Ranking-Based Affect Estimation of Motion Capture Data in the Valence-Arousal Space

William Li
Simon Fraser University
dla135@sfu.ca

Omid Alemi
Simon Fraser University
oalemi@sfu.ca

Jianyu Fan
Simon Fraser University
jianyuf@sfu.ca

Philippe Pasquier
Simon Fraser University
pasquier@sfu.ca

ABSTRACT

Affect estimation consists of building a predictive model of the perceived affect given stimuli. In this study, we are looking at the perceived affect in full-body motion capture data of various movements. There are two parts to this study. In the first part, we conduct groundtruthing on affective labels of motion capture sequences by hosting a survey on a crowdsourcing platform where participants from all over the world ranked the relative valence and arousal of one motion capture sequences to another. In the second part, we present our experiments with training a machine learning model for pairwise ranking of motion capture data using RankNet. Our analysis shows a reasonable strength in the inter-rater agreement between the participants. The evaluation of the RankNet demonstrates that it can learn to rank the motion capture data, with higher confidence in the arousal dimension compared to the valence dimension.

CCS CONCEPTS

• **Computing methodologies** → *Modeling methodologies*;

KEYWORDS

Affective computing, motion capture, machine learning

ACM Reference Format:
William Li, Omid Alemi, Jianyu Fan, and Philippe Pasquier. 2018. Ranking-Based Affect Estimation of Motion Capture Data in the Valence-Arousal Space. In *MOCO: 5th International Conference on Movement and Computing, June 28–30, 2018, Genoa, Italy.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3212721.3212813

1 INTRODUCTION

In the recent growing interest of developing technology to recognize people's affective states [15], more and more studies have shown that body expressions are effective in conveying emotion [3, 40]. Combined with the increase in the volume of sheer amounts of data, there is an increasing demand for the development of affect recognition systems which in turn has potential impacts in clinical and entertainment contexts. Thus, we developed an affect ranking system using the valence-arousal (VA) model of human emotion, and used full body motion capture data as input, which does not contain any information regarding facial expressions or voice. When considering the three aspects of movement, functional (the task of the movement, such as picking up a cup), executional (the pattern of movement, such as using the left or right hand to pick up the cup), and expressive dimensions (the emotions behind the movement) [1], we are essentially measuring the expressive dimension of full body movements.

The ground-truthing experiment was conducted using a ranking system in which we ask participants to rank the relative valence and arousal for different pairs of movements. This results in a complete relative ranking of the movements. The machine learning models are trained on this ranking.

The contribution of this paper is a first step in the processing and affect estimation of a large amount of motion capture data, leading to future off-line and on-line applications. The main off-line applications involve database labelling, which is especially useful for the development of movement databases [32]. A valid and reliable ranking model for movement expressivity would allow us to automatically label existing motion capture data according to the valence-arousal model. In on-line scenarios, such a model could be used in interactive arts or therapy contexts. Such a system can also be used in generating movement with user-specified valence and arousal [1]. Our goal is therefore to estimate affect expressed by movement using the VA model, specifically a meaningful rank on the VA spectrum relative to the other data.

For the rest of the paper, we start by outlining the related work in affect classification. After that, we describe the data and the processing used in our study, followed by experimental methods, participants, results, and analysis for each iteration separately. Lastly, we end with concluding remarks and future work.

2 PREVIOUS WORK

2.1 Affect Estimation

In affect estimation, considerations that come into play include the intended emotion of the mover, and the perceived emotion of the mover [24, 29]. Malandrakis et al. [27] have shown that there can be a difference even in award-winning movies in the intended and experienced emotions. How well the intended emotions are portrayed plays an important role in the movies. With their experiment, they used award-winning films and expert annotators to narrow the gap between the intended and expressed emotions. In our experiment, we are able to direct the movers, so we assume the intended affect is identical to the perceived affect.

MOCO, June 28–30, 2018, Genoa, Italy

ACM ISBN 978-1-4503-6504-8/18/06…$15.00
https://doi.org/10.1145/3212721.3212813

In the field of affective computing, facial expressions are often examined in the determination of affective states [13, 19]. However, Inderbitzin et al. [21] have shown that it is possible to perceive VA states from movement even on faceless generated characters, regardless of viewing angle. They have even identified some canonical parameters that control the expression of emotions in locomotive behavior, such as upright upper body postures being perceived as more emotionally positive and vice-versa for forward leaning postures. Other documented sources also suggest that humans convey emotions through body movement and postures [10, 11]. Analysis of head pose and movement is able to achieve 71.2% accuracy in recognizing depression [2]. Furthermore, studies in movement have shown certain features in expressive movement, such as portrayal of strength, can be linked to specific emotions, such as fear or anger [11, 41].

In affect estimation based on body movement, there have been many studies in using dance with mixed results ranging from barely above random chance to close to human levels of accuracy [8, 22, 33]. Kapur et al. developed classifiers that achieved comparable accuracy as observers using dance movements [23]. However, as Kleinsmith points out, dance is often exaggerated to convey affect [24].

