Detecting emotions from connected action sequences

Daniel Bernhardt & Peter Robinson

Computer Laboratory, University of Cambridge, UK.
firstname.lastname@cl.cam.ac.uk

Abstract. In this paper we deal with the problem of detecting emotions from the body movements produced by naturally connected action sequences. Although action sequences are one of the most common form of body motions in everyday scenarios their potential for emotion recognition has not been explored in the past. We show that there are fundamental differences between actions recorded in isolation and in natural sequences and demonstrate a number of techniques which allow us to correctly label action sequences with one of four emotions up to 86% of the time. Our results bring us an important step closer to recognising emotions from body movements in natural scenarios.

1 Introduction

Inferring emotions from human body motion in natural environments can be very difficult as body movements are virtually unconstrained. This makes it difficult to train emotion recognisers which are robust enough to tolerate this kind of real-world variability while still picking up subtle emotion-communicating cues. In this paper we describe an approach to recognising emotions from natural action sequences. We refer to these sequences as connected actions.

The first contribution of this work is the description of an end-to-end system which is able to detect emotions from connected action sequences. In order to perform emotion recognition effectively, we first need to build a solid understanding about the underlying constraints of human movements. We show that an increased refinement of action models can boost our ability to recognise the emotions communicated through connected actions.

Secondly, we highlight the differences between actions recorded in isolation and as naturally connected sequences. Although obtaining and working with isolated data is often easier, we show that results achieved on isolated data are not necessarily transferable to cases where actions appear in connected sequences. In order to bridge the gap between the two, we describe ways to adapt isolated models to the connected cases. As a result we hope to bring emotion recognition one step closer to naturally-occurring scenarios.

2 Background

Human action and activity recognition has been studied extensively in the past with connection to unusual event detection, crime prevention and the like. In those cases it is the action itself which is the focus of the recognition effort. Actions, however, can also provide a valuable context for emotion recognition in natural environments. Imagine yourself as a human judge faced with an impoverished video recording of a human subject. The subject you are watching is stretching the right arm backwards and moving it forcefully forwards again. If this subject is involved in a conversation with another person, you might interpret the movement as communicating a hostile stance. If the person, however, was moving a piece of paper this movement can be easily interpreted as a throwing action. Only if the movements were extremely forceful would we be likely to associate the motion itself with an aggressive emotion. This example illustrates that our emotional interpretation of human body motion is based on our understanding of the action which is being performed.

This is the problem pattern recognition algorithms face when classifying emotional content from body movements. Algorithms which have no prior model of movement patterns are likely to register large differences between examples of the same emotion category but in very different actions such as running, walking and knocking. Clearly, this extreme kind of variation will render any attempt to discover the underlying patterns due to emotional changes extremely difficult. In this paper we therefore use explicit models of action patterns to aid emotion classification.

In some cases authors discussing the recognition of emotions from body movements manage to side-step the above problem by only considering one type of action such as knocking [2] or a prescribed arm lowering action [3]. In other cases researchers have focused on stylised body motions. Those are motions which usually arise from laboratory settings where subjects are instructed to act an emotion freely without any constraints. Authors of those studies often find that under those circumstances subjects produce stereotypical expressions [1, 4]. These produce strong patterns which are easier to detect with statistical pattern classification techniques.

In many ways the analysis of connected actions is similar to that of connected speech. Indeed, emotion recognition from speech has been a prominent problem since the early days of affective computing. Many different sets of low level acoustic features have been proposed over the years to capture emotional information in recorded speech. However, one recent study by Lee and Narayanan suggests that major improvements in emotion discrimination can in fact be achieved by making the recognition algorithms aware of higher level lexical and discourse structure [5]. Our work builds on these results by adding structural knowledge about common *action patterns* to the emotion recognition framework.

Our motion data comes from a motion-captured corpus of actions recorded at the Psychology Department, University of Glasgow [6]. It contains samples of knocking, throwing, lifting and walking actions recorded both in isolation and as naturally connected sequences. 15 male and 15 female untrained subjects were recruited and actions recorded in 4 emotional styles: neutral, happy, angry and sad. For the performance of isolated actions subjects were instructed fairly carefully, e.g. which hands to use for actions and how far to stand from certain props. Connected actions were naturally less constrained.

