Impact of Varying Vocabularies on Controlling Motion of a Virtual Actor

Klaus Förger[1], Timo Honkela[2], and Tapio Takala[1]

[1] Department of Media Technology,
[2] Department of Information and Computer Science,
School of Science,
Aalto University,
Espoo, Finland
{klaus.forger,timo.honkela,tapio.takala}@aalto.fi

Abstract. An ideal verbally controlled virtual actor would allow the same interaction as instructing a real actor with a few words. Our goal is to create virtual actors that can be controlled with natural language instead of a predefined set of commands. In this paper, we present results related to a questionnaire where people described videos of human locomotion using verbs and modifiers. The verbs were used almost unanimously for many motions, while modifiers had more variation. The descriptions from only one person were found to cover less than half of the vocabulary of other participants. Further analysis of the vocabularies against the numerical descriptors calculated from the captured motions shows that verbs appeared in closed areas while modifiers could be scattered to disconnected clusters. Based on these findings, we propose modeling verbs with a hierarchical vocabulary and modifiers as transitions in the space defined by the numerical qualities of motions.

Keywords: motion capture, natural language, virtual actors

1 Introduction

Animations and computer games have characters that act out scenes which an animator has designed. When creating these scenes, animators need believable human motion and ways to control the motion. To satisfy this need for human motion, many collections of captured motion have been made available [1]. Word based searching can be used to find suitable motions without a need to browse through the whole database. This way of searching corresponds to an ideal situation in which an actor would be always ready to act out motions based on short descriptions. In this paper, we concentrate on the effects of varying vocabularies on the motion searches.

In addition to words, human motion databases could also be searched by giving example motions or giving numerical requirements as search expressions. However, we limit the scope of the paper to collections of human motion where every motion clip is annotated with at least one written search term. The annotations can be the instructions given to an actor or opinions of persons viewing the motions. A potential problem is that a third person might not use or even understand the same vocabulary which was used in the annotations.

To find out how much variation there is in the vocabularies of people describing human motion, we constructed a questionnaire containing several different kinds of human locomotion. We asked people to describe the animated motion with one verb and from zero up to three modifiers which were adjectives or adverbs. Data from the questionnaire shows that variation between vocabularies of different people is large enough to cause potential misunderstandings.

We also present further analysis of the vocabularies against the numerical descriptors calculated from the captured motions. This analysis shows that verbs appear in closed areas whereas modifiers can be scattered to disconnected clusters. Based on these finding, we discuss what are the best ways to model the vocabularies.

2 Related Work

Controlling virtual actors with natural and unrestricted language requires creating links between the describing words and physical motions. A simple approach for creating the links is manual annotation which means writing labels for every motion. The task can be made easier by calculating descriptor values which reflect the quality of the motion [2]. The motion descriptors allow generalizing annotations as we can assume that two motions that are numerically close to each other are likely to be annotated in the same way. In this paper, we use motion descriptors when comparing motions.

Many methods and systems designed for controlling virtual characters assume that there is a small selection of allowed commands [3–5]. More fine grained control of both style and length of motions performed by a virtual character could be desired. This can be achieved with real-time interaction rules between two virtual characters, as the rules are based on continuous parameters [2]. However, the set of parameters can feel artificial to the end user, especially if the parameters are derived from the numerical qualities of the motions. Motion analysis frameworks such as Laban Motion Analysis (LMA) assume that the user knows a set of expert terms for describing human motion such as the Laban notation [6]. It has been found that systems allowing the use of natural language can reduce the amount of expertise and time needed in controlling virtual actors [7]. A challenge in natural language processing is that people can have subjective views on the meaning of words [8]. Our interests are in finding out how much manual annotation and analysis of motion is needed to enable controlling a virtual actor with natural language.

The assumption that, a small amount of motion classes is enough, does not appear only in systems that control virtual characters. Commonly used motion databases are often based on a selection of words given to the actors who perform the motions [9]. This can result in databases with plenty of motions, but where all the motions belong to stereotypical categories. A reason for taking shortcuts in annotation is that manual annotation can take a lot of time and effort [10].

As a motion database with annotations by several persons was not available, we decided to create one.

