Detecting emotions from everyday body movements

Daniel Bernhardt
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue
Cambridge, CB3 0DF, UK
+44 1223 763694
db344@cam.ac.uk

Peter Robinson
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue
Cambridge, CB3 0DF, UK
+44 1223 334637
pr10@cam.ac.uk

ABSTRACT

The work presented in this paper focuses on the development of a computational model for describing and detecting affective content in everyday body movements. The approach is based on the extraction and analysis of dynamic motion qualities as opposed to limiting itself to the denotative meanings of body posture and gestures. To this end, a database of affective everyday motions such as knocking and walking has been analysed. Our approach makes use of a segmentation technique which can divide complex motions into a set of automatically derived motion primitives. The parsed motion is then analysed in terms of dynamic features which are shown to encode affective information. In order to adapt our algorithm to personal movement idiosyncrasies we developed a new approach for deriving unbiased motion features. We demonstrate that the resulting performance of our algorithm is similar to that of humans who took part in a comparable psychological experiment.

1. INTRODUCTION

The human body is a complex hierarchical structure which has evolved to enable us to perform sophisticated tasks. At the same time, movements and posture of our limbs, head and torso communicate affect and inter-personal attitudes. To a large extent our functioning as socially intelligent individuals relies on our ability to decode the affective and expressive cues we perceive through facial or body gestures. Research suggests that our responses to avatars in Immersive Virtual Environments (IVEs) are governed by our expectations about the presence and correct exhibition of those expressive cues.

This paper describes a novel framework for analysing everyday or *non-stylised* body motion in order to detect affect. This is very different from analysing *stylised* body motions. In a stylised motion the entirety of the movement encodes a particular emotion. Stylised motions normally originate from laboratory settings, where subjects are asked to freely act an emotion without any constraints. They also arise from stylised dance. This paper, however, concerns itself with the more subtle aspects of human movement. We will examine how affect is communicated by the manner, in which everyday actions are performed. We hope that this work will open up a multitude of opportunities for intelligent human-machine interaction which is not viable with approaches that assume stereotypical, stylised body motions.

2. MOTIVATION AND BACKGROUND

One of the natural applications for our emotion-sensing technology is in the field of IVEs. A major goal for virtual reality research is to give the user of an IVE a sense of presence – the feeling of "being there" in the virtual environment [1]. In many cases it is beneficial or inevitable to populate a virtual environment with virtual humans (agents), most notably for the investigation of social interaction, co-operative tasks or tutoring. In those cases, numerous authors have argued that we deal with a special form of presence as relating to the sense of being with other social beings, called social presence or copresence. It is believed that a strong sense of copresence is the result of an intelligently acting and responsive agent [2]. One important aspect of this is the agent's ability to decode and elicit cues of non-verbal communication (NVC) such as appropriate body posture and gesture as well as facial actions [3]. Our work aims to establish a framework which can help to give virtual humans the emotional intelligence they need to decode affective cues which the users of an IVE elicit. If the agent elicits the appropriate reactions, this will help to enhance the sense of presence a user experiences.

In this paper we are only discussing the detection of affect from *non-stylised everyday* motions. Other authors have discussed the use of cues from other aspects of bodily NVC such as posture [4] and stylised gestures [5]. They report good results, but as we shall discuss later, detecting affect from the kind of non-stylised motions we are interested in is rather more difficult. In particular, we need to find a language which allows us to describe human body motion. One approach regards body language as analogous to natural language. Ray Birdwhistell argued that complex motions can be broken down into an ordered system of isolable elements which he called kinemes [6]. The notion of a universal set of kinemes or motion primitives is a compelling one as it gives structure to the otherwise vast complexity of human motion — a goal which the Facial Action Coding System [7] has achieved so successfully for the face. Furthermore, research in neurobiology suggests that the execution of complex motor behaviours in vertebrates might be based on such a combination of basic motor primitives [8]. In our approach it is the segmentation into motion primitives which will help us to discard the structural information of motions, leaving the essentially dynamic cues which we will use to distinguish different affects.

3. MOTION ANALYSIS

For this work we used a motion-captured database recorded at the Psychology Department, University of Glasgow [9]. It gave us access to a collection of knocking, throwing, lifting and walking motions performed by 30 individuals (15 male and 15 female) in neutral, happy, angry and sad affective styles. Most of our quoted results are based on the approximately 1200 knocking motions from the database.

