Automatic Multimodal Descriptors of Rhythmic Body Movement

Marwa Mahmoud
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
mmam3@cl.cam.ac.uk

Louis-Philippe Morency
USC Institute for Creative Technologies
12015 Waterfront Drive
morency@ict.usc.edu

Peter Robinson
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
pr10@cl.cam.ac.uk

ABSTRACT

Prolonged durations of rhythmic body gestures were proved to be correlated with different types of psychological disorders. To-date, there is no automatic descriptor that can robustly detect those behaviours. In this paper, we propose a cyclic gestures descriptor that can detect and localise rhythmic body movements by taking advantage of both colour and depth modalities. We show experimentally how our rhythmic descriptor can successfully localise the rhythmic gestures as: hands fidgeting, legs fidgeting or rocking, significantly higher than the majority vote classification baseline. Our experiments also demonstrate the importance of fusing both modalities, with a significant increase in performance when compared to individual modalities.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Motion

Keywords

Rhythmic body movements; Multidimensional tracklets; Multimodal fusion

1. INTRODUCTION

Nonverbal communication plays a central role in human-human communication. The ability to read nonverbal cues is essential to understanding, analyzing, and predicting the actions and intentions of others. Nonverbal cues include facial expressions, hand gestures, body posture and tone of voice. These nonverbal cues may indicate expression of emotions and mental states or even some medical conditions such as pain [22, 9], depression [13, 8] and anxiety [14]. Observation of those behavioural cues usually help clinicians in diagnosing. Moreover, manual labelling of those cues is the common practice for experimental psychologists in studying different behavoiurs.

ICMI'13, December 9–13, 2013, Sydney, Australia

ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522895 .

Over the past few years, there has been an increased interest in machine recognition of people's nonverbal cues, such as facial expressions and gestures. Automatic analysis of such cues can help as a tool for experimental psychologists. Also, it can assist physicians in diagnosing by providing quantitative measures after or during face to face sessions or telemedicine sessions or even in systems like a virtual coach. Recent research directions look into automatic detection of cues associated with psychological disorders, like depression, but most of the work focuses on the facial cues [8] as the main channel of nonverbal signal. Recent studies show that body movements and gestures are significant visual cues that can complement facial expressions [10] and they can be utilised in automatic detection of human internal states [6, 4].

In this paper, we study rhythmic body movement, as one of the indicators of psychological distress. They include a range of gestures, such as self-adaptors/self-grooming, fidgeting gestures (legs and body) and rhythmic torso movements (rocking). We noticed that these gestures share a common feature, which is the repetitive rhythmic motion. Automatic detection and localisation of this rhythmic body movement can help as an assistive tool in automatic analysis and diagnosis of more than one psychological disorder, such as Anxiety, Post-Traumatic Stress Disorder (PTSD) or Autism.

The main contributions of this paper are:

1. Proposing an automatic multimodal rhythmic gesture descriptor that detects rhythmic body motion by tracking multi-dimensional tracklets

2. Fusing depth and intensity features in a classification system to localise the rhythmic gesture as: hands fidgeting, legs fidgeting or rocking. Classification results were significantly higher than the majority vote classification baseline

3. Demonstrating the importance of fusing intensity and depth modalities in capturing gesture dynamics. Multimodal classification yields significantly better results when compared to individual modalities

In the following sections, we present our approach and results. we present the related work in automatic analysis of body motion in section 2. The significance of rhythmic body movements in psychological distress is explained in section 3. Sections 4 and 5 describe our proposed rhythmic descriptor and results. Conclusions and future directions are presented in section 6.

