least square errors (LSE) method to fit the $h(n)$ by minimizing the variance σ^2 of the noise term. Note that at least k linear equations are required to obtain a solution for $h(n)$, and thus the length of each motion trajectory segment should be at least $2k + l$.

The FIR digital filter well encodes the self-causality property of a human motion trajectory segment. To avoid the possible instability caused by the value of k, we do not directly use the $h(n)$'s as the self-causality features, and instead $h(n)$ is transformed into the frequency domain with z-transform to obtain the frequency responses for feature extraction. More specifically, the z-transform of $h(n)$ is calculated as

$$H(z) = \sum_{n=1}^{k} h(n) z^{-(n+l)}. \tag{2}$$

The frequency domain is generally more robust than the time domain for extracting features, and the self-causality property conveyed by the motion trajectory segment s is then represented as five evenly sampled frequency responses (magnitudes and phases),

$$f_1(s) = [|H(e^{j0})|, |H(e^{j\frac{\pi}{4}})|, \cdots, |H(e^{j\pi})|, \\ \angle H(e^{j\frac{\pi}{4}}), \cdots, \angle H(e^{j\frac{3\pi}{4}})]^T, \tag{3}$$

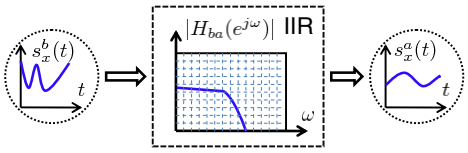
which is an 8-dimensional vector. Note that for $z = e^{j0}$ and $z = e^{j\pi}$, we do not calculate their phases since they are always equal to zeros. It is possible for us to sample more frequencies to constitute longer-length $f_1(s)$, but our offline experiments show that five frequencies are generally enough to achieve near-optimal recognition performance.

An intuitive explanation of the effectiveness for this encoding scheme is that, different category of human motion trajectory segment may have a FIR digital filter with distinctive impulse response $h(n)$. For example, the $h(n)$'s for static, constant-speed, constant-acceleration motion trajectory segments could be expressed as $s(t) = s(t-1)$, $s(t) = 2s(t-1) - s(t-2)$, and $s(t) = s(t-1) - 3s(t-2) + s(t-3)$, respectively.

2.4. Pair-causality Analysis

The pair-causality describes the interaction properties of two persons. The pair-causality information includes two parts. The first part is the *strength* of one person's affect on another one, and the second part is *how* one person affects another one. In [22], Zhou *et al.* used the Granger Causality Test (GCT) [16] to obtain two quantities, causality ratio and feedback ratio, for measuring the causality and feedback strength between two persons based on their tracked concurrent motion trajectories.

More specifically, for a concurrent human motion trajectory pair of $s_a = [s_a(1), s_a(2), \cdots, s_a(t), \cdots]$ and $s_b =$



$$H_{ba}(z) = \frac{\sum_{n=1}^{k} \gamma(n) z^{-n-l}}{1 - \sum_{n=1}^{k} \beta(n) z^{-n-l}}$$

Figure 3. Digital filter representation for pair-causality analysis of the human motion trajectory segment pair.

$[s_b(1), s_b(2), \cdots, s_b(t), \cdots]$, we assume that the interaction between two trajectories is a stationary process, *i.e.*, the prediction functions $P(s_a(t)|s_a(1:t-l), s_b(1:t-l))$ and $P(s_b(t)|s_a(1:t-l), s_b(1:t-l))$ do not change within a short time period, where $s_a(1:t-l) = [s_a(t-l), s_a(t-l-1), \cdots, s_a(1)]$ and so for $s_b(1:t-l)$.

To model $P(s_a(t)|s_a(1:t-l), s_b(1:t-l))$, we can use k-th order linear predictor, namely,

$$s_a(t) = \sum_{n=1}^{k} \beta(n) s_a(t-n-l) + \gamma(n) s_b(t-n-l) + \epsilon_a(t), \tag{4}$$

where $\beta(n)$'s and $\gamma(n)$'s are the regression coefficients, and $\epsilon_a(t)$ is the Gaussian noise with standard deviation $\sigma(s_a(t)|s_a(1:t-l), s_b(1:t-l))$. These model parameters can be derived based on the concurrent human motion trajectory segment pair of s_a and s_b. Similarly, we use the linear predictor to model $P(s_a(t)|s_a(1:t-l))$ as in (1), and the standard deviation of the noise signal is denoted as $\sigma(s_a(t)|s_a(1:t-l))$.