The skeletal structure of the recorded bodies is represented by 15 joints, positioned relative to a world frame. In order to obtain a rotation- and scale-invariant representation, we transform the joint positions into a body-local coordinate system and normalise them with respect to body size. Let f stand for the dimension of time, measured in frames. We denote the time-varying signal of normalized joint positions as the matrix $\boldsymbol{\Psi}$. We can also represent the motion in terms of the joint rotations over time, $\boldsymbol{\Theta}$. A particular body configuration at frame f can be represented as a row vector, denoted as $\boldsymbol{\psi}_f$ or $\boldsymbol{\theta}_f$. The gth positional or rotational degree of freedom at frame f is written as $\psi_{f,g}$ or $\theta_{f,g}$ respectively.

3.1 Motion Segmentation

The goal of motion segmentation is to parse high-dimensional body movements into a sequence of more basic primitives. In general, this is a hard problem which is of interest to researchers from many different areas, including gesture recognition and robotics. Our approach is based on the work by Fod et al. [10]. It makes use of an objective function $E(f)$ which is a measure for the overall motion energy (activation) at time frame f. In many ways this concept of energy is analogous to that employed in the segmentation of speech into phonemes or words [11]. Let $\dot{\theta}_{f,g}$ denote the angular speed of the gth rotational degree of freedom at time frame f. Then we can define the body's motion energy as a weighted sum of the rotational limb speeds.

$$E(f) = \sum_{k=1}^{n} w_k \dot{\theta}_{f,g}^2 \qquad (1)$$

In essence, E will be large for periods of energetic motion and will remain small during periods of low motion energy. Figure 1 shows E for repeated knocking. Figure 2 illustrates how the observed energy peaks coincide with actions such as arm raises or individual forward and backward movements during the knock.

Local minima in E can be observed whenever the trajectory of the right arm changes direction. We can use these insights to segment a complex motion as follows.

1. Compute E for the whole motion sequence.
2. Threshold the signal at a threshold t. Mark all frames f for which $E(f) > t$.
3. Find all connected regions of marked frames and regard them as individual motion segments.
4. Extend the segments to the preceding and succeeding local minima of E.

Obviously, our choice of t has a major impact on the nature of the segments. Fod et al. use empirically derived thresholds. If this method is to be used in a general framework, however, we need an automatic way of finding an optimal t. We propose the following solution. For every pair (E, t_n) we obtain a number of segments by thresholding E at t_n. Let $numseg_E(t_n) = s_n$ be the function which computes the number of segments s_n for any such pair. Figure 3 shows *numseg* for the motion in Figure 1 and sampled at various thresholds. Our goal is to find a threshold which will exhibit all major motion segments (energy peaks) while filtering out small scale motions due to low-level signal noise. We note that noise is mainly registered during periods of low energy (e.g. between frames 250–300 and 450–500 in Figure 1). Let t_0 be an empirical noise threshold. Then the optimal threshold t_{opt} is defined as the threshold which maximises the number of major motion segments.

$$t_{opt} = \arg\max_t \{numseg_E(t)\} \qquad (2)$$

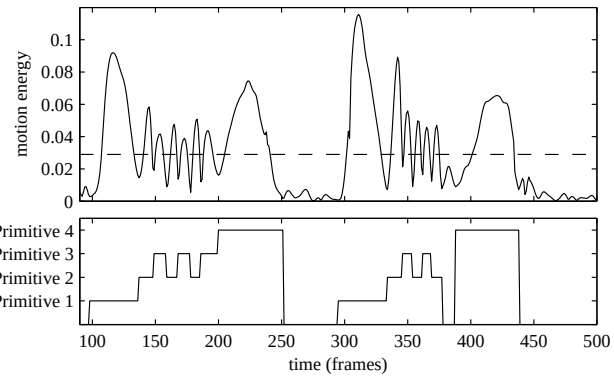

Figure 1. Objective function *E(f)* (top) with automatically calculated optimal segmentation threshold t_{opt} for part of a repeated knocking motion. The bottom shows the parse into four motion primitives and periods of no motion.