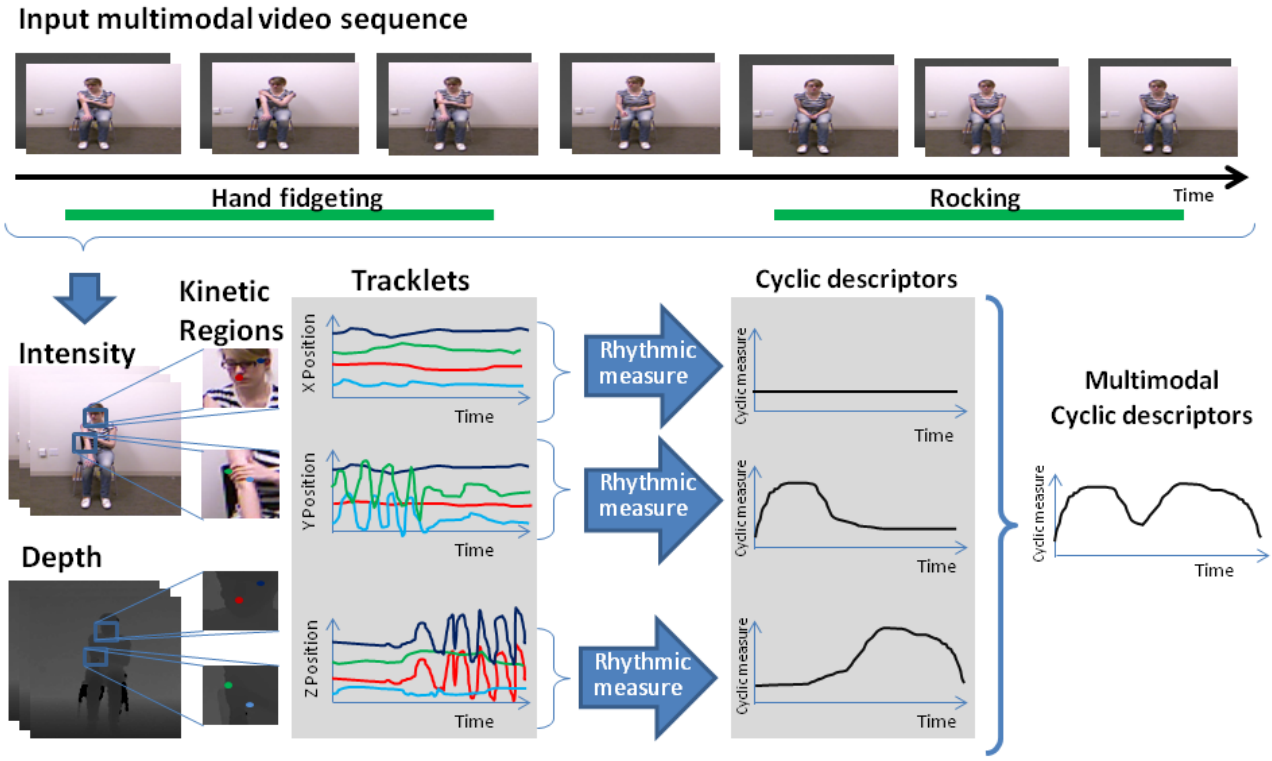

Figure 1: Overview Diagram shows the main steps in our cyclic descriptor. Multidimensional tracklets are extracted from intensity and depth images. A rhythmic measure is applied on every tracklet to generate our multimodal cyclic descriptor.

2. RELATED WORK

A lot of research studied periodic and cyclic motion associated with human activities by analysing colour images in video sequences. The difference between periodic and cyclic motion was nicely defined by Xiong and Quek [26] as: for a repeating motion, if its period p is a constant, this motion can be called periodic motion; If p is not perfectly constant over time, the motion is cyclic. Rhythmic body movement we are studying falls under the cyclic motion category with high variation in the cycles.

Heisele and Woehler [15] recognise human pedestrian motion by tracking motion parellel to the image plane. Xiog et al. tracked hands short oscillatory motion during natural speech using wavelet analysis [26]. Tsaiet al. [25] recognised cyclic motion by manually tracking the joints of the body and using Fourier transforms and autocorrolation for cycle detection. Self-similarity and autocorrellation were widely used for the detection of the cyclic motion [25, 24]. Autocorrelation analysis does not work for the non-uniform rhythmic gestures of the body because of the irregularity in the periodicity of the motion. Naveda and Leman [20] used Periodicity Transform to analyse periodicity of dance movements. Anaylsis of dance movements depends on the synchrony between the movements and the music beat, which limits the variability in the frequency of the motion. Moreover, most of the work uses only intensity information which makes it hard to capture motion perpendicular to the image plane.

With the availability of cheap consumer depth sensors, recent studies looked at ways of fusing depth and intensity information to achieve better results in tracking of rigid and non-rigid objects, especially in the face tracking field. Incorporating 3D depth information provided by consumer depth cameras with intensity information proved to improve tracking of the head and face [19, 1]. Tracking of body parts such as the hand [21, 17] using depth information was proved successful as well. However, most of this work fails during hand occlusion with other body parts, such as hands touching face or hands crossing.

