A Pyramidal Layered HMM for Multiview Human Behavior Recognition in Asynchronous Video Streams

Amir Farid Aminian Modarres
Iran University of Science and Technology
School of Computer Engineering
Tehran, Iran

Mohsen Soryani
Iran University of Science and Technology
School of Computer Engineering
Tehran, Iran

ABSTRACT

Extracted features which are obtained from a multiview video stream form a special case of a multi-sensor observation sequence. If the sensors are not synchronous, the observed features of views are not aligned together and this makes some difficulties in classification applications. A new architecture for hidden Markov model, namely pyramidal layered hidden Markov model, is proposed in this paper to handle this situation. This is accomplished by means of separate decoding in each view stream in bottom layer and then fusion of the aligned decoded symbols in top layer. Structure and algorithms of the new structure are introduced and are then used for human behaviour recognition in multiview video sequences. Considering collected information from all views of a multiview human action recognition system, one expects the recognition rate to increase and some problems like occlusion to be rectified. Several experiments have been performed in this paper. The experimental results show high performance, about 93.8% in average, in multiview human behavior recognition, as well as accuracy improvement compared to similar methods. The results are also compared with other contributions on three different multiview behavior datasets.

General Terms

Video Processing, Classification, Recognition, Hidden Markov Models

Keywords

Pyramidal Layered HMM, PLHMM learning and decoding, Dynamic Time Warping, Human Behavior Recognition

1. INTRODUCTION

Human action recognition is an active research area in video processing and exploitation, in which a sequence of human motions in video streams is modelled and classified. As any typical classification problem, human action recognition has two essential phases: extraction of significant features which have good discriminability and training a conform model.

Recently, many efforts have been made to utilize the benefits of using multiview video sequences and perform human behavior recognition in multiview scope. A multiview video sequence is a combination of two or more video streams, which are obtained from multiple cameras arranged around a scene in different views capturing the occurred events simultaneously. Researchers hope that by means of multiview video streams, the performance of recognition system is increased and also some problems such as occlusion are rectified.

Data fusion, which is the integration of multiple data from different sensors or different related processes and producing accurate information from them, can be done in three different levels: sensor, feature and decision level. Obviously data fusion must be used in one multiview action recognition scenario [1]. To perform data fusion in senor level in multiview video sequences, the raw imaging data of all views must be combined, which in turn needs to find related pixels in all views using projective geometry concepts and three dimensional scene reconstructions. In the feature level fusion, features are extracted from video stream of each view independently and then combined to a multi-dimensional feature space. The learning and classification are done in this new space. At last, in the decision level fusion, feature extraction and classifier training are done independently in each view and the decisions of all classifiers are combined to an ensemble classifier to make the final decision. A new architecture is proposed for Hidden Markov Models (HMM) in this paper which can classify the observation raised from a multiview video stream in the feature level fusion.

Assuming that feature extraction and selection have been done before, a new extension of HMM is proposed which is named ‘pyramidal layered hidden Markov model (PLHMM)’, for classification and recognition of human behavior in multiview video streams. After a brief review of current HMM-based extensions and their advantages and disadvantages, the structure of the model was introduced. The word 'pyramidal' was assigned to PLHMM because of its pyramidal structure that has been spread in the bottom and aggregated at the top. The learning and inference algorithms of PLHMM have been explained. It is very important to note that these algorithms are a mixture of traditional algorithms available for HMMs, which are customized for PLHMM. Therefore, no additional difficulties have been arising, unlike the similar proposed extensions of HMM.

After introducing the PLHMM and its algorithms, we test its ability to learn and classify human behaviors in multiview video sequences as a multi sensor problem in three different famous multiview datasets: MuHAVi, IXMAS and HumanEvaI. The experiments are designed to specially show the high performance of PLHMM in situations where incoming information from multiple sensors is not synchronized. Also, the best parameters of PLHMM are discovered and the final results are compared with other works.

This paper is organized as follows. The next section includes a brief review of some of extended architectures of HMM. The new PLHMM architecture is introduced in section 3 and its learning and decoding algorithms will be described in section 4. Experimental results of applying PLHMM on multiview human behavior recognition application are discussed in section 5, and finally section 6 concludes the paper.

2. HMM EXTENSIONS

HMM and its extensions have been widely used in action recognition. HMM and its extensions are a particular case of temporal or dynamic graphical models (DGMs) [2]. Some advantages of DGMs that are relevant to the problem of human behavior modelling are: the ability of handling incomplete data as well as uncertainty; trainability of them; conditional independency encoding; existence of efficient algorithms for doing predictive inference; suggestion of a framework for combining prior knowledge and data; and finally modularity and parallelizability [2].