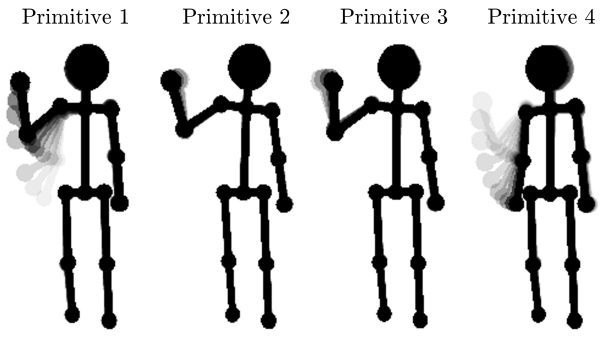

Figure 2. Four phases of a knocking motion exhibiting distinct peaks of motion energy. Each phase is detected as a segment and labelled with one of four automatically derived motion primitives. The primitives coincide with the semantically meaningful basic actions "Raise arm", "Knock", "Retract", "Lower arm".

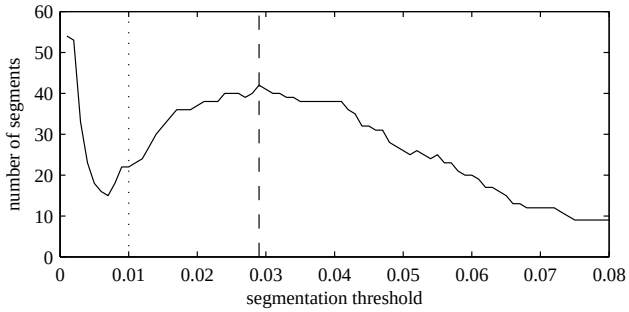

Figure 3. $numseg_E(t_n)$ for a repeated knocking motion. The diagram also shows t_{opt} (dashed) and t_0 (dotted).

3.2 Motion Primitives

Ideally, we would like to group the extracted segments into semantically meaningful clusters representing primitive motions. One approach to define such primitives would be to use a comprehensive list as devised by Birdwhistell or Bull to transcribe their anthropological and psychological observations [6][12]. Due to their generality, however, these sets are large. Many of the listed primitives are irrelevant for any particular context. Indeed, context often governs the affective and social meaning of movements [6]. We therefore adopt a more context-dependent approach to the definition of motion primitives. It is based on the clustering of a set of example motions which are representative for a certain context. For our current scenario the context is very specific (knocking) and therefore the number of motion primitives is rather small. In more complex scenarios such as "everyday activities" or "interpersonal conversations" we would expect to require a larger set of primitives to represent all observed movements well.

Consider the database of affective knocking motions described above. After segmenting the movements, we need to find a representation for the segments which allows us to compare and cluster them. We therefore consider the joint angles of the motions and time-normalise them. This is done by resampling each segment at 25 equally spaced intervals. We also subtract the segments' means in order to capture the relative motion rather than the absolute body configurations. Next, we wish to group the segments into semantically distinct categories. We hypothesised that the knocking motions can be divided into four basic phases: lift arm, repeatedly knock and retract, lower arm. We therefore used a simple k-means clustering algorithm with k=4. In a completely unsupervised scenario without any prior knowledge of the number of motion primitives, we would choose a clustering technique which automatically determines an optimal number of clusters such as hierarchical or Markov clustering. The following steps summarise our algorithm to compute a set of motion primitives from a set of example motions:

1. Segment the set of motions as described in Sect. 3.1.
2. Time-normalise all segments. Subtract sample means.
3. Cluster the normalised segments.
4. The clusters (or cluster centroids) represent the motion primitives.

Having defined our primitives, we can now parse a new motion by following steps 1 and 2 as outlined above and replacing steps 3 and 4 by an assignment to the closest cluster centroid (most similar primitive). Figure 1 illustrates how a repeated knocking motion (energy curve shown on top) has been parsed into a sequence of primitives (bottom). The motion is parsed in a semantically meaningful fashion. Figure 2 shows that primitives 1 and 4 correspond to the larger scale motions of raising and lowering the right arm while primitives 2 and 3 capture the smaller scale knocking motions. We will now turn to the analysis of the dynamic and affective parameters of the segmented motions.