3. RHYTHMIC GESTURES IN PSYCHOLOGICAL DISTRESS

Rhythmic gestures are a set of gestures that are understudied in the field of automatic detection of nonverbal visual cues. From research in psychology, certain rhythmic gestures were seen with greater frequency in clinical populations.

Fidgeting - which includes gestures such as tapping or rhythmically shaking hands or feet - is often seen and reported in both anxiety and depression [14]. Depressed patients also often engage in "self-adaptors" [12] , such as rhythmically touching, hugging or stroking parts of the body or self-grooming, such as repeatedly stroking their hair [14].

Scherer et al. [23] studied indicators of psychological distress, including Depression, Anxiety and PTSD. They demonstrated that subjects with psychological conditions exhibit on average longer self-touches and fidget with both hands (e.g. rubbing, stroking) and legs (e.g. tapping, shaking). Their studies of body gestures depend on manual annotation

since there are no automatic behavior descriptors currently available that robustly detect these behaviors.

Moreover, individuals with autistic spectrum disorder exhibit rhythmic body behaviours, namely self-stimulatory behavior [3, 16]. As defined in [16], self-stimulatory behavior consists of repetitive, stereotyped behavior that has no apparent functional effects on the environment, such as body rocking, hand-waving, and head-weaving [16].

The importance of these rhythmic gestures as indicators for different psychological disorders was the main motivation to study automatic descriptors to detect and localise these behaviours. Detection of these gestures is challenging because they consist of cycles that repeat over time in a similar pattern but can vary in amplitude and frequency. They are rhythmic movements - typically at a frequency of 0.5 - 2.5 Hz - similar in nature to rhythmic movements described by Dyken et al. [11]. Besides detection, localising the body area that exhibits this motion is vital for psychologists, as different types of rhythmic movements can be indicators for different types of disorders.

We study rhythmic gestures in three areas:

1. Hands fidgeting: This includes self-adaptors and hand fidgets, such as hand tapping, stroking, grooming, playing with face, fingers or the hair, and similar fidgeting behaviors

2. Legs fidgeting: This is similar to the hand fidgets and includes behaviors such as leg shaking and foot tapping

3. Rocking: This includes forward-backward body rocking and side-to-side rocking

4. RHYTHMIC GESTURE DESCRIPTOR

In order to develop a descriptor for these rhythmic body movements, we are faced by a set of challenges that rise mainly from the high variability in the gestures being tracked. We want to be able to track gestures that varied from thumbs twiddling to body rocking. The main challenges in designing robust rhythmic gesture descriptors are:

- **Irregularity in the periodicity of the motion**. In one gesture, cycles usually differ in frequency and amplitude. Moreover, we want our descriptor to detect different types of gestures that vary among themselves as well, for example the signal characteristics of hands fidgeting is different than legs shaking. Individual differences is another factor that affects the nature of the rhythmic movement

- **Non-rigid tracking**. Tracking of the motion involves tracking of different body parts that are non-rigid and deformable especially during the rhythmic motion. For example, hand fidgeting can exhibit different articulation especially during the rhythmic motion. Moreover, tracking of body parts using joints positions only will not generate accurate measures due to body parts occlusions and different body parts articulation during the motion

- **Multidimensional motion**. The rhythmic motion we are interested in involves motion in three directions. Extracting motion information from intensity images only can miss some motion components that are perpendicular to the image plane. Meanwhile, tracking body parts from depth images only can miss motion parallel to the image plane that does not involve significant change in the depth value, like motion involving body parts occluding each other such as hand rubbing arm

To meet these challenges, we propose a rhythmic motion descriptor. Figure 1 outlines the main building blocks of our approach. We address the non-rigid tracking by extracting tracklets and defining their kinetic region around different body parts. We handle the irregularity in the periodicity of the motion by introducing a rhythmic measure that does not depend on exact matching of cycles. To capture the multidimensional motion, multimodal fusion of depth and intensity tracklets is performed. These main steps are described in detail in the following sections.

4.1 Multidimensional tracklets

We avoid tracking of the non-rigid deformable body parts, by extraction of multidimensional tracklets. Tracklets are points of interest that are tracked over space and time.