Looking at non-dance-based systems, Castellano et al. have attempted to infer emotional states using video analysis on movement qualities such as amplitude, speed, and fluidity. Their system was able to discriminate between "high" and "low" arousal emotions and "positive" and "negative" [9]. Pollick et al. conducted a study to compare the performance of their automatic system with human recognition. In their study, they used 3D positioning measurements of the arm in knocking, lifting, and waving motions with two affective states, neutral and angry. They concluded that the automatic system was able to discriminate between the two states more consistently than humans [34]. Samadani et al. developed a system for both full body as well as hand-arm improvisiation movements to discriminate between 4 affective states using HMMs with good results [38].

Nicolau et al. developed a system for estimation of affect modalities in the Valence-Arousal space using multi-modal inputs (based on facial expression, shoulder gesture, and audio cues). Their approach claims to be unique in that it performs *continuous* affect prediction according to the valence-arousal model. They compare both Support Vector Machines (SVM) and bi-directional Long-Short-Term Memory Neural Networks (BLSTM-NN), concluding that BLSTM-NN performs better [31]. However, we have decided not to use BLSTM-NN due to the fact that they are using different sets of input data (extracting data from video and audio as well as mainly focusing on facial expressions); in our case we are using motion capture data with no facial expressions. Furthermore, the lack of a benchmark and standard skeleton markers due to the use of different datasets and body markers in the aforementioned studies makes it difficult to compare and evaluate different systems.

2.2 Ranking and Rating

Most of the previous research have used a rating system. However, Yannakakis et al. point out some limitations to using ratings in 2015 [42]. Firstly, inter-personal differences including cultural background and experiences can lead to different perceptions of affects. What appears to be happy to one person might appear neutral to others. Similarly, Baveye et al. in 2014 [4] point out that using ratings require the participant to understand the full range of the valence and arousal scale of the data, which is usually not feasible. Secondly, Yannakakis et al. argue that using adjectives such as "moderately" and "extremely" are not numbers, and thus any method that treats them as numbers such as average values or t-tests are fundamentally flawed. This also ties in with the third issue they point out that ratings are not always linear.

Therefore, we have chosen to use a ranking system for classification. Ranking approaches are easier in terms of cognitive load and have a higher inter-rater agreement [4]. Yannakakis et al. [42] also claim that using rankings eliminates the cultural and subjective biases in the annotation. However, a disadvantage to using rankings is that the ranks do not indicate the distance between them. For example, we know video A has higher valence than video B, but the amount by which it is higher is not clear.

There are three approaches to the learning to rank: pointwise, pairwise, and listwise [25]. In the pointwise approach, the model predicts a score for a single input item. In the pairwise approach, the model is given two input items, and ranks them accordingly. In the listwise approach, a set of items is given to the model, in which it outputs a ranked list of the input items. As in a similar experiment by Fan et al. [12], we will be going with a pairwise approach, as it produced the best results.

There are a number of systems used in the literature for pairwise ranking, including SortNet [36], RankNet [7], RankSVM [20], and RankBoost [16]. We use RankNet as it is an efficient model for working for high-dimensional motion capture data.

2.3 Model of Affect

For our study, we will be using Russell's model of affect [37]. A potential drawback of Russell's model is that some researchers such as Fontaine et al. [14] are starting to believe more dimensions are needed to describe the emotional space, such as the PAD (Pleasure-Arousal-Dominance) model put forth by Mehrabian et al. [28], which includes the addition of the dominance dimension. Dominance refers to whether an affect is controlling, submissive, or otherwise influenced by something else. For example, if a person is subjecting themselves to the command of another person or feel pressured or otherwise controlled by another entity, this would be submissive on the dominance dimension. We did not include the dominance dimension in our experiments because all of our movement are performed individually without another body or objects in the scene. Thus the concept of dominance does not make sense in the context of our motion capture data.

Another potential drawback is that Schacter et al. [39] claimed that the physiological reactions contribute to the emotional experience by facilitating a cognitive recognition of a physiologically stimulating event, which then defines the emotional experience. The emotion is the result of a combination of the cognition of a physiological event and the participants' reception of adrenaline. For our studies, we do not address nor do we have control over either the actors' or viewers' physiological states. With our actors, we are assuming that they are able to act the specified emotional state despite their internal physiological states. With the viewers, we are assuming their physiological states do not drastically affect their perception of the movements while watching the animation as it is a low-stimulus activity.

Other than dominance, there is not yet a well-established set of dimensions in addition to valence and arousal that are considered necessary in order to describe affect. Determining such a set of

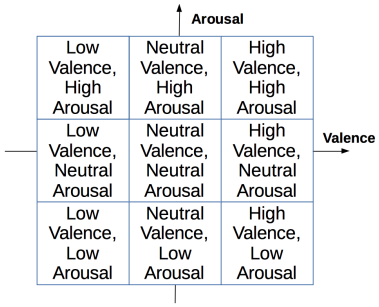

Figure 1: Valence and arousal combinations

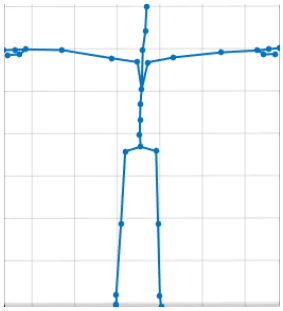

Figure 2: MoCap skeleton

standards is outside the scope of this study. Therefore, we will be using just valence and arousal as the basis for our model of affect, similar to other previously cited studies. However, we will keep in mind that as research in the dimensions of affect progresses, our experiments will potentially need to be replicated or our models tweaked to account for additional dimensions or new models of affect. To our knowledge, automatic systems in affect estimation using motion capture data has not yet been attempted in the context of a dimensional model of affect.