According to the GCT theory [16] and pair-trajectory analysis in [22], we could obtain two measurements on the causality strength, namely,

1. **Causality ratio**: $r_c = \frac{\sigma(s_a(t)|s_a(1:t-l))}{\sigma(s_a(t)|s_a(1:t-l), s_b(1:t-l))}$, which measures the relative strength of the causality.

2. **Feedback ratio**: $r_f = \frac{\sigma(s_b(t)|s_b(1:t-l))}{\sigma(s_b(t)|s_a(1:t-l), s_b(1:t-l))}$, which measures the relative strength of the feedback.

The causality ratio and feedback ratio could well characterize how strong one person affects the motion of another one, however cannot encode the underlying mechanism that drives the motions of these two persons. In other words, the causality ratio and feedback ratio only leverage on the information from the noise terms, namely the standard deviations, but do not reflect the information characterized by $\beta(n)$ and $\gamma(n)$, which are the model parameters and essentially characterize how one person affects another one.

If we regard this relationship in (4) as a digital filter with the input signal as $s_b(t)$ and the output signal as $s_a(t)$, as illustrated in Figure 3, we can calculate the z-transforms for both sides, and then we obtain the following equation by ignoring the noise term,

$$X_a(z) = \sum_{n=1}^{k} \beta(n) X_a(z) z^{-n-l} + \gamma(n) X_b(z) z^{-n-l}, \tag{5}$$

where $X_a(z)$ and $X_b(z)$ are the z-transforms for the output and input signals respectively. Denoting the impulse response of this digital filter as $H_{ba}(z)$, and based on the relation of $H_{ba}(z) = \frac{X_a(z)}{X_b(z)}$, we have,

$$H_{ba}(z) = \frac{\sum_{n=1}^{k} \gamma(n) z^{-n-l}}{1 - \sum_{n=1}^{k} \beta(n) z^{-n-l}}, \tag{6}$$

from which we can observe that this digital filter is an infinite impulse response (IIR) digital filter.

Similar to self-causality analysis, the magnitudes and the phases of the z-transform function at a set of evenly sampled frequencies are used to describe the digital filter (*i.e.*, the style of the pair-causality). More specifically, we use the magnitudes of the frequency responses at 0, $\pi/4$, $\pi/2$, $3\pi/4$, π and the phases of the frequency responses at $\pi/4$, $\pi/2$, $3\pi/4$ (constant values for frequency 0 and π), namely,

$$f_{ba} = [|H_{ba}(e^{j0})|, |H_{ba}(e^{j\frac{\pi}{4}})|, \cdots, |H_{ba}(e^{j\pi})|, \\ \angle H_{ba}(e^{j\frac{\pi}{4}}), \cdots, \angle H_{ba}(e^{j\frac{3\pi}{4}})]. \tag{7}$$

Similarly we can define the feature vector f_{ab} by considering s_a as the input signal and s_b as the output signal for an IIR digital filter which characterizes how the person in the trajectory segment s_a affects the motion of the person in the trajectory segment s_b.

The causality ratio and feedback ratio characterize the strength of one person's affect on another one, while the extracted frequency response features f_{ab} and f_{ba} convey how one person affects another one. Intuitively, these features are mutually complementary, and hence we combine them to form the description vector for pair-causality representation. Also, the relative distance Δd_{ba} and relative speed Δv_{ba} of two interacting persons are very useful for discriminating activities such as *walk-in-group* and *gathering*, therefore we add them to form the description vector,

$$f_2(s_a, s_b) = (f_{ab}, f_{ba}, r_c, r_f, \Delta d_{ba}, \Delta v_{ba})^T, \tag{8}$$

which is a 20-dimensional feature vector with 5×2 magnitude values, 3×2 phase values, the causality and feedback ratio values, as well as the relative distance and speed.

