Evidence Feed Forward Hidden Markov Models for Visual Human Action Classification

MICHAEL S. DEL ROSE
U.S. Army Tank Automotive Research, Development, and Engineering Center
Warren, MI 48309-5000, U.S.A.
Email: mike.delrose@us.army.mil

CHRISTIAN C. WAGNER
Department of Industrial and Systems Engineering, Oakland University
Rochester Hills, MI 48309, U.S.A.

Abstract

Predictions of peoples actions based on visual data is a fairly easy job for people, harder job for animals, and virtually impossible for machines, although many classification systems can predict a limited number of actions. This is due to the many different movements people make while performing the action. Take, for example, a visit to the local store. If we were to sit and watch people walk up and down isles, we would see a unique style of movement from each person. There may be close similarities, but the actual position of the body parts in relation to time would all be unique. People tend to merge these together and look at the overall movement, focusing on only one thing at a time, making an assumption, and validating the assumption. Animals do the same thing but with less a priori knowledge, or less understanding, of the movements. Algorithms that are written for classification of human movement often look at the specific details of movements. It is much harder to generalize an algorithm while testing it on a procedural machine.

A new type of Hidden Markov Models (HMM) is developed to help generalize the movements of people. These HMMs, called Evidence Feed Forward HMMs (EFF-HMM), add a link to the normal HMM through the observations. This seems to contradict the laws of causality strictly defined in standard HMMs. However, as will be seen below, looking at the probabilities associated with the observation to observation linkages, no rules are broken.

Keywords: action recognition, artificial intelligence, hidden markov model, visual human motion classification.

1. Introduction

Automating the ability of a machine to predict the actions of a human is an import mission for several technologies, like online search algorithms, media archival databases, autonomous driver systems, and security surveillance systems to name a few. There are a large number of active research projects in algorithm development for action recognition systems, however there is still a long way to go before commercialization can happen.

The EFF-HMM is one such research area that has showed promise in overcoming many of the faults of current research. It is designed to be robust with the inability to distinguish a person moving right to left or left to right. It looks at the details of the movements and their relation to other movements.

This paper is sectioned into several sections to include Introduction (section I), Related Research

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **12 APR 2011**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **Evidence Feed Forward Hidden Markov Models for Visual Human Action Classification (PREPRINT)**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Michael S. Del rose’ Christian C. Wagner**	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000, USA Department of Industrial and Systems Engineering, Oakland University Rochester Hills, MI 48309, USA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000, USA**	10. SPONSOR/MONITOR’S ACRONYM(S) **TACOM/TARDEC/RDECOM** 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
Submitted for publication in Journal of Artificial Intelligence

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**11**	

(section II), EFF-HMM Theory (section III), Example (section IV), and finally Conclusion (section V). The related Research section will briefly describe some current algorithms that try to tackle this problem. The EFF-HMM Theory section will explain the EFF-HMM theory with an example. It will also answer the three problems all HMMs should answer. The Example section will apply the EFF-HMM to a well known data set. Finally, the conclusion section will wrap up the discussion.

2. Related Research

There are several research efforts involved in human action recognition that provide a wide variety of techniques, from standard statistical approaches to computational intelligent designs. A large amount of the research uses visual cues of human actions without any traditional artificial or computational intelligent techniques. These algorithms rely on simplicity at the cost of fusing input data. They often use less than typical data inputs; that is, inputs that would not necessarily be used by human observers. They rely almost exclusively on the pre-processing of the data while using statistical or non-traditional artificial and computational intelligent algorithms to determine the behavior. Many research projects in this area use their own form of plotting space-time data from the image sequences and calculating the closest distance to pre-determined or automatically determined events to decide what action the human is performing. M. Dimitrijevic et al. [1] developed a template database of actions based on five male and three female people. Each human action is represented by three frames of their 2D silhouette: the frame when the person first touches the ground with one of his/her feet, the frame at the midstride of the step, and the end frame when the person finishes touching the ground with the same foot. The three frame sets were taken from seven camera positions. For classifying the action, they use a modified Chamfer's distance calculation to match to the template sequences in the database.

D. Weiland et al. [2] use motion history volumes to determine human gestures by extending the 2D pixel representation with time to a 3D representation with time. This is accomplished by using multiple cameras around the person and subtracting out any background information. Classes are created manually for each action or gesture. Mahalanobis distance with principle component analysis is used to identify action from the appropriate class.

Some of the traditional artificial and computational intelligence techniques are used for classifying human action, many with a spin towards specific motions. J.-Y. Yang et al. [3] uses neural networks to determine human actions. They reduce the errors associated with normal human motion capturing by placing tags on body parts for tracking. They also strap a tri-axial accelerometer to the subject's wrist to monitor three degrees of motion on the specific body part. Tri-axial data is captured at pre-determined time intervals. This data is the input into a neural network specifically designed to determine if the event is static, like standing or sitting, or the event is dynamic, like walking and running. Once the event is determined to be either static or dynamic, another neural network is used on the same data, either a static event neural network or a dynamic event neural network, and the action is classified. The results are promising for the limited actions the system is designed to detect: standing, sitting, walking, running, vacuuming, scrubbing, brushing teeth, and working on a computer.