All HMM extension models could be transformed into a traditional HMM in practice [3]. One of the advantages that may be expected from using such models over a large usual HMM is that these models are less likely to suffer from over-fitting, since the individual sub-components are trained independently on smaller amounts of data. A consequence of this is that a significantly smaller amount of training data is required for the LHMM to achieve a performance comparable with the HMM.

A stochastic context-free grammar has been proposed to compute the probability of a temporally consistent sequence of primitive actions which are recognized by HMMs [4]. An entropic-HMM is proposed to segment the observed video activities into semantic states [5].

The standard Baum–Welch procedure has been extended in Hierarchical HMMs (HHMMs) structure [6]. Since the original approach is very time consuming, a linear-time inference algorithm for HHMMs was introduced [7]. The HHMM is also used in some complex behavior recognitions such as multi-person activity recognition [8].

Embedded HMMs that are extended architectures of HMMs are proposed and are used on two-dimensional data such as image processing [9]. In such architectures, one HMM models one dimension of the data while its state variables correspond to the other dimension of the data.

Layered LHMM is a technique proposed to use learning at multiple levels [10]. A learning algorithm is used to determine how the outputs of the base classifiers should be combined, so this is an example of ensemble classification. This is a more sophisticated technique than cross-validation and has been shown to reduce the classification error due to the bias in the classifiers [2]. Rather than training the models at all levels at the same time, the parameters of the HMMs at each level can be trained independently; provided that the previous level has been already trained, in a bottom-up fashion. The inputs (observations) of each level are the classification outputs of the previous level. At the lowest level the observations are the feature vectors extracted directly from sensor signals.

A layered structure provides several valuable properties. These properties make it feasible to decouple different levels of analysis for training and inference. Each level of LHMM is trained independently, with different feature vectors and time granularities. Once the system has been trained, inference can be carried out at any level of the hierarchy. Another advantage is that the layers at the bottom of the LHMM, which are more sensitive to changes in the environment such as the type of sensors and sampling rate can be retrained separately without altering the higher layers of the LHMM [2].

Several extensions of HMM have also been introduced in the past years, such as factorial HMM [11], mixed HMM [12] and profile HMM [13]. Many of these extended HMM models are also frequently used in human action recognition. More complex models, such as Parameterized-HMMs [14], Variable-length HMMs [15] and Coupled-HMMs [16-18] have been used to recognize more complex activities such as the interaction between two people. Many other complex Bayesian networks have also been used for the modelling and recognition of human activities [19-22].

The proposed architecture classifies the observation sequences in a bottom-up layered fashion similar to previous model LHMM, but the difference is that the PLHMM can receive more general asynchronous multi-sensor observation sequences.

3. PLHMM STRUCTURE

To construct an automatic multiview human behavior recognition system, a powerful classifier which can receive and operate on multiview video sequences is required. Although present classifiers can perform this job anyway, using HMM or its extensions such as LHMM and HHMM may cause some restrictions on the applicability of the system. One of the most important limitations is that the observations should be prepared as a matrix in which each column belongs to the observation vector in one time slice. In a multiview video sequence, the observation vector contains the symbols sensed from all cameras. This means that the cameras should be synchronized necessarily. However this is not true in most multiview human action datasets, which could limit the application of the system.

The output of a typical multi sensor system is demonstrated in Figure 1. As it is shown in this figure, one multi-sensor sample consists of many symbol sequences, each coming from a distinct sensor or different process. It is very important to note that the lengths of these sequences are not necessarily equal because the sampling rate of the sensors may be different. In addition, if the sensors are not synchronized, perceived symbols of the sensors in a specific position are not related and belong to diverse time instances.

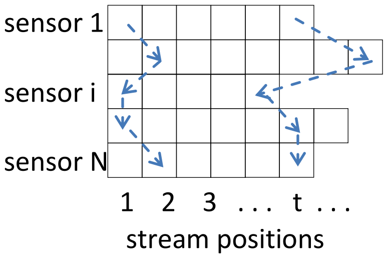

Fig 1: The typical structure of a general multi-sensor observation sequence. The dotted lines show the probable dependencies in the observation of the sensors.

To perform training and classification in this situation, the traditional classifiers like HMM cannot be implemented without any consideration. To overcome this problem, PLHMM is proposed which can receive the outcomes of many sensors simultaneously and after a preliminary decoding, warp the sequences together to perform the classification in the next step.