3 DATASET

3.1 Motion Capture Data

As part of the efforts of the MovingStories project, an open source MoCap database[1] [32] has been created. For this study, we are using some of the recordings in this database that have been labelled according to the circumplex model of affect [37]. The data are in the form of MoCap bvh files and accessible in the MoCap database[2]. The key point for this experiment is that only the skeletal information is retained in the end. There are no facial expressions and nothing that explicitly indicate gender, body type, or ethnicity. Three professional actors, one male and two female, performed in the data collection stage. Two of the actors performed 9 different types of movements: walking in a figure eight pattern, hugging, static improvisation, free improvisation, sitting down, pointing while sitting, walking with sharp turns, improvisation while facing another actor, and lying down. Improvisation refers to a movement that is at the discretion of the mover. In other words, they are given the direction of acting in a certain affect, but are free to carry out whatever movement they wish that they believe would illustrate that affect. The difference between free and static improvisation is that in static improvisation the only additional restriction is that they must remain standing in the same spot. The third mover performed only the two walking movements. There are 9 takes for each movement, corresponding to the 9 different possible VA combinations shown in Figure 1, covering more emotional states than similar existing datasets (e.g. 4 emotions in the library presented by Ma et al. [26]). Existing labels were created by dividing the Russell's model [37] into low, neutral, and high along both the valence and arousal axis. Using this model, anger would be classified as low on the valence axis but high on the arousal axis.

[1]http://moda.movingstories.ca/
[2]http://moda.movingstories.ca/projects/29-affective-motion-graph

The data were recorded with a Vicon motion capture system[3] and mapped to a skeleton representation with 30 joints as shown in Figure 2. Eighteen of the motion capture files were recorded at 60 frames-per-second, while the rest were recorded at 120 frames-per-second. Therefore, we downsample 120fps files to 60fps. Sequences vary in length from 1000 frames to 8000 frames. Each frame contains the rotations for each of the joints in the Euler representation, as well as the spatial location and orientation of the skeleton root.

3.2 Pre-Processing

We convert the rotational data from Euler representation to the exponential maps [18] as suggested in the literature for training neural networks on motion capture data. After removing empty dimensions as well as orientation and translation of the skeleton's root to eliminate bias due to geometrical translation, we are left with a 46-dimensional vector per frame. We further concatenate a window of consecutive frames into one feature vector to flatten the time dimension. We experiment with window sizes of 1, 3, and 12 frames. The intent behind using these window sizes is to test a variety of window sizes and determine if there is a trend in the performance of the model relative to the window size. However, we acknowledge that at these window sizes, the duration captured is very short. It is closer to a snapshot of the movement rather than fully capturing the temporal aspect of the movement. Based on previous works, it is not clear at this time what sort of high-level features would be useful in recognizing affect in full-body motion capture data that would persist over a long period of time. Depending on the models, long window periods can also significantly increase processing time, which would be detrimental to most practical applications. Hence we test using small window sizes. Converting to real time using the frame rates, these window sizes account for 0.1 second or less of the movements. Therefore, the window sizes are not realistic to the perception or reflex of humans. From the perspectives of a machine, these window sizes are enough to see a trend in the performance of the models. Furthermore, building systems using smaller window sizes will also be more advantageous in any application that relies on real-time affect estimation in order to give a prediction as fast as possible.

4 DATA COLLECTION

We conduct our survey on the CrowdFlower platform. Full documentation for this platform can be found at the CrowdFlower

[3]https://www.vicon.com/

website[4]. We chose this platform due to a similar successful study in affective rankings. Baveye et al. [5] conducted a study where 9800 videos were ranked on CrowdFlower in terms of valence and arousal. Their goal was to provide rankings for the affective video database LIRIS-ACCEDE. We obviously cannot use this database as our videos are motion capture data, but as our goals are similar, we have chosen to use the same experimental protocol that they have.

The CrowdFlower platform was chosen because it reaches several crowdsourcing services, which also allows for a good distribution of demographics [4]. However, the number of labor channels have been reduced in the last few years. It is no longer disclosed by CrowdFlower where they may distribute their tasks nor is it possible to choose specific channels. The advantages of using crowdsourcing platforms like CrowdFlower include supported infrastructure in both the survey design and payment. For example, CrowdFlower offers templates for a variety of surveys that are all fast and easy to implement. This reduces the potential problems that may come up from the researchers having to host their own online survey such as website or server issues. Another concern with gathering large number of responses is the payment. Most people would not want to answer a long survey for free. Using a well-known crowdsourcing platform also ensures that financial transactions are trustworthy and transparent. We choose CrowdFlower also because it is possible to limit the survey to users that have shown to give quality responses on other studies conducted on CrowdFlower.