Rule based and fuzzy systems are a few other common types of artificial and computational intelligent technique used to identify patterns and have been adapted to analyze human events. H. Stern et al. [4] created a prototype fuzzy system for picture understanding of surveillance cameras. His model is split into three parts, pre-processing module, a static object fuzzy system module, and a dynamic temporal fuzzy system module. The static fuzzy system module takes in the pre-processed data and outputs the number of people involved in the scene: a single person, two people, three people, many people, or no people. The dynamic fuzzy system determines the intent of the person, or people, based on their global temporal movements. Although this requires only a basic understanding of human intent by using global movements of people and their interactions based on global positions, it is included in many application research programs within the U.S. Department of Defense: Near Autonomous Unmanned System [5], Army Research Lab Collaborative Technology Alliance [6], and Mobile Detection Assessment and Response System [7].

Of all the visual human action recognition networks constructed, HMMs, or some variant to standard

HMMs, are the most widely used. HMMs keep a network of body poses related to each other and provided a way of learning parameters that best fit a set of training data with known classifications. Campbell, Becker, and Azarbayejani [8] used HMMs to recognize eighteen Tai Chi moves. Each move was represented by a series of vectors formed by the 3D position of the head and the hands. Yu and Ballard [9] use HMMs to distinguish similar action based on head and eye movements. Gehrig and Schulz [10] used HMMs to recognize ten kitchen actions based on the movement of twenty four points on the upper body. They looked at skeletal data and calculated the correct movements of people and reduce the number of body parts down to thirteen with similar results.

Wilson and Bobick [11] use a Parametric Hidden Markov Model (PHMM) to recognize gestures. The PHMM has an additional parameter used to represent meaningful variations of gestures across the set of all gestures. This gives PHMMs the ability to distinguish between gesture meanings with similar hand movements.

Oliver et al. [12] developed a real time system that detects and classifies interactions between people using a Coupled Hidden Markov Model (CHMM). They used synthetic environments to model person to person interactions and thus creating their CHMM. Data from a static camera was used and moving objects were segmented and tracked. Data describing the location, heading, and relative location to other people were inputted into the synthetically created CHMM for analysis and classification of the interaction type. Results show they outperformed standard HMMs. This is not a far stretch since standard HMMs work on single automatons where CHMMs work on coupled automatons, thus HMMs cannot outperform CHMMs in this particular environment.

Multi-Observation Hidden Markov Models (MOHMM) are discussed in both [13] and [14] from Xaing and Gong. In [13] they use MOHMMs to create breakpoints in the video content of an activity. Blobs above a certain threshold in each frame are segmented from the pixel change history. Several functions of these blobs are used in the feature vector to classify the video with the MOHMM. In [14] an MOHMM was used to detect piggybacking of people off someone else's security card to open a secured, card access only door. Piggybacking is when someone follows another person through a security door without using his/her security card to open it. The framework of the system allowed for continual changes based on changes in peoples' movements, thus unsupervised learning is used to continually update the model.

Continuous HMMs (cHMMs) are used in the work of Antonakaki et al. [15]. Their work classifies abnormal behavior of people based on both their short term behavior and the global trajectory of each subject. A short term behavior is a behavior that can be classified in twenty five frames, or one second. A one class support vector machine (SVM) is used to distinguish abnormal behavior from the short term behavior sequence. For trajectory data, a one class cHMM is used to determine if the person's movement is abnormal. Both are used to determine the final results.

Layered Hidden Markov Models (LHMMs) are used in Oliver et al. [16] to detect specific activities in an office environment. They employ a two level cascade of HMMs with three processing layers. The first layer captures video, audio and keyboard/mouse activity to create the first level feature vector. The middle layer has two HMMs, one for creating an audio feature vector and one for creating a video feature vector. The top layer uses the results of these HMMs along with keyboard/mouse activity and the derivative of the sound localization component as the final feature vector. The results from this top layer determine the activity in the office. They claim the LHMM makes it feasible to decouple different levels of analysis for training and inferences. By using a single HMM it would need a large parameter space, thus need a large amount of data to train. Also, a single HMM would not be robust enough to move to a different office without retraining, unlike the LHMM claims.

3. EFF-HMM Theory

The EFF-HMM was designed to better classify visual actions of people based on their movements. This is accomplished by adding a linkage through the observations in a standard HMM. The EFF-HMM is more than an extension to standard HMMs, like Parametric HMMs or Hierarchical HMMs, because it associates the previous observations and previous state to the current observation in the sequence by assigning a probability associated with this and integrating it into the state-observation probability described in standard HMMs. This association changes the thinking of

HMMs by describing the HMM process in terms of both the hidden state and the causes for the observation. HMMs are widely used throughout the machine learning community in the development of classification systems. However, HMMs require the assumption of independence, the understanding that events are causal, and the requirements that only hidden nodes are affected by previous hidden nodes. The assumption of causality is relatively true in real world, but if you look at the reason for making the model is to “model the event” not create real world, then we can relax the causality rule, at least how we look at causality in modeling with HMMs. It is not the intention of EFF-HMMs to relax the causality rule. Only the way the HMMs look at causality. For example, suppose it is desired to model the famous weather example using EFF-HMMs. That is, if we have knowledge of only the observations of a single person entering a building with or without an umbrella, can we predict what the weather is outside? Figure 1 shows an HMM developed for this problem.