2.5. Group-causality Analysis

Human group activities characterize the behaviors of a group of persons, and the information conveyed may be beyond the self-causality and pair-causality features. For example, the group activity *gathering* generally involves three

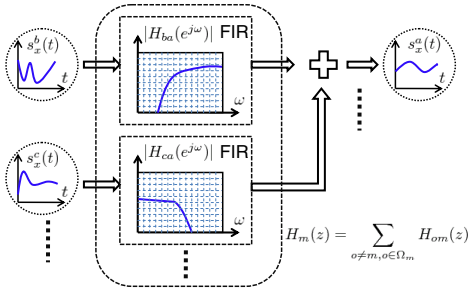


Figure 4. Digital filter representation for group-causality analysis of a group of human motion trajectory segments.

or more persons walking towards the same point. Therefore, we need features which can describe the interaction properties among multiple persons. In this subsection, we present the prediction model to describe the affects of all other persons to a certain person, namely,

$$s^m(t) = \sum_{o\neq m, o\in\Omega_m} \sum_{n=1}^{k} h^o(n) s^o(t-n-l) + \xi^m(t), \quad (9)$$

where Ω_m is the index set for all concurrent motion trajectory segments with s^m within a video clip, $h^o(n)$'s are the regression parameters, and $\xi^m(t)$ is the Gaussian noise term. Here, the trajectory $s^m(t)$ is predicted by utilizing all the other trajectories, and the model parameters could be estimated using least square errors approach based on all the motion trajectory segments concurrent with s^m.

From the perspective of digital filters, the whole system can be regarded as the integration of a set of FIR digital filters. Among the filter set, each filter describes the causality relation as

$$\hat{s}^o(t) = \sum_{n=1}^{k} h^o(n) s^o(t-n-l), \;\; o \neq m, \quad (10)$$

where $\hat{s}^o(t)$ is unknown beforehand. Its z-transform function is then,

$$H_{om}(z) = \sum_{n=1}^{k} h^o(n) z^{-n-l}, \;\; o \neq m. \quad (11)$$

Finally the outputs from all these FIR digital filters are summed to output the signal $s^m(t)$ along with the noise signal $\xi^m(t)$, namely,

$$s^m(t) = \sum_{o\neq m, o\in\Omega_m} \hat{s}^o(t) + \xi^m(t). \quad (12)$$

Figure 4 shows the whole system.

The number of input signals does not convey information for characterizing the human group activity, and hence for group-causality, we use the sum of the frequency responses from all the FIR digital filters as the feature space for group-causality representation. Denote

$$H_m(z) = \sum_{o\neq m, o\in\Omega_m} H_{om}(z), \quad (13)$$

and the group-causality features related with human motion trajectory segment s^m are finally represented as

$$f_3(s^m) = [|H_m(e^{j0})|, |H_m(e^{j\frac{\pi}{4}})|, \cdots, |H_m(e^{j\pi})|, \angle H_m(e^{j\frac{\pi}{4}}), \cdots, \angle H_m(e^{j\frac{3\pi}{4}})]^T. \quad (14)$$

3. Classification with Localized Causalities

A video clip containing a human group activity example usually consists of a large number of short human motion trajectory segments. Based on self-causality analysis, we can extract an 8-dimensional feature vector for each segment. Based on pair-causality analysis, we can extract a 20-dimensional feature vector for a concurrent segment pair, and based on group causality, generally we can extract an 8-dimensional feature vector for each segment. As the number of trajectory segments and the number of concurrent trajectory segment pairs may be different in different video clips, we use the bag-of-words approach to construct three visual word dictionaries based on three types of causality features. The feature extraction as well as the formulation of the bag-of-words representation procedure is illustrated in Figure 5. Note that in this work, for the tracked long