For the CrowdFlower platform, each worker has a discrete trust level associated with his or her account ranking from level 1 to level 3. Participants can raise their trust level by completing jobs on the CrowdFlower platform successfully and without failing. A participant fails a study if they ever drop below a certain level of accuracy (70% by default) on pre-defined test questions. However, the exact algorithms for determining user trust levels are internal to CrowdFlower and not visible to us. We are only able to specify the trust level to which our survey is available. Every job on CrowdFlower will have a pre-survey quiz to ensure participants understand the task. Only participants that achieve at least 70% will be allowed to participate in the survey. Furthermore, throughout the survey, there will be a random test question on every survey page to ensure the participant stays focused throughout. Participants gain trust levels by passing these quizzes and test questions and consequently lose trust levels by failing. If they feel a particular test or survey question was unfair, they are able to challenge it and provide a written response that we can monitor in real time and change the questions accordingly while the survey is still live for other participants. We have chosen to limit our survey to only the highest trust level participants at level 3. This is to ensure as much as possible that our responses come from participants who are experienced with the platform and have shown themselves to be trustworthy (ie. do not click through surveys randomly) from past studies on CrowdFlower.

5 METHODS

5.1 Participants

The participants of the survey were users of the CrowdFlower platform from all over the world, including but not limited to countries such as USA, Brazil, Ukraine, Poland, Turkey, Russia, France, Egypt, Mexico, and India. However, as many users choose not to disclose their nationality, we do not have the complete statistics on the locations of the participants. All participants are completely anonymous and only trackable in our experiment via worker ID. The participants are paid $0.02 CDN per comparison once they pass the pre-survey quiz. If a participant drops below 70% accuracy on test questions, they will be kicked from the survey, but still be paid for the comparisons they have done so far. Their responses up to that point will be used. Lastly, participants are free to quit the survey at any time. However, they will only be paid for each page of comparisons for which they have clicked "Submit". After the survey, participants have the option of providing feedback in the clarity of the instructions, easiness of the task, compensation, and overall experience.

There was a total of 1263 trusted annotators from 65 countries, with the majority of the workers coming from Venezuela (24.5%), Brazil (6.7%), Serbia (6.6%), Turkey (6.0%), Russia (5.6%), and Bosnia (5.1%). The trusted annotators had an accuracy of 93.6% on the quiz and the test questions throughout the survey. There were a total of 103 untrusted annotators who were kicked from the survey and 109 people who failed the pre-survey quiz, and thus did not participate in the study. There were a total of 19848 submitted comparisons by trusted annotators. Annotators spent an average of 15 seconds per comparison. This is a reasonable amount of time as the animation clips are being played side by side simultaneously and are 10 to 25 seconds long.

5.2 Procedure

Similar to Baveye et al. [5], we are using a quick-sort algorithm to rank our motion capture animation clips. The reason for this is to cut down the number of comparisons needed to significantly reduce the cost. We have 9 takes for each movement for a total of 181 motion capture clips. Surveying participants on every possible pairwise combination would result in $N(N-1)/2$ comparisons. The idea behind the quick-sort algorithm is that at the first iteration, a specific MoCap is chosen as the pivot. Every other MoCap is compared to the pivot element, creating two subgroups. The assumption is that everything in the subgroup that was considered to have a lower affect than the pivot is also lower in affect than everything in the subgroup higher in affect than the pivot. However, we never compared MoCap from the two subgroups directly. Everything was only compared with the pivot. Then two pivots are chosen within the subgroup in the next iteration. Everything in the first subgroup is then compared with the first subgroup pivot, and vice-versa for the second pivot. This results in an average of $O(nlogn)$ comparisons rather than the full $N(N-1)/2$ for every possible pair-wise comparison. We continue dividing until the size of every subgroup is no more than 5-10. We treat this last subgroup size as having equal affect. We stopped at this subgroup size because we have many improvisation movements that appear to be difficult for people to distinguish a relative rank. Many comparisons at this stage were decided with 3:2 votes, no better than random. We ended up using 9 pivots for valence and 8 pivots for arousal. Using this method, we paid $346.55, saving about $2800-$3000 for this experiment.

The survey begins with an explanation of the concept of valence and arousal with accompanying examples of emotions on both dimensions. The scale and accompanying example for valence is shown in Figure 3. Likewise, the example for arousal is shown in Figure 4. The overlapping terms are to provide additional examples

[4] https://success.crowdflower.com/hc/en-us

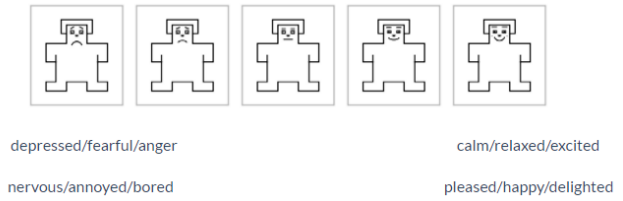

Figure 3: Valence scale example

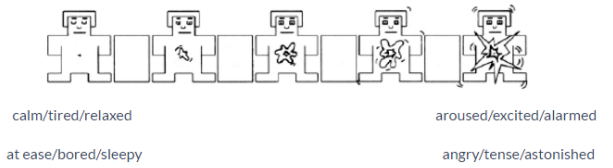

Figure 4: Arousal scale example

of affects that are considered to be low or high valence and arousal. Due to space constraints they were placed on different lines as to maintain a clear distinction of the low VA examples and the high VA examples. The characters in the images are from the SAM (Self-Assessment Manikin) scale [6], a popular set of standard images to convey the spectrum of affect used in research. Even though this is a ranking experiment and not a rating experiment, these images were chosen in addition to the adjectives shown to help the particpant understand the concept of affect.