Since the motion we are tracking can happen either parallel or perpendicular to the image plan, we extract tracklets in three dimensions: X,Y and Z. From the intensity images, tracklets are defined by extracting local feature-based keypoints from the intensity images- such as Speeded Up Robust Features (SURF) or Scale-Invariant Feature Transform (SIFT) keypoints - that are tracked over time to estimate their motion trajectories. We reinitialise and update the keypoints during the tracking in order to handle new features that can appear during the motion. Tracking these keypoints provides motion trajectories of the motion in the X and Y directions. For the Z direction, we do not use the same feature based keypoints extractions because interest points detection is not as reliable with the smooth depth signal. That is why we sample the depth image at a high density to maximise the performance of the tracking. Depth tracklets are defined by the depth value at every other pixel position. Implementation details of tracklet extraction are explained in section 4.4.

4.2 Rhythmic measure

As described above, irregularity of the rhythmic motion is one of the main challenges we faced. Moreover, individual differences add to the possible variability of the signal tracked. Due to the nature of this rhythmic signal, techniques that depend on exact matching of cycles like Fourier transforms and autocorrelation fail to detect cycles in our tracklets.

To handle this, we define a rhythmic measure that checks the similarity among the cycles in the extracted tracklets. The main idea is to define a set of constraints that are rigid enough to differentiate between rhythmic and non rhythmic motion in the tracklets and flexible enough to handle the high variability in and among different gestures tracked.

First, we extract local maximas and minimas in every tracklet to define the cycles. Since we are interested in prolonged rhythmic motion, we assume that the repetitive motion starts to be rhythmic if it repeats for more than three cycles. Thus, we can avoid erroneously capturing short repetitive gestures such as hand gestures accompanying speech. We analyse every tracklet in a sliding window W_t every four consecutive peaks.

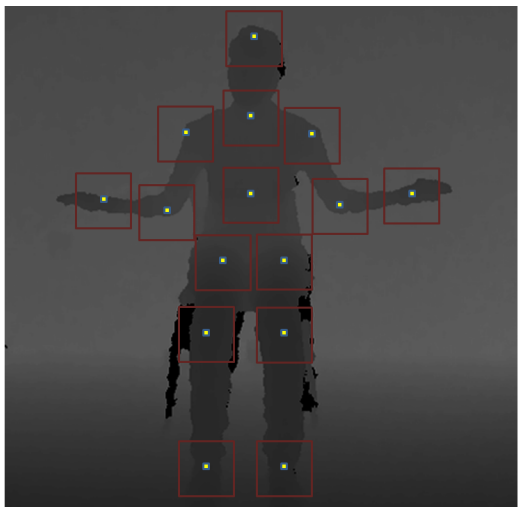

Figure 2: Kinetic regions around 15 joints positions obtained from Microsoft Kinect SDK skeleton tracker. Multimodal tracklet features are extracted from these regions and fused to localise the rhythmic motion

A similarity measure test is then applied on every window in the tracklets to identify the rhythmic movement. To reinforce the oscillatory nature of the repetitive movement, we check if one local minima exists between every two consecutive peaks (maximas) in the window. Since the motion we are interested in varies in frequency between 0.5 Hz to 2Hz, tracklet segments with frequencies below and above this range are considered non-rhythmic. Finally, a maximum difference between the size of the three cycles in a given window is set to one second, which means there should not be more than one second difference between any two cycles in the same window.

If a window W_t in a tracklet ϕ_i passes the rhythmic measure test, we output a 'cycle' at this pixel position in the frames in W_t.

4.3 Kinetic regions for multimodal fusion

Since we are interested in tracking motion related to different body parts, we define kinetic regions, which are 15 regions of 50x50 pixels around body joint positions. They are used for localising the detected rhythmic motion. Tracklets that fall inside these regions are assigned an attribute representing the kinetic region they belong to. The result of the rhythmic measure of all the tracklets in one kinetic region represents one feature in our feature vector. Features are extracted from depth and intensity tracklets in the 15 kinetic regions and fused for multimodal classification.

4.4 Implementation details

As a proof of concept, we implement our approach in detecting rhythmic motion and use a non-linear classifier to test its performance in localising rhythmic body movements.

For the tracklet extraction in the X and Y directions, keypoints are extracted using SURF [2] features and updated every 100 frames to add new local features generated from the motion. Keypoints are tracked using Pyramidal OpenCV implementation [5] of Lucas-Kanade optical flow [18]. Tracklets in the Z dimension are obtained from the depth values in the depth map obtained from Kinect sensor at every other pixel position.

