Improving the Accuracy of the CogniLearn System for Cognitive Behavior Assessment

Amir Ghaderi
University of Texas at Arlington
Texas, USA
amir.ghaderi@mavs.uta.edu

Srujana Gattupalli
University of Texas at Arlington
Texas, USA
srujana.gattupalli@mavs.uta.edu

Dylan Ebert
University of Texas at Arlington
Texas, USA
dylan.ebert@mavs.uta.edu

Ali Sharifara
University of Texas at Arlington
Texas, USA
ali.sharifara@uta.edu

Vassilis Athitsos
University of Texas at Arlington
Texas, USA
athitsos@uta.edu

Fillia Makedon
University of Texas at Arlington
Texas, USA
makedon@uta.edu

ABSTRACT

HTKS [9] is a game-like cognitive assessment method, designed for children between four and eight years of age. During the HTKS assessment, a child responds to a sequence of requests, such as "touch your head" or "touch your toes". The cognitive challenge stems from the fact that the children are instructed to interpret these requests not literally, but by touching a different body part than the one stated. In prior work, we have developed the CogniLearn system, that captures data from subjects performing the HTKS game, and analyzes the motion of the subjects. In this paper we propose some specific improvements that make the motion analysis module more accurate. As a result of these improvements, the accuracy in recognizing cases where subjects touch their toes has gone from 76.46% in our previous work to 97.19% in this paper.

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); •Computing methodologies → Artificial intelligence; Computer vision; Machine learning;

Keywords

Computer Vision; Deep Learning; Human Computer Interaction (HCI); Head-Toes-Knees-Shoulders (HTKS); Cognitive Assessment

1. INTRODUCTION

HTKS [8, 9] is a game-like cognitive assessment method, designed for children between four and eight years of age. During the HTKS assessment, a child responds to a sequence of requests, such as "touch your head" or "touch your toes". The cognitive challenge stems from the fact that the children are instructed to interpret these requests not literally, but by touching a different body part than the one stated. For example, a child may be instructed to touch her toes in response to the request "touch your head". HTKS has been shown to be related to measures of cognitive flexibility, working memory, and inhibitory control. At the same time, HTKS has also been shown to be useful in predicting academic achievement growth for prekindergarten and kindergarten children.

In our prior work, we have developed the CogniLearn system [3], that can be used to record the motion of human subjects as they play the HTKS game, and that also analyzes that motion so as to assess how accurately the subjects executed the requested tasks. In CogniLearn, a *Microsoft Kinect V2* camera is used for recording human motion. Then, we use the DeeperCut method [6] to perform body pose estimation in each frame. Finally, using the body pose estimates from DeeperCut we use a classification module that determines whether the subject touched his or her head, shoulders, knees, or toes. The CogniLearn system compares the part that was touched with the part that should have been touched based on the rules of the game, and assesses the overall accuracy score of the person playing the game.

The rest of the paper is organized as follows: In Section 2 we discuss related work in this area. In Section 3 we describe the proposed improvements to the prior version of the CogniLearn system. A quantitative evaluation of these improvements is offered in the experiments (Section 4).

2. RELATED WORK

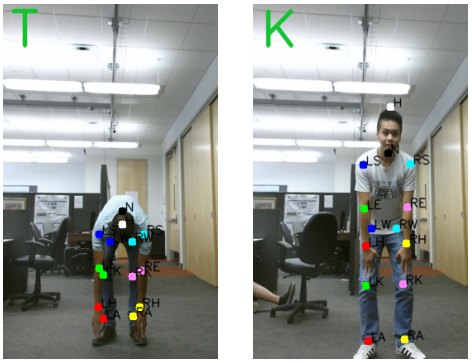
Several deep-learning methods have been proposed in recent years for video analysis and activity recognition [1, 4, 2], offering significantly improved accuracy compared to previous approaches[7, 10]. Deep learning methods have also been used in supervised or unsupervised manner in different tasks in computer vision [6, 5], oftentimes producing state-of-the-art results.

In [3] we have introduced the CogniLearn system, which is used for automated video capture and performance assessment during the HTKS assessment. CogniLearn is designed to provide meaningful data and measures that can benefit therapists and cognitive experts. More specifically, the motion analysis and evaluation module provides systematic feedback regarding the performance of the HTKS tasks to the human experts. In this paper, we build upon the CogniLearn system, and we suggest some specific improvements in the motion analysis module, that lead to higher recognition accuracy.

PETRA '17 June 21-23, 2017,Island of Rhodes,Greece

DOI: http://dx.doi.org/10.1145/3056540.3064942

Figure 1: A sample human body pose estimation on a frame using DeeperCut [6].