The general architecture of a PLHMM may have many layers but in this paper a two-layered PLHMM is introduced. As seen in Fig. 2, the lower layer has many groups of HMMs; each group will receive its observations from a distinct sensor. The fact that PLHMM can receive many observation sequences from different sources is the main contribution and advantage of it, especially when operating on the outcomes of a multi-process system. This is the main difference between

PLHMM and LHMM, in which there is just one group of HMMs in each layer.

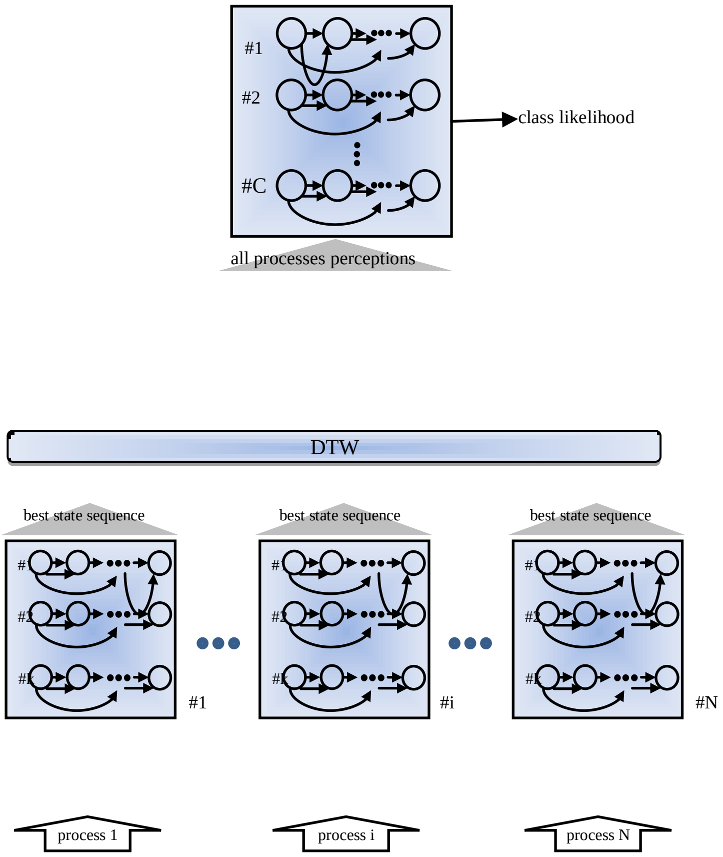

Fig 2: A PLHMM structure with two layers. It was used as a classifier for multiview human action recognition.

Each HMM group in the lower layer belongs to one view of multiview video sequences. The sequence of the most probable states of the winner HMM in each HMM group are produced as the output of that group (Fig. 3). As it is shown in Fig. 3, the input observation matrix of one HMM group, which includes *M* features in *T* time instances, formed the matrix$\{O_{m,t}\}_{M\times T}$, that is applied to all HMMs. The output of this appliance is vector $\{s^{w}_{t}\}_{1\times T}$, which denotes the most probable state string of the winner HMM. The term s^{w}_{t} stands for the most probable state in time *t* of the winner HMM marked by *w*.

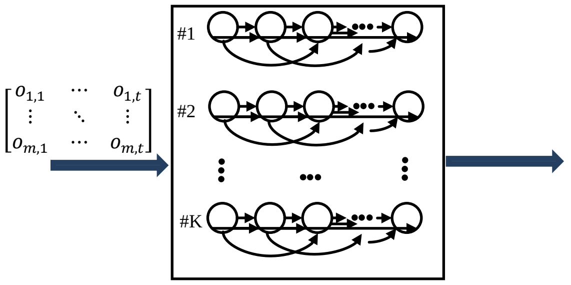
$\begin{bmatrix} o_{1,1} & \cdots & o_{1,t} \\ \vdots & \ddots & \vdots \\ o_{m,1} & \cdots & o_{m,t} \end{bmatrix}$
#1
#2
#K

Fig 3: A group of HMMs and their input and output structures.

The decoded state symbols of the lower layer are gathered, aligned and fed to the upper layer as the observations of it. The details of preparing the output of the lower layer for the top layer are described in the next section. The upper layer of the PLHMM consists of one group of HMMs. The final classification is done in the upper layer. Because of supervised learning manner of PLHMM, the number of HMMs in all the layers must be the same as the number of classes.

4. PLHMM TRAINING AND DECODING ALGORITHMS

In this section, training and decoding algorithms of the mentioned PLHMM are discussed. Fortunately, each part of the overall approach is one of the known and common algorithms and this helps the authentication of the overall algorithm.