For every tracklet obtained from the keypoints tracking in the intensity image and depth pixel values for the depth image, a set of signal smoothing steps are applied. A median filter is used to suppress narrow impulses while preserving smoother regions of the trajectory [25]. To normalise the signal, we subtract the mean value of the signal from each value at time t. Then the local maximas and minimas are extracted to detect the signal peaks.

In every sliding window in the tracklets, the rhythmic measure test explained in section 4.2 is applied to test if this window has a rhythmic movement at this specific pixel position or not. Applying a sliding window analysis on the motion trajectory level - rather than a unified sliding window on the frame level - allows for tracking of rhythmic movements of different frequencies. As described in section 3, the frequency of the rhythmic movements we are tracking can vary from 0.5 to 2 Hz , which means that the window of the repetitive motion of four peaks can span from 2 to 8 seconds (60 to 240 frames). We added a constraint on the minimum peak value in the Z trajectory to avoid the noisy low value spikes in the depth input from Microsoft Kinect. This is set empirically to 100. If a more accurate depth sensor is used, this constraint can be removed.

If a window W_t in a trajectory ϕ_i (where i is the pixel position in the Z trajectory or a keypoint position in the X and Y trajectories) passes the similarity measure test, we output a 'cycle' - which is a boolean value - at this pixel position in the frames in W_t.

Kinetic regions are defined by 15 regions of 50x50 pixels around 15 joint positions obtained from Microsoft Kinect SDK skeleton tracker (figure 2). The joints we used are Head, Neck, Torso, Left/right shoulder, Left/right elbow, Left/right wrist, Left/right hip, Left/right knee and Left/right ankle. Multimodal feature vectors are then calculated for every kinetic region, as the total number of positive outputs from all the tracklets in this region. Three feature vectors, representing X, Y and Z tracklets are then fused into a multi-class non-linear Support Vector Machines (SVM) classifier. A Radial Basis Function (RBF) kernel is used. We use LibSVM [7] for SVM's implementation.

5. EVALUATION AND RESULTS

We evaluate our approach on a dataset of acted gestures that include different rhythmic gestures. We compare our approach with majority vote baseline classification. We also compare classification results using single modality features to multimodal features, namely depth and intensity features.

5.1 Dataset description

We collected a dataset of acted rhythmic gestures. Twenty participants (12 males and 8 females) were recruited via university mailing list. They were asked to perform a set of rhythmic gestures imitating the gestures found in the Distress Assessment Interaction Corpus (DAIC) [23], that are believed to be correlated with psychological distress. The recorded gestures can be organised in three categories:

- Hands fidgeting: This category includes self- adaptors such as repetitive hand-touching-face gestures (hand scratching forehead, chin and cheeks). It also includes

Predicted / Ground Truth	Non-Rhythmic	Hands	Legs	Rocking	Multimodal classification per gesture		
					Precision	Recall	F1-score
Non-Rhythmic	**26683**	11512	6099	5510	53%	54%	54%
Hands	10235	**21266**	3618	1968	57%	57%	57%
Legs	11457	3871	**16125**	997	62%	50%	55%
Rocking	1513	933	22	**19806**	70%	89%	78%
Total recognition rate - Accuracy (Multimodal classification)					**59**%		
Total recognition rate - Accuracy (Majority vote baseline)					35%		

Table 1: Confusion matrix showing classification results of fusing intensity and depth features. The last two rows compare recognition rate (accuracy) of our approach to majority vote classification baseline. Total accuracy of our mutlimodal classification is significantly higher than the baseline classifier.

hand rubbing arm, hand scratching on the other hand, and thumb twiddling

- Legs fidgeting: This category includes foot/leg shaking and foot tapping
- Rocking: This category includes body rocking back and forth and body rocking side to side

Although participants were told to act the rhythmic gestures, each participant performed the gesture in his/her style. Each gesture was performed for a period of around 20 seconds. Moreover, after each category, participants were asked to perform other gestures that involved moving hands and legs in a way similar to the rhythmic gestures but not in a repetitive manner. The reason to include this set of non-rhythmic gestures is to test our system ability to recognise the rhythmic gestures among other similar gestures.

For recording, we used Microsoft Kinect sensor to capture colour and depth data. Each video is $\sim$3.5 minute long (μ=7081 frames). After the data collection, each frame was labelled with one of four labels: non-rhythmic , hands fidgeting, legs fidgeting or rocking. The labeled dataset will be made publicly available for the research community.