The participant then takes the pre-survey quiz to ensure they understand the concepts. Animations chosen for the quiz are obvious examples of difference in valence or arousal in the sense that they were extreme comparisons such as high valence high arousal compared with low valence low arousal. An explanation is provided for each question in case the participant picks the wrong answer. The animation itself shows both motion captures simultaneously. The particpant is able to zoom in or out and pan around the scene as they wish to explore the 3D nature of the movement. An example of the interface is shown in Figure 5.

The animation clips used range in length from about 10 to 25 seconds of three different professional actors as described in Section 3.1. Each animation shows a mover performing one of our recorded movements mentioned in Section 3.1. There are 181 different motion capture clips. We collect a total of 5 responses for each comparison. Similar to Baveye et al. [5], we choose an odd number of comparisons so that each pair is guaranteed to have a distinction as to which has the higher affect. However, we choose 5 comparisons as opposed to the 3 comparisons from Baveye to reduce the likelihood of pairs getting close votes by chance or pairs getting ranked in the wrong order. A pair can be ranked incorrectly if two people happened to not pay attention on a particular question. With 5 comparisons, three people need to be not paying attention for a pair to be ranked incorrectly. All of the above procedure is performed twice, once for valence and once for arousal.

Table 1: Distribution of ranks for the 181 motion capture sequences

	Arousal						
180	0	0	1	3	6	21	Valence
150	0	1	1	10	9	9	
120	0	4	6	11	8	1	
90	0	5	16	4	5	0	
60	4	16	6	2	2	0	
30	26	4	0	0	0	0	
Ranks	30	60	90	120	150	180	

6 EXPERIMENT AND RESULTS

6.1 Inter-annotator Reliability

For the ranking experiment, we first look at the inter-annotator reliability. Unreliable responses are filtered by the trust level settings, pre-survey quiz, and random test questions. In the experiment of Baveye et al. [5], they collected 3 responses for each comparison and they measured Inter-annotator reliability using percent agreement, Krippendorff's alpha, Fleiss' kappa, and Randolph's kappa. We conduct the same metrics except Fleiss' kappa in our experiment as a comparison. The results are presented in Table 2. Krippendorff's alpha is a flexible metric for measuring inter-annotator reliability in that it allows for comparions being made by any number of participants and missing data. Randolph's kappa is an alternative to Fleiss' kappa, allowing for more flexibility in the distribution of responses [35]. We did not include Fleiss' kappa because it assumes there will be a certain number of responses for each category. Thus we felt Fless' kappa was not suitable for our experiment.

Similar to Baveye et al., our reliability results indicate that agreement is better than what would have been expected by chance. The agreement from the experiment of Baveye et al. were similar to Malandrakis et al. [27] and Mohammad et al. [30]. Our percent agreement Krippendorff's alpha, and Randolph's kappa were found to be lower than the experiment by Baveye et al., about 5% percent agreement and 0.05 in both Krippendorf's alpha and Randolph's kappa. This suggests it is harder to rank the valence and arousal of motion capture sequences than videos, which is not surprising. Videos contain facial expressions and sound. In the case of the LIRIS-ACCEDE database presented by Baveye et al., the videos are excerpts extracted from movies. Agreement in participant responses may be in part due to recognition of those movies, in which case the participants have the context of the entire movie from which to draw their sense of perceived affect. For example, if all participants recognize the video as an excerpt from a happy movie, they would perhaps be more likely to rank it as higher valence than an excerpt from a sad movie, even if the excerpts might be similar in affect.

In Table 1 we present a distribution of the rankings. As an example, the cell in the bottom left illustrates the number of sequences with a rank of between 0 and 30. The interval size of 30 was chosen to evenly divide all the motion capture clips into sufficient bucket sizes in order to see an overall pattern or trend in the distribution. The sparseness of the top left and bottom right of the graph indicates that there were little to no motion captures that were ranked as either high valence and low arousal or low valence and high arousal. The rankings are concentrated along the $y = x$ line, indicating that valence rises with arousal. This suggests that