There are methods for creating new motions with different styles by using a selection of parameters which may be stylistic and emotional [11] or related to the trajectories of the motions [12]. These methods enrich a motion database as they create new motions by blending existing ones. We decided to use motion blending as it allows producing motions between stereotypical classes.

Three questions of interest were left open by the related works. How sufficient annotations from a single person are when building natural language descriptions? Do people describe the same motions with several synonyms? Do people have different opinions about the meaning of the used words? To answer these questions, we created a motion collection and a questionnaire which are presented in the next sections.

3 Motion Data Generation

To study natural language descriptions of human motion, we first needed a collection of motions to be described. We chose locomotion as it appears commonly in animations and it also allows displaying many motion styles. In order to create a set of motions that would have variation in both verbs and modifiers, we decided to use a mix of acted motions and interpolations between those motions. We recorded short locomotion sequences with two actors using Optitrack motion tracking system. The actors were asked to perform walking and limping with styles 'sad', 'slow', 'regular', 'fast' and 'angry'. Running was recorded with only the styles 'slow' and 'fast' as the limited capture area made recording running challenging. To make the motions easy to interpolate, the actors were instructed to always start from the same position with their right leg and to perform the motions towards the same direction.

The blended motions were produced with three steps which were initial alignment, time warping and interpolation. In the first step, the supported and lifted phases of the feet were detected and aligned among the motions. The second step was time warping the motions to make them synchronized. The aligned frames between the supported and lifted phases were matched and the rest of the frames were re-sampled to get a smooth frame rate. As the last step, the coordinates of the root joints were interpolated linearly and the joint rotations were interpolated as quaternions with the slerp algorithm [13].

We used two-way and three-way interpolation to create the blended motions. In the two-way case, three new motions are created with steps of 25%. In the three way case, we created all the two-way combinations, three motions with the percentages 70%-15%-15% and one motion with an even split of 33%-33%-33%. Ideally, we would have created blends from all possible combinations of the original motions, but that would have resulted in too many to be viewed reasonably. Also, some motions like fast running and slow walking were too different to be interpolated. We ended up creating blends between the motions that had a similar style and also between motions that had the same intended verb. The combinations used in the blends are shown in Figure 1.

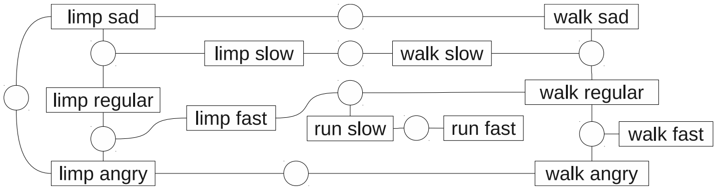

Fig. 1. The boxes show the original captured motions with the instructions given to the actors, the circles represent the combinations used in motion blending.

4 Questionnaire and Methods for Analysis

The idea of the questionnaire was to collect verbs and modifiers that describe the motions. The questionnaire was web based and all the motions were shown as videos with a stick figure character as shown in Figure 2. The duration of the videos ranged from 3 seconds (fast running) to 12 seconds (slow limping). Finnish language was used in the questions and the answers. The participants were gathered through work contacts and social media.

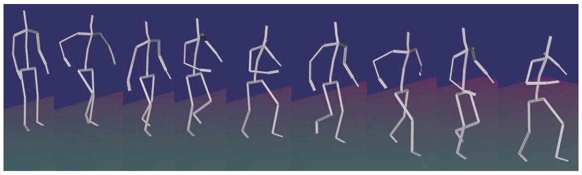

Fig. 2. An example of the stick figure representation portraying an angry walk.

The task given to the participants was to describe the seen motion with one verb or phrase (such as 'swimming' or 'mountain climbing') and from zero up to three modifiers (such as 'colorfully' or 'very colorfully'). To make the answering easier we divided the videos into three sets and the participants could answer as many sets as they liked. Set A included all unmodified motions and had 24 videos, set B had 40 motions which were 50%-50% interpolations and the set C had the rest 60 motions. The total amount of videos was 124.

Our first research question is: Is the collective vocabulary used by a group of annotators larger than the target vocabulary given to actors of the motions and larger than the vocabulary of a single annotator? An answer to this question helps in deciding how much effort should be put into developing better search terms for motion databases. A way to find an answer to this question is to calculate how much of the collective vocabulary would be covered by the terms given to the actors and the words used by a single participant.