5.2 Classification results

We use a non-linear SVM classifier to detect and localise rhythmic gestures in the dataset of acted rhythmic gestures. We classify the gestures as one of four classes : Non-rhythmic motion, hands rhythmic motion, legs rhythmic motion or torso rhythmic motion (rocking). Depth and intensity features obtained from our rhythmic descriptor are fused into a four-class non-linear Support Vector Machines (SVM) classifier. A Radial Basis Function (RBF) kernel is used. The optimal parameters for the SVM are automatically obtained using a 4-fold testing hold-out validation, by holding 5 participants out for testing at each iteration. Experiments are performed in a user-independent approach, as none of the participants in the test set are used for validation or testing. The penalty parameter c and the RBF kernel parameter γ are varied from 10^k with k=-2, -1, 0, 1, 2. The optimal parameters obtained for c and γ are 1 and 0.01 respectively.

Table 1 shows the confusion matrix and classification recognition rates (precision, recall and F1-score) of our multimodal classifier. Our classifier achieves a total accuracy rate of 59%, which is significantly higher than a majority vote baseline classifier. The baseline classification recognition rate is 35%. We performed a paired- sample t-test to compare the total recognition rate per participant for the 20 participants between our approach and the majority vote baseline. A p value < 0.001 was obtained, which indicates that our approach recognition rates are significantly higher than the baseline.

The F1-scores for the non-cyclic, hands, legs and rocking classes are 54%, 57%, 55% and 78% respectively. Our descriptor achieves a high F1 score of 89% for the rocking rhythmic motion. Rocking usually involves rhythmic motion in many kinetic regions around head, neck, shoulders, torso and may be hands. This means that the rhythmic motion will definitely be captured by our tracklets. On the other hand, hands and legs categories include gestures involving rhythmic motion around fewer kinetic regions, such as thumbs twiddling that involves rhythmic motion in the hands region only or foot tapping that involves rhythmic motion in the kinetic regions around foot and knee. Some of this motion will be difficult to by captured in the depth map because of the relatively low resolution of the Kinect sensor to capture motion on the fingers level. This means that classification for these gestures mainly depends on the intensity features, which might affects the recognition rate. Using a higher resolution depth sensor can increase the recognition rate for these classes. Moreover, the mis-classification instances might be caused by the fact that rhythmic motion in some body parts can be propagated to other body parts. For example, foot tapping gesture can include slight rhythmic motion in the torso as well.

To compare single modality to multimodal feature fusion, we evaluate the classification using depth features only and intensity features only. A Radial Basis Function (RBF) kernel SVM is also used. We validate the SVM parameters using the same 4-fold testing hold-out validation approach. Figure 3 shows a a bar-chart comparing the average classification accuracies of the baseline majority vote classification, classification using depth features only, classification using intensity features only and multimodal classification. Single modality classification approaches achieve similar recognition rates: 49% for the depth only classification and 50% for the intensity only classification. Paired- sample t-test performed among all the pairs showed significant difference between multimodal approach and single modality (either depth or intensity). There was no significant difference between depth only classification and intensity only classification. This indicates that depth and intensity tracklets complement each other and demonstrates that multimodal fusion of depth and intensity features significantly increase the performance of recognition and localisation of different rhythmic body movements.

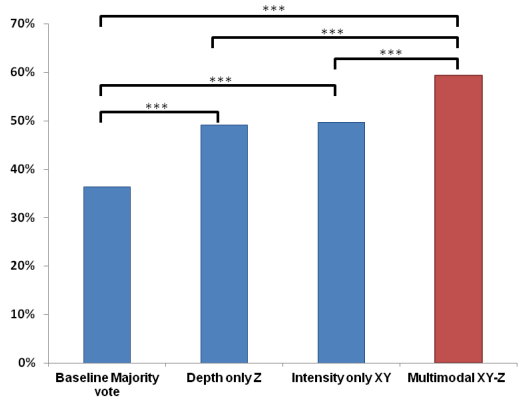

Figure 3: Comparing average classification accuracies shows significant increase in classification performance when we combine intensity and depth modalities. [*** p<0.001)]

6. DISCUSSION AND FUTURE WORK

We presented an automatic descriptor that can detect and localise rhythmic body movements, that are believed be associated with different clinical conditions, such as depression, anxiety, PTSD and autism. We propose the use of multidimensional tracklets to extract intensity and depth motion features around kinetic regions. We also propose the use of a similarity measure, rather than exact matching of cycles in the rhythmic motion to detect rhythmic gestures that varies in frequency and amplitude. Using a non-linear SVM classifier, our multimodal approach can successfully detect rhythmic body movements and localise them in three categories: hands rhythmic motion, legs rhythmic motion or body rocking. We compare our multimodation classification approach with a majority vote baseline classification and shows that our approach significantly outperforms the baseline classifier. Comparing our multimodal approach to single modality classification shows significant increase in the recognition rate when using both depth and intensity features, demonstrating the importance of fusing depth and intensity features in tracking gesture dynamics.