3 System Overview

Our goal is to classify each of the action sequences in our corpus into one of the four emotion classes. Note that we do not in general know the order of actions that make up a sequence, nor do we know where the action boundaries are. We will present a solution to this segmentation problem in Section 4.2. Currently all our system assumes is that it knows, and has models for, each of the action categories and emotions it could be faced with. We describe in detail how we build those models in Sections 4 & 5.

Importantly, both the action and emotion models are initially trained on *isolated* samples. In other domains such as speech recognition it is often believed that models need to be trained on data stemming directly from connected samples [8]. Within the scope of this research, our decision to initially base our models on isolated data has a number of advantages:

1. A number of systems that analyse isolated actions have been built and discussed in the past [2, 3]. We are building on their insights to derive our action and emotion models.
2. Starting out with isolated models allows us to evaluate their performance for connected actions. We will discuss how isolated models can be adapted to perform better on connected actions. The gained insights are very illuminating in understanding the differences between emotions expressed through isolated and connected actions.

Our recognition framework works as follows. Given a set of action categories $\mathcal{A}$ (e.g. $\mathcal{A} = \{\text{knocking}, \text{throwing}, \text{lifting}, \text{walking}\}$) and emotion classes $\mathcal{E}$ (e.g. $\mathcal{E} = \{\text{neutral}, \text{happy}, \text{angry}, \text{sad}\}$), we classify an action sequence $S = (s_1, s_2, \ldots, s_n)$ with $s_i \in \mathcal{A}$ as an emotion $e \in \mathcal{E}$ as follows:

1. Train a set of action models $\Lambda = \{\lambda_a\}$ on samples of isolated actions.
2. Train emotion models $M_{a,\mathcal{E}}$ on isolated samples of action category a and emotion set $\mathcal{E}$.
3. Adapt Λ and $M_{a,\mathcal{E}}$ to the patterns observed in connected actions yielding the adapted models $\hat{\Lambda}$ and $\hat{M}_{a,\mathcal{E}}$.
4. Segment S into its component actions $(s_1, \ldots, s_n)$ using $\hat{\Lambda}$.
5. For each $s_i \in S$, find the most likely emotion class e_i using $\hat{M}_{s_i,\mathcal{E}}$.
6. Combine $(e_1, \ldots, e_n)$ into an overall emotion class e for the whole sequence.

Note that we explicitly model the difference of emotional appearance in different actions by training emotion models $M_{a,\mathcal{E}}$ dependent on both emotions *and* action. This allows us to deal with the cases introduced in our initial example. In the next sections we discuss how we define Λ and $M_{a,\mathcal{E}}$ and how to adapt them to $\hat{\Lambda}$ and $\hat{M}_{a,\mathcal{E}}$ respectively.

4 Action analysis

In the absence of any context information, isolated actions are defined and identified by the spatio-temporal trajectories of body joints. Formally, we represent an action category a as a set of joints and a description of their movements over time, λ_a. We are using Hidden Markov Models (HMMs) to solve the isolated action recognition problem. HMMs have been applied successfully to this kind of temporal pattern recognition problem in the past [8] and we are able to draw on an extensive body of knowledge documenting their use.

HMMs are particularly suitable for modelling temporally evolving systems which produce observable outputs. At each discrete point in time, the system can be in one of a finite number of hidden states. The transitions between states over time are governed by a matrix $\boldsymbol{A}$ of transition probabilities. At every time frame the system outputs an observation vector. The probability of observing a particular output is conditioned only on the current hidden state. Because joint movements exhibit complex trajectories in position and speed we model the system's output as a vector of *continuous* observation variables, parameterised by the mean and standard deviation of a normal distribution. The observation densities are represented in two matrices $\boldsymbol{O}_m$ and $\boldsymbol{O}_s$ capturing the mean and standard deviation of the observation variables in each state.

4.1 Model parameters

The essence of an action is its pattern of posture and movement changes over time. We are therefore using the following quantities as our HMM observation variables.