4. Affect Recognition

Angry movements in the analysed database tend to look energetic and forceful while sad knocks appear relatively slow and slack. Similar observations are true for the other classes of motions such as throwing and walking. This role of dynamic movement qualities such as velocity and acceleration in affect recognition has been stressed by other authors [5]. Never before, however, has the analysis of dynamics been attempted at the level of motion primitives. We propose this solution as a more flexible and well-founded alternative to the use of fixed or sliding windows as used before.

We are employing four statistical measures as features for affect recognition. They are computed over a whole motion segment such as an arm raise. For the analysed knocking motions only the right arm exhibits significant movement. Therefore all dynamic features are currently based on the right arm. We define the features as follows.

- Maximum distance of hand from body (d_h)
- Average hand speed ($\bar{s}_h$)
- Average hand acceleration ($\bar{a}_h$)
- Average hand jerk ($\bar{j}_h$)

We can also compute analogous features d_e, $\bar{s}_e$, $\bar{a}_e$, $\bar{j}_e$ based on the elbow motion. For any person p and motion segment m this gives us the feature vector $\boldsymbol{\varphi}_{p,m} = (d_h, \bar{s}_h, \bar{a}_h, \bar{j}_h, d_e, \bar{s}_e, \bar{a}_e, \bar{j}_e)$.

4.1 The problem of individual movement bias

In Sect. 5 we shall show that our data does indeed reveal some global correlation between the above features and the different emotion classes. Figure 4(a), however, shows that the between-class variability of the two very different emotion classes sad and angry is smaller than we would hope. The hand speed distribution for sad knocks (black) overlaps heavily with that of angry knocks (white). In order to be separable through a pattern recognition approach, the two distributions should show a large between-class variability while exhibiting a small within-class variability. This exemplifies the problem of individual movement bias. Different people tend to display the same emotion in very different ways, thus impeding classification.

Our approach to this problem is a normalisation procedure based on the following intuition. It seems a reasonable assumption that a person's motion idiosyncrasies influence his or her movements in a consistent fashion — after all we expect them to be governed by gender, physical build and other constant factors. Even dynamic

factors such as mood might be changing slowly enough to be assumed temporarily constant. We therefore propose to model individual motion bias as an additive constant signature $\overline{\boldsymbol{\varphi}}_{\mathbf{p}}$ which influences the motion features introduced above. We obtain an estimate of the unbiased motion features $\hat{\boldsymbol{\varphi}}_{\mathbf{p,m}}$ by subtracting the personal bias.

$$\hat{\boldsymbol{\varphi}}_{\mathbf{p,m}} = \boldsymbol{\varphi}_{\mathbf{p,m}} - \overline{\boldsymbol{\varphi}}_{\mathbf{p}} \tag{3}$$

An important problem is how to estimate $\overline{\boldsymbol{\varphi}}_{\mathbf{p}}$. If we do not "know" a person, i.e. have no history of his or her movements, we may need to take an a priori guess, maybe conditioned on gender or other cues. However, if we have a history, we can compute $\overline{\boldsymbol{\varphi}}_{\mathbf{p}}$ from all the observed motions. In our case, we take an average over all the knocking motions in the database in order to learn about a person's motion bias. Note that this operation does not tell us anything about affect-specific factors as all motions are treated equally and different affects are represented at equal frequencies in the database. Figure 4(b) illustrates how this normalisation improves the between-class variability for the two shown classes considerably. Sect. 5 gives a more rigorous account of the improvements achieved when taking movement bias into consideration.

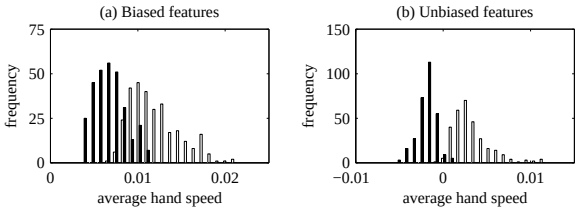

Figure 4. Biased and unbiased feature distributions for sad knocks (black) and angry knocks (white).

