RECOGNITION OF HUMAN ACTIVITIES USING LAYERED HIDDEN MARKOV MODELS

Serafeim Perdikis[2], *Dimitrios Tzovaras*[1], *Michael Gerasimos Strintzis*[1,2]

[1] Informatics and Telematics Institute, P.O. Box 361, 57001 Thermi-Thessaloniki, Greece

[2] Information Processing Laboratory, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, 540 06 Thessaloniki, Greece

ABSTRACT

Human activity recognition has been a major goal of research in the field of human - computer interaction. This paper proposes a method which employs a hierarchical structure of Hidden Markov Models (Layered HMMs) in an attempt to exploit inherent characteristics of human action for more efficient recognition. The case study concerns actions of the arms of a seated subject and depends on the assumption of a static office environment. The first layer of HMMs detects short, primitive motions with direct targets, while every upper layer processes the previous layer inference to recognize abstract actions of longer time granularities. The problem of unsupervised learning within the LHMM framework is also addressed, through automatic segmentation of raw data and hierarchical clustering of motion samples. Finally, the idea of context - aware HMM modeling is also introduced and future directions for its application are proposed. The results demonstrate the efficiency, the tolerance on noise interpolation and the high degree of person - invariance of the method.

Index Terms— Human Activity Recognition, Learning Theory and Modeling, HMM, Unsupervised Learning, Context - Aware Systems

1. INTRODUCTION

Automatic Human Activity Recognition (HAR) has received great attention by researchers involved in human - computer interaction, due to the continuous need for smarter and more user - friendly interfaces. HAR implementations presented so far vary widely in terms of the medium of surveillance (e.g. camera, motion tracker), the target of recognition (e.g. indoor or outdoor activity), the human model and the mathematical model.

As far as the mathematical model is concerned, activity recognition methods can generally be classified into those who employ a state - space model (Bayesian Networks [1], Finite State Machines, Hidden Markov Models [2]) and those who rely on pattern recognition techniques (Support Vector Machines, Neural Networks, Dynamic Time Warping, Bayes and K - means classifiers [3]).

State - space models and especially Hidden Markov Models (HMMs) have been preferred in most cases for solving the activity recognition problem, due to their efficiency in capturing spatio - temporal dynamics of signals [4]. In this paper a layered HMM structure (LHMMs) is applied to replace the typical single - layer HMM classifier, thus facilitating the learning and inference procedures. Every upper layer processes the inferential results of the previous one in order to detect actions at a higher level of abstraction and of longer temporal granularities. By decomposing the inherent structure of human activity, the method manages to reduce the training requirements of the HMMs, thus enhancing the efficiency and robustness of the recognition system.

The paper is organized as follows: in Section 2 the basic ideas behind the proposed method are explained. In Section 3 implementation issues are thoroughly discussed. An unsupervised learning technique is described in Section 4 and in Section 5 the idea of context - aware HMMs is introduced. In Section 6 experimental results are presented and commented on.

2. METHOD DESCRIPTION

The key feature of the HMM recognition framework is the property that given a HMM λ, a probability $P(O|\lambda)$ can be assigned to the generation of any observation sequence O. Observation sequences can be denoted $O = O_1 O_2 \dots O_t \dots$, where $O_t = \{Feature_1, Feature_2, \dots\}$ the feature vector at time slot t.

The classical single - layer approach for HMM activity recognition suffers certain limitations. Modeling actions of relatively long duration leads to long observation sequences that burden the training process and reduce the efficiency of the recognition. Besides that, extraction of a large number of activity features (e.g. multi - sensorial environments) augments the training data, encumbering the inference process.

These drawbacks can be overcome by implementing a layered structure of HMMs. Layered representations of state - space models such as Finite State Machines, Bayesian Networks [5] and HMMs [6] have been considered in order to overcome the limitations of the single - layer approaches. Lay-