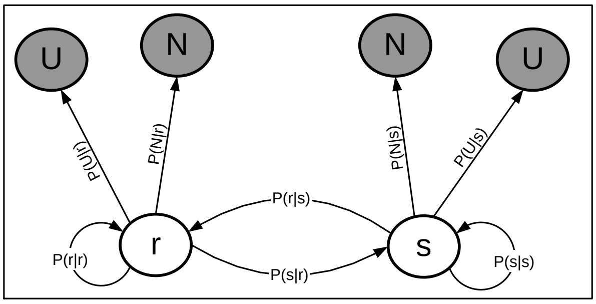


Fig. 1. Weather example using HMMs. The nodes (states) in the hidden layer show the weather and are represented by Rainy (r) and Sunny (s). The states in the observation layer represent seeing an Umbrella (U) or No umbrella (N). Probabilities from state to state are shown.

The HMM accounts for the person carrying the umbrella based on the rain (like P(U|r), however it does not take into account the probability associated with the actions of the single person yesterday along with yesterdays weather and how that effects the decision to carry (or not carry) an umbrella today. Figure 2 shows the diagram of an EFF-HMM on this same problem. As a note, the linkages between observations are only showed as bi-directional for clarity and do not infer the associated probability to be the same in both directions.

Suppose that the person carrying an umbrella did not carry one the previous day, but it rained. Wouldn’t this add more weight to the probability of carrying an umbrella today (provided the person did not like getting wet) thus increasing the probability whether it rained or not? The answer should be yes. The observations and hidden state one the previous day should have an effect on the current observation, thus having an effect on the probability of the current state with the current observation. This is what the EFF-HMM provides; a way to account for the observation to observation linkages. Notice that causality with respect to the observations are looked at differently than represented in a standard HMM model. The underlying reason the observations have an effect on the next observation is based on the event by the single person carrying (or not carrying) an umbrella and not looked at as just an observation. Thus, causality is still adhered to.

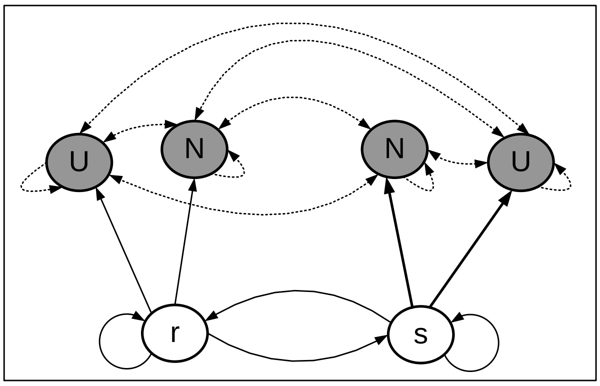


Fig. 2. Weather example using Evidence Feed Forward HMMs. Linkages between the evidence (marked U and N) and the other evidences, shown with dashed lines, represent the probability of the evidence occurring based on the previous evidence and previous weather state.

As another example that relates to human action recognition, suppose you have several video clips of people pitching in baseball. There are both left handed pitchers and right handed pitches in this data set. Now, since the EFF-HMM has a the observation to observation linkages, then the person pitching right handed will have the same movements (as far as the EFF-HMM is concerned) as the person pitching left handed. Arm movement changes pretty much the same, only in the opposite direction. EFF-HMMs, in theory, should not distinguish between the two and be able to classify both as pitching.

3.1. *Evidence Feed Forward Hidden Markov Model Derivation*

To define the variables, let N be the total number of hidden nodes (or states) and M be the total number of observations. Let T be the total number of transitions (or time). Let Q_t be the current hidden state at time t, $1 \leq t \leq T$; $Q_t = S_i$ means the state at time t is S_i where $1 \leq i \leq N$. Often the hidden state, S_i, is represented only as i for brevity. The observations at time t are represented by O_t. $O_t = V_h$ means the observation at time t is V_h where $1 \leq h \leq M$. Again, like the hidden state, the observation state is often represented only as h.

The probability of transitioning from one hidden state to another is captured by a_{ij} (or $a_{i,j}$), where i is the current state and j is the next state, $1 \leq i,j \leq N$. Let A be an NxN matrix which captures all the a_{ij}'s.

Let b_{jh} (of $b_{j,h}$) be the probability of observing V_h while in state S_j. Let B be an NxM matrix which holds the observation probabilities.