The second research question is: Can the variation in the collective vocabulary be decreased by finding synonyms? An answer to this question is important as joining search terms requires only a small change in a motion database. This question requires qualitative grouping and analysis of the vocabulary. Comparison of the distributions of the words over the motion samples can also help recognizing synonyms. We use FinnWordNet [14] as the source of the definitions of the words. The translations of the Finnish words into English are also based on the FinnWordNet as it contains professionally made translations.

The third research question is: Do people have different opinions about the meaning of the used words? If there are large variations in how people use the same words, it would make building an optimal motion search much harder as the subjectivity would have to be taken in to account. Answering this question calls for plotting the distributions of the describing words on a space that is defined by numerical qualities of the motions.

To form a space which is based on the qualities of the motion, we calculate describing values called motion descriptors which include coordinates, velocities, accelerations and rotations as quaternions of each joint. From the velocities we used both absolute values and the velocities separately along the x, y and z axes. Also, we included the distances between pairs of body parts in a set that includes hips, neck, head, elbows, hands, knees and feet. To remove the variation caused by physical differences between the actors, we removed the personal means of descriptors as that has been found to help classification of motions [15].

5 Results

The participants of the questionnaire consisted of 9 females and 13 males with ages between 21 to 70 years. For the participants, the previous experiences with human motion were mainly linked to sports related hobbies. All 22 participants completed the set A, 10 also completed the set B and 2 participants did all the three sets of videos. Varying inflections which do not affect the meaning in this context such as 'walk' and 'walking' were cleaned from the data.

In the analysis, we have two points of view to the vocabularies. The first is the plain vocabulary where all the used words are considered equally important. The second is the shared vocabulary in which a word used by N persons is N times more important than a word used by one person. The distribution of the shared vocabulary is shown in Figure 3. From the figure we can see that 88 unique verbs and 233 unique modifiers were used by the participants. It also shows that the most common words explain a large part of the word usage, but there is also a long tail of rarely used words. For example nine most used verbs explain 50% of the shared vocabulary, but in order to reach to 90% one must consider 65 verbs.

Coverage of the words which were given to the actors and the words used by an average annotator are shown in Figure 4. Analysis of the vocabularies in Figure 4 is limited to the 24 videos in the set A as we needed to have annotations from all the participants to make a fair comparison. For the other analyses all the motion sets were used. Acted verbs plotted in Figure 4 have only coverage of 3% in the plain vocabulary as the three verbs given to actors are only a small part of the total 88 used unique verbs. However, when considering the shared vocabulary the three words have coverage of 29%. This comes from the fact that walking (kävelee) was used by all the 22 participants, running (juoksee) by 19 and limping (ontuu) by 14, while the total sum of usage counts was 190.

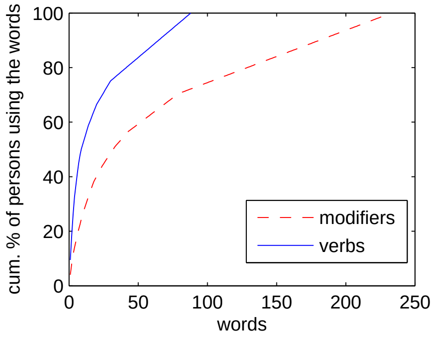

Fig. 3. Cumulative percentage of coverage of the shared vocabulary. The words are sorted from most used to the least used.

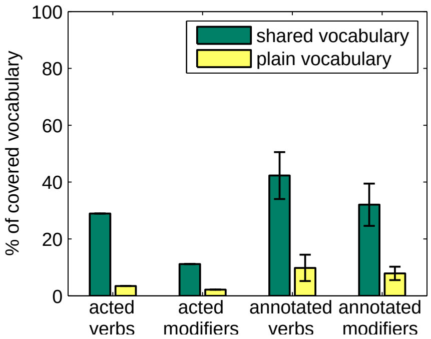

Fig. 4. Coverage of the plain vocabulary and shared vocabulary for verbs and modifiers given to the actors and average coverage of annotations from a single person. Standard deviation is shown for the averages.

