INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 11(4):553-566, August 2016.

A Hybrid Model for Concurrent Interaction Recognition from Videos

M. Sivarathinabala, S. Abirami

M. Sivarathinabala*
Department of Information Science and Technology,
Anna University, Chennai, India.
*Corresponding author: sivarathinabala@gmail.com

S. Abirami
Department of Information Science and Technology,
Anna University, Chennai, India.
abirami_mr@yahoo.com

Abstract: Human behavior analysis plays an important role in understanding the high-level human activities from surveillance videos. Human behavior has been identified using gestures, postures, actions, interactions and multiple activities of humans. This paper has been analyzed by identifying concurrent interactions, that takes place between multiple peoples. In order to capture the concurrency, a hybrid model has been designed with the combination of Layered Hidden Markov Model (LHMM) and Coupled HMM (CHMM). The model has three layers called as pose layer, action layer and interaction layer, in which pose and action of the single person has been defined in the layered model and the interaction of two persons or multiple persons are defined using CHMM. This hybrid model reduces the training parameters and the temporal correlations over the frames are maintained. The spatial and temporal information are extracted and from the body part attributes, the simple human actions as well as concurrent actions/interactions are predicted. In addition, we further evaluated the results on various datasets also, for analyzing the concurrent interaction between the peoples.
Keywords: Pose prediction, interaction recognition, layered HMM

1 Introduction

Human interaction analysis involves human activities that are happening between two or more persons to understand the interaction. The automation is required in video surveillance to detect activities from the videos and infer some useful information without the intervention of human beings. The type of human activities detected mainly depends upon the domain in which surveillance system has been employed. Human activity recognition may be used for behavior pattern observation or for suspicious activity detection. The activities can either be detected or reported during the event when it takes place or it can be predicted in advance. A system or framework to understand human interaction from surveillance videos involves the following key components: a) Low level components for Background modeling, Feature extraction and Object tracking, b) Middle level components for Object classification c) High level components for Semantic interpretations (ie, understanding actions, interactions between two / multiple people). Significant works have been progressing in the literature for each level in this framework. This work mainly focuses on higher level components in order to predict the human interactions. Human Interaction Recognition (HIR) involves activity recognition of humans to understand their behaviors. Human activity recognition has been presented by ([2], [18]) as, single layered and hierarchical approaches like space- time volumes, space- time trajectories, space-time features and state based models such as HMM and DBN. From these approaches, human activity such

as shaking hands, punching, pushing, pointing, picking up the object, throwing are recognized. There exist some limitations such as difficulty in recognizing the interactions that happened between multiple people in varying time difference in when he/she re-enters in the scene, difficulty in distinguishing the poses when transformation and scaling has been performed, variation in the performance of body part detection, difficulty to recognize the interactions in complex environments.

2 Related works

Activity analysis involves two fold actions: (i) Analysis of motion patterns (ii) Understanding of High level descriptions of actions/interactions happening place among humans or in an environment. Activities can be recognized [2] in single layered approaches and in hierarchical approaches. In Single layer approaches human activities could be recognized based on the image sequences and it is suitable for gesture/action recognition. In contrast, Hierarchical approaches recognize high level activities which are complex in nature. It has been observed from the literature that the hierarchical approach well suits to recognize high level activities (Interactions). Thus, many researchers proposed a level of HMM differently as Hierarchical HMM, Semi- HMM, 2D- HMM, factorial HMM, Coupled HMM, Asynchronous IO HMM etc., This paper attempts to analyze the Interactions between two or more persons in a new fashion of Hidden Markov Model .

Hidden Markov Model (HMM) plays a vital role in determining activities which are vision based. The research works ([11], [13], [13], [12], [19])proves that HMM is one of the most appropriate models for person activity prediction. [9] stated that, a layered hidden Markov model (LHMM) can obtain different levels of temporal details while recognizing human activity. The LHMM is a cascade of HMMs, in which each HMM has different observation probabilities processed with different time intervals. [22] has proposed layered hidden Markov model (LHMM) as a statistical model which is derived from the hidden Markov model (HMM).

In HMM, the activities are recognized with less temporal correlated frames and over fitting problem occurs during the calculation of observation probability. The activities that take place in long temporal difference cannot be identified with good accuracy. Moreover, the use of single variable state representation makes more difficult to model the complex activities involving multiple interacting agents. In order to solve the over fitting problem in the HMM, Human action recognition [8] has been done using three layered HMM. Their system was capable of recognizing six distinct actions, including Raising the right arm, Raising the left arm, Stretching, Waving the right arm, Waving the left arm, and Clapping. From the observations made in the survey process, most of the previous work centered on identification of particular activities in the particular scenario and less effort has been done to recognize interactions.