Let $c_i(h,k)$ be the probability contribution of being in state S_i and currently observing V_h with the next observation being V_k for $1 \leq h,k \leq M$. In this case, probability contribution is the amount of contribution to the b term that this probability will provide. Let C be the MxMxN matrix holding the $c_i(h,k)$ terms. Since this is a probability contribution, there must be a normalizing factor so all probability cases sum to one. The C terms, which will be shown later, have the constraints of $\sum_{k=1}^{M} c_i(h,k) = 1, for\ all\ i\ and\ h.$ For this to contribute to the B term and not throw off the probabilities, a normalizing term that includes both B and C terms must be incorporated. This term will be referred as the $Sum_h(i,j)$ and is computed as:

$$Sum_h(i,j) = \sum_{k=1}^{M} [P(O_t = k|q_t = j) * P(O_t = k|O_{t-1} = h \land \mathrm{q}_{t-1} = \mathrm{i})]\ for\ each\ i,j, and\ h$$

Let the probabilities of starting in state S_i be represented as π_i. π is a 1xN vector which holds all the π_i values. $\pi = [\pi_1 \quad \pi_2 \quad \pi_3]$ for a three hidden node Evidence Feed Forward HMM.

Similar to the standard HMM, the Evidence Feed Forward HMM is represented by λ but with an added parameter. λ has the parameters A, B, C, and π and can be written as λ=(π,A,B,C). Either λ or (π,A,B,C) can represent the EFF-HMM. Given the model λ at time t in the current state S_i observing V_h, what is the probability of being in state S_j observing V_k? The calculation of transitioning from state i to state j is a_{ij}, observing V_h from state j is b_{jh}, and being in state i observing V_h with next observation being V_k is $c_i(h,k)$. The normalizing factor is $Sum_h(i,j)$. This probability is $a_{ij} \cdot [b_{jk} \cdot c_i(h,k)]/Sum_h(i,j)$ for t >1 and $\pi_i \cdot b_{ih}$ for t = 1.

There are three typical problems that all HMMs should solve and thus the EFF-HMM should also solve:

1. What is the probability of the observation sequence O = (O1,O2,…,OT) given the model λ? This is asking to find P(O|λ).
2. What is the most optimal hidden state path given the observations O and the model λ? This is asking to solve P(Q|O,λ) where Q = (Q1,Q2,…,QT).
3. Given a number of observations, what are the best parameters of λ which maximizes P(O|λ)? This is the learning problem.

To solve the first problem, "What is the probability of the observation sequence $O = (O_1,O_2,\ldots,O_T)$ given the model λ?", a forward algorithm procedure is developed to compute $\alpha_i(t) = p(O_1,O_2,\ldots O_t,Q_t = i|\lambda)$. When t = T, P(O|λ) is found by summing all the α_i's at time T. The forward algorithm procedure is:

1. $\alpha_i(1) = \pi_i b_i(O_1)$ for all i, $0 \leq i, t \leq T$, and $b_i(O_1) = b_{ih}$ for some h which $O_1 = V_h$. Notice that there is no c term contribution because this is the initial starting state calculation for α_i so there is no observation to observation in the calculation of the initial probabilities.
2. $\alpha_j(t+1) = \sum_{i=1}^{N} \alpha_i(t) a_{ij} \left[\left(b_j(O_{t+1}) * c_i(O_t, O_{t+1}) \right) / Sum_h(i,j) \right]$, where $c_i(O_t,O_{t+1})$ is $c_i(h,k)$ for $O_t = V_h$ and $O_{t+1} = V_k$, $Sum_h(i,j)$ is described in eq. 1, and N is the total number of hidden states.
3. $P(O|\lambda) = \sum_{i=1}^{N} \alpha_i(T)$.

From the final part of step 2 where $\alpha_i(T) = P(O_1,O_2,\ldots,O_T,Q_T = i|\lambda)$, we find the probability of the observation sequence and the final state Q_T = i given the model. By summing up all the α_i's we get the final probability of P(O|λ) shown in step three.

A backwards algorithm procedure can also be developed to find P(O|λ). In the backwards algorithm, the starting state is at time T and the algorithm is worked backwards towards time 1. The variable β_i is defined as $\beta_i(t) = P(O_{t+1},O_{t+2},\ldots,O_T \mid Q_t = i, \lambda)$.

1. $\beta_i(T) = 1$
2. $\beta_i(t) = \sum_{j=1}^{N} a_{ij}\left(\left(b_j(O_{t+1}) * c_i(O_t, O_{t+1})\right) / Sum_h(i,j)\right) * \beta_j(t+1)$
3. $P(O|\lambda) = \sum_{i=1}^{N} \beta_i(1)\,\pi_i b_i(O_1)$.

The second common HMM problem which the Evidence Feed Forward HMM should solve is computing the optimal path of hidden states from the observations given the model. Optimal path could mean many different things. One way to look at optimal paths is to find the best probability moving from one state/observation to another, considering only the path which is maximized for only the current transition. This has the possibility of transitioning to a state in which leaving from is impossible. Here, finding the optimal path is considered the same as finding the best probability for the entire series and not individual maximums. By optimal path it is assumed that one is looking for the path that gives the best probability of the state sequence given the observations and the model; maximizing P(Q|O,λ). This solution requires the use of both the forward and backwards algorithms. To accomplish this, two new variables are created: δ and PATH. δ is defined as the running probability of paths at time t. PATH is the current path found from computing δ.

1. $\delta_1(i) = \pi_i b_i(O_1)$. PATH = [].
2. $\delta_t(j) = \max_{1\le i\le N}\left[\delta_{t-1} a_{ij}\left[\left(b_j(O_t) * c_i(O_{t-1}, O_t)\right) / Sum_h(i,j)\right]\right]$. The state for which this is maximized is added to PATH.
3. Final step is finding the state which maximizes $\delta_T(i)$ for $1 \le i \le N$. Include this to PATH.