For future work, we would like to explore further classification of the rhythmic motion as separate gestures. This can be done by defining categories inside our current three rhythmic motion categories. For example, instead of defining a gesture as rhythmic hand motion, we classify it as either hand-touching-face, hand-on-hand fidgeting or thumbs twiddling. This will probably need the use of higher resolution depth sensors to capture motion in small areas of the body. We also plan to test our approach on a spontaneous dataset of participants with clinical conditions, such as the Distress Assessment Interaction Corpus (DAIC) [23]. Future work also includes testing our multimodal features using different classifiers that incorporate temporal features. Because of the continuous nature of the rhythmic gestures, this can lead to better performance.

7. ACKNOWLEDGMENT

We acknowledge funding support from Yousef Jameel Scholarships. This work is also supported by DARPA under contract (W911NF-04-D-0005) and U.S. Army Research, Development, and Engineering Command. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

8. REFERENCES

[1] T. Baltrusaitis, P. Robinson, and L. Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2610–2617. IEEE, 2012.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

[3] G. Berkson and R. Davenport. Stereotyped movements of mental defectives: I. initial survey. *American Journal of Mental Deficiency*, 1962.

[4] D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. In *Affective Computing and Intelligent Interaction*, pages 59–70. Springer, 2007.

[5] J.-Y. Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 2001.

[6] G. Castellano, S. D. Villalba, and A. Camurri. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective computing and Intelligent Interaction (ACII)*, pages 71–82. Springer, 2007.

[7] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[8] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction (ACII).*, pages 1–7. IEEE, 2009.

[9] K. D. Craig, K. M. Prkachin, and R. V. Grunau. *The facial expression of pain.* Guilford Press, 1992.

[10] B. de Gelder. Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3475–3484, 2009.

[11] M. E. Dyken, D. C. Lin-Dyken, and T. Yamada. Diagnosing rhythmic movement disorder with video-polysomnography. *Pediatric neurology*, 16(1):37–41, 1997.

[12] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98, 1969.

[13] H. Ellgring. *Non-verbal communication in depression.* Cambridge University Press, 1989.

[14] L. A. Fairbanks, M. T. McGuire, and C. J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of abnormal psychology*, 91(2):109, 1982.

[15] B. Heisele and C. WOEHLER. Motion-based recognition of pedestrians. In *14th International Conference on Pattern Recognition. Vol. 2*, 1998.

[16] M. E. Kaufman and H. Levitt. A study of three stereotyped behaviors in institutionalized mental defectives. *American Journal of Mental Deficiency*, 1965.

[17] C. Keskin, F. Kirac, Y. Kara, and L. Akarun. Real time hand pose estimation using depth sensors. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1228–1234. IEEE, 2011.

[18] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *7th international joint conference on Artificial intelligence*, 1981.

[19] I. Marras, J. Alabort-i Medina, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Online learning and fusion of orientation appearance models for robust rigid object tracking. In *10th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013.

[20] L. Naveda and M. Leman. A cross-modal heuristic for periodic pattern analysis of samba music and dance. *Journal of New Music Research*, 38(3):255–283, 2009.

[21] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *British Machine Vision Conference (BMVC)*, pages 101–1, 2011.

[22] K. M. Prkachin. The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51(3):297–306, 1992.

[23] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency. Automatic Behavior Descriptors for Psychological Disorder Analysis. In *The 10th IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[24] S. M. Seitz and C. R. Dyer. View-invariant analysis of cyclic motion. *International Journal of Computer Vision*, 25(3):231–251, 1997.

[25] P.-S. Tsai, M. Shah, K. Keiter, and T. Kasparis. Cyclic motion detection for motion based recognition. *Pattern recognition*, 27(12):1591–1603, 1994.

[26] Y. Xiong and F. Quek. Hand motion gesture frequency properties and multimodal discourse analysis. *International Journal of Computer Vision*, 69(3):353–371, 2006.