The layered structure of PLHMM imposes a layer fashion in the learning and decoding algorithms. The general overview of algorithms is: the multi-sensor observation sequences are fed to the lower layer first, and after appropriate processing, the output of the lower layer is given to the top layer. In the learning stage, all layer HMMs are trained bottom-up and the parameter sets of them will be estimated. In the classification phase, the lower layer extracts some features from the observed sequence and passes them to the top layer, which is responsible for assigning a class tag to the sample.

4.1 LEARNING STAGE

The lower layer of the proposed PLHMM in Fig. 2 receives a training sample. Each training sample is a multi-sensor observation sequence, consisting of *N* input observation sequences, which are shown in Fig.1 as rows. The observation sequence of the *i*th process is considered as $\mathbf{O}_i=\{O_{f,t}\}_{Fi\times Ti}$ in which $O_{f,t}$ is the *f* th observation feature in the *t*th time instant and F_i and T_i stand for the size of the feature vector and time duration of the i^{th} process, respectively. Therefore, each process may have its own feature numbers and time duration and this is a very powerful ability of PLHMM that can receive some parallel observation sequences with different vector sizes and sampling rates. This ability is available because of independent operation of the HMM groups in the lower layer of the PLHMM.

Considering the observation of the i^{th} process in the i^{th} HMM group of the lower layer, the traditional algorithms such as expectation-maximization can be used as training algorithms to estimate the HMM parameter set in each HMM of each HMM group of the lower layer directly. First, usage of all training samples of class k, $\boldsymbol{\lambda}^k_i$ which is the parameter set of the k^{th} HMM in the i^{th} group of the bottom layer, is obtained by EM algorithm:

$$\boldsymbol{\lambda}_i^k = \arg\max_{\boldsymbol{\lambda}} \sum_i P\left(\boldsymbol{O}_i^k / \boldsymbol{\lambda}\right) \qquad (1)$$

In (1) $\boldsymbol{O}^k_i$ denotes all observation sequences of the i^{th} sensor which belongs to class k. When all the HMM parameters of the lower layer has been estimated, the training samples are launched again to the model in order to discover the best matched state sequences of HMMs. The Viterbi algorithm is used as a decoding algorithm for this purpose. This process is done for each view of the lower layer groups:

$$\boldsymbol{S}_i^k = \arg\max_{\boldsymbol{S}} P\left(\boldsymbol{S} / \boldsymbol{O}_i, \lambda_i^k\right) \qquad (2)$$

In (2) $\boldsymbol{S}$ stands for the sequence of states and $\boldsymbol{S}^k_i$ is the most probable state sequence of the k^{th} HMM in the i^{th} group of the bottom layer.

The outcome of the lower layer that is used as the observation sequence in the upper layer is a matrix of the form $\hat{\boldsymbol{S}}=\{\hat{S}_{i,j}\}_{N\times T}$, in which $\hat{S}_{i,j}$ is the state symbol of the i^{th} HMM group in the lower layer in the j^{th} time instant. Also, T is the time length of the overall observation sequences after performing the warping step which is described below.

A time warping algorithm is applied to align the length of the winning decoded sequences. The dynamic time warping (DTW) algorithm can be utilized for this purpose. Suppose that $\boldsymbol{S_i}=[s_{i,1}, \dots, s_{i,Ti}]_{1\times Ti}$ and $\boldsymbol{S_j}=[s_{j,1}, \dots, s_{j,Tj}]_{1\times Tj}$ are two state symbol series of lower layer of PLHMM. The DTW algorithm [23] finds the matrix $\boldsymbol{P}= [\boldsymbol{p}_i\boldsymbol{p}_j]$ for sequences $\boldsymbol{S_i}$ and $\boldsymbol{S_j}$ in which $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ are T×1 vectors of indices of two state symbol series, in such a way that the following expression is minimized:

$$\sum_{t=1}^{T} \left\| s_{i,p_{i,t}} - s_{j,p_{j,t}} \right\|^2 \qquad (3)$$

In (3) ||.|| is a distance measure which returns similarity between the two components of the observation sequences. Also, T denotes the total number of steps needed to align two observations or the length of warped strings.

Since the lower layer of PLHMM may have more than two HMM groups, the aligning algorithm should be able to receive and unify the size of many sequences altogether. T is set to the average length of all strings, and then by using (3), short strings are stretched and long ones are compressed.