(a) (b)

Figure 2: Results using the full method described in this paper, i.e., when both Rule 1 and Rule 2 are used. On the left, we see a frame where the hands touch the toes. On the right, we see a frame where the hands touch the knees. The green letter on the top left of each frame is the classification output of the system, where "T" stands for "toes", "K" stands for "knees".

3. OUR METHOD

We use DeeperCut [6] to estimate the location of human body parts in each color frame of the video. Figure 1 shows a video frame where we have superimposed the body part locations estimated by DeeperCut. Each color frame of a test video sequence is provided as input to the DeeperCut method. The output of the algorithm is the image location of 12 body parts: head, shoulder(right and left), elbow(right and left), wrist(right and left), hip, knee(right and left), ankle(right and left).

After we obtain the body part locations from DeeperCut, we perform an additional step, in order to estimate whether the human, at that frame, is touching his or her head, shoulders, knees, or toes. As a first step, we define a distance D between hands and head, hands and shoulder, hands and knees, and hands and ankles. Using $\|\cdot\|$ to denote Euclidean norms, this distance is defined as follows:

$$D(\text{head}) = \frac{\|\text{lh} - \text{head}\| + \|\text{rh} - \text{head}\|}{2} \quad (1)$$

$$D(\text{shoulders}) = \frac{\|\text{lh} - \text{ls}\| + \|\text{rh} - \text{rs}\|}{2} \quad (2)$$

$$D(\text{knees}) = \frac{\|\text{lh} - \text{lk}\| + \|\text{rh} - \text{rk}\|}{2} \quad (3)$$

$$D(\text{ankles}) = \frac{\|\text{lh} - \text{la}\| + \|\text{rh} - \text{ra}\|}{2} \quad (4)$$

In the above definitions, head stands for the (x, y) pixel location of the center of the head in the color frame, as estimated by DeeperCut. Similarly, lh and rh stand for the locations the left and right hand, ls and rs stand for the locations of the left and right shoulder, lk and rk stand for the locations of the left and right knee, and la and ra stand for the locations of the left and right ankle.For example, $\|\text{lh} - \text{head}\|$ denotes the Euclidean distance between the left hand and the center of the head.

Based on these D values, one approach for estimating the body part that is being touched is to simply select the body part for which the D score is the smallest. This was the approach used in [3]. However, when the person touches the toes or knees, this approach does not work well. When a person bends down to touch the knees or toes with the hands, the head inevitably also gets near to the knees or toes. In that case, two issues may arise. The first one is that the accuracy of the body joint estimator is decreased. The second issue is that the detected location for the head is near the detected locations for the knees or toes. As a result, for example, when the hands are touching the toes, it frequently happens that the distance of hands to the head is estimated to be smaller than distance of the hands to the toes. These two issues can lead to inaccuracies. As we see in Table 1, in the original CogniLearn results of [3], 9.33% of toe frames are classified as head frames, and 14.00% of toe frames are classified as knee frames.

In this paper, we propose two heuristic rules to improve the classification accuracy of toe frames:

Rule 1: If the distance between the head and the hip is less than a predefined threshold, we can immediately conclude that the hands are touching the toes.

Rule 2: Sometimes, when the hands are touching the head, the distance between the hands and the head is estimated to be longer than the distance between the hand and the shoulders. To address this issue, we add a constant bias value to the distance between hands and shoulders, before comparing it with the distance between the hands and the head.

In the experiments, we demonstrate that these two rules significantly improve the classification accuracy on toe and head frames, while only minimally affecting the classification accuracy on frames where the hands touch the shoulders or knees.

4. EXPERIMENTS

For our experiments, we use the same dataset that was used in the original CogniLearn paper [3]. The dataset includes color videos from 15 participants, whose ages are between 18 and 30 years (while the HTKS assessment has been designed for children between the ages of 4 and 8, at this time we still do not have recorded data available from children of that age). In total, the dataset contains over 60,000 video frames. Figure 2 shows examples of test frames correctly recognized by our algorithm. The green letter in top left of the images shows the classification output of our system ("T" stands for "toes", "K" stands for "knees").

Our method is applied on each color frame separately. The goal

of our method is to classify each frame into one of four classes, corresponding respectively to whether the human is touching his or her head, shoulders, knees, or toes. Ground truth information is provided for 4,443 video frames, and we use those frames as our test set. The ground truth specifies, for each frame, which of the four classes belongs to. Accuracy is simply measured as the percentage of test frames for which the output of our system matched the ground truth.

We should emphasize that the results that we present are user-independent. None of the 15 subjects appearing in the test set is used to train any part of our models. The only module that uses training is DeeperCut, and we use the pretrained model that has been made available by the authors of [6].