- global body speed
- body-local joint positions
- body-local joint speeds
- body twist (angle between shoulder-shoulder and hip-hip vectors)

These quantities are easy to calculate from the 3D joint position data available directly from the motion capture corpus. Body-local measures are derived by a simple transform placing a coordinate system at the pelvis joint of the subject. This gives us a representation invariant to absolute body position and orientation. The transition matrices for each action model impose a left-to-right structure, thus strictly enforcing a traversal from the first to the last hidden state. The only complication arises for walking motions. They are *cyclic* in nature and therefore the action model for walking allows a transition from the last back to the first hidden state. The HMM parameters $\boldsymbol{A}$ and $(\boldsymbol{O}_m, \boldsymbol{O}_s)$ are estimated from the isolated action samples using the standard Baum-Welsh algorithm [8]. The number of hidden states for each model was chosen empirically and is in each case less than 10.

4.2 Parsing connected actions

We use the isolated action models to build a connected action recogniser. This problem is very similar to connected speech recognition from individual word models [8]. We can therefore make use of the extensive literature available on the subject.

One popular technique developed by Rabiner and Levinson to solve this problem is Level Building (LB) [7]. Given a sequence of observations it uses an efficient Dynamic Programming approach to find the most likely sequence of actions and according segmentation boundaries. This approach is very similar to the Viterbi algorithm which finds the most likely hidden state sequence of a single HMM given an observation sequence. Indeed, LB uses the Viterbi algorithm repeatedly at every level (see Rabiner and Levinson [7] for details). In order to be able to compare the segmentations achieved through LB to some ground truth, we also hand-segmented each of the sequences. We will make use of this manual segmentation in our evaluation in Section 6.

4.3 Isolated and connected action differences

By playing back videos of our motion corpus, we quickly realised that actions did not appear the same in isolation and in connected sequences. Because many constraints were placed on the subjects for the isolated recordings, they tended to appear more controlled and uniform. In the connected case we observe actions blending into each other, making it hard to identify unique transition points between individual actions. Anticipatory effects were particularly strong. For example, isolated knocking actions uniformly started with a succinct arm lift before the knock. When knocking is preceded by a walking action, however, we can observe the arm lift to commence at various points during the walking action and long before the knock itself starts.

We were interested in finding quantitative evidence for the difference in appearance of isolated and connected actions. Here we focus on the amount of *variation* observed across different subjects and repetitions of the same action. For each action sample, we computed a set of features $\mathcal{F}$ capturing the temporal evolution of each joint as its mean and standard deviation in position, speed, acceleration and jerk. These quantities were shown by Bernhardt and Robinson in [2] to capture the static and dynamic qualities of body motions well. We then compute the sample standard deviations $\sigma_{I,f}$ and $\sigma_{C,f}$ over all isolated and connected action samples respectively. For every feature f, $\sigma_{I,f}$ and $\sigma_{C,f}$ tell us how much f varies across different instances of the same action. Finally, we partition $\mathcal{F}$ into $\mathcal{F}_C$ and $\mathcal{F}_I$ such that

$$f \in \mathcal{F}_C \Leftrightarrow \sigma_{C,f} > \sigma_{I,f} \tag{1}$$

$$f \in \mathcal{F}_I \Leftrightarrow \sigma_{C,f} \leq \sigma_{I,f} \tag{2}$$

That is, $\mathcal{F}_C$ contains the features which show relatively large variation across connected samples while the features in $\mathcal{F}_I$ show larger variation across isolated

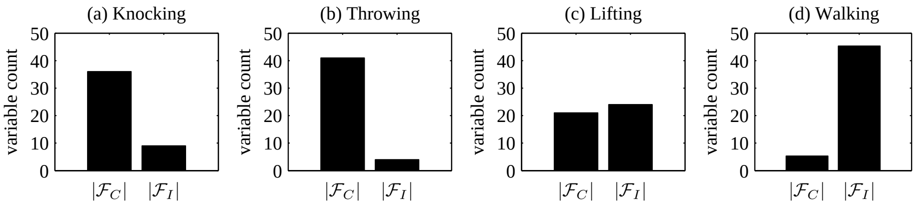

Fig. 1. Difference in variation between isolated and connected actions.

samples. Figure 1 shows $|\mathcal{F}_C|$ and $|\mathcal{F}_I|$ for every action category. If the isolated and connected cases had similar dynamic characteristics, we would expect each pair of bars to be of roughly the same height. We see, however, that only for lifting actions the variation is relatively similar in both cases. For all other actions, the dynamics differ substantially between isolated and connected cases. The observed differences in knocking, throwing and walking actions make it necessary to *adapt* the action models to the connected cases. It is important to note at this point that, although we discussed these dynamic differences in the light of action models, emotion recognition is likely to suffer similarly from these changes in appearance. We will see quantitative evidence for this in Section 6.