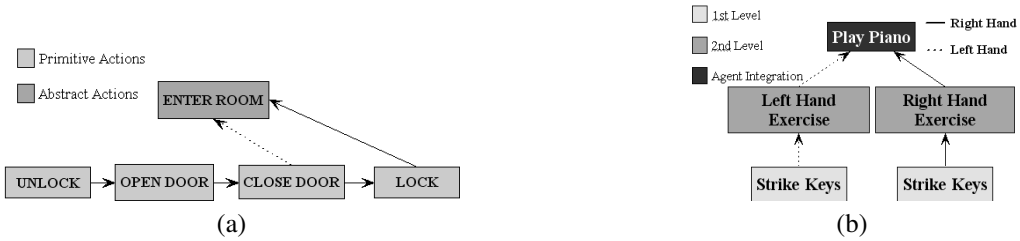


Fig. 1. Strucuture of human activity

ered HMMs (LHMMs) can improve the training process in two fashions. First, the high - complexity processing of low - level data is restricted to the first layer only, permitting the upper layers to process simple discrete input symbols, based on the previous layer inference. Additionally, LHMMs can achieve efficient segmentation of the parameter space, by integrating the inferential results of multiple HMMs in the same layer.

The contribution of this paper lies mainly on the demonstration of the applicability of LHMMs for the Activity Recognition problem, when a person's actions (e.g. putting a stamp), rather than his "state" or "situation" (e.g. phone conversation, [6]) has to be detected. Recognition of abstract actions proves to be a challenging problem, since the order of the series of events is of great importance. In order to achieve this goal, a decomposition of the structure of human action is necessary. Eventually, the application of LHMMs becomes feasible thanks to the innovative idea of exploiting two typical characteristics of the human activity:

Hierarchical and chronological structure of activity
Human actions can be classified into hierarchical levels of abstraction. The lowest level of human activity hierarchy is occupied by simple, short motions with single, direct targets, referred to as Primitive Motions (PMs). Every upper layer contains more abstract motions (AMs), that take place in longer time intervals, accomplish more complicated goals and reveal complex intentions. Actions at some level are composed by a sequence of actions of the previous level. In this manner, actions in successive levels are connected to each other, because every action can be described as the result of the execution of simpler actions at the previous level over some period of time. An example of this structure is shown in Figure 1(a).

Distribution of activity to multiple cooperative agents
Another inherent characteristic of human activity is the execution of composing actions by different motion agents. When a single human is considered, the role of motion agents is played by the human limbs. For instance, walking consists of periodical movements of the two legs. When a whole team is taken into consideration, then every member can be seen as an agent, whose action contributes to the fulfilment of the team's objective. The knowledge about the activity of every single cooperative agent is crucial for a reliable inference about the type of the overall activity. Figure 1(b) presents an example of how the combination of agent inferential results differentiates the final inference.

The above observations inspire a layered structure of HMM model for activity recognition. More specifically, the LHMMs recognition method is based on the following ideas: A set of N motion agents $A = \{A_1, A_2, \ldots, A_N\}$ is defined for the activity in question. A set of M_i Primitive Motions is defined for every agent A_i $(1^{st}level)$: $PM^{A_i} = \{PM_1^{A_i}, PM_2^{A_i}, \ldots, PM_{M_i}^{A_i}\}$. A set of R_i Abstract Motions is defined for every agent A_i $(2^{nd}level)$: $AM^{A_i} = \{AM_1^{A_i}, AM_2^{A_i}, \ldots, AM_{R_i}^{A_i}\}$. More layers can be added as the level of abstraction of the described actions increases.

For every layer L of an agent A_i, a bank of HMMs is assigned performing a mapping of the layer's observation sequences O_L to the actions X^L contained in this layer: $f_L : O^L \rightarrow X^L$. For the first layer, the observation sequences O^1 are sequences of feature vectors extracted by the raw input data, while the actions X^1 belong to the set PM^{A_i}. The mapping procedure f_L at every layer L implements the classical HMM recognition framework. For the second and every upper layer, the observation sequences consist of the inferential results of the previous layer over some period of time. Thus, successive outputs of some layer form the (discrete) input vectors of the next one.

At some level an integration procedure takes place, so that the overall activity can be inferred by the partial inferential results of every single agent alone. The agent integration process concerns the detection of meaningful, simultaneous, and cooperative actions among the defined activity agents. Figure 2 depicts graphically the proposed method.

The advantages emerging by the application of the method include: a) the restriction of continuous observation sequences, that require laborious processing, to short sequences at the 1st layer only, through the introduction of levels of abstractions, and b) the segmentation of long feature vectors to multiple shorter ones thanks to the introduction of multiple motion agents.

3. ACTIVITY RECOGNITION IN OFFICE ENVIRONMENT

The functionality of the proposed method has been tested under a simple implementation scheme containing two layers.