The first step of calculating the δ value is to assign each δ value the probability of starting in each of the states. The recursion step continues to keep the best value throughout the model. The final step finds the final hidden node which is the maximum of all the δ_i's at time T.

The third and final problem that the Evidence Feed Forward HMM should be able to solve is the learning problem. To learn, assume there are a number of observations with known results to train on. These observations are used to calculate new parameters that maximize the probability of the observations given the model, P(O|λ). To do this, re-estimation of the parameters for the EFF-HMM must increase P(O| λ) where:

$$\bar{\pi}_i = expected\ times\ in\ state\ i\ at\ time\ t = 1$$

$$\bar{a}_{ij} = \frac{expected\ transitions\ from\ state\ i\ to\ state\ j}{expected\ transitions\ from\ state\ i}$$

$$\bar{b}_{jh} = \frac{expected\ times\ in\ state\ j\ observing\ V_h}{expected\ times\ in\ state\ j}$$

$$\bar{c}_i(h,k) = \frac{expected\ times\ in\ state\ i\ observing\ V_h\ at\ time\ t\ and\ observing\ V_k at\ time\ t+1\ for\ all\ t}{expected\ times\ in\ state\ i\ observing\ V_h}$$

It should be noted that $\bar{c}_i(h,k)$ contributes to the probability of being in a hidden state and observing V_k, that is $\bar{b}_{jk}$. It is treated separately by optimizing the probability associated with it given the constraints to be discussed. The final probability associated with $\bar{c}_i(h,k)$ after optimization from the learning process is used to increase or decrease the value of $\bar{b}_{jk}$. It provides a way of including outside influences based on the observations to the HMM.

First, define the variable $\gamma_i(t)$ to be the probability of being in state i at time t for the sequence of observations O and the model λ. That is:

$$\gamma_i(t) = P(Q_t = i|O,\lambda),$$
$$\gamma_i(t) = \frac{P(Q_t = i, O|\lambda)}{P(O|\lambda)},$$
$$\gamma_i(t) = \frac{P(Q_t = i, O|\lambda)}{\sum_{j=1}^{N} P(Q_t = j, O|\lambda)} \qquad (1)$$

The variable $\alpha_i(t)$ was defined in the forward algorithm as $\alpha_i(t) = P(O_1,O_2,\ldots,O_t, Q_t = i|\lambda)$ and $\beta_i(t)$ was defined in the backwards algorithm as $\beta_i(t)$ =

$P(O_{t+1},O_{t+2},\dots,O_T|Q_t = i,\lambda)$. Multiplying these together will give $\alpha_i(t)\cdot\beta_i(t) = P(O_1,O_2,\dots,O_t, Q_t = i|\lambda)\cdot P(O_{t+1},O_{t+2},\dots,O_T|Q_t = i,\lambda)$ which is the same as $P(Q_t = i, O|\lambda)$. Thus, we can calculate $\gamma_i(t)$ from equation 1 as

$$\gamma_i(t) = \frac{\alpha_i(t)\cdot\beta_i(t)}{\sum_{j=1}^{N}\alpha_j(t)\cdot\beta_j(t)}$$

Now, define the new variable $\xi_{ij}(t)$ to be the probability of being in state i at time t and state j at time t+1 given the observation sequence O and the model λ. That is

$$\xi_{ij}(t) = P(Q_t = i, Q_{t+1} = j|O,\lambda)$$

The equation for $\xi_{ij}(t)$ can be re-written as

$$\begin{aligned}\xi_{ij}(t) &= \frac{P(Q_t = i, Q_{t+1} = j, O|\lambda)}{P(O|\lambda)}\\ &= \frac{P(Q_t = i, Q_{t+1} = j, O|\lambda)}{\sum_{i=1}^{N}\sum_{j=1}^{N}P(Q_t = i, Q_{t+1} = j, O|\lambda)} \qquad (2)\end{aligned}$$

For illustration, Figure 3 shows a graph from Rabiner (17) with modifications to represent the EFF-HMM sequence flow. The forward algorithm computes the hidden state $Q_t = i$ to be equal to $\alpha_i(t)$. The backwards algorithm computes the hidden state $Q_{t+1} = j$ as $\beta_j(t)$. It is also known that the probability of going from state $Q_t = i$ to $Q_{t+1} = j$ with our observations is equal to (the probability of transition from state i to state j) times (the probability of observing O_{t+1} at state j) with increasing or decreasing contribution based on (the probability of being in state i observing O_t and next observing O_{t+1}); i.e. $P(Q_t = i,Q_{t+1} = j,O|\lambda) = \alpha_i(t)\cdot a_{ij}\,[(b_{jk}*c_i(h,k))/Sum_h(i,j)]\cdot\beta_j(t)$. Equation 2 for $\xi_{ij}(t)$ can now be written as:

$$\xi_{ij}(t) = \frac{\alpha_i(t)a_{ij}\left[\left(b_{jk}c_i(h,k)\right)/Sum_h(i,j)\right]\beta_j(t+1)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i(t)a_{ij}\left[\left(b_{jk}c_i(h,k)\right)/Sum_h(i,j)\right]\beta_j(t+1)}$$