Many research works ([20], [21], [3], [13]) have been done using graphical models such as HMM and CRF. A conventional HMM can have many mathematical structures and has proved its simplicity in recognizing temporal events. Its only limitation in conventional HMM is, it cannot capture high temporal correlated frames since the output depends only on the current states. Our work is different in recognizing interactions (i.e., actions between two or more persons) in the group. When there are multiple persons in the scene, one person may stand still (or) not doing any actions (or) not interacting with any other persons in the group. That particular person's activity may be considered as abnormal activity. In order to identify the abnormal activity in the particular frame, we are interested in concurrent interactions between a group of peoples. As a result, a three layered HMM has been designed in our work to recognize the actions and interactions in the group. In the first layer, pose of the persons has been identified and in the second layer ie., the action layer in which actions of the individual persons has been recognized

and in the third layer, the actions of the first person and the second person are coupled in order to identify the interaction of the people. This three layered HMM able to handle temporal correlations between the frames.

This paper has been motivated to design a human interaction analysis system which can overcome these limitations and to design an automated smart surveillance system which could predict actions/interactions taking place in public environments. As motivated by the above challenges the following contributions arose in our work: a) Learning model has been designed and implemented in order to learn the complex activity/interaction, b) joint form of LHMM and CHMM provides concurrent interactions between the persons c) proposed model has been validated using different interaction recognition datasets and self generated data set.

3 Preprocessing and feature extraction

Pre-processing includes two phases called Background modeling and Foreground segmentation. This phase highlights the portion of the frame which is under motion. Background modeling [5] is a process which tries to model the current background of the scene precisely and aids to segment the foreground objects from the background. Foreground segmentation has been performed to obtain the foreground image from the actual image.

Feature extraction starts with contour extraction [19] and the region properties of the frame have been considered, the number of objects in each frame has been found out. Object centroid and object area have been calculated from the region properties as the next set of features. The human body parts need to be modeled before performing the pose estimation. Here, to employ a silhouette based approach ([6], [14], [15], [13]) in body part modeling, convex hull technique has been adapted. Here, the convex hull points have been obtained for the whole blob to construct the skeleton. A minimum of 5 dominant points such as head, left hand, right hand, left leg and right leg that lie on the convex hull polygon has been chosen in our observation. Body part modelling [4] differs for each and every pose since the pose stand may completely vary from the pose sit. In order to predict every pose precisely, the height of the human is divided into four quadrants, upper most quarter, upper middle quarter, lower middle quarter, and lower most quarters. From the contour (sketch) of the human body, the location of the hands, head and legs have been exactly predicted using the convex hull co-ordinates that lie in the arrangement of the four regions mentioned above. This type of body part modelling can better suit to any kind of pose prediction.

4 Human interaction recognition

In our work, three layers of HMM have been framed in order to identify the group activity in varying time intervals. Figure 1 shows the overview of the Human Interaction Recognition System. This layered representation provides outputs at different levels and decompose the levels in different time granularity. Hidden Markov Model (HMM) [7] is the most successful approach to modeling and classifying dynamic behaviors. Layered Hidden Markov Model (LHMM) and Coupled Hidden Markov Model (CHMM) are combined together to enhance the robustness of the interaction analysis system by reducing the training parameters. Effective learning process has been carried out by combining max-belief algorithm and Baum-Welch algorithm. Max-belief algorithm is used to derive the most likely sequence of results in observation activities and the Baum-Welch algorithm reduces the inference errors.

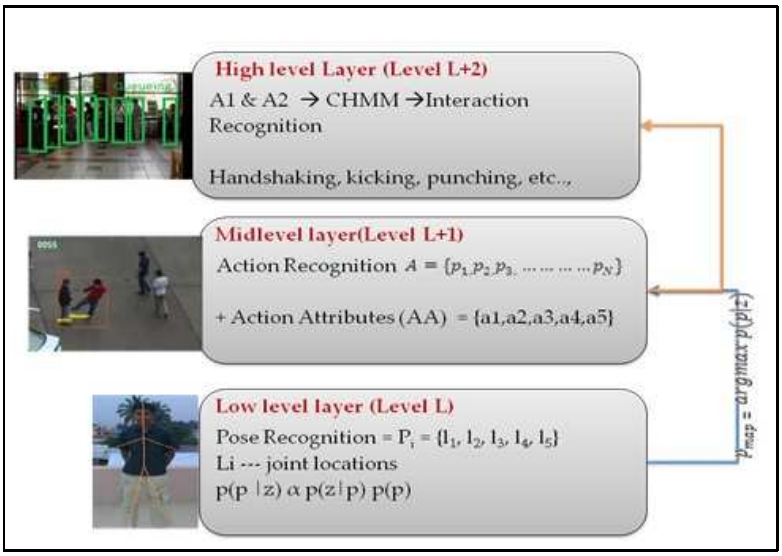

Figure 1: Overview of human interaction recognition system

5 Proposed learning model

The visual features are extracted from the videos and the feature vector has been derived from the observations and shown in figure 2. Feature vector observation includes spatial information $S_i(t)$ and temporal information T_N where t corresponds to the time stamp within the set of frames and T_N represents the trajectory information for N number of frames. Z is the feature set and it is represented as $Z = S_1(t), S_2(t), S_3(t), ...S_N(t), T_1, T_2, T_3, ...T_N$. Here, spatial representations have been given as joint locations.