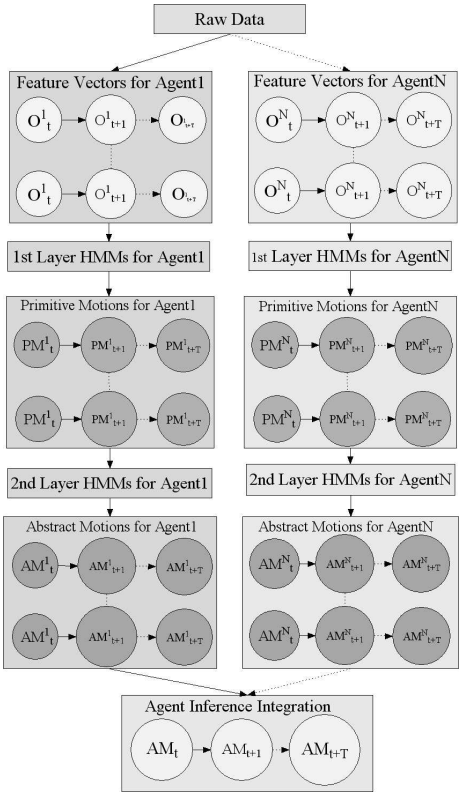


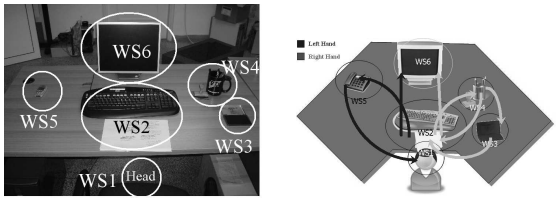
Fig. 2. Layered HMM method block diagram

The target of recognition concerns actions of the arms in an office environment and was limited to the following actions: Pick Up Phone, Adjust Screen, Switch Screen On/Off, Take Pen and Put Stamp.

The implementation of the method relies on the assumption of a static office environment, where the positions of all objects on the desk and the subject's seat are relatively fixed. With regard to the analysis in Section 2, two cooperative agents are defined, namely the two arms of the subject, denoted LA and RA for the left and right arm respectively. The static office environment is divided into 6 workspaces $WS_i, i = 1, 2, \ldots, 6$ as shown in Figure 3(a). Workspaces can be viewed as the surrounding space of one or more objects.

The reason for introducing the static environment and the workspace definition is, that this scheme enables the bounding of the PM set for both agents to transitions between two workspaces. Formally, PMs can be denoted $LAWS_iWS_j$ or $RAWS_iWS_j, i \neq j$ respectively. Finally, 8 PMs have been defined for the first level of abstraction, 3 for the left and 5 for the right arm. In Figure 3(b), PM transitions are represented as arrows in the static environment.

According to the method description, every AM of the second level is formed by a sequence of PMs of the previous level, following the natural structure of human activity. With respect to that, the final form of the implementation scheme



(a) (b)

Fig. 3. Workspaces in static office environment and PM definition

is presented in Figure 4. It is important to underline that the distinction of the actions Adjust Screen and Switch Screen On/Off can only be achieved after the agent integration procedure dictated by the method's formulation.

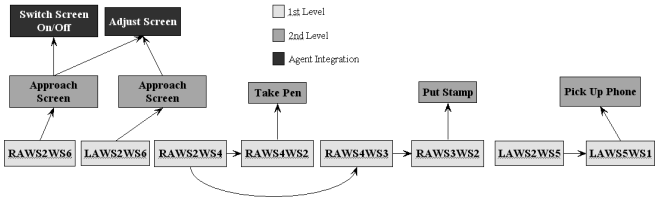


Fig. 4. Implementation scheme diagram

The subject's arms are modeled with two distinctive body spots, the wrist and the elbow (Figure 5(a)). The trajectories of these spots are captured by a wearable magnetic motion tracker (Ascension MotionStar©, Figure 5(b)). Consequently, the raw data produced at every single time - slot t for both agents contains 3D positions of the associated body spots: $R_t = \{x_W^t, y_W^t, z_W^t, x_E^t, y_E^t, z_E^t\}$, where W stands for wrist and E for elbow. The motion features extracted are the 3D position and the vectorial velocity, so a feature vector at time slot t can be denoted: $O_t = \{x_W^t, y_W^t, z_W^t, x_E^t, y_E^t, z_E^t, V_{x_W}^t, V_{y_W}^t, V_{z_W}^t, V_{x_E}^t, V_{y_E}^t, V_{z_E}^t\}$.