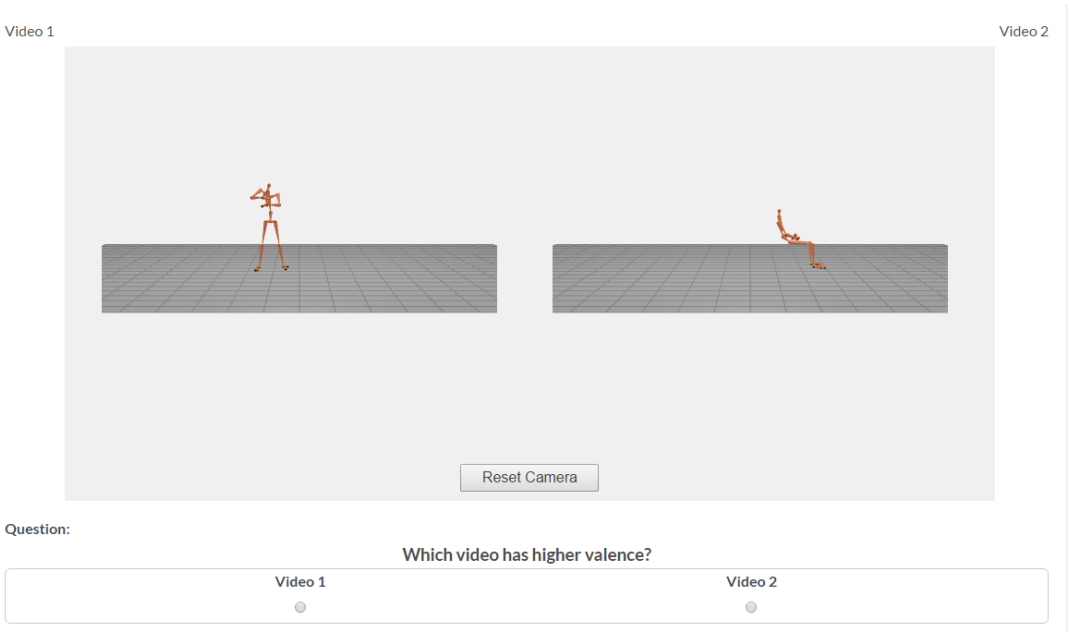

Figure 5: Screenshot of video in survey

	Valence	Arousal
Percent agreement	80.9%	81.2%
Krippendorff's alpha	0.096	0.157
Randolph's kappa	0.370	0.419

Table 2: Inter-annotator reliability

either it is difficult for the actors to portray valence and arousal independently through motion capture, or people in general have difficulty perceiving valence and arousal independently.

6.2 Learning to Rank Method

We use RankNet [7] to train a model to rank motion capture segments based on their valence and arousal levels, that match the rankings provided by the survey. RankNet is a pairwise rank-learning model that consists of a neural-network and a probabilistic loss-function that aims at minimizing the number of rankings in the wrong order. It can be trained using stochastic gradient descent.

6.2.1 RankNet. Burges et al. [7] proposed a probabilistic cost for training systems to learn ranking functions using pairs of training examples. From the training examples they attempt to learn a ranking function that does not map to a particular rank value. For example, if A is ranked higher than B, the system just needs to be able to determine that A is ranked higher, but it does not assign a value to the rank of A or B when learning the training samples. Burges et al. then chose to implement their probabilistic cost function in a neural network. The cost function is a function of the difference of the ranking outputs of two consecutive training samples, ie. what the system thinks is the difference in ranks between two samples versus the actual difference in ranks. In the case of consecutive samples, the true difference in ranks would always be 1. A forward propagation is performed on the first sample, storing the activation of each node in the network and the gradient value. Then the forward propagation is performed on the second sample, again storing the activation and the gradient. The cost then is the difference between the gradients of the two samples. Through learning to minimize the difference of the gradient, RankNet models the training samples in a monotonically increasing order.

6.2.2 Performance Analysis. We evaluate the ranking performance of RankNet using the Goodman-Kruskal gamma [17]. The Goodman-Kruskal gamma is a measure of rank correlation between two variables, the ground truth and the predicted ranks in our case, and is identified by G as follows:

$$G = \frac{(N_s - N_d)}{N_s + N_d} \tag{1}$$

where N_s represents the number of cases that are ranked in the same order on both variables, and N_d represents the number of cases that are ranked in the reverse order. This measure ignores ties. In our case, we did not have any ties. A G close to 1 shows strong agreement in rankings between the two variables, while a G close to -1 shows a strong disagreement in the rankings. A G close to 0 shows the rankings are independent and not associated with each other.

We use a neural network with two hidden layers, with 600 and 300 Rectifier Linear Units (ReLU) each, respectively and chosen from cross-validation to give the best results. We train the model on different subsets of the data based on the movement type and the performer. We also experiment with different window sizes. The windows of frames are concatenated in a single feature vector and then fed into RankNet. This is to cover different lengths of temporalities of movement. We stop learning after 3000 epochs. This value was chosen because we notice that after 3000 epoches the

Table 3: The Goodman-Kruskal Gamma for Valence Rankings ($p < 0.0001$)

	Window Sizes (all performers)		
	W = 1	W = 3	W = 12
Walking	0.664982	**0.695**	0.668
Sitting	**0.543481**	0.535	0.537
Hugging	0.205258	0.217	**0.245377**
Pointing	**0.548836**	0.544	0.548
Lie Down	0.600258	**0.602**	0.589
Free Improv	0.603071	**0.608**	0.574696
Improv Facing	**0.731329**	0.714	0.713
Improv Static	0.274803	**0.291**	0.262
All Movements	**0.582195**	0.561	0.536

Table 4: The Goodman-Kruskal Gamma for Arousal Rankings ($p < 0.0001$)

	Window Sizes (all performers)		
	W = 1	W = 3	W = 12
Walking	0.757263	**0.763**	0.741
Sitting	**0.852446**	0.846	0.789
Hugging	**0.938629**	0.914	0.878
Pointing	**0.903560**	0.884	0.853
Lie Down	**0.883156**	0.878	0.821
Free Improv	**0.863757**	0.862	0.826
Improv Facing	**0.878537**	0.859	0.831
Improv Static	**0.866419**	0.834	0.832
All Movements	**0.6271973**	0.623	0.606

loss value was no longer changing. We use 10-fold cross-validation on our training data to report the model accuracy. The G values are summarized in Table 3 For valence rankings, and in Table 4 for arousal rankings. The results from models that are trained on the data from only one performer each time is reported in Table 5.