Let the layered HMM1 models the observations of the person 1 and layered HMM2 models the observation of person 2 respectively. The Observations of the person 1 and 2 have been given as input to the layered HMM1 and layered HMM2 respectively. Let $P_1, P_2, ...P_n$ represents the pose of the person that has been shown in Pose layer (P-HMM) and defined as layer 1i and layer 1j for person 1 and 2 respectively. The Action layer (A-HMM) for person 1 and 2 is defined in layer 2i and layer 2j. Pose as the feature vector has been given as input to this layer and action of the person has been identified. In order to provide high level information action attributes has been added to the action. After identifying the action attributes, it has been manually labeled. Action and attributes of person 1 and action and attributes of person 2 has been coupled together to recognize the new interaction. The Interaction layer (I-HMM) has been considered as layer 3 and interactions are represented as $I_1, I_2, I_3, ...I_n$.Because of coupling the A1- HMM and A2-HMM are coupled so that interactions between two persons have been recognized.

6 Layers in HIR model

6.1 Pose layer (low level layer)

Pose is defined as the preliminary motion sequences that are obtained from the observations. Let the input observations be $o_1, o_2, o_3, ...o_n$as n represents the number of observations. High dimensional pose vector, $P_i = L_1, L_2, L_3, L_4, L_5$ where L_i represents the joint locations of each part of an image frame.$Pi \in \mathbb{R}^2$ where $\mathbb{R}^2$ represents the 2d space. $P_i \in p$ where p represents the pose of the body.$P(p|z)$ represents the inference framework to estimate the posterior probability

where p represents the pose and z represents the feature set. The desired posterior probability has been calculated using likelihood and prior. $P(p|z) \propto p(z|p)p(p)$. Maximum a posteriori solution can give a high likelihood. In the layer 1, the pose of the person in the particular time interval has been identified and the output has been given as input to the next level.

$$p_{map} = argmax(p(p|z)) \tag{1}$$

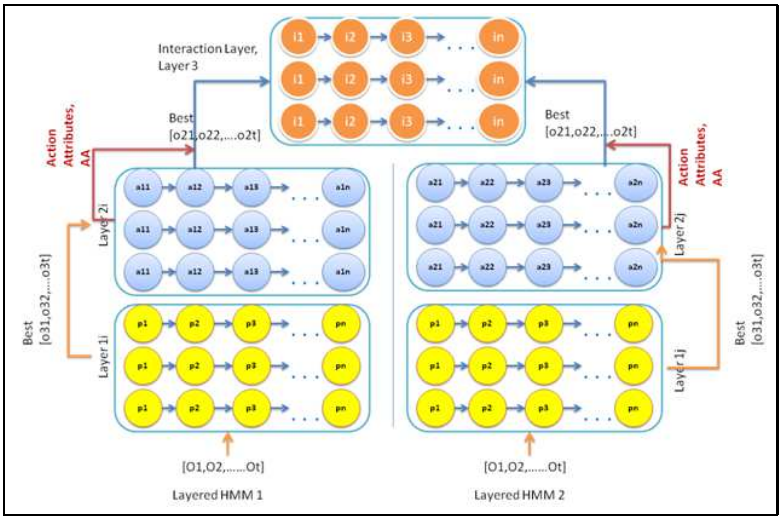

Figure 2: HMM model for HIR system

6.2 Action layer (mid level layer)

Action is defined as the stream of successive pose that happened in the particular time interval. Actions may be defined as the event that takes place for a single person.

$$Action, A = p_1, p_2, p_3, ...p_N \tag{2}$$

The Action of the individual person includes pose and action class labels. ie.,$Action, A \in P, A_i(t)$ Maximizing the log likelihood probabilities, the inferential result has been calculated and the actions such as walking, running, jogging, loiter and get hurt has been identified. After training, the likelihood of the action class in the layer 2i and layer 2j is calculated separately. Let $a_t = a_1^t, a_2^t, ...a_{pn}^t \in \mathbb{R}^{pn}$ denotes the pose vector in a continuous space of dimension equal to the number of individual actions.This layered approach directly outputs the probability p_k^t for each individual action model M_k Where $k = p_1, p_2, p_3, ...p_N$ as input to action HMM where $a_k^t = p_k^t$ for all k. To calculate the probability of model M_k for the given sequence x_1^t is computed in the following manner: $x_1^t = x_1, x_2, ...x_t$ represents the sequence of model.