4.4 Adapting action models

The above differences make it necessary to adapt action models to the appearance of connected actions. We chose to adapt the models statistically using a small set of representative sequences $\mathcal{S}$ and an associated set of weak labels $\mathcal{L}$. By weak labels we refer to a sequence of action categories such as "knock, walk, lift, throw" as exhibited by the sequences but without any explicit action boundaries. We can then use a bootstrapping approach to iteratively refine the set of action models Λ as follows:

1. Start with an initial set of models $\hat{\Lambda}^0$, $i = 0$
2. Segment all sequences in $\mathcal{S}$ by LB using $\hat{\Lambda}^i$
3. Retrain a new set of models $\hat{\Lambda}^{i+1}$ using the action samples of sequences which agree with $\mathcal{L}$
4. If the number of correctly segmented sequences increased $i = i + 1$, goto 2.

To start off the bootstrapping loop we initialise $\hat{\Lambda}^0$ to the isolated action models. Successive iterations then improve the model parameters based on the connected samples which were segmented correctly. For our data, we found that the bootstrapping iterations converge very quickly and we obtain a converged set of models after two iterations. While $\hat{\Lambda}^0$ only segmented 45% of $\mathcal{S}$ correctly, $\hat{\Lambda}^2$ improved this to 87%. Subsequent bootstrapping iterations decrease the number of correctly segmented sequences slightly.

5 Emotion recognition

Our emotion recognition framework is based on Bernhardt and Robinson's framework for classifying isolated knocking motions [2]. We extend their approach by training individual classifiers $M_{a,\mathcal{E}}$ for each supported action category a, thus allowing emotion recognition based on a variety of actions. From each action time series we extract a rich feature vector which captures the static and dynamic information of the action. The features include mean and standard deviations of posture, as well as joint speed, acceleration and jerk [2] calculated over the whole action. We then normalise each feature to ensure similar orders of magnitude for each feature dimension. This aids robust training for pattern recognition algorithms.

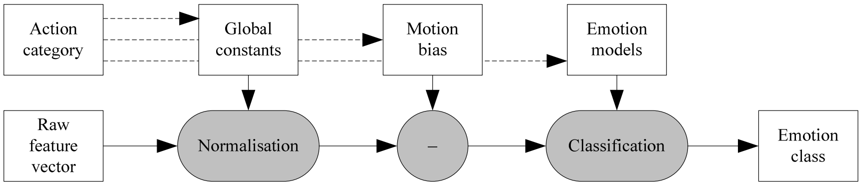

Fig. 2. Emotion recognition pipeline. Grey components denote operations performed on the data components shown in white.

As the next step, we then subtract an *individual movement bias*. This extra normalisation step has been shown to remove a major source of confusion for the classification of emotions [2]. It accounts for the fact that different subjects tend to exhibit different motion idiosyncrasies, thus confounding the subtle dynamic differences between different emotions. The unbiased feature vector is then fed to a Support Vector Machine-based classifier $M_{a,\mathcal{E}}$ which classifies it into one of the emotions in the emotion set $\mathcal{E} = \{\text{neutral, happy, angry, sad}\}$. This classification pipeline is shown in Figure 2. The inputs shown at the top of the pipeline need to be calculated prior to a classification from representative training data. In detail, those are

1. the global normalisation constants calculated per action category
2. the personal motion bias constants calculated per action category and person
3. the emotion classifiers trained for each action category and on all emotion classes.

In order to find a unique emotion label for a sequence S we classify each component action s_i using the classifier $M_{s_i,\mathcal{E}}$. We treat each of the classification results from different component actions as independent evidence towards the overall emotion classification. Therefore, we arrive at a combined emotion class by taking a majority vote. Ties are resolved by assigning one of the candidate classes randomly.