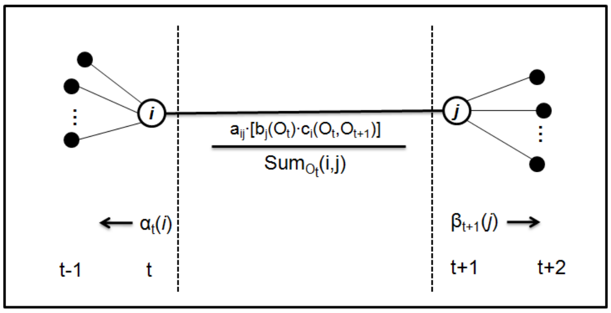


Fig. 3. Illustration of sequences required for computation of the EFF-HMM. Modified from Rabiner[17].

Notice that if you sum $\gamma_i(t)$ across all of t, $\sum_{t=1}^{T}\gamma_i(t)$, you get the total expected number of times in state i. Also, if you sum $\xi_{ij}(t)$ across all of t, $\sum_{t=1}^{T-1}\xi_{ij}(t)$, you get the total expected number of transitions from state i to state j. An equations for re-estimating the parameters of the model with the given observations to maximize $P(O|\lambda)$ can be written.

$$\bar{\pi}_i = expected\ times\ in\ state\ i\ at\ time\ t = 1$$

$$\bar{\pi}_i = \gamma_i(1)$$

$$\bar{a}_{ij} = \frac{expected\ transitions\ from\ state\ i\ to\ state\ j}{expected\ transitions\ from\ state\ i}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_{ij}(t)}{\sum_{t=1}^{T-1}\gamma_i(t)}$$

$$\bar{b}_{jh} = \frac{expected\ times\ in\ state\ j\ observing\ V_h}{expected\ times\ in\ state\ j}$$

$$\bar{b}_{jk} = \frac{\sum^{T}{}_{\substack{t=1\\ s.t.O_t=V_k}}\gamma_j(t)}{\sum_{t=1}^{T}\gamma_j(t)}$$

$$\begin{aligned}&\bar{c}_i(h,k)\\ &= \frac{\begin{array}{c}expected\ times\ in\ state\ i\ observing\ V_h\\ at\ time\ t\ and\ observing\ V_k at\ time\ t+1\ for\ all\ t\end{array}}{expected\ times\ in\ state\ i\ observing\ V_h}\end{aligned}$$

$$\bar{c}_i(h,k) = \frac{\sum^{T-1}{}_{\substack{t=1\\ s.t.O_t=V_h,\\ O_{t+1}=V_k}}\gamma_i(t)}{\sum^{T-1}{}_{\substack{t=1\\ s.t.O_t=V_h}}\gamma_i(t)}$$

4. Example

The Weizmann Human Action data set [18] uses single monocular cameras to record ten actions from nine people. Each action from each person has a different number of frames associated with it. The frames were recorded in color with pixel dimensions of 144x180. These actions are:

- Bend over to grab an object on the ground
- Jumping jacks several times
- Jumping forward several times
- Jumping up several times
- Run
- Side shuffle
- Skip
- Walk
- Wave with one hand facing forward
- Wave with two hands facing forward

To compare the EFF-HMM algorithm to the better classifiers on this data, the processing of the image should be identical, or as close as possible. However, it is not possible to exactly duplicate the input parameters since this is an HMM where the comparative algorithm uses SVMs and a decision system. If the same data inputs were to be used, the EFF-HMM would require more training data to accurately model the action; the more observations used to describe the action, the more data needed for training the EFF-HMM to accurately classify.

The first step is to identify the arms, legs and head. This is roughly achieved by using a silhouette of the person performing the action in each frame. By finding the centroid of the silhouette and using it as the center of four quadrants separating the image, the arms and legs can be extracted. See Figure 4. Each quadrant, represented by Q1 through Q4, is assumed to hold either an arm or a leg; Q1 holds the left arm, Q2 holds the right arm, Q3 holds the right leg, and Q4 holds the left leg. This is not an exact way of determining the arm and leg positions, but for the EFF-HMM classification system it will work fine.

The EFF-HMM classification system is concerned with the position of the arms, legs, and head in each frame. Lateral distances from the centroid are used, that is only the x value in the point position that represents the body parts. For example, the maximum distance in Q1 represents the position of the left arm in that frame. This is performed for each quadrant to get all arms and legs positions. The head is determined by the maximum height of the image.

The final step to calculating our inputs into the EFF-HMM is to reduce the number of data points. As mentioned earlier, if you increase the number of observations than to accurately classify action sequences a large data set would be needed. The Weizmann Human Action data set is not large, only nine people performing ten actions. To reduce the number of observations (i.e. input parameters) the distances of the arms are combined together per frame. This is performed on the legs as well. To further reduce the number of data points, a pair-wise comparison is performed on the arms, legs, and head to determine if the distances increased, decreased or did not change from consecutive frames in the sequence.