4.1 Results

Table 1 shows the confusion matrix reported in the original CogniLearn paper [3]. As we can see in that table, shoulder and knee frames are recognized at rather high accuracies of 99.63% and 98.17% respectively. However, head and toes frames are recognized with lower accuracies, 94.47% and 76.46% respectively. This paper was primarily motivated by the need to improve the accuracy for those two cases.

Table 1: Confusion matrix reported by [3]. Rows correspond to ground truth labels, and columns correspond to classification outputs.

		Recognized				
		Head	Shoulder	Knee	Toe	Sum
Real	Head	**94.47**	5.53	0.00	0.00	100
	Shoulder	0.12	**99.63**	0.25	0.00	100
	Knee	0.00	0.54	**98.17**	1.29	100
	Toe	9.33	0.21	14.00	**76.46**	100

In Table 2 we report the results from the method proposed in this paper (i.e, when we apply both Rule 1 and Rule 2 from Section 3. As we can see, the accuracy for all four categories is more than 94.7%. The accuracy for head frames is marginally improved compared to [3]. The accuracy for shoulder and knee frames is slightly worse compared to [3]. At the same time, the accuracy for toe frames is now 97.19%, significantly higher than the accuracy of 76.46% reported in [3]. Finally, in Table 3 we show results using a partial implementation of our method, applying only Rule 1, and not Rule 2. We note that the overall accuracy is mostly similar to what we get when we combine Rules 1 and 2. Overall, Rule 1 is by far the biggest contributor to the improvements we obtain over the original results of [3]. At the same time, the accuracy for head frames improves from 93.21% to 94.78% when we use Rules 1 and 2, compared to using only Rule 1. Rule 2 was explicitly designed to reduce the percentage of head frames that were classified as shoulder frames. Indeed, using Rule 2 (together with Rule 1) reduces that percentage from 4.96% (obtained using only Rule 1) to 3.39%. Table 4 shows the overall classification accuracy. In that table, the overall accuracy is defined as the average of the accuracies over the four different classes. The overall accuracy improves from the 92.18% rate of [3] to 96.75% when we add Rule 1, and to 97.11% when we also add Rule 2.

Figure 3 shows some sample test frames. More specifically, from each of the four classes we show an example that was classified correctly, and an example that was classified incorrectly. We note that separating the head from the shoulder class can be quite challenging at times, because the distribution of hand positions does not vary much between the two classes. Separating knees and toes can also be difficult, because in frames belonging to both classes the knees are typically occluded, and there is significant overlap between the arms and the legs. This leads to errors in the estimated positions of the hands and the knees.

Table 2: Confusion matrix obtained using the full method described in this paper, i.e., when both Rule 1 and Rule 2 are added to the method of [3]. Rows correspond to ground truth labels, and columns correspond to classification outputs.

		Recognized				
		Head	Shoulder	Knee	Toe	Sum
Real	Head	**94.78**	3.39	0.26	1.57	100
	Shoulder	0.50	**99.25**	0.12	0.12	100
	Knee	0.00	0.60	**97.22**	2.18	100
	Toe	0.76	0.00	2.05	**97.19**	100

Table 3: Confusion matrix obtained by adding Rule 1 to the method of [3]. Rows correspond to ground truth labels, and columns correspond to classification outputs.

		Recognized				
		Head	Shoulder	Knee	Toe	Sum
Real	Head	**93.21**	4.96	0.26	1.57	100
	Shoulder	0.37	**99.39**	0.12	0.12	100
	Knee	0.00	0.60	**97.22**	2.18	100
	Toe	0.76	0.00	2.05	**97.19**	100

Table 4: Comparisons in accuracy between the original results of [3], the results obtained by adding Rule 1 to the method of [3], and the results obtained by adding both Rule 1 and Rule 2 to the method of [3]

	Overall	**H**	**S**	**K**	**T**
Original[3]	92.18	94.47	**99.63**	**98.17**	76.46
Rule 1	96.75	93.21	99.39	97.22	97.19
Rules 1,2 combined	**97.11**	**94.78**	99.25	97.22	**97.19**

5. CONCLUSIONS AND FUTURE WORK

We have propose a method for improving the accuracy of the original CogniLearn[3] system in recognizing, for each video frame, whether the human is touching the head, shoulders, knees, or toes in that frame. The experiments have shown that our improvements lead to significantly better accuracy, especially for frames where the human touches the toes. In those cases, the accuracy increased from the 76.46% rate in [3] to 97.19%.