5.1 Adapting emotion classifiers

In Section 4.3 we described in detail how the appearance of actions differs when we move from isolated to connected actions. As for our action models, our initial emotion classifiers $M_{a,\mathcal{E}}$ are trained on the appearance of emotions in *isolated* cases. We may, however, wish to adapt our classifiers to better capture the appearance of emotions in connected actions. A number of adaptation methods are possible, each varying in the associated cost. We will describe each of them here and evaluate their performance in Section 6.

Clearly, the cheapest adaptation method is to simply reuse the isolated emotion models $M_{a,\mathcal{E}}$. We would expect these models still to perform better than random as emotion appearances should not change so extremely as to render the isolated models entirely useless. At the other end of the scale lies a total retraining of the classifiers on connected action data. In essence, we need to recompute all three inputs to the classification pipeline listed in the previous section: global constants, personal bias and emotion classifiers. This is likely to give us the best results. These two extremes represent recalculating either none or all of the three inputs to the pipeline. Apart from these two extremes we explore two intermediate adaptation strategies: recalculating only the first or the first two inputs to the pipeline using the connected data. We call the derived models $\hat{M}^0_{a,\mathcal{E}}$, $\hat{M}^1_{a,\mathcal{E}}$, $\hat{M}^2_{a,\mathcal{E}}$ and $\hat{M}^3_{a,\mathcal{E}}$ according to how many inputs are recomputed.

6 Experimental results

Having discussed approaches to adapt both action and emotion models from isolated to connected data, it is now time to evaluate how much difference these changes actually make on real data. To this end we conducted an experiment using the full set of data in our corpus. At a high level we treated the isolated actions as training data and evaluated our algorithms on the connected action sequences. We had around 4000 isolated action samples and 220 action sequences. The latter all followed the same order: walking, lifting, walking, knocking, walking, throwing. We ignored this knowledge, however, when segmenting the data. The sequence data was also used to adapt the action and emotion models as outlined in Sections 4.4 & 5.1. With a number of adaptation approaches in hand, we asked ourselves the following questions:

1. What recognition rates are achievable with our classification approach?
2. Are there evidence for differences in the expression of emotions through isolated and connected actions?
3. If so, which of the described adaptation schemes provide the best improvements?

To answer these questions our experiment measured the effects of two independent factors: level of adaptation of action models and level of adaptation of emotion classifiers. The dependent variable we measured in each case was the rate of correct emotion classifications for whole sequences. As cases for adapted

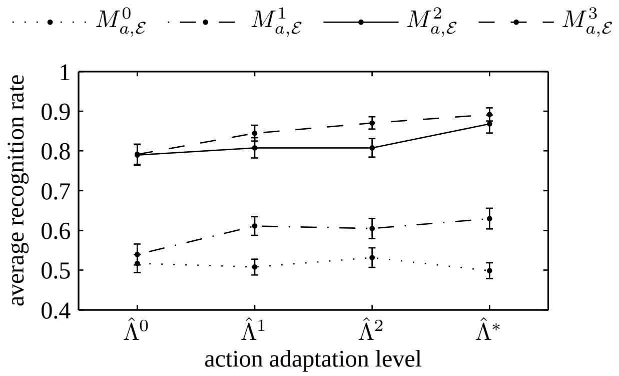

Fig. 3. Recognition rates for various levels of action and emotion model adaptation.

action models, we considered the segmentations achieved after iterations 0, 1 and 2 of the bootstrapping algorithm presented in Section 4.4. Each of the iterations produced action models of increasing adaptation levels ranging from no adaptation for $\hat{\Lambda}^0$ to good adaptation for $\hat{\Lambda}^2$. As a gold-standard we also considered an ideal set of action models $\hat{\Lambda}^*$ which produces the segmentation we had produced manually. As cases for the emotion classifiers we considered $\hat{M}^0_{a,\mathcal{E}}$ to $\hat{M}^3_{a,\mathcal{E}}$. As for the action models, $\hat{M}^0_{a,\mathcal{E}}$ represents no adaptation while $\hat{M}^3_{a,\mathcal{E}}$ represents a gold-standard achieved by totally retraining the emotion models on the connected data. For the last condition we used 10-fold cross validation to prevent training and testing on the same samples.

Table 1. Average emotion recognition rates for whole action sequences for no adaptation (left) and good adaptation (right). Emotions appear in the order neutral, happy, angry, sad.