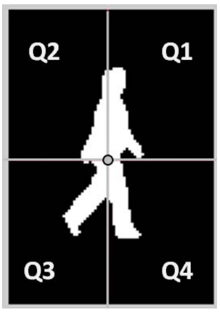
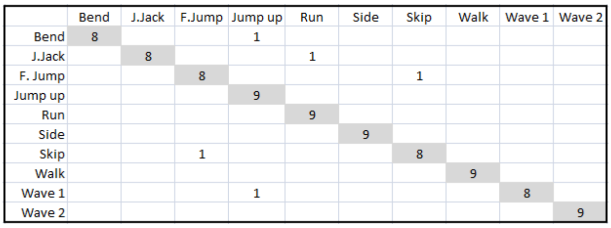


Fig. 4. Silhouette of a person walking taken from a single frame in an action set. The center circle represents the centroid of the silhouette. The four sections are quadrants represented by Q1 through Q4.

Similar motions on a particular body part were grouped and EFF-HMMs were computed. For example, walking and running have the same leg motion, taking out the frequencies, so these were lumped together for legs. Waving with two arms and jumping jacks have the same arm motion, so these were grouped together. Separate EFF-HMMs were calculated for legs, arms, and head motion in the training phase.

For the testing phase, the legs, arms and head were classified against each trained EFF-HMM, respectively. A decision system performed the final classification. The leave-one-out cross validation procedure was performed and the confusion matrix is shown in Table 1. The correct classification rate is 94.4% correctly classified. This is comparable with other classification systems on this data set, as will be discussed in later.

Table 7.1. This table shows the results of the leave-one-out cross validation procedure. The columns represent the classification and the rows represent what they should be in.



	Bend	J.Jack	F.Jump	Jump up	Run	Side	Skip	Walk	Wave 1	Wave 2
Bend	8			1						
J.Jack		8			1					
F. Jump			8				1			
Jump up				9						
Run					9					
Side						9				
Skip			1				8			
Walk								9		
Wave 1				1					8	
Wave 2										9

The five misclassified events can be further analyzed to determine the cause of the classification errors. For the single Bend action which was classified as a Jump Up action, the pre-processing of the image frames are not representative to the Bend action set. Image processing errors showed consecutive frame maximum heights increasing and decreasing quite often where they should have just decreased. This is representative of the Jump Up action.

The Jumping Jack misclassified person had several more jumping jacks than the rest of the people performing them. This caused the leg distances to change similar to the Run action. Also, head movement was similar to running. The hands peaked at a higher frequence than other people performing this action. When the hands peaked, it was classified as a head position, since it is the greatest y value, and when they came down past the head, the head took the role of head position, thus creating several frequencies of movements of up-down.

Figure 5 shows the silhouette of a Skip action and the Jump Forward action of the same person. These were taken from the same person and can be seen as very similar. The only difference is in the third quadrant where the leg comes out further in the skip than the Jump Forward. If there were image processing errors in the third quadrant of the Jump Forward or the Skip actions then these would likely cause a misclassification. This is what happened in both the Jump Forward action misclassified person and the Skip action misclassified person.

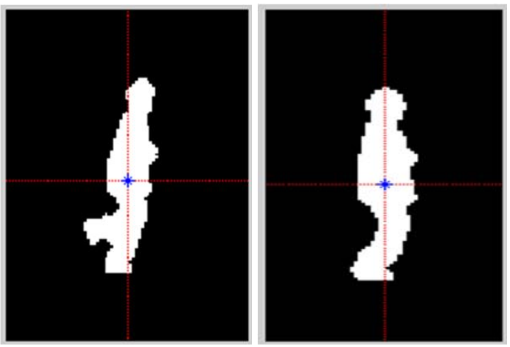


Fig. 5. Single frame silhouette of a Skip action (right) and Jump Forward action (left) of the same person.

The final misclassified action is in the Wave 1 action (wave with one hand). It was misclassified as a Jump Up action. Looking at the image frames there were two problems with this. The first is the image processing of the legs. Throughout the image frames, the legs were poorly processed and often times had extending features in the silhouettes which caused the distances in the legs to alternate increasing and deceasing at a high frequency. For the Wave 1 action, the legs should be stationary. The more important problem in the image processing portion is the position of the hand as it waved. Again, due to image processing problems, the top of the hands where increasing and decreasing at a high frequency on both the up movements and the down movements. This was viewed as the head moving up and down at a high frequency since the highest y value is considered the head. Motion of the head was classified as running motion head movement.

4.1. *Comparing EFF-HMM with comparable classifier*

The work of [18] provides a good comparison. The input data set was developed as close to the same as could be provided from the different types of classification systems. The model to compare against is the SW-SVM model since it provides similar functionalities.

Pre-processing of the SW-SVM classifier input data is generally the same except instead of defining a single position for the arms, legs, and head, it defines three (x,y,z) for each arm, each leg, and the head. Thus, instead of three input variables as the EFF-HMM uses, the SW-SVM system uses fifteen variables. The tracking from one frame to the next is also performed as in the EFF-HMM, but a listing of twenty curves is matched up for each variable. These twenty curves are determined using the Smith-Waterman (SW) dynamic programming technique. For EFF-HMM, there are a possible three values each variable can take on (Increase, Decrease, and No Change). Finally, a multi-class Support Vector Machine (SVM) is used to classify the action; thus giving this algorithm the name SW-SVM.