Let us define forward variable $\alpha(i,t) = P(x_1^t, q_t = i)$ which is the probability of having generated the sequence x_1^t being in state i at time t. In asynchronous HMM,$\alpha(i,t)$ can be replaced by a corresponding factor ([9], [23]). Assume$\sum_{j=1}^{NS} p(q_t = j) = 1$ where NS is the number of states for all models. Probability $p(q_t = i|x_1^t)$ of state i is given as

$$\begin{aligned} p(q_t = i|x_1^t) = \frac{p(q_t = i|x_1^t)}{p(x_1^t)} &= \frac{p(q_t = i|x_1^t)}{p(q_t = j|x_1^t)} \\ &= \frac{\alpha(i,t)}{\sum_{j=1}^{NS} \alpha(j,t)} \end{aligned} \tag{3}$$

From this, the probability of model M_k can be computed as

$$p_{k^t} = \sum_{i \in M_k} (p(q_t = i|x_1^t)) = \frac{\sum_{i \in M_k} \alpha(i,t)}{\sum_{j=1}^{NS} \alpha(j,t)} \tag{4}$$

Where i is the state of the model M_k, i ∈ states of all models and NS denotes the total number of the states. In this work, individual action recognition vectors with the action attributes as observations given to the interaction level HMM.

Attribute selection criteria

Human interactions can be recognized with the help of Action Attributes (AA). The purpose of the attributes is to provide high level knowledge about actions.

$$Action Attribute set, AA = a_1, a_2, a_3, a_4, a_5 \tag{5}$$

Attributes have been manually labeled where low level features are given a class label. Here, in this work, five attributes have been manually labeled. Stretching arm, withdrawing arm, stretching legs, withdrawing leg, hand contact and body contact are defined as the attributes. The presence or absence of each attribute is approximated by the confidence value (0 or 1). Attribute classifiers are learned from training data sets. In the multi-level modeling approach [24] , the Knowledge of actions and attributes gives interaction in a more accurate way.

6.3 Interaction layer (high level layer)

Interaction is defined as successive actions between two persons that are integrated together. I represent the possible interactions between all possible co-existing pairs A1 and A2 where A1 and A2 denote the action of the person 1 and 2 respectively. A1 and A2 are coupled together to identify the new interaction

$$I(1,2) = [((A1 + AA)(t1).....(A1 + AA)(tz))U((A2 + AA)(t1).....(A2 + AA)(tz))] \tag{6}$$

where Iij represents the interactions between 2i layer, 2j layers respectively, and t1,t2...tz represents the time frames from 1 to z. The Interaction layer (I-HMM) is defined as the third layer in which the observations have been done using log-likelihood of the action layer and its inferential results. The interactions such as handshaking, pushing, hugging, fighting and meeting have been recognized. The knowledge of both the layers has been coupled to recognize the interaction between two or more persons. This layer uses spatio-temporal constraints as features. A CHMM model (λ) is defined by the following parameters. [16]

$$\pi_{o^C}(i) = P(q_1^c = S_i) \tag{7}$$

$$a_{i|j,k}^C = P(q_t^C = S_i | q_{t-1}^{A1} = S_j, q_{t-1}^{A2} = S_k) \tag{8}$$

$$b_t^C(i) = P(O_t^C || q_t^C = S_i) \tag{9}$$

Where $C \in (Action1, Action2)$and q_t^C Represent the state of coupling nodes in the C^{th} stream at time T. In Coupled HMM, the output observed from layer 2i and layer 2j are given as inputs. The observed sequence has been given as, $O = A_{1T}, A_{2T}$ Where $A_{1T} = a_{11}, a_{12}, a_{13}, a_{1T}$ are the observations of the first person and $A_{2T} = a_{21}, a_{22}, a_{23}, a_{2T}$ are the observed sequence of second person. The observation a_{11} consists of A1 + AA. Here, the observation sequence consists of actions of each person (A) and action attribute (AA) of each person. The state sequence

has been given us, $S = X_1^T, X_2^T$ where $X_1^T = X_{11}, X_{12}, X_{13}, ...X_{1T}$ $X_1 \in 1, ...M$ are the state sequence of first observations and $X_2^T = X_{21}, X_{22}, X_{23}, ...X_{2T}$ $X_2 \in 1, ...M$are the state sequence of second observations. State Transition probabilities of the first chain of observations have been represented as, $P(X1_{t+1}|X1_t, X2_t)$ and for the second chain as P $(X2_{t+1}|X1_t, X2_t)$. P (X1) and P (X2) are the prior probabilities of first and second chain respectively. $P(A1_t|X1_t)$ and $P(A2_t|X1_t)$ are the observation densities assumed to be multivariate Gaussian with mean vectors μ_x , μ_y and covariance matrices Σx,Σy. Expectation Maximization (EM) Algorithm finds the maximum likelihood that estimates the model parameters by maximizing the following function (10). Parameter λ , contains parameters of transition probability, prior probability, and parameters of observation densities.

$$\begin{aligned} M(\lambda) = P(X1)P(X2)\pi_{t-1}^{N}P(A1_t|X1_t)P(A2_t|X2_t)P(X1_{t+1}|X1_t, X2_t) \\ P(X2_{t+1}|X1_t, X2_t), 1 \leq t \leq N \end{aligned} \tag{10}$$

7 Concurrent interaction recognition

The interactions that happened between groups of people simultaneously at a particular time interval are defined as concurrent interactions. Example of concurrent interaction is hugging and handshaking in a single scenario. Here, in this work four persons have been considered and concurrent interactions between them have been identified.