0.15	0.06	0.00	0.79	**0.88**	0.03	0.03	0.06
0.07	**0.47**	0.06	0.40	0.12	**0.75**	0.11	0.02
0.01	0.39	**0.47**	0.13	0.00	0.14	**0.86**	0.00
0.00	0.00	0.02	**0.98**	0.24	0.01	0.00	**0.75**
average rate: 0.52				**average rate: 0.81**			

In order to get the most comprehensive picture possible, we decided to adopt a factorial design. By investigating the cases in all possible combinations we get a set of 4×4 classification results. These results are visualised in Figure 3. Each line represents a series of results obtained for different action models Λ and for the same emotion classifier M. Because our voting algorithm resolves ties by making a random choice we also indicate the spread with error bars. As a general trend we observe that the recognition rate increases with the levels of adaptation. In Table 1 we provide the confusion matrices for the pair of unadapted models ($\hat{\Lambda}^0$, $\hat{M}^0_{a,\mathcal{E}}$) and for the combination of best-adapted models ($\hat{\Lambda}^2$, $\hat{M}^2_{a,\mathcal{E}}$) short of the gold-standard. The significant increase in recognition rate confirms our initial intuitions.

7 Discussion and future work

Our experimental results clearly show that using unadapted models trained on isolated data on connected samples produces suboptimal emotion recognition results. This is clear evidence that the appearance of actions change as we move from strictly controlled, isolated samples to more natural, connected sequences. As we have managed to show, this does not only impact the recognition of actions. Emotion recognition performance improves *both* as we adapt our action models *and* our emotion models. A change in the appearance of actions therefore degrades emotion recognition in two ways. Firstly, a change in appearance impedes our ability to recognise actions reliably which has a knock-on effect on emotion recognition as we choose the wrong emotion models $\hat{M}_{a,\mathcal{E}}$. Secondly, the change in movement dynamics as we move to connected actions means that our emotion models are simply not representative anymore.

We have shown that our adaptation approaches are effective. As we expected, the recognition rate achieved with unadapted models is significantly better than chance at 52%. This rate can be improved to 81%, however, by using our well-adapted models. Note that for the latter case we did not need to retrain the actual emotion classifiers, but the adaptation stemmed from appropriate pre-processing of the feature vectors. This means that we do not need connected motion sequences labeled by emotion. Using our gold-standard adaptations the sequences can be classified at a rate of 86%. It seems, however, that the adaptation step from $\hat{M}^1_{a,\mathcal{E}}$ to $\hat{M}^2_{a,\mathcal{E}}$ brings the biggest improvement. This suggests that there is no clear pattern with which individuals' behaviour changes when they go from isolated to connected action displays — we simply need to recompute the personal motion bias of connected actions. This highlights once more how important the modelling of individual differences is for the recognition of emotions from body motions — both for isolated [2] and connected actions.

On a larger scale we conclude that data collected under very constrained laboratory conditions is not necessarily representative of data occurring in more natural scenarios. Of course, our connected data was only recorded under laboratory conditions as well and it is therefore likely that truly natural data will show effects beyond of what we observed. Repeating this experiment on data collected in a natural environment will be an interesting goal for future research. Nevertheless we believe that we have taken an important step towards being able to deal with real-world scenarios. It is encouraging to note that although there are changes in appearance, methods previously developed for isolated data are in fact applicable to connected samples if they are adapted appropriately.

References

1. T. Balomenos, Amaryllis Raouzaiou, Spiros Ioannou, Athanasios I. Drosopoulos, Kostas Karpouzis, and Stefanos D. Kollias. Emotion analysis in man-machine interaction systems. In *1st International Workshop on Machine Learning for Multimodal Interaction*, pages 318–328. Springer, 2004.
2. Daniel Bernhardt and Peter Robinson. Detecting affect from non-stylised body motions. In *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pages 59–70. Springer, 2007.
3. Ginevra Castellano, Santiago D. Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pages 71–82. Springer, 2007.
4. Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007.
5. Chul M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303, 2005.
6. Yingliang Ma, Helena M. Paterson, and Frank E. Pollick. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior Research Methods*, 38:134–141, 2006.
7. L. Rabiner and S. Levinson. A speaker-independent, syntax-directed, connected word recognition system based on hidden markov models and level building. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(3):561–573, 1985.
8. Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.