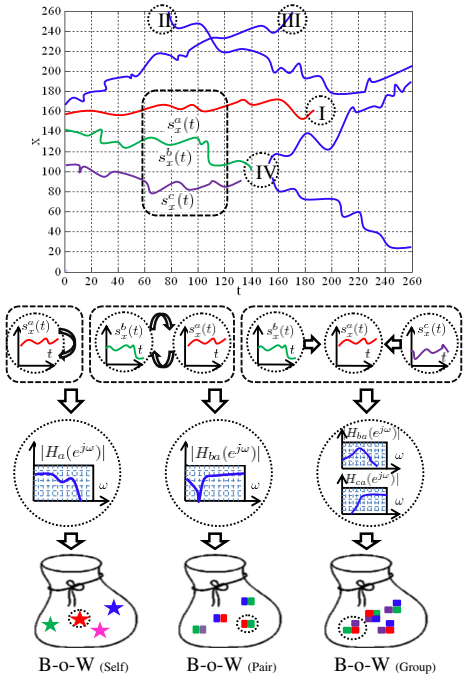


Figure 5. An example illustration of the process to extract localized causality features.

motion trajectories, we further extract the short trajectory segments (about $4k$ frames) by sliding a temporal window. Then, each video clip is represented as three histogram vectors based on these three visual word dictionaries. Finally, direct classifiers like Nearest Neighbor (NN) and other machine learning algorithms such as Support Vector Machine (SVM) [6] are used for human group activity classification.

4. Experiments

In this section, we first introduce the details of the group activity database we collected, demonstrate the effectiveness of the proposed three types of localized causality features, and further compare these features with the low-level visual features extracted at the space-time interest points [9], which have been verified effective in many applications.

4.1. Group Activity Database Construction

The human group activity database was collected by monitoring an outdoor scene (an university car park). The video camera, Panasonic-NV-DX100EN with frame rate of 25fps and image size of 720×576 pixels, was mounted and zoomed such that the captured persons are in proper scale for ease of human tracking and detection. The database was collected in five different sessions, with different actors, or different number of actors, or at the different time. In each session, six categories of group activities were staged by the actors and each activity contains 10-20 instances with 4-8 actors, and each session spans several minutes. For each video sequence, we manually crop the entire video into small video clips, each of which contains a human activity instance of 8-15 seconds and is assigned the corresponding human activity category. The whole database includes 476 labeled video samples in total. A summary of the segmented video samples is given in Table 1.

To track the human motion trajectory, each human actor is considered as a $2D$ blob, and then the task is to locate the position sequence of each blob in the $2D$ image. Our tracking method is based on the *CONDENSATION* [8] algorithm with manual initializations. In our implementation, about 100 particles are used to track each person for the tradeoff of accuracy and computational cost. The tracking fails easily due to the mutual occlusions, and manual re-initializations are used for simplicity although motion segmentation algorithm can be developed for automatic re-initializations [7, 19]. Finally, for each video clip, a set of short human motion trajectory segments are obtained, and we down-sample the frame rate by a factor of 2 for further process.

4.2. Localized Causality Feature Visualization

Figure 5 shows a detailed example of the tracked human trajectory segments. It could be observed that there exist several reasons for broken trajectory segments, *e.g.*, tracker's failure (marked by I), subject entering the view (marked as II) and out of the view (marked as III), and the manual re-initialization after occlusion (marked as IV). Therefore, it is intractable to extract the complete trajectory throughout the whole video clip. In Figure 6, we further illustrate the examples of the extracted human motion trajectories, the three types of filter responses from the localized causality analysis, as well as the three histogram representations based on the bag-of-words models. Each of the subfigure on the right of the sampled video frames shows the tracked trajectories of the human group activities. Note that there exist distinctive patterns associated with different categories of group activities. For *walk-in-group* and *run-in-group* activities, all the trajectories are nearly parallel; for *stand-and-talk* and *fighting* activities, the trajectories are more random, and for *fighting* activity, there exist lager variations among trajectories; for *gathering* activity, all the trajectories converge to a common point; and for *ignoring* activity, each trajectory points to a different direction. The subfigures below show examples of the filter responses in terms of magnitude (2nd row) and phase (third row) for self-causality (left), pair-causality (middle) and group-causality (right) features. The last row shows the histogram representations for three types of causality features based on the bag-of-words models.