(a) (b)

Fig. 5. Arm model and wearable motion tracker

The bank of HMMs of the first layer was trained with 100 PM samples taken by 5 different subjects using the Baum - Welch parameter estimation algorithm [7]. Instead of a single HMM per PM, 5 HMMs per PM are trained, in order to capture variations on the execution time of the actions. In the inference phase, the first layer of every agent emits every 150 msec a discrete symbol associated to the PM taking

place at that time. The final inference procedure combines partial, classical HMM - based inferences of the 5 HMMs associated to every PM, in a simple voting process over 10 time slots (150 msec). Furthermore, raw data input undergoes an Euclidean distance segmentation process before being fed to the first layer HMMs, so that only segments of the testing sequence where motion has been detected are taken into consideration. The segmentation procedure increases the speed of inference and eliminates false alarm errors.

The concatenation of the first layer inference symbols over longer periods of time, form the observation sequences of the second layer HMMs. It should be noted that in case of an absolutely accurate inference at the first layer, simple Finite State Machines instead of HMMs could be used at the second layer to detect the desirable sequence of PMs that form an AM. In fact, the first layer's inference proved to be prone to wrong decisions over short time intervals. For this reason, the symbols emitted by the first layer are treated as observation symbols of the "hidden" state, which represents the actual PM currently executed. Accordingly, second layer HMMs have been trained in a heuristic manner using direct specification of the HMM parameters, so that non - zero probabilities are attributed to first - layer inference sequences, either perfect or containing minor mistakes, and zero probability otherwise. Thus, significant enhancement in the system's robustness is achieved, through the "correction" of first layer inferential results.

4. UNSUPERVISED LEARNING IMPLEMENTATION SCHEME

The supervised learning fashion dominates state of the art HMM modeling, but it is responsible for certain limitations of the according implementations. More particularly, the system's autonomy and adaptiveness is limited by its inability to retrain itself when changes in the recognition environment occur. Besides that, human intervention in the segmentation process is known to be error - prone. Inaccurate segmentation leads to non - representative training sequences that harm the recognition process by misleading the model's parameter estimation algorithm.

Supervised learning has been usually imposed by the fact that there exists no explicit way for a machine to realize which parts of a raw data sequence are of interest. In addition to that, automatically extracting important segments with high accuracy has also proved to be a non - trivial problem in most applications. In the following, an unsupervised learning technique applicable in the LHMM recognition framework of Section 3 is presented, addressing both of the above issues efficiently.

The proposed method for unsupervised learning consists of two distinct phases. In the first pre - processing phase, an automatic segmentation algorithm is employed to detect all primitive motions in a given raw data sequence. In the second phase, extracted motion samples are grouped to form the HMM training sets using hierarchical cluster analysis.

Automatic segmentation is achieved thanks to the observation that any human motion can be represented by a major fluctuation in the moving limb's speed diagram. Therefore, the main concept of the segmentation algorithm is the detection of major fluctuations in the speed diagram extracted by the raw data (motion trajectories). The main steps of the segmentation process are shown in Figure 6.

In the motion clustering phase, every extracted motion is labeled by a vector V containing coordinates of the first and last point of the respective motion trajectory: $V = \{x_A, y_A, z_A, x_B, y_B, z_B\}$. The resulting vectors are clustered using the Ward's linkage clustering method. This method proved to be the most efficient among the agglomerative methods for hierarchical clustering that were tested (single-,average- and complete-linkage). Ward's linkage minimizes at every step the increase in the total within - cluster error sum of squares (ESS) as a result of joining two clusters. For a set X the ESS is described by:

$$ESS(X) = \sum_{i=1}^{N_X} |x_i - \frac{1}{N_X} \sum_{j=1}^{N_X} x_j|^2$$

The linkage function is described by the expression:

$$D(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)]$$

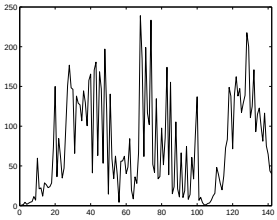
Termination of the algorithm is determined setting a minimum cluster distance, above which fusion of clusters stops and a final number of clusters is obtained.