After these steps, the warped symbol sequences are combined to form the $\hat{\boldsymbol{S}}$ matrix, as the observation of the top layer:

$$\hat{\boldsymbol{S}} = \begin{bmatrix} s_{1,1}^k & \cdots & s_{1,T}^k \\ \vdots & \ddots & \vdots \\ s_{N,1}^k & \cdots & s_{N,T}^k \end{bmatrix} \qquad (4)$$

In (4) the i^{th} row is the aligned state symbol sequence of the k^{th} HMM of the i^{th} HMM group and the j^{th} column forms an observation vector in the j^{th} time instance for the top layer in the k^{th}class.

When the observation sequence of the upper layer is made, its training must be done in the same manner as the bottom layer. So, the EM algorithm is used for the training of the top layer HMMs. Thus if the input sample belongs to class c, the EM algorithm is applied to yielding λ^c, the parameters of the c^{th} HMM of the top layer. Accordingly all PLHMM parameters are obtained in the learning algorithm.

4.2 DECODING STAGE

To use a PLHMM for a classification or recognition purpose, it is needed to estimate the likelihood of generating a multi-sensor sequence by HMMs in the top and bottom layers of the PLHMM.

Similar to the training algorithm, the evaluation algorithm is done in a layered fashion. More precisely, the testing sample is given to the lower layer and after decoding, it is passed to the top layer to do the final classification. The decoding process in the lower layer is done again by Viterbi algorithm. This algorithm was run in each i^{th} group of HMMs in the lower layer to determine the winner HMM in the i^{th} group as well as the most probable state transitions of it. Similar to the training algorithm, the decoded strings of the lower layer may not have equal lengths, so the warping algorithm has been performed to unify their time duration and production.

Now, the $\hat{\boldsymbol{S}}$ matrix, which is the lower layer outcome matrix used as the upper layer observation, is passed to all HMMs of the top layer and the likelihood of production of it was estimated using available simple algorithms such as Viterbi. Finally, the class which is assigned to the HMM with the highest probability is selected as the winner class and as the output of the classifier.

In the use stage, it is possible that the winner HMMs of all HMM groups of the lower layer, do not belong to the same class. This means that one or more HMM groups in the lower layer mistake the real class of the observation sequences and it causes a noisy observation for the top layer. However, if the noise is not dominant, the classification in top layer possibly corrects it and detects the true class. In the field of multiview human behavior recognition, if due to some problems such as occlusion, the true behavior is not recognized in some views, there is still a possibility that the final decision of PLHMM is correct. In fact the whole architecture produces an ensemble classifier which confirms the good recognition results of multiview human behavior recognition.

5. EXPERIMENTAL RESULTS

Various experiments about the proposed PLHMM will be discussed in this section. Multiview human action recognition is used as a multi-sensor problem to test the performance of PLHMM. The body posture graph (BPG) descriptor is used as the features or the observations for the recognition purpose. The BPG is a posture-based descriptor of body silhouette which is generated in three steps: performing a silhouette modelling, generating a skeleton-like graph and producing a special adjacency matrix for it [24]. This descriptor produces a 45×1 vector of binary features for one frame in each view of a multiview behaviour dataset. Observation sequences for one view are made by placing these vectors altogether according to the time.

5.1 EXPERIMENT A

There are many parameters in a PLHMM that must be defined before using it as a classifier. These parameters include: the number of layers, the number of HMM groups in each layer, the number of HMMs in each group of HMMs and finally the number of states of each HMM. According to the PLHMM structure descriptions, the number of HMM groups are obtained easily. There is just one HMM group in the top layer and the number of HMM groups in the lower layer is equal to the number of available views of the used dataset. Also, it can be inferred that the number of HMMs in each HMM groups of top and bottom layers must be the same as the number of action classes in the dataset. In this experiment the proper number of HMM states in each layer is investigated.

The MuHAVi multiview behavior dataset [25] was selected to test the performance of the proposed PLHMM in this experiment. The MuHAVi dataset originally has eight views and 17 actions repeated by 7 actors. Each action may be broken into primitive actions. The time duration of one action in different views may vary and the sequences of views are not synchronous.

To determine the situation in which the best accuracy is achieved, many tests with different PLHMM state numbers are set up. The recognition results are collected in the Table 1. These results show that, the best recognition rate of PLHMM in MuHAVi dataset is 93.91%, when each HMM in both layers has 6 states.

Table 1. Accuracy results of multiview human actions classification with different state numbers of PLHMM on MuHAVi dataset.

accuracy		the number of states in each HMMs of the bottom layer		
		4	6	8
the number of states in each HMMs of the top layer	4	90.97	92.65	91.60
	6	92.44	**93.91**	93.28
	8	92.44	93.07	92.02

In fact, decoding in lower layer groups results in segmentation of the action sequences into some primitive action components. In other words, one can presume that the lower layer HMMs learns the primitive actions and the top layer combines these symbol-coded actions and puts them in a meaningful behavior category.