$$CI(1, 2, 3, 4) = I(1, 2) + I(3, 4) \tag{11}$$

The training of CHMM differs from standard HMM in the expectation step (E) while they are both identical in the maximization step (M) which tries to maximize the equation (10).The expectation step of CHMM is defined in terms of forward and backward recursion. For the forward recursion, we define a variable for both observation chains at t=1,

$$\alpha_{t=1}^{person1(A1)} = P(A1|X1)P(X1) \tag{12}$$

$$\alpha_{t=1}^{person2(A2)} = P(A2|X2)P(X2) \tag{13}$$

Then the variable α is calculated incrementally at any arbitrary moment t as follows.

$$\begin{aligned} \alpha_{t+1}^{person1(A1)} = P(A1_{t+1}|X1_{t+1}) \int\int (\alpha_t^{person1(A1)})(\alpha_t^{person2(A2)} \\ P(X1_{t+1}|X1_t, X2_t), dX1_t dX2_t \end{aligned} \tag{14}$$

$$\begin{aligned} \alpha_{t+1}^{person2(A2)} = P(A2_{t+1}|X2_{t+1}) \int\int (\alpha_t^{person1(A1)})(\alpha_t^{person2(A2)} \\ P(X2_{t+1}|X1_t, X2_t), dX1_t dX2_t \end{aligned} \tag{15}$$

In the backward direction, there is no split in the calculated recursions which can be expressed as:

$$\begin{aligned} \beta_{t+1}^{person1(A1),person2(A2)} = P(O_{t+1}^{N}|S_t) = \int\int P(A1_{t+1}^{N}, A2_{t+1}^{N}|X1_{t+1}, X2_{t+1}) \\ P(X1_{t+1}, X2_{t+1}|X1_t, X2_t)dX1_{t+1}dX2_{t+1} \end{aligned} \tag{16}$$

After combining both forward and backward recursion parameters, an interaction will be tested on the trained model, generating the equivalent interaction that most likely fit the model. The generated interaction sequence is determined when there is a change in the likelihood.

8 Results and discussion

8.1 Datasets and experimental setup

The proposed system has been implemented in MATLAB version13a, with no special requirements in hardware. The proposed work has been analyzed and evaluated using four video datasets. They are UT-Interaction dataset, BEHAVE dataset, Self generated dataset and KTH dataset. The UT-Interaction dataset consists of different two person interaction patterns like shake hands, hug, point, kick, punch and push. The BEHAVE dataset consist of the outdoor environment and it consists of activities like group formation, crossing each other, depart, approach, move closer, move farther, etc. The generated dataset consists of single person activities and interactions that happen between two persons are taken under indoor environment. The activities covered in this dataset are walking, running, jogging, loitering and getting hurt. In Kth dataset, the walking, jogging and running scenarios have been taken for evaluation. In Experiment 1, The UT-Interaction dataset [12] contains videos of continuous executions of 6 classes of human-human interactions: shake-hands, point, hug, push, kick and punch. There are a total of 20 video sequences whose lengths are around 1 minute. The videos are taken with the resolution of 720*480 and 30fps. High level interactions have been recognized using four participants. Results from this dataset has been shown in figure 3,4,5.

Figure 3: Activity recognition - approach and group formation, Departing, Collide and Divert

Figure 4: Interaction recognition (Kick, Hug and Push)

Figure 5: Concurrent interaction recognition (Hug and Punching, Pushing and Approaching, Handshaking and Pushing)

In Experiment 2, training and testing has been carried out using Behave dataset [17] comprises of two views of various scenario's of people acting out various interactions. The data is captured

at 25 frames per second. The resolution is 640 * 480. The following ten interactions have been recognized in the Behave dataset. In Group, Approach, Walk Together, Meet, Split, Ignore, Chase, Fight, Run Together and Following are the interactions. The results from this dataset have been shown in figure 6.

Figure 6: Concurrent interaction recognition
(in group and crossing each other, walk together and walking)

8.2 Performance analysis

In this research, metrics such as Accuracy, Precision, Recall, Positives such as True Positive (TP), False Positive (FP) and Negatives such as True Negative (TN), and False Negative (FN) have been used to measure the performance of the system. The Performance metrics have been calculated for the poses, activity and interactions.