However, there are a few factors of which we should make note. For the walking movements, we can consider a single walking cycle to be regarded as a single instance of that specific movement type and we have a few cycles of those per affect and mover. For all other movement types we only have a single instance. So however many of the frames for validation we set aside, those frames belong to the same performance. Thus, any conclusions about the generalization of the model beyond this dataset is unreliable. In other words, with the data we have, we are unable to test our machine learning models properly, except maybe for the walking movements.

The fact that many of the highest gammas came from a window size of 1 suggests that this method, or at the very least this configuration of neural network, is more effective for postures than movement. The high gamma for performer 3 suggests that she was able to act out her intended affect most effectively. However, this may also be due to performer 3 only acting in the walking movements, which as was just mentioned is the most reliable movement. Furthermore, we see that arousal is much more consistent in having a higher gamma in the window size of 1 as well as having higher gamma than valence across the board. This is again in accordance with the suggestion that arousal is easier to recognize and learn, both by human viewers as well as a machine. Looking at the different movements in Table 3, we see that that the hugging and improv static movements have resulted in a much lower gamma than the others. Compared to the others, these two types have the least amount of movement, suggesting that changes in postures or some other movement characteristics that we are unaware of are needed in order to recognize valence with a machine.

7 CONCLUSION

We first conducted a groundtruthing experiment on CrowdFlower, a crowdsourcing platform, where we surveyed participants from all over the world. We then constructed a machine learning model using RankNet to predict the relative ranks for pairs of motion capture clips in a variety of movements. These results and contributions are a first step in future experiments in affect estimation with large amounts of motion capture data, leading to use cases such as database labelling and movement generation.

Table 5: The Goodman-Kruskal Gamma for Valence and Arousal Rankings - Individual Performers ($p < 0.0001$, $W = 3$)

	P1	P2	P3
Valence	0.496008	0.587459	**0.647168**
Arousal	0.638740	0.667621	**0.750769**

In the CrowdFlower survey, participants were given the definition of valence and arousal as well as training examples. They then watched pairs of motion capture clips and were required to answer which within the pair had higher valence or arousal. These rankings were used in a quicksort algorithm to establish the relative ranks of 181 motion capture clips, containing 9 different movements. The survey showed better than random percent agreement but slightly lower compared to video-based affect recognition surveys. This is to be expected as it is not surprising that the lack of facial expression would result in an increased difficulty to distinguish affect.

Our experiments with RankNet show that it is possible to build machine learning models to learn the relative ranking of motion capture clips, with Goodman-Kruskal gamma of 0.62 to 0.93 in case of arousal rankings, and 0.24 to 0.73 in case of the valence rankings. The performance highly depends on the movement type, as well as the performers in that, having consistent movement patterns in the training data (i.e., same movement type and same performer) improve the chances that RankNet can effectively learn from them.

Another observation is that ranking predictions for valence were more or less lower than those for arousal. This is consistent with the findings from the survey in which higher number of people failed the valence pre-survey quiz, as well as the lower post-survey ratings for valence.

We understand the current limitations of our learning to rank approach. First, there is only one repetition of each combination of movement type/performer/affective state. This limits the amount of variation in the data and the ability of the model to generalize. We plan to gather more data and perform larger ground-truthing studies. Second, RankNet does not take into account the temporality of the data. As we see in the results, the model achieve higher precision, in most cases, with a window size of 1, which means that the model is essentially learning the postures alone and not the dynamic qualities of the movement.

We are interested in exploring training more effective machine learning models such as Recurrent Neural Networks. Furthermore, as mentioned in Section 6.2.2, the hugging and improv static movements have resulted in significantly lower gamma compared to the other movements. An area of exploration is establishing a set of rules or characteristics for different movement types that will determine whether the movement would be easy for affect estimation by machines. It may also be worth considering time-series analysis techniques instead of neural networks to try to account for the temporal aspect of movement.

ACKNOWLEDGMENTS

We would like to thank the Social Sciences and Humanities Research Council for funding in this experiment.

REFERENCES

[1] Omid Alemi, William Li, and Philippe Pasquier. 2015. Affect-expressive movement generation with factored conditional Restricted Boltzmann Machines. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 442–448.

[2] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parkerx, and Michael Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *Proceedings of the 2013 Conference on Affective Computing and Intelligent Interaction*. IEEE Computer Society, 283–288.

[3] Michael Argyle. 1988. *Bodily Communications*. Methuen & Co. Ltd.

[4] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2014. From crowdsourced rankings to affective ratings. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*. IEEE, 1–6.

[5] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[6] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 89–96.

[8] Antonio Camurri, Ingrid LagerlÄűf, and Gualtiero Volpe. 2003. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies* 59, 1-2 (2003), 213–225. https://doi.org/10.1016/S1071-5819(03)00050-8

[9] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. 2007. Recognising Human Emotions from Body Movement and Gesture Dynamics. In *Affective Computing and Intelligent Interaction*. Lecture Notes in Computer Science, Vol. 4738. Springer Berlin Heidelberg, 71–82. http://dx.doi.org/10.1007/978-3-540-74889-2_7

[10] Beatrice de Gelder. 2006. Towards the neurobiology of emotional body language. *Nature Reviews Neurosciences* 7, 3 (2006), 242–249. https://doi.org/10.1038/nrn1872

[11] Marco de Meijer. 1989. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior* 13, 4 (1989), 247–268. https://doi.org/10.1007/BF00990296

[12] Jianyu Fan, Kıvanç Tatar, Miles Thorogood, and Philippe Pasquier. 2017. Ranking-based emotion recogntion for experimental music. *Intern* (2017).