Our first research question was related to how much the words given to the actors or the words of a single annotator cover of the overall vocabulary. The answer based on Figure 4 is that in the best case the words given to the actors cover a third of the vocabulary. Therefore, we can say that the set of words given to the actors of the motions would not enable making motion searches with natural language. The vocabulary of an average annotator does work better as it covers nearly 50% of the verbs in the shared vocabulary. Still, there is room for improvement. The coverage of the plain vocabulary is less than 10% which shows that having only one annotator will cause missing many rarely used words.

For finding synonyms, we used dictionary definitions of words and their translations to English as provided by FinnWordNet [14]. The words 'ontuu', 'nilkuttaa', and 'linkuttaa' are synonyms based on dictionary definitions and they all translate to 'limps' in English. This also shows that they could be considered to be alternative labels for the exactly same motions. From the modifiers we could not find synonyms as easily as from the verbs. Modifiers such as 'nopeasti' – 'fast' and 'kiirehtien' – 'hurriedly' can be considered to be similar, but whether they are synonyms is uncertain based on the data from the questionnaire.

For seeing the relationship between the numerical qualities of the recorded motions and the words used in the descriptions, we plotted the nine most frequent verbs (Fig. 5) and nine most frequent modifiers (Fig. 6) onto the PCA (principal component analysis) space based of the motion descriptors. To make the figures more readable we added small offsets to the overlapping pies to separate them. Web based versions of the two figures that also show the related animated motions are available at: http://research.ics.aalto.fi/cog/mglt/

Figure 5 shows that for many motions vast majority of the annotators are unanimous about the verbs. The three alternative words for 'limping' appear in the same area of the map and cause division between the annotators, but joining those words as synonyms would clean up the division. Two subjective divisions which cannot be accounted to synonyms are visible in the verbs. The first is between 'jogging' and 'running'. It seems that the participants could not agree where to draw a line between the two actions. The second subjective division is between 'walking' and 'limping'. While 'walking' has an area that is almost unanimously 'walking', almost all of the 'limping' motions have also a small share of 'walking' in them.

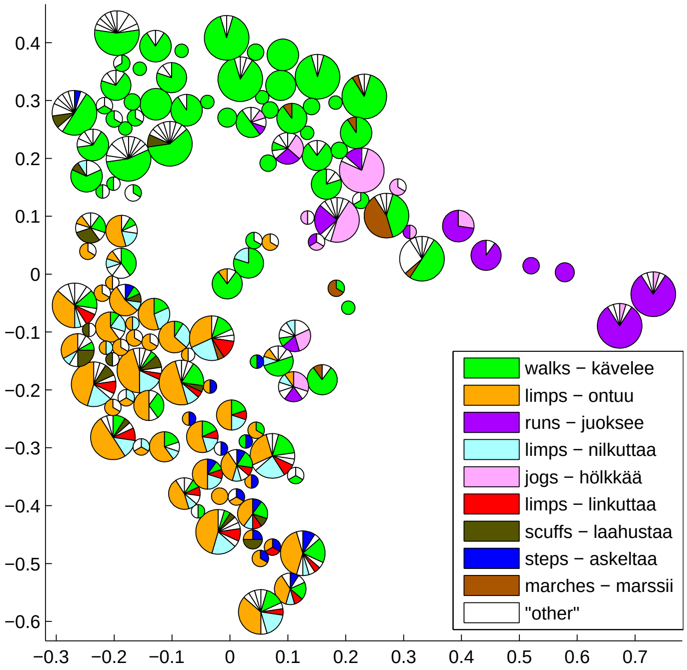

Fig. 5. Distributions of most common verbs for each motion mapped on the first and second normalized PCA components. The surface area of the pies is proportional to the number of answers and the position of the pies reflect the style of the motions.

Modifiers plotted in Figure 6 show that the participants were less unanimous in their answers than with verbs. There are even cases where almost all the participants gave different modifiers. Part of the variation can be explained by the fact the participants could give up to three modifiers. Still, even limiting the analysis to the first given modifiers, there would be no videos where one word would cover more than 50% of the answers if the video got more than two answers.