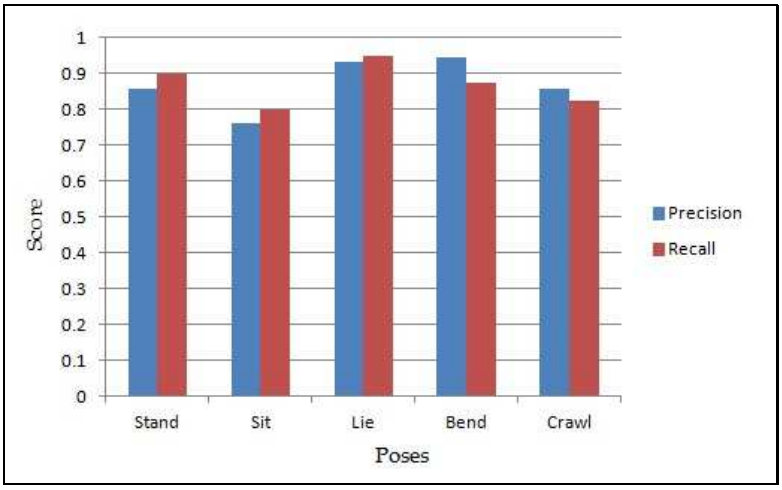

Figure 7: Pose recognition rate

Precision and recall values for each pose such as stand, sit, lie, bend and crawl have been shown in figure 7. Stand and sit poses has more precision and recall values. Lie, bend and crawl poses has more precision values than recall values. Figure 8 shows the precision- recall curve for activity recognition rate. The performance for the activities such as walk, run, jog, loiter and get hurt has been shown. Walk, run and loiter activities have high recall value than precision value. The activities walk, run and loiter are slightly confusing due to temporal difference. The other activities such as jog and get hurt have high precision values.

Figure 9 show the precision-recall curve for interaction recognition rate. The interactions such as handshaking, hugging, kicking, punching and pushing have been predicted. Handshaking, punching and pushing interactions have high precision values. Hugging and kicking have low precision values than other interactions. The recognition rate of poses, activity, interaction

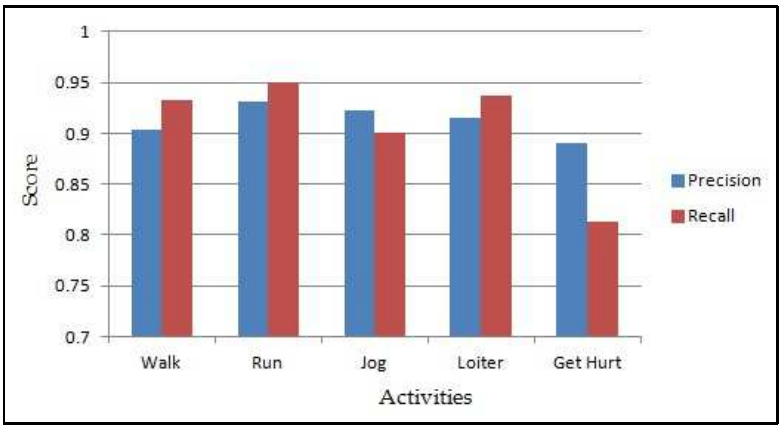

Figure 8: Activity recognition rate

and concurrent interaction in different datasets has been shown in table 1. These concurrent interactions have been identified using the proposed learning model. Single person action has been recognized from Kth dataset and recognition rate has been obtained as 87.62%.

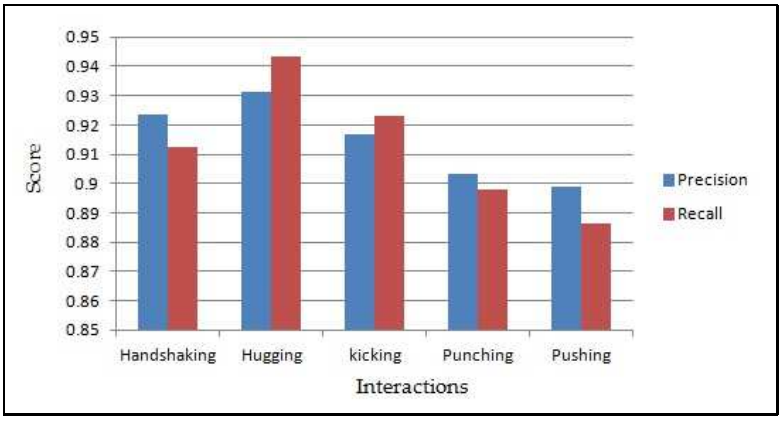

Figure 9: Interaction recognition rate

Table 2 shows the comparison of the previous works that are carried out in this activity recognition. In Learning Methodologies, the joint form of LHMM and CHMM outperforms other learning models. The spatial relations and the levels of temporal granularity have been considered. Usually, the HMM can handle temporal variations in the video. In this paper, the interaction between neighboring states also been considered. CHMM has the capacity to underlying synchronization of two different processes. The two actions have been coupled using coupling probabilities with proper weights and the concurrent interactions have been recognized.