On the other hand, the small number of states in each layer makes the learning process more accurate when little training data is available.

5.2 EXPERIMENT B

To indicate the better performance of PLHMM against the former architecture LHMM and the standard HMM, different LHMM and HMM models on MuHAVi dataset is trained. According to the results of Table 2, the best recognition rate of LHMM is 90.34% and the HMM with 10 state has the best accuracy of 89.71% due to the results of Table 3. Noting the best accuracies of Table 1, 2 and 3, the first conclusion is that the recognition using new architecture was improved significantly. This is an expected result, because the PLHMM performing separate decoding stages before the final recognition in each view of the multiview dataset, which helps the better characteristics extraction from each distinct view. In other words, the PLHMM performs the classification in each view independently and this is the reason why asynchronous cameras are not affecting the accuracy results.

It is important to note that to make it possible to implement the multiview behaviour recognition system with LHMM and HMM, the observation sequences of all views are manually aligned and by combining them one multi-feature observation sequence is made. In other words, fusion is made before decoding, in comparison with the PLHMM which performed it after initial decoding step between two layers. Obviously, when the sensors are not synchronous such as in the MuHAVi dataset, the learning and classification of such models caused some trouble and their accuracy was decreased. The results of followed experiment prove this opinion.

Table 2. Accuracy results of the classification of multiview human actions with different state numbers of LHMM.

Accuracy		the number of states in each HMMs of bottom layer		
		8	10	12
the number of states in each HMMs of top layer	6	87.82	89.29	88.24
	8	89.50	**90.34**	89.71
	10	88.87	89.92	88.45

Table 3. Accuracy results of the classification of multiview human actions with different state numbers of standard HMM.

the number of states in HMM	8	9	10	11	12
Accuracy	86.98	87.82	89.71	88.66	87.82

The second issue is that the best recognition results of LHMM occur when it has 10 and 8 states in each HMM of bottom and top layers, respectively. Reminded that the learning scheme of LHMM is supervised and each layer of it has 10 HMMs, concluding that the LHMM has totally 100 and 80 HMM states in bottom and top layer, respectively. On the other hand, the best model of PLHMM has 60 HMM states in each HMM group of both layers in PLHMM. The parallel nature of processing in each HMM group of the lower layer prevents the learning algorithm to be affected by over-fitting. Also, fewer states of the top and bottom layers caused better learning with more generalization ability of PLHMM compared with LHMM.

The overall result is that by using PLHMM, not only the recognition rate of the system increases, but also a better and more accurate training may occur when training samples are limited.

5.3 EXPERIMENT C

To prove that the proposed architecture is applicable to different scenarios, the multiview human action recognition was performed on three different datasets. In addition to MuHAVi dataset, in this experiment Inria Xmas Motion Acquisition Sequences (IXMAS) and HumanEva-I datasets are used. The IXMAS dataset contains 11 actions, each performed 3 times by 10 actors (5 males / 5 females). The acquisition was achieved using 5 cameras [26]. The HumanEva-I dataset contains 7 calibrated video sequences (4 grey scale and 3 colour) that are synchronized with 3D body poses obtained from a motion capture system. The database contains 4 subjects performing 6 common actions [27].

The high accuracy rate of PLHMM in the last row of Table 4 shows that the PLHMM has successfully overcome the multiview human behavior recognition. Although the recognition rate is not the same in three datasets, as a whole the result is reliable. Final results of the proposed recognition system are compared with the results presented by other researchers of these datasets. It is observed that the proposed approach succeeds to overcome others' results.

Table 4. Accuracy results of classification of multiview human actions with PLHMM on three different datasets, compared with other researchers' results.

	MuHAVi	IXMAS	Human EVA I
(Karthikeyan *et al.*, 2011) [28]	88.23	-	-
(Wu *et al.*, 2010) [1]	69.88	-	-
(Weinland *et al.*, 2007) [29]	-	81.27	-
(Junejo *et al.*, 2008) [30]	-	72.70	-
(Vitaladevuni *et al.*, 2008) [31]	-	87.00	-
(Weinland *et al.*, 2010) [32]	-	83.50	-
(Junejo *et al.*, 2011) [33]	-	74.60	-
(Jingen *et al.*, 2011) [34]	-	82.80	-
(Ning *et al.*, 2008) [35]	-	-	95.00
our approach	**93.91**	**91.71**	**95.83**