Many of the words are limited to a part of the PCA space. Verbs in Figure 5 form connected areas while modifiers can have disconnected distributions. For example the modifier 'slowly' appears mostly in the left side of Figure 6 where are the verbs 'walking' and 'limping', but also a few times near the center where the motions are described as 'jogging' or 'running'. The greater variation of modifiers is visible as the greater amount of the class 'other' than in the verbs.

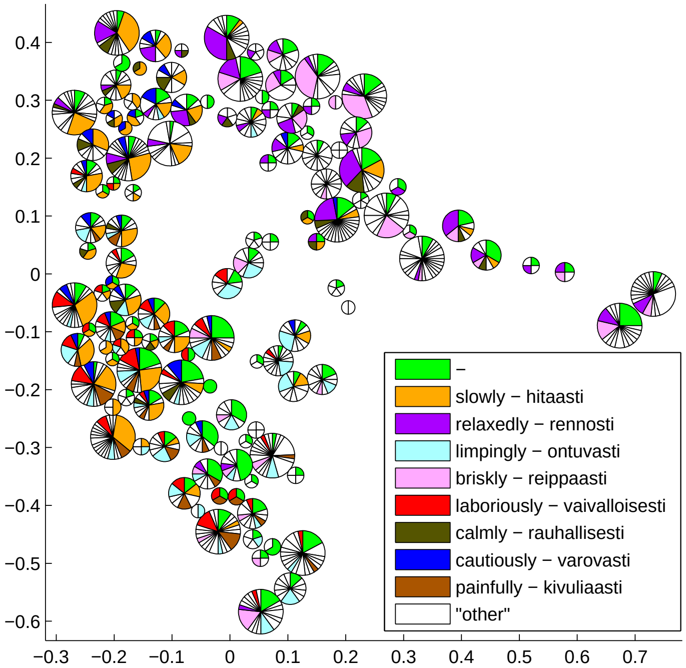

Fig. 6. Distributions of most common modifiers for each motion mapped on the first and second normalized PCA components. The surface area of the pies is proportional to the number of answers and the position of the pies reflect the style of the motions.

6 Discussion

How do the results of the questionnaire guide building a virtual actor that could be controlled with natural language? The first lesson is that relying only on the words given to the actors is not likely to cover the required vocabulary. Having one person annotate all the motions works better. However, the annotations of a single person are not enough in the cases where the borders between different verbs are subjective or when several synonyms exist. Modifiers are more challenging than verbs as the participants were far from unanimous and the modifiers did not always form continuous areas in the descriptor space.

For the verbs, hierarchical style of description could be beneficial as that would allow using words in a general sense and in a more specific sense. For example a parent category 'walking' could be divided into subcategories 'limping' and 'walking'. This way part of the subjectivity could be taken into account without needing more than one annotator. In practice, this could be achieved by giving the annotators two motions and a task to describe the motions with one verb.

Verb-modifier combinations could act as the most specific level of the description hierarchy. However, this would mean annotating a large amount of verb-modifier combinations. A more practical approach to handle modifiers could be to treat them as transitions in the descriptor space instead of areas of the space.

For a user instructing a virtual actor, this would mean first saying a verb and then saying a modifier to adjust the style of motion towards a desired direction. This approach could fix the problems caused by discontinuities in the distributions of modifiers. For example starting from walking and moving repeatedly towards a faster motion style would end up in a running motion. To find out what transitions correspond to which modifiers, a comparative task such as 'motion A is more X than motion B' should be given to the annotators.

While the questionnaire could always be made better, the main factor that speaks for the questionnaire is that the participants were able to freely select the words they used. If a selection of possible words had been given, it would have distorted the vocabularies of the participants. The decision to analyze the vocabularies as words-per-person instead of words-per-video makes our results more general. The counts for words-per-video are closely tied to the selection of videos, but the counts for words-per-person should not change dramatically even if part of the videos would be shown more times than others. One shortcoming in the questionnaire is the lack of repetitions. From data with repetitions, we could analyze how much of the variation in the descriptions is caused by difficulty of deciding between possible alternatives.