Table 1. Recognition Rate across Datasets (in terms of Accuracy)

Datasets	Poses	Single - person Action	Two- person Interaction	Concurrent Interaction
Kth	87.62	-	-	-
UT-interaction	-	-	83.90	81.54
BEHAVE	-	-	-	71.56
Self-generated	91.55	93.47	84.23	88.36

Table 2. Comparison of Activities recognized with previous works

Previous Works	Learning Model Proposed	Concentrated on	Activities Recognized	Recog. Rate(%)
Sunyoung Cho et al.,2013 [25]	Visual and textual information as features,Graphical model using structured learning with latent variables	Activity Recognition	High five, handshake, hug and kiss.	78.4
M.J.Marin Jimenez et al.,2013 [26]	Dictionary learning, support vector machines with φ^2 Kernel.	Activity recognition	Hug, kiss and hand shake.	78
Hejin Yuan et al.,2015 [27]	Semi supervised learning-k-means clustering, Skeleton features. (Cumulative Skeleton Image(CSI), Silhouette History Image(SHI))	Activity recognition - single person - basic actions	Weizmann dataset (10 basic actions) Bend, jump, jump, jack, wave, wave1, wave 2, side, walk, run.	90
Fadime Sener et al .,2015 [4]	Multiple instance learning process. Shape and motion features	Two person Interactions.	UT Interactions dataset and TV interaction Dataset	75.60
Proposed System	Three layer HMM along with coupled HMM	Concurrent Interaction with four persons - complex activities	In Group, walking together, chase, fight, following, handshaking and hugging.	88.36

8.3 Discussion

Our interaction recognition system is based on Hierarchical Hidden Markov Model (HHMM) which combines layered and coupled HMM that tries to find the concurrent interactions. The three layers such as Pose layer, action layer and interaction layer has been modeled. In order to identify the interaction, action of the two persons has been modeled independently. As a preprocessing step, we have been deployed body part modeling and location based trajectory tracking to aid the localization of the people in frames. The action sequence is composed of parallel states presenting the poses and each pose is composed of the specific number of observations (M= 5 in our case). The action sequence of person 1 and 2 has been modeled individually in the layered fashion. Interaction Sequence is composed of the action sequence of person 1 and the action sequence of person 2. (M=10). After training LHMM, the observation sequence from the databases, the system tries to find a corresponding sequence of poses based on the learning during the training phase. The generated pose sequence is the sequence that achieves the maximum likelihood estimation with the poses. Thus the observed pose is given as input

to the next layer HMM. In the second layer HMM, the system tries to find the corresponding action sequences. The action of person1 is identified from the maximum likelihood estimation of the action sequences. In the same way, the action of person 2 also identified using two layered HMM. Coupled HMM (CHMM) couples both the action sequence of person1 and person2 and the system tries to find the interaction in a better way. CHMM in lag1 condition can couples the observation channels. Each channel has its own action sequence. From both the action sequences the next state emission probability has been generated. Based on all previous works, the specific activity has been recognized using the specific learning model. HMM is the most successful framework in speech and video applications and it is well suited for computing with uncertainties. Here, in this work to demonstrate the concurrent interactions, the HMM learning model has been extended in the joint form of layered HMM and coupled HMM. Layered HMM models the non-causal symmetric influences and CHMM to model the temporal and asymmetric conditional probabilities between observation chains.

Conclusion

In this work, a hybrid learning framework is designed to recognize the concurrent interactions between multiple peoples. The spatial and temporal information and body part attributes are considered as features. The poses and actions are recognized in a layered fashion. The actions of multiple persons are coupled to recognize the interactions and concurrent interactions. The joint form of LHMM and CHMM has been used for providing concurrency. The interactions between neighboring persons has also been recognized. Here, the activity recognition has been done for the continuous events and this could be extended in future to discrete event recognition mechanisms also. Further work will focus on identifying interactions and behavior in different person to person interaction contexts that will allow the system to recognize the interactions under different conditions.This system can act as a smart surveillance to recognize the actions/interactions of multiple people without human intervention in the environments such as meeting hall, discussion groups, public places, banking sectors, where multiple people could interact with each other.

Acknowledgment

The work reported in this paper has been supported by Anna University,Chennai by providing Anna Centenary Research Fellowship. We also acknowledge the anonymous reviewers for comments that lead to clarification of the paper.

Bibliography

[1] Alexandros Andre Chaaraoui, Pau Climent-Perez, Francisco Florez-Revuelta(2013); Silhouette-based human action recognition using sequences of key poses, *Pattern Recognition Letters*, 34(15): 1799-1807.

[2] Aggarwal, J. K.and Ryoo, M. S. (2011); Human activity analysis: A review, *ACM Computing Survey*, 43(3): 16:1–16:43.

[3] Arnold Wiliem,Vamsi Madasu, Wageeh Boles and Prasad Yarlagadda (2012); A suspicious behaviour detection using a context space model for smart surveillance systems, *computer vision and Image Understanding*, 116(2): 194-209.

[4] Fadime sener and Nazli Ikizler-cinbis (2015); Two Person Interaction Recognition via spatial Multiple Instance Embedding, *Journal of Visual Communication and Image Representation*, 32: 63-73.

[5] Gowsikhaa.D, Abirami.S and Baskaran.R. (2012); Automated human behavior analysis from surveillance videos: a survey, *Artificial Intelligence Review* , DOI 10.1007/s10462-012-9341-3, 1-19.

[6] Gowsikhaa.D, Manjunath and Abirami S. (2012); Suspicious Human activity detection from Surveillance videos, *International Journal on Internet and Distributed Computing Systems*, 2(2): 141-149.