The data set is a subset of the original by removing the Skip Action. If the skip action was removed from the EFF-HMM classifier, the results would be only three of eighty-one misclassified, 96.3%. For the SW-SVM classifier, it correctly matched 79 of the 81 actions, thus having an accuracy of 97.5%, one better than the EFF-HMM.

This is comparable with the EFF-HMM since the EFF-HMM did a little worse, but the number of inputs were greatly reduced (from 15 to 3) and the values each input could have was also greatly reduced (from 20 to 3).

5. Conclusion

The Evidence Feed Forward Hidden Markov Model is more than a standard Hidden Markov Model. It provides observation to observation linkage in the algorithm. The linkages were developed through analysis and proven mathematically. It was originally designed to provide a way of classifying visual activities better, like the differences in pitching and throwing from the outfield. The idea is that if there existed a way to link one observation frame to another, then there may be some patterns that the EFF-HMM could recognize better than if there were no observation linkages, like how a standard HMM would classify. This was extended for more than just visual data and has shown to work in other classification areas.

The convergence of EFF- HMMs is slightly longer then the convergences of standard HMMs when changes in parameters are measured against the log-likelihood of the testing data to its classification. This is due to the complexity of EFF- HMMs compared with standard HMMs.

This new HMM has worked well for classifications of items based on the observation to observation link that is not available in other types of HMMs. When comparing to other autonomous classifications systems, it performs well with less data as shown on the Weizmann Human Action data set [18].

6. References

1. Dimitrijevic, M., Lepetit, V., Fua, P., “Human Body Pose Detection Using Bayesian Spatio-Temporal Templates,” *Computer Vision and Image Understanding,* Vol. 104, No. 2, 2006, pp.127-139.
2. Weinland, D., Ronfard, R., Boyer, E., “Motion History Volumes for Free Viewpoint Action Recognition,” *Proceedings from IEEE International Workshop on Modeling People and Human Interaction,* 2005.
3. Yang, J. Y., Wang, J. S., Chen, Y. P., “Using Acceleration Measurements for Activity Recognition: an Effective Learning Algorithm for Constructing Neural Classifiers,” *Pattern Recognition Letters,* Vol. 29, 2008, pp. 2213–2220.
4. Stern, H., Kartoun, U., Shmilovici, A., “A Prototype Fuzzy System for Surveillance Picture Understanding,” *Proceedings from Visual Imaging and Image Processing Conference,* Sept 2001, pp. 624-629.
5. Del Rose, M., Stein, J., “Survivability on the ART Robotic Vehicle,” *Proceedings from the Seventeenth Ground Vehicle Survivability Symposium,* 2006.
6. Army Research Lab (ARL), Collaborative Technology Alliance (CTA): www.arl.army.mil/www/default.cfm?page=392.
7. Mobile Detection Assessment and Response System (MDARS), www.spawars.navy.mil/robots/land/mdars/mdars.html.
8. Campbell, L., Becker, D., Azarbayejani, A., “Invariant Features for 3-D Jester Recognition,” *Proceedings from IEEE Automatic Face and Gesture Recognition (AFGR),* 1996, pp. 157-162.
9. Yu, C., Ballard, D., “Learning to Recognize Human Action Sequences,” *Proceedings from International Conference on Development and Learning*, 2002, pp. 28-33.
10. Gehrig, D., Schulz, T., “Selecting Relevant Features for Human Motion Recognition,” *Proceedings from International Conference on Pattern Recognition*, 2008, pp. 1-4, doi:10.1109/ICPR.2008.4761290.
11. Wilson, A., Bobick, A., “Parametric Hidden Markov Models for Gesture Recognition,” *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI),* Vol. 21, 1999, pp. 884-899.
12. Oliver, N. M., Rosario, B., Pentland, A. P., “A Bayesian Computer Vision System for Modeling Human Interaction,” *IEEE Transactions in Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, Aug. 2000, pp. 831-843.
13. Xiang, T., Gong, S., “Activity Based Video Content Trajectory Representation and Segmentations,”

Proceedings from British Machine Vision Conference, 2004, pp. 177-186.

14. Xiang, T., Gong, S., "Incremental Visual Behaviour Modelling," *Proceedings from European Conference on Computer Vision*, 2006, pp. 65-72.
15. Antonakaki, P., Kosmopoulos, D., Perantonis, S. J., "Detecting Abnormal Human Behaviour Using Multiple Cameras," *Signal Processing Journal*, Vol. 89, 2009, pp. 1723 – 1738.
16. Oliver, N., Horvitz, E., Garg, A., "Layered Representations for Human Activity Recognition," *Proceedings from IEEE International Conference on Multimodal Inferences (ICMI)*, 2002, pp. 3-8.
17. Rabiner, L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol. 7, 1989, pp.257-286.
18. Gorelick, M., Blank, M., Shechtman, E., Irani, M., Basri, R., "Actions as Space Time Shapes," *Proceedings from the Tenth IEEE International Conference on Computer Vision, 2005.*