6. CONCLUSION

A new HMM topology for multiview human action recognition was proposed in this paper. The structure of PLHMM can receive several observation sequences from different sources in parallel and classify them by merging the available information, even if they are not synchronized. There are some novelties in the structure and algorithms of PLHMM. It is a general extension of LHMM which can receive multi-sensor observation sequence. The PLHMM processes this sequence in parallel and fuses useful information of it in feature level. Also, the algorithms of the proposed architecture were presented and theorized. These algorithms utilized previous algorithms in a layered manner, so the verification of the algorithms becomes easy. In the proposed training algorithm, the DTW algorithm is used to align more than two strings together. The decoding in the lower layer helps the PLHMM to successfully detect primitive actions in a behavior sequence and improve the performance of its classification. Also it causes each HMM to have a few states, which in turn leads to a better learning in the lack of huge training data.

Multiview human action recognition problem is chosen as a multi sensor problem and the recognition ability of the proposed PLHMM on it is tested. Then its best variable setting was found. Some experiments are performed to show the high performance of PLHMM compared to other architectures. The accuracy of PLHMM was also compared to other works in three different datasets. It was showed that the proposed architecture has overcome many difficulties in multiview sensor problem.

7. REFERENCES

[1] Wu, C., Khalili, A.H., Aghajan, H., 2010. Multiview activity recognition in smart homes with spatio-temporal features. Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '10), p.142-149. [doi:10.1145/1865987.1866010]

[2] Oliver, N., Garg, A., Horvitz, E., 2004. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, **96**(2):163–180. [doi:10.1016/j.cviu.2004.02.004]

[3] Rabiner, L., 1989.A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2):257-286. [doi:10.1109/5.18626]

[4] Ivanov, Y.A., Bobick, A.F., 2000. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(8):852-872. [doi:10.1109/34.868686]

[5] Brand, M., Kettnaker, V., 2000. Discovery and segmentation of activities in video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(8):844-851. [doi:10.1109/34.868685]

[6] Fine, S., Singer, Y., Tishby, N., 1998. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, **32**(1):41–62. [doi:10.1023/A:1007469218079]

[7] Murphy, K.P., Paskin, M.A., 2002. Linear time inference in hierarchical HMMs. Advances in neural information processing systems 14: proceedings of the 2001 conference, p.833. [doi:10.1.1.20.8131]

[8] Guo P., Miao Z., 2010. Multi-person activity recognition through hierarchical and observation decomposed HMM. Multimedia and Expo (ICME), IEEE International Conference on, p.143-148. [doi:10.1109/ICME.2010.5582559]

[9] Nefian, A.V., Hayes, M.H., 1999. An embedded HMM-based approach for face detection and recognition. Acoustics, Speech, and Signal Processing, IEEE International Conference on, p.3553-3556. [doi:10.1109/ICASSP.1999.757610]

[10] Oliver, N., Horvitz, E., Garg, A., 2002. Layered representations for human activity recognition. Multimodal Interfaces, Proceedings. Fourth IEEE International Conference on, p.3-8. [doi:10.1109/ICMI.2002.1166960]

[11] Chen, C.H., Liang, J.M., Hu, H.H., Jiao, L.C., Yang, X., 2007. Factorial hidden Markov models for gait recognition.International conference on Advances in Biometrics (ICB'07), p.124-133. [doi:10.1007/978-3-540-74549-5_14]

[12] Altman, R.M., 2007. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102**:201-210. [doi:10.1198/016214506000001086]

[13] Huang, R., Pavlovic, V., Metaxas, D.N., 2007.Shape analysis using curvature-based descriptors and profile hidden Markov models. Biomedical Imaging: From Nano to Macro, 4th IEEE International Symposium on, p.1220-1223. [doi:10.1109/ISBI.2007.357078]

[14] Wilson, A.D., Bobick, A.F., 1998. Recognition and Interpretation of Parametric Gesture. Proceedings of the Sixth International Conference on Computer Vision (ICCV '98), p.329–336. [doi:10.1109/ICCV.1998.710739]

[15] Galata, A., Johnson, N., Hogg, D., 2001. Learning Variable-Length Markov Models of Behavior. *Computer Vision and Image Understanding*, **81**(3):398-413. [doi:10.1006/cviu.2000.0894]

[16] Brand, M., Oliver, N., Pentland, A., 1997. Coupled hidden Markov models for complex action recognition. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, p.994-999. [doi:10.1109/CVPR.1997.609450]