4.2 Machine Learning

We can use the biased or unbiased motion features to train a classifier which distinguishes the four emotions neutral, happy, angry and sad. We decided to use support vector machines (SVMs) with a polynomial kernel as they tend to exhibit good generalisation performance. The suitability of SVMs for this domain was demonstrated by Kapur et al. [5]. In order to solve the general problem of recognising the affect of a motion sequence, we train a family of binary SVMs $M^{z}_{x,y}$. The classifier $M^{z}_{x,y}$ aims to find the maximum margin between affect classes *x* and *y* for motion primitives of type *z*. Once these binary classifiers have been trained, we can classify a new motion as follows:

1. Segment the motion into a list of primitives as described in Sect. 3.
2. Let the first segment in the list be of primitive type *z*. Apply all pairwise SVMs $M^{z}_{x,y}$. Classify the segment according to a majority vote.
3. Remove the first segment from the list and repeat from step 2 until the list is empty.
4. Classify the whole motion by a majority vote of individual segment classifications.

5. EXPERIMENTAL RESULTS

With the conducted experiments we aimed to answer three questions:

1. What recognition rate can be achieved with our approach?
2. How does movement bias (see Sect. 4.1) affect the recognition performance?
3. How do our results compare to related results found in the literature?

We used the knocking motions from our database to run Leave-One-Subject-Out cross-validation (LOSO-CV) tests. Overall, we used approximately 1200 motion samples with an equal proportion for each of the considered emotions neutral, happy, angry and sad. For each iteration of the cross-validation the system was therefore trained on around 1160 samples and validated on 40 samples. In different tests we found that the system does considerably better if we add some of the remaining 40 samples to the training set or perform a subject-independent cross-validation. In contrast to those tests, the figures we quote here are representative for the generalisation performance of our system for an unknown person.

The confusion matrices for LOSO-CV using biased and unbiased features are shown in Table 1. Note that angry and sad knocks are classified more reliably than neutral and happy ones. The most significant factor which negatively affects recognition rates (sensitivity) is the confusion between neutral and happy knocks. In answer to question 2 above, we find that using unbiased features improves the overall recognition rate considerably from 50% to 81%. Our informal observations from Sect. 4.1 have hence been confirmed.

Table 1. Confusion matrices for LOSO-CV using biased features (left) and unbiased features (right). All average and affect-specific sensitivities are above chance level (0.25).

Truth	classified as			
	neu	hap	ang	sad
neu	**0.38**	0.23	0.13	0.27
hap	0.28	**0.41**	0.18	0.13
ang	0.18	0.20	**0.59**	0.03
sad	0.21	0.14	0.02	**0.62**
	average sensitivity: **0.50**			

	classified as			
	neu	hap	ang	sad
neu	**0.74**	0.20	0.01	0.05
hap	0.28	**0.65**	0.06	0.01
ang	0.01	0.06	**0.92**	0.00
sad	0.07	0.01	0.00	**0.92**
	average sensitivity: **0.81**			

We can obtain a measure for the more objective recognition efficiency η if we normalise the achieved sensitivity by the sensitivity expected by chance (sometimes referred to as generality).

$$\eta = \frac{\text{achieved sensitivity}}{\text{sensitivity expected by chance}} \tag{4}$$

In our case we would expect a classifier which assigns one of the four affect classes at random to achieve a sensitivity of 25%. Therefore the efficiencies of our classifiers for biased and unbiased features are $\eta_b = 2.0$ and $\eta_{ub} = 3.24$ respectively. We

can use these measures to compare our results to those of related experiments in the next section.

6. DISCUSSION AND FUTURE WORK

For our discussion we consider the results of two other related experiments. We were using part of a database which was created by Pollick et al. for psychological work. In one particular study they examined how accurately human subjects could classify affect from knocking motions displayed as point-light or full video stimuli [13]. The only major difference from our experimental setup was their forced choice between five rather than our four emotional states (afraid being the additional class). They report that humans achieved a recognition rate of 59% for point-light and 71% for full video stimuli. These figures illustrate that, although performing significantly above chance level, even humans are far from perfect at classifying affect from non-stylised body motions. We can calculate the efficiency $\hat{\eta}$ achieved by humans as defined in Eq. 4. For point-light and video displays humans exhibit efficiencies of $\hat{\eta}_{pl} = 2.95$ and $\hat{\eta}_{v} = 3.55$ respectively.