7 Conclusions and Future Work

In this paper, we presented results from a questionnaire in which participants were asked to describe videos of human locomotion with one verb and from zero up to three modifiers which were adjectives or adverbs. We analyzed the vocabulary as such and also in connection with numerical motion descriptors calculated from the motions. The results show that the original words given to the actors of the motions did not cover the used vocabulary of the participants viewing the motions. The vocabulary of a single annotator had better coverage, but the data would not help in cases where several synonyms exist for a verb or when the exact definition of a verb is not shared between the participants. The results also show that the modifiers used in describing the motions contain more variation than the verbs.

The main use case we considered was a virtual actor that can be controlled with natural language. Based on our results, we conclude that just linking each motion with the describing words would not allow controlling a virtual actor accurately. The linking would not take into account that meaning of verbs can be subjective and that modifiers are used variedly. The improvements we are planning include building a hierarchical vocabulary for verbs and modeling modifiers as transitions in the space defined by the numerical qualities of the motions. Realizing these improvements requires changing the annotation method from annotation of one motion at a time to annotation where similarities and differences are described between two motions.

Acknowledgments. This work has received funding from the Hecse doctoral program and the project Multimodally grounded language technology (254104) which is funded by the Academy of Finland.

References

1. Ahad, M., Tan, J., Kim, H., Ishikawa, S.: Action dataset - A survey. Proc. of SICE Annual Conference 2011 (SICE 2011), pp. 1650-1655. (2011)
2. Förger, K., Takala, T., Pugliese, R.: Authoring Rules for Bodily Interaction: From Example Clips to Continuous Motions. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 341-354. Springer, Heidelberg (2012)
3. Blumberg, B., Galyean, T.: Multi-level direction of autonomous creatures for real-time virtual environments. In: Mair, S.G., Cook, R. (eds.) Proc. of SIGGRAPH 1995, pp. 4754. ACM, New York (1995)
4. Perlin, K., Goldberg, A.: Improv: a system for scripting interactive actors in virtual worlds. In: Proc. of SIGGRAPH 1996, pp. 205-216. ACM, New York (1996)
5. Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., Welbergen, H., Werf, R. J.: The Behavior Markup Language: Recent Developments and Challenges. In: Pelachaud, C., Martin, J., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS vol. 4722, pp. 99-111. Springer, Heidelberg (2007)
6. Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. IEEE International Workshop on Robot and Human Interactive Communication 2005 (ROMAN 2005), pp. 294-299. IEEE (2005)
7. Talbot, C., Youngblood, G.: Spatial Cues in Hamlet. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. Conf. (eds.) IVA 2012. LNCS, vol. 7502, pp. 252-259. Springer, Heidelberg. (2012)
8. Honkela, T., Raitio, j., Lagus, K., Nieminen, I., Honkela, N., Pantzar, M.: Subjects on objects in contexts: Using GICA method to quantify epistemological subjectivity. In Proc. of International Joint Conference on Neural Networks (IJCNN 2012), pp. 2875-2883. (2012)
9. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing, Vol. 28, Issue 6, pp. 976-990. (2010)
10. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently Scaling up Crowdsourced Video Annotation. International Journal of Computer Vision, Vol. 101, Issue 1, pp. 184-204. Springer US. (2012)
11. Rose, C., Bodenheimer, B., Cohen, M. F.: Verbs and Adverbs: Multidimensional Motion Interpolation Using Radial Basis Functions. Computer Graphics and Applications, IEEE, vol.18, no. 5, pp. 32-40. (1998)
12. Kovar, L., Gleicher, M.: Automated Extraction and Parameterization of Motions in Large Data Sets. In: Marks, J. (ed.) Proc. of SIGGRAPH 2004, pp. 559-568. ACM, New York. (2004)
13. Shoemake, K.: Animating rotation with quaternion curves. ACM SIGGRAPH computer graphics, Vol. 19(3), pp. 245-254. (1985)
14. Lindén, K., Carlson, L. FinnWordNet - WordNet på finska via översättning. (In English: FinnWordNet - Finnish WordNet by Translation). LexicoNordica - Nordic Journal of Lexicography, vol. 17, pp. 119-140. (2010)
15. Bernhardt, D., Robinson, P.: Detecting affect from non-stylised body motions. In: Paiva, A., Prada, R. (eds.) Affective Computing and Intelligent Interaction, pp. 59-70. Springer, Heidelberg. (2007)