4.3. Classification with Localized Causalities

The human group activity classification experiments are conducted based on the leave-one-session-out strategy. For the parameter l, k and the sizes of three visual word dictionaries, we evaluate all the proper combinations of these parameters and set them to be optima in the experiments, and finally $l = 4$, $k = 4$, and the dictionary size is set to be 20 for each type of causality feature. For the nearest neighbor classifier, we use the commonly used χ^2 distance for dissimilarity measure,

$$d(x, y) = \frac{1}{2} \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}, \tag{15}$$

where x and y are the concatenated vector from the three histogram vectors based on three types of localized causality features. For SVM, we use the following kernel based on the χ^2 distance,

$$k(x, y) = \exp\{-\frac{1}{\gamma} d(x, y)\}, \tag{16}$$

where the kernel parameter γ is tuned to be optimal. We use the non-linear multi-class SVM toolbox *LIBSVM* in [6] for model training and final classification.

Figure 8 lists the classification accuracies in terms of confusion matrices by using different types of causality features, as well as their combination, based on both NN and

Table 1. A summary on the collected human group activity database.

Human activity category	walk-in-group	run-in-group	stand-and-talk	gathering	fighting	ignoring	Total
No. of sessions	5	5	5	5	5	5	5
No. of segments	94	65	88	86	74	69	476

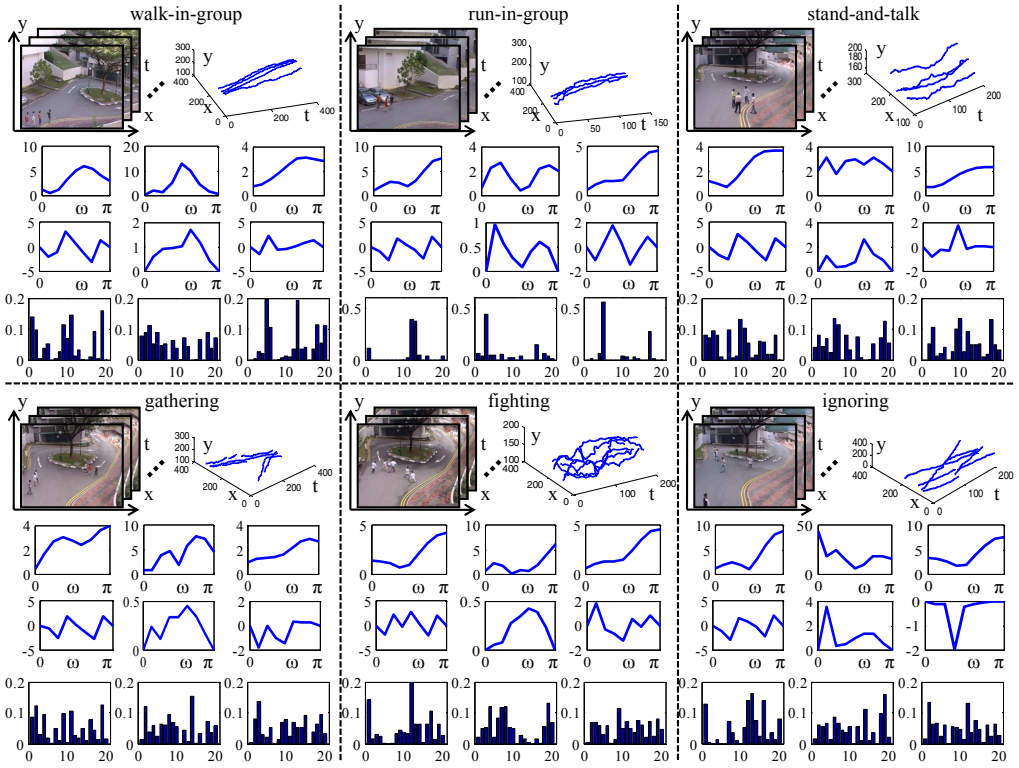


Figure 6. An illustration of the captured video sequences (left 1st row), tracked motion trajectory segments (right 1st row), examples of the filter responses for both magnitudes (2nd row) and phases (3rd row) and bag-of-words representations (4th row). Note that for the 2-4th rows, from left to right, the subfigures correspond to self-causality, pair-causality and group-causality features respectively.