Our project of automatically capturing and analyzing performance in the HTKS test is still in its initial stages. A high priority for us is to obtain data from children between the ages of 4 and 8, as that is the target age group for the HTKS test. Also, we plan to explore using the depth modality of the Kinect camera in addition to the color modality that we have used in [3] and in this paper. Finally, we should note that the HTKS assessment includes a "self-correction" category, in which the subject has started doing an incorrect motion and then self-corrected [9]. In the near future we plan to work on developing methods for identifying such self-correction cases, so that our assessment fully matches the formal HTKS description.

Acknowledgments

This work was partially supported by National Science Foundation grants IIS-1055062, CNS-1338118, CNS-1405985, and IIS-

1565328. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[2] M. Fouladgar, M. Parchami, R. Elmasri, and A. Ghaderi. Scalable deep traffic flow neural networks for urban traffic congestion prediction. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.

[3] S. Gattupalli, D. Ebert, M. Papakostas, F. Makedon, and V. Athitsos. Cognilearn: A deep learning-based interface for cognitive behavior assessment. In *intelligent user interfaces*, 2017.

[4] S. Gattupalli, A. Ghaderi, and V. Athitsos. Evaluation of deep learning based pose estimation for sign language. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2016.

[5] A. Ghaderi and V. Athitsos. Selective unsupervised feature learning with convolutional neural network (S-CNN). In *International Conference on Pattern Recognition(ICPR)*, 2016.

[6] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *arXiv preprint arXiv:1605.03170*, 2016.

[7] D. Leightley, J. Darby, B. Li, J. S. McPhee, and M. H. Yap. Human activity recognition for physical rehabilitation. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 261–266. IEEE, 2013.

[8] M. M. McClelland and C. E. Cameron. Self-regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives*, 6(2):136–142, 2012.

[9] M. M. McClelland, C. E. Cameron, R. Duncan, R. P. Bowles, A. C. Acock, A. Miao, and M. E. Pratt. Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in psychology*, 5, 2014.

[10] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2834–2841, 2013.

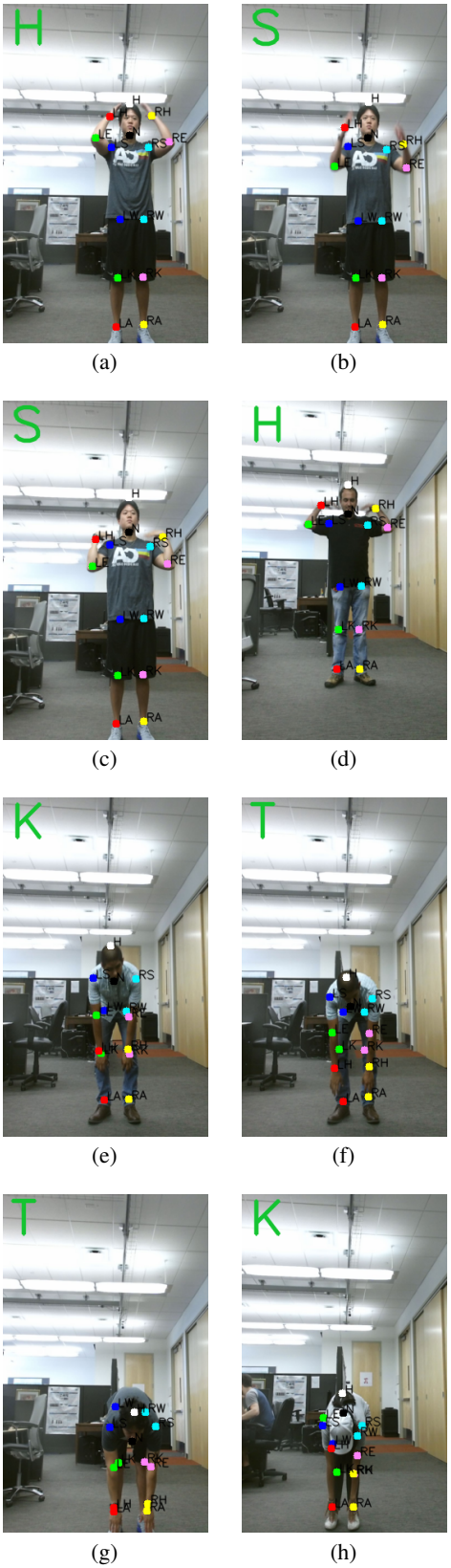


Figure 3: Example test frames, with the classification output superimposed. The classification output is correct for the examples on the left column, and incorrect for the examples on the right column. The ground truth is: "head" for row 1, "shoulders" for row 2, "knees" for row 3, "toes" for row 4.