[17] Binder, J., Koller, D., Russell, S.J., Kanazawa, K.,1997.Adaptive Probabilistic Networks with Hidden Variables. *Machine Learning*, **29**:213-244. [doi:10.1023/A:1007421730016]

[18] Ren, H.B., Xu, G.Y., 2002. Human action recognition with primitive-based coupled-HMM. Pattern Recognition Proceedings, 16th International Conference on, p.494-498. [doi:10.1109/ICPR.2002.1048346]

[19] Fernyhough, J., Cohn, A.G., Hogg, D.C., 1998. Building qualitative event models automatically from visual input. Computer Vision, Sixth International Conference on, p.350-355. [doi:10.1109/ICCV.1998.710742]

[20] Intille, S.S., Bobick, A.F., 1999. A framework for recognizing multi-agent action from visual evidence. Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence (AAAI '99/IAAI '99), p.518-525.

[21] Madabhushi, A., Aggarwal, J.K., 1999. A Bayesian approach to human activity recognition. Visual Surveillance, Second IEEE Workshop on, (VS'99), p.25-32. [doi:10.1109/VS.1999.780265]

[22] Hoey, J., 2001. Hierarchical unsupervised learning of facial expression categories. Detection and Recognition of Events in Video Proceedings, IEEE Workshop on, p.99-106. [doi:10.1109/EVENT.2001.938872]

[23] Zhou, F., De la Torre, F., 2009. Canonical time warping for alignment of human behavior. *Advances in Neural Information Processing Systems (NIPS)*, **22**:2286-2294.

[24] Aminian-Modarres, A. F., Soryani, M., BPG: A new graph-based posture descriptor for human behavior recognition. *IET Computer Vision*, accepted for publication (acceptance date: 15-Feb-2013). [doi:10.1049/iet-cvi.2012.0121]

[25] Singh, S.,Velastin, S.A., Ragheb, H., 2010.MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods. Advanced Video and Signal Based Surveillance (AVSS), Seventh IEEE International Conference on, p.48-55. [doi:10.1109/AVSS.2010.63]

[26] Weinland, D., Ronfard, R., Boyer, E., 2006. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, **104**(2–3):249-257. [doi:10.1016/j.cviu.2006.07.013]

[27] Sigal, L., Balan, A.O., Black, M.J., 2010. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion.*International Journal of Computer Vision*, **87**(1-2):4-27. [doi:10.1007/s11263-009-0273-6]

[28] Karthikeyan, S., Gaur, U., Manjunath, B.S., Grafton, S., 2011. Probabilistic subspace-based learning of shape dynamics modes for multi-view action recognition. Computer Vision Workshops (ICCV Workshops), IEEE International Conference on, p.1282-1286. [doi:10.1109/ICCVW.2011.6130399]

[29] Weinland, D., Boyer, E., Ronfard, R., 2007. Action Recognition from Arbitrary Views using 3D Exemplars. Computer Vision, ICCV 2007, IEEE 11th International Conference on, p.1-7. [doi:10.1109/ICCV.2007.4408849]

[30] Junejo, I.N., Dexter, E., Laptev, I., Perez, P., 2008. Cross-View Action Recognition from Temporal Self-similarities. Proceedings of the 10th European Conference on Computer Vision: Part II (ECCV '08), p.293-306. [doi:10.1007/978-3-540-88688-4_22]

[31] Vitaladevuni, S.N., Kellokumpu, V., Davis, L.S., 2008.Action recognition using ballistic dynamics. Computer Vision and Pattern Recognition, CVPR 2008, IEEE Conference on, p.1-8. [doi:10.1109/CVPR.2008.4587806]

[32] Weinland, D., Özuysal, M., Fua, P., 2010.Making action recognition robust to occlusions and viewpoint changes. Proceedings of the 11th European conference on computer vision conference on Computer vision: *Part III* (ECCV'10), p.635-648. [doi:10.1007/978-3-642-15558-1_46]

[33] Junejo, I.N., Dexter, E., Laptev, I., Perez, P., 2011. View-Independent Action Recognition from Temporal Self-Similarities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **33**(1):172-185. [doi:10.1109/TPAMI.2010.68]

[34] Jingen, L., Shah, M., Kuipers, B., Savarese, S., 2011.Cross-view action recognition via view knowledge transfer. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, p.3209-3216. [doi:10.1109/CVPR.2011.5995729]

[35] Ning, H., Xu, W., Gong, Y., Huang, T., 2008. Latent Pose Estimator for Continuous Action Recognition. Proceedings of the 10th European Conference on Computer Vision: Part II (ECCV '08), p.419-433. [doi:10.1007/978-3-540-88688-4_31]