Application of an hierarchical clustering method for this problem is imposed by the fact that the number of clusters is initially unknown. Assuming that meaningful motions in a long activity sequence are frequently repeated, corresponding clusters are expected to contain a large number of motion samples. The size of the final clusters could be a reliable criterion for distinguishing meaningful and random motions. Selection of target - motions from the group of meaningful motions and annotation of the training sets remain the only jobs to be inevitably done manually.

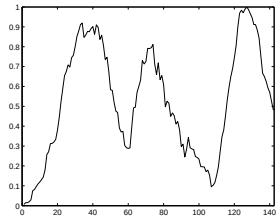
Multiple advantages arise from the application of the proposed unsupervised learning scheme. Firstly, the system is able to capture the inherent structure of human motion by detecting the PMs in a sequence, thus providing useful assistance during the designing phase of the system. Furthermore, the static workspace limitation can be suspended due to the capability of the system to retrain itself when objects move around in the workspace. Finally, automatic segmentation guarantees enhanced reliability as far as integrity of the resulting training sequences is concerned.

5. CONTEXT - AWARE MODELING

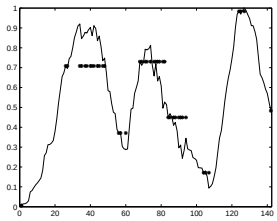
The idea of context - aware modeling refers to the capability of a system to switch among a number of kindred models de-



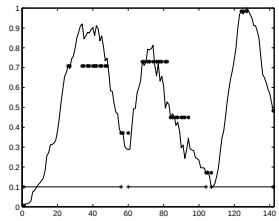
(a) Normalized speed



(b) Normalized speed after smooth filtering



(c) Detection and quantization of extrema



(d) Extraction of segments

Fig. 6. Extraction of the three PM segments contained in a sample of the AM "Put Stamp"

pending on certain predefined conditions. In other words, it is assumed that any procedure can be modeled more efficiently, if different models are adopted for capturing all differentiations that can occur.

The application of the above idea into the HMM activity recognition framework seems promising, as many variables of the problem can be thought to have certain effect on the final recognition rate. These variables mainly concern biometric features of the acting subject. For instance, body proportions (e.g. height, length of arms) can largely affect the way in which a subject moves around in a static environment. Different execution times can also be expected, as individuals tend to adopt a different average speed in everyday activities.

The functionality of a context - aware system depends on the selection of proper differentiating variables, which prerequisites deep knowledge of the procedure's inherent characteristics and on the building of an efficient model - selection method based on the integration of those variables' values.

The model selection method can be formulated as a MAP (Maximum a posteriori) decision rule according to the Bayes decision theory. For every modeled motion a set C of corresponding HMMs is trained. Model selection is conditional on the predefined set of N differentiating variables: $DV = \{DV_1, DV_2, ..., DV_N\}$. According to MAP decision rule, every time a motion segment must be classified, the most probable HMM C_i of every motion's bucket is selected:

$$\hat{C} = argmax_i(P(C_i|DV_1, DV_2, ..., DV_N))$$

Using Baye's theorem it holds that:

$$P(C|DV_1, DV_2, ..., DV_N) = \frac{P(C)P(DV_1, DV_2, ..., DV_N|C)}{P(DV_1, DV_2, ..., DV_N)}$$

The denominator is independent of C, so the decision rule can be rewritten as:

$$\hat{C}_i = argmax_i(P(C_i)P(DV_1, DV_2, ..., DV_N|C_i))$$

Assuming that each differentiating variable is conditionally independent of every other ($P(DV_i|C, DV_j) = P(DV_i|C), i \neq j$), the final form of the decision rule would be:

$$\hat{C} = argmax_i(P(C_i)\prod_{j=1}^{N} P(DV_j|C_i))$$

An HMM class C_i is trained with motion samples taken from a specific individual. In other words, HMM classes C_i are assigned to specific individuals that act like prototypes. It is obvious that subjects for training should be carefully chosen, so as to represent adequately the possible combinations of differentiating variables that can occur in the general population. Both HMM priors and conditional variables' distributions can be approximated with relative frequencies from the training set. Let $N(C_i)$ the number of training sequences of HMM class C_i and N_C the total number of training sequences of all HMMs of a motion. Prior probabilities $P(C_i)$ can be calculated as: $P(C_i) = \frac{N(C_i)}{N_C}$. If equal number of training sequences are taken from each subject, then all prior probabilities are equal. In a similar manner, conditional distributions can be estimated by: $P(DV_j|C_i) = \frac{N(DV_j, C_i)}{N(C_i)}$, where $N(DV_j, C_i)$ the number of training sequences in class C_i characterized by a specific value for variable DV_j. Appropriate quantization is necessary in case of continuous differentiating variables.