SVM classifiers. From Figure 8(a) and 8(b), we can observe that, on one hand, self-causality features carry limited information for discriminating the activities *walk-in-group*, *ignoring* or *gathering* from each other since when specific to a person, his/her motion trajectory segment may be similar for these three activities, however, they could be easily differentiated by the pair-causality features. On the other hand, the pair-causality features cannot effectively discriminate the activity *walk-in-group* from *run-in-group* since the pair interaction properties of these two activities are similar, however, the self-causality features can properly deal with this case since the motion models of walking and running for a single person are different. When these three types of localized causality features are combined, the classification performance could be further boosted (see Figure 8(d)). We can also note that the SVM classifier generally offers higher performance than NN classifier.

We further compare the localized causality features with the state-of-the-art low level visual features, *i.e.*, the features extracted at the space-time interest points (STIP) [9]. This method extracts the HOG (Histograms of Oriented Gradients) and HOF (Histograms of Optical Flow) features computed within a $3D$ video patch around each detected STIP. The patch is partitioned into a grid with $3 \times 3 \times 2$ spatio-temporal blocks, and the 4-bin HOG descriptors and 5-bin HOF descriptors are then computed for all blocks and are concatenated into a 72-element and 90-element descriptors respectively. Finally they are concatenated into a $162D$ description vector. We model the video clips based on the bag-of-words method by first clustering all the description vectors from the training set using K-means method. We vary the dictionary size from 10 to 2000 and compare with our proposed features also in a leave-one-session-out strategy. For SVM classifier, the kernel used is the same as in Eqn. (16). From the comparison results shown in Figure 7, it is observed that STIP based low level features yield very low performance compared with our proposed localized causality features, since no interaction information among a group of persons is explicitly exploited for the features extracted at the space-time interest points.

5. Conclusions and Future Works

In this paper, we introduced the video database collected for the rarely studied problem of recognizing human group activities in surveillance videos. Three types of causality

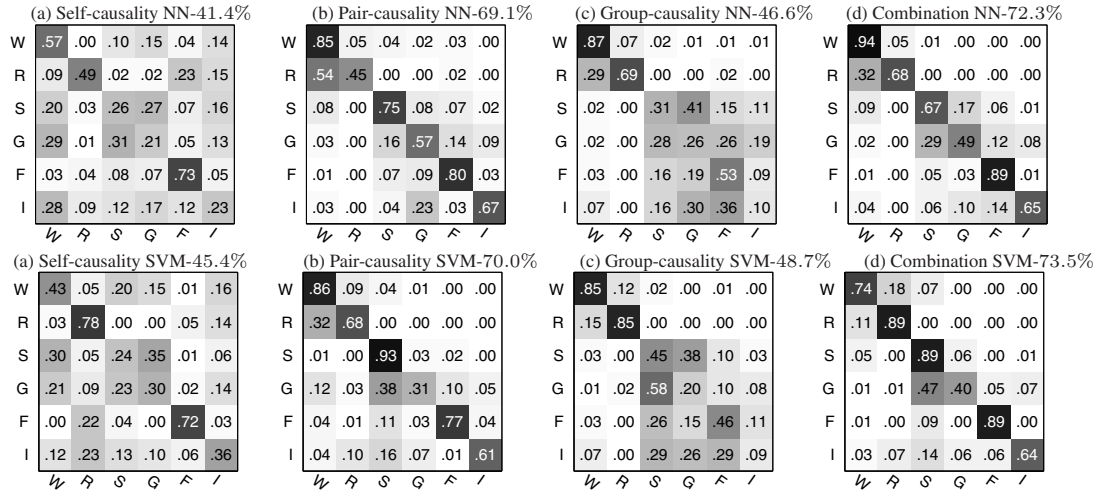


Figure 8. The confusion matrices of the human group activity classification results with different features and classifiers. For ease of display, each activity is denoted by its first character, *i.e.*, 'W' for walk-in-group, 'R' for run-in-group, 'S' for stand-and-talk, 'G' for gathering, 'F' for fighting, and 'I' for ignoring. For better viewing, please see the pdf file.