[13] Beat Fasel and Juergen Luettin. 2003. Automatic facial expression analysis: a survey. *Pattern recognition* 36, 1 (2003), 259–275.

[14] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological science* 18, 12 (2007), 1050–1057.

[15] Nickolaos Fragopanagos and John G Taylor. 2005. Emotion recognition in human-computer interaction. *Neural Networks* 18, 4 (2005), 389–405.

[16] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research* 4, Nov (2003), 933–969.

[17] Leo A Goodman and William H Kruskal. 1954. Measures of association for cross classifications. *J. Amer. Statist. Assoc.* 49, 268 (1954), 732–764.

[18] Sebastian Grassia. 1998. Practical parameterization of rotations using the exponential map. *Journal of graphics tools* 3, 3 (1998), 29–48.

[19] Shan He, Shangfei Wang, Wuwei Lan, Huan Fu, and Qiang Ji. 2013. Facial Expression Recognition Using Deep Boltzmann Machine from Thermal Infrared Images. In *Proceedings of the 2013 Conference on Affective Computing and Intelligent Interaction*. IEEE Computer Society, 239–244.

[20] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. (2000).

[21] Martin Inderbitzin, Aleksander Väljamäe, Jose Maria Blanco Calvo, Paul FMJ Verschure, and Ulysses Bernardet. 2011. Expression of emotional states during locomotion based on canonical parameters. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 809–814.

[22] Shihoko Kamisato, Satoru Odo, Yoshino Ishikawa, and Kiyoshi Hoshino. 2004. Extraction of Motion Characteristics Corresponding to Sensitivity Information Using Dance Movement. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 8, 2 (2004), 168–180. http://dblp.uni-trier.de/db/journals/jaciii/jaciii8.html#KamisatoOIH04

[23] Asha Kapur, Ajay Kapur, Naznin Virji-Babul, George Tzanetakis, and Peter Driessen. 2005. Gesture-Based Affective Computing on Motion Capture Data. In *Affective Computing and Intelligent Interaction*. Lecture Notes in Computer Science, Vol. 3784. Springer Berlin Heidelberg, 1–7. https://doi.org/10.1007/11573548_1

[24] Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective Body Expression Perception and Recognition: A Survey. *IEEETransactions on Affective Computing* 4, 1 (2013), 15–33. https://doi.org/10.1109/T-AFFC.2012.16

[25] Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[26] Yingliang Ma, Helena M Paterson, and Frank E Pollick. 2006. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods* 38, 1 (2006), 134–141.

[27] Nikolaos Malandrakis, Alexandros Potamianos, Georgios Evangelopoulos, and Nancy Zlatintsi. 2011. A supervised approach to movie emotion tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2376–2379.

[28] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14, 4 (1996), 261–292.

[29] Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology*. M.I.T. Press.

[30] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[31] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on* 2, 2 (2011), 92–105. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5740839

[32] Michael Nixon, Ulysses Bernardet, Sarah Alaoui, Omid Alemi, Ankit Gupta, Thecla Schiphorst, Steve DiPaola, and Philippe Pasquier. 2015. MoDa: an open source movement database. In *Proceedings of the 2nd International Workshop on Movement and Computing*. ACM.

[33] Hanhoon Park, Jong-Il Park, Un-Mi Kim, and Woontack Woo. 2004. Emotion Recognition from Dance Image Sequences Using Contour Approximation. In *Structural, Syntactic, and Statistical Pattern Recognition*. Lecture Notes in Computer Science, Vol. 3138. Springer Berlin Heidelberg, 547–555.

[34] Frank E Pollick, Vaia Lestou, Jungwon Ryu, and Sung-Bae Cho. 2002. Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research* 42, 20 (2002), 2345 – 2355. https://doi.org/10.1016/S0042-6989(02)00196-7

[35] Justus J Randolph. 2005. Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online submission* (2005).

[36] Leonardo Rigutini, Tiziano Papini, Marco Maggini, and Franco Scarselli. 2011. SortNet: Learning to rank by a neural preference function. *IEEE transactions on neural networks* 22, 9 (2011), 1368–1380.

[37] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161–1178.

[38] Ali-Akbar Samadani, Rob Gorbet, and Dana Kulic. 2014. Affective Movement Recognition Based on Generative and Discriminative Stochastic Dynamic Models. *Human-Machine Systems, IEEE Transactions on* 44, 4 (2014), 454–467.

[39] Daniel L Schacter, Daniel T Gilbert, and Daniel M Wegner. 2009. *Introducing psychology*. Macmillan.

[40] Jan Van den Stock, Ruthger Righart, and Beatrice De Gelder. 2007. Body expressions influence recognition of emotions in the face and voice. *Emotion* 7, 3 (2007), 487–494.

[41] Harald G Wallbott. 1998. Bodily expression of emotion. *European journal of social psychology* 28, 6 (1998), 879–896.

[42] Georgios N Yannakakis and Héctor Perez Martínez. 2015. Ratings are Overrated! *Frontiers in ICT* 2, 13 (2015). https://doi.org/10.3389/fict.2015.00013