Future work of the authors will concentrate on extending the LHMM recognition framework with context awareness.

The influence of various biometric features on the final recognition rate will be explored.

6. RESULTS AND CONCLUSIONS

Table 1 presents the results acquired by testing the classifier described in Section 3 with 5 subjects (different from those used for training) performing uninstructed 5 repetitions of the target actions. Recognition rates were noted over the 25 samples of each action. The results demonstrate a recognition rate over 80% for all the actions in question.

In the previous case, measurements are assumed to be noiseless due to the accuracy of the magnetic tracker. However, noisy measurements are expected in applications where motion capture is performed using computer vision techniques. For this reason, the effects of Gaussian noise interpolation on the raw data were investigated, revealing the immunity of the LHMM system to noise with SNR > 10 (Figure 7).

Table 1. Recognition rates

Abstract Motion	Recognition rate
Pick Up Phone	100%
Adjust Screen	100%
Take Pen	80%
Put Stamp	92%

The unsupervised learning technique of Section 4 was tested with a long sequence containing 25 repetitions of each AM. The segmentation algorithm achieved a total 97,5% success rate. The clustering algorithm managed to put in the right clusters all the correctly segmented PMs, creating one training set for each PM. A few almost empty clusters were also created containing falsely segmented sequences, thus demonstrating the effectiveness of the "cluster size criterion" for detecting meaningful primitive motions. A context - aware LHMM classifier was also tested depending only on the motion execution time as a conditional feature. Preliminary results show a satisfactory overall recognition rate, revealing a perspective on important amelioration when more biometric features are taken into consideration.

Concluding, the presented implementations demonstrate that LHMMs can be successfully employed for the recognition of human actions, achieving more efficient training, reliable inference and improvement of the system's robustness. Additionally, training and testing of the classifier with different persons reveals a certain person - invariance of the classifier. Finally, the possibility of performing unsupervised learning with such systems and the perspective of enhancing the LHMM structure with context awareness are adequately illustrated.

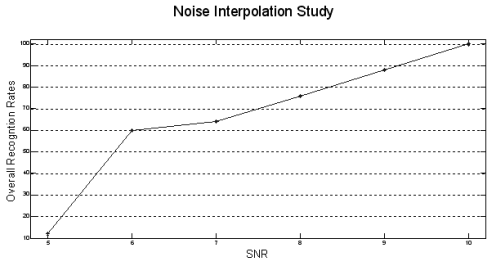


Fig. 7. Performance of the proposed method in the presence of measurements noise

7. REFERENCES

[1] P.E. Rybski and M.M. Veloso, "Robust real-time human activity recognition from tracked face displacements," in *EPIA '05, 12th Portuguese Conference on Artificial Intelligence*, 2005.

[2] Feng Niu and Mohamed Abdel-Mottaleb, "View-invariant human activity recognition based on shape and motion features," in *ISMSE '04: Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04)*, Washington, DC, USA, 2004, pp. 546–556, IEEE Computer Society.

[3] P.C. Ribeiro and J. Santos Victor, "Human activity recognition form video: modelling, feature selection and classification architecture," in *HAREM '05, International Workshop on Human Activity Recognition and Modelling*, 2005.

[4] Daiki Kawanaka, Takayuki Okatani, and Koichiro Deguchi, "HHMM based recognition of human activity," *IEICE - Trans. Inf. Syst.*, vol. E89-D, no. 7, pp. 2180–2185, 2006.

[5] K. M. Kitani, Y. Sato, and A. Sugimoto, "Deleted interpolation using a hierarchical Bayesian grammar network for recognizing human activity," in *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, Washington, DC, USA, 2005, pp. 239–246, IEEE Computer Society.

[6] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *ICMI '02: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, Washington, DC, USA, 2002, p. 3, IEEE Computer Society.

[7] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.