One of the major contributions of our work derives from the fact that classifying affect from non-stylised motions is harder than from stylised ones. This is demonstrated by the experiments performed by Kapur et al. [5]. They recorded stylised emotions and compared the accuracy of various machine learning techniques as well as human performance. For the task of distinguishing four basic emotions from point-light displays, humans achieved a recognition rate of 93% ($\eta = 3.72$). This is considerably higher than human performance reported by Pollick et al. for non-stylised movements ($\hat{\eta}_{pl} = 2.95$). For SVMs the recognition rate was lower at 83.6% ($\eta = 3.34$). These results are summarised in Table 2.

We have shown that using unbiased dynamic features based on motion primitives boosts the recognition rate considerably. Our computational approach exhibits a better efficiency than humans for classifying affect in non-stylised movements from point-light displays. The performance of our approach is also comparable to that of Kapur et al. This is significant since their stylised motion data contained solely affective information. For our non-stylised motions, on the other hand, only certain subtle aspects communicate affect while most of the motion signal is governed by the independent semantic meaning of the motion.

Table 2. Comparison of our and related results.

experiment	Kapur et al. [5]		Pollick et al. [13]		Our results	
motions	**stylised**		**non-stylised**		**non-stylised**	
classifier	human	SVM	human		SVM	
features	biased		Pt.-l.	video	biased	Unbiased
#emotions	4	4	5	5	4	4
sensitivity	93%	84%	59%	71%	50%	81%
efficiency	**3.72**	**3.34**	**2.95**	**3.55**	**2.00**	**3.24**

We are currently working on extending our approach in various ways. In the version described here we only consider the right arm for extracting affect-related dynamic features. Incorporating features from other body parts will help us to analyse motions such as walking, which are not primarily based on arm movements. Furthermore, the torso and head can be expected to hold valuable cues even for heavily arm-based actions. We are also investigating better ways to estimate the personal movement signature $\overline{\boldsymbol{\varphi}}_{\mathbf{p}}$. Early results suggest that our current formulation is relatively consistent across different types of motion. Ultimately we would like to be able to estimate a person's movement signature from as few and unconstrained example motions as possible.

7. REFERENCES

[1] Heeter, C. *Being there: the subjective experience of presence.* Presence: Teleoperators & Virtual Environments 1992, 262–271.

[2] Garau, M., Slater, M., Pertaub, D., Razzaque, S. *The Responses of People to Virtual Humans in an Immersive Virtual Environment.* Presence: Teleoperators & Virtual Environments 2005 14:1, 104–116.

[3] Gillies, M., Slater, M. *Non-verbal Communication for Correlational Characters.* Proceedings of The 8th Annual International Workshop on Presence, London, September 2005.

[4] Kleinsmith, A., De Silva, P. R., Bianchi-Berthouze, N. *Grounding Affective Dimensions into Posture Features*. Affective Computing and Intelligent Interaction 2005, Beijing, October 22-24, 2005, 263–270.

[5] Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., Driessen, P.F. *Gesture-Based Affective Computing on Motion Capture Data.* Affective Computing and Intelligent Interaction 2005, Beijing, October 22-24, 2005, 1–7.

[6] Birdwhistell, R.L. *Kinesics and Context: Essays on Body Motion Communication.* University of Pennsylvania Press (1970).

[7] Ekman, P., Friesen, W.V. *Facial action coding system.* Consulting Psychologists Press (1978).

[8] Mussa-Ivaldi, F. A., Giszter, S.F., Bizzi, E. *Linear combinations of primitives in vertebrate motor control.* Proceedings of the National Academy of Sciences USA, 1994, 7534–7538.

[9] Ma, Y., Paterson, H.M., Pollick, F.E. *A motion capture library for the study of identity, gender, and emotion perception from biological motion.* Behavior Research Methods 38 (2006), 134–141.

[10] Fod, A., Mataric, M.J., Jenkins, O.C. *Automated derivation of primitives for movement classification.* Autononmous Robots 12(1), 2002, 39–54.

[11] Wang, D., Lu, L., Zhang, H.J. *Speech segmentation without speech recognition.* IEEE International Conference on Acoustics, Speech and Signal Processing, 2003, 468–471.

[12] Bull, P.E. *Posture and Gesture.* Volume 16. Pergamon Press (1987)

[13] Pollick, F.E., Paterson, H.M., Bruderlin, A., Sanford, A.J. *Perceiving affect from arm movement.* Cognition 82, 2001, 51–61