[7] Junji Yamato, Jun Ohya and Kenichiro Ishii (1992); Recognizing Human Action in Time-Sequential Images using Hidden Markov Model, *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi:10.1109/cvpr.1992.223161, 379-385.

[8] Matthew Brand, Nuria Oliver, and Alex Pentland (1997); Coupled Hidden Markov Models for Complex Activity Recognition, *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, DOI: 10.1109/CVPR.1997.609450, 994 - 999.

[9] Nuria Oliver, Ashutosh Garg and Eric Horvitz (2004); Layered Representations for learning and inferring office activity from multiple sensor channels,*Computer Vision and Image Understanding*, 96: 163-180.

[10] Roberto Melfi, Shripad Kondra and Alfredo Petrosino (2013); Human activity modeling by spatio temporal textural appearance, *Pattern Recognition Letters*, 34(15): 1990-1994.

[11] Ryoo M.S. (2011); Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos, *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, DOI: 10.1109/ICCV.2011.6126349, 1036-1043.

[12] Ryoo, M.S, and Aggarwal, J.K. (2010); UT Interaction Dataset, *Proc. of ICPR Contest on Semantic Description of Human activities.*

[13] Sangho Park and J.K. Aggarwal (2004); Semantic-level Understanding of Human Actions and Interactions using Event Hierarchy, *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, DOI:10.1109/CVPR.2004.160, 1-12.

[14] Sang Min Yoon , Arjan Kuijper (2013); Human action recognition based on skeleton splitting, *Expert systems with Applications*, DOI:10.1016/j.eswa.2013.06.024, 40(17): 6848-6855.

[15] Sivarathinabala M. and Abirami S. (2014); Motion Tracking of Humans under Occlusion using Blobs, *Proceedings of Advanced Computing, Networking and Informatics- Volume 1, Smart Innovation, Systems and Technologies*, 27: 251-258.

[16] Shih-Kuan Liao, Baug-Yu Liu,(2010); An edge-based approach to improve optical flow algorithm, *Proceedings of Third International Conference on Advanced Computer Theory and Engineering*, 6: 45-61.

[17] Shizhong and Joydeep Ghosh (2001); A New formulation of Coupled Hidden Markov Models, doi=10.1.1.607.5700rep=rep1type=pdf.

[18] S. J. Blunsden and R. B. Fisher (2010); The BEHAVE video dataset: ground truthed video for multi-person behavior classification, *Annals of the BMVA*, 4: 1-12.

[19] Teddy Ko (2010); A Survey on Behavior Analysis in Video Surveillance Applications. *Proceedings of IEEE, Applied Imagery Pattern Recognition Workshop*, 1-8.

[20] Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers (2010); Combined Region and Motion-Based 3D Tracking of Rigid and Articulated Objects, *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 32(3): 402-415.

[21] Weilun Lao, Jungong Han, and Peter H. N. deWith (2010); Flexible Human Behavior Analysis Framework for Video Surveillance Applications. *International Journal of Digital Multimedia Broadcasting*,ID: 920121, 1-9.

[22] Weiyao Lin, Ming-Ting Sun, Radha Poovendran and Zhengyou Zhang (2010); Group Event Detection with a Varying Number of Group Members for Video Surveillance, *IEEE Transactions on Circuits and Systems for Video Technology*, 20(8): 1503.00082.

[23] Weiming Hu, Guodong Tian , Xi Li , Stephen Maybank (2013); An Improved Hierarchical Dirichlet Process-Hidden Markov Model and Its Application to Trajectory Modeling and Retrieval, *Int J Comput Vis*, DOI 10.1007/s11263-013-0638-8, 105:246-268.

[24] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, Iain McCowan, and Guillaume Lathoud (2004); Modeling Individual and Group Actions in Meetings: a Two-Layer HMM Framework, *the Second IEEE Workshop on Event Mining: Detection and Recognition of Events in Video, In Association with CVPR*, 1-8.

[25] Gildas Morvan, Daniel Dupont,Jean-Baptiste Soyez, Rochdi Merzouki (2012); Engineering hierarchical complex systems: an agent-based approach, The case of flexible manufacturing systems, *Chapter - Service Orientation in Holonic and Multi-Agent Manufacturing Control, series Studies in Computational Intelligence*, 402: 49-60.

[26] Cho, Sunyoung and Kwak, Sooyeong and Byun, Hyeran (2013); Recognizing Human-human Interaction Activities Using Visual and Textual Information, *Pattern Recogn. Lett.*, 34(15):1840-1848.

[27] Manuel J. Marin-Jimenez, Enrique Yeguas, Nicolas Perez de la Blanca (2013); Exploring STIP-based models for recognizing human interactions in TV videos, *Pattern Recognition Letters*, 34: 1819 -1828.

[28] Hejin Yuan (2015); A Semi-supervised Human Action Recognition Algorithm Based on Skeleton Feature, *Journal of Information Hiding and Multimedia Signal Processing*, 6(1): 175-181.