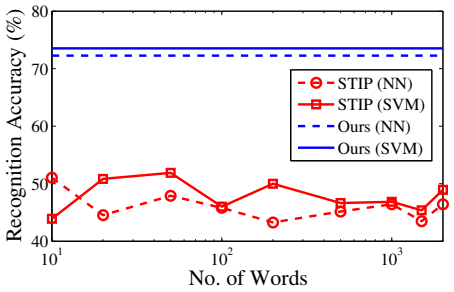


Figure 7. Classification comparison between our proposed localized causality features and the local features exacted at the STIP's.

features, characterizing the dynamic interaction properties within, between and among human motion trajectory segments, were proposed to obtain a fixed-length representation for video clip with variable length and number of persons. To the best of our knowledge, the group activity database introduced in this work is the first one available for extensively studying the human group activities, and also it is the first work to extensively study the human group activity classification problem. We plan to further exploit this topic in two aspects: 1) to design salient point based trajectory for human group activity representation to avoid the requirement of manual initializations in human tracking process, and 2) to extend the current work to a more general framework for automatic group activity detection, segmentation, and classification within a long video.

Acknowledgment

This work is supported by NRF/IDM Program, under research Grant NRF2008IDM-IDM004-029.

References

[1] The BEHAVE Website: *http://homepages.inf.ed.ac.uk/rbf/BEHAVE/.*

[2] The CAVIAR Website: *http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.*

[3] E. Andrade, S. Blunsden, and R. Fisher. Modelling Crowd Scenes for Event Detection, *ICPR*, 2006.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes, *ICCV*, 2005.

[5] M. Brand, N. Oliver, and A. Pentland. Coupled Hidden Markov Models for complex Action Recognition, *CVPR*, 1997.

[6] C. Chang and C. Lin. LIBSVM: A Library for Support Vector Machines, Software available at: *http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[7] I. Haritaoglu, D. Harwood, and L. Davis. W4: Real-time Surveillance of people and their activities, *TPAMI*, 2000.

[8] M. Isard1 and A. Blake. CONDENSATION - Conditional Density Propagation for Visual Tracking, *IJCV*, 1998.

[9] I. Laptev, M. Marszatek, C. Schmid, and B. Rozenfeld. Learning Human Actions from Movies, *CVPR*, 2008.

[10] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams, *TPAMI*, 2001.

[11] D. Moore, I. Essa, and M. Haye. Exploiting human actions and object context for recognition tasks, *ICCV*, 1999.

[12] J. Nascimento, M. Figueiredo, and J. Marques. Segmentation and Classification of Human Activities, *Proceedings of International Workshop on Human Activity Recognition and Modelling*, 2005.

[13] S. Park and J. Aggarwal. Segmentation and Tracking of Interacting Human Body Parts under Occlusion and Shadowing, *IEEE Workshop on Motion and Video Computing*, 2002.

[14] A. Prati, S. Calderara, and R. Cucchiara. Using Circular Statistics for Trajectory Shape Analysis, *CVPR*, 2008.

[15] P. Ribeiro and J. Victor. Human Activity Recognition from Video: Modeling, Feature Selection and Classification Architecture, *Proceedings of International Workshop on Human Activity Recognition and Modelling*, 2005.

[16] C. Sims. Money, income, and Causality, *American Economic Review*, 1972.

[17] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking, *TPAMI*, 2000.

[18] P. Turaga, A. Veeraraghavan, and R. Chellappa. From Videos to Verbs: Mining Videos for Activities using a Cascade of Dynamical Systems, *CVPR*, 2007.

[19] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body, *TPAMI*, 1997.

[20] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A Scalable Approach to Activity Recognition Based on Object Use, *ICCV*, 2007.

[21] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video, *CVPR*, 2004.

[22] Y. Zhou, S. Yan, and T. Huang. Pair-Activity Classification by Bi-Trajectory Analysis, *CVPR*, 2008.