A Refined 3D Dataset for the Analysis of Player Actions in Exertion Games

Chrysoula Varia*, Georgios Tsatiris*, Kostas Karpouzis*
*School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
chrysvaria@gmail.com, gtsatiris@image.ntua.gr,
kkarpou@cs.ntua.gr

Stefanos Kollias*†
†School of Computer Science
University of Lincoln
Lincoln, UK
skollias@lincoln.ac.uk

***Abstract*—Modeling and accurately analyzing human activities plays an important role, considering the rise of modern applications in human-computer interaction and, more recently, exertion games. Especially in serious exergames aimed at tutoring (e.g. sports) or rehabilitation and physiotherapy, the need for accurate detection of the human body and its motion is uncompromising. However, modern human skeleton tracking techniques suffer from a variety of issues, such as jittering and sensitivity to original conditions. In this study we show how a simple yet effective fairing pipeline on an inherently noisy dataset can produce data capable for precise experimentation with state-of-the-art human action modeling algorithms.**

***Index Terms*—exertion games, skeleton tracking, human action analysis, player actions**

I. Introduction

Exertion games form a very active research field with applications of almost interdisciplinary fashion, such as interaction design, gaming, sports, health and rehabilitation and, naturally, entertainment [1]. Many challenges arise in the field, mainly from the necessity of capturing and analyzing human motion, and transferring it into the application. In that manner, we investigate the exergames domain, mainly from the standpoint of recognizing and categorizing human activities.

Modeling of human actions find increasing use in the field of digital games and human-computer interaction [2], due to the commercial success of low-cost depth sensors, such as the Microsoft Kinect and the Intel RealSense series. Recent applied examples of serious games utilizing human actions can be found in the literature [3] and showcase the potential of human motion analysis, especially in healthcare-related games. The fact that there are many recent and prominent attempts at modeling actions for exertion gaming applications [4] shows the growing research interest in the field.

II. Human pose acquisition techniques and issues

One of the most widely applied methods to capture human motion, using vision-based techniques, is to detect the human pose in the frames of a video sequence. The pose of a subject is usually characterized by their skeleton joints, i.e. points in the frame that correspond to parts of the human body, such as the ankles, the wrists, the elbows, the head and the base of the neck.

In the context of depth sensing, skeleton joints are extracted from depth videos, using methodologies such as the ones documented in [5] and [6]. In [5], the authors proposed a methodology that is based on an object recognition approach, designing representations for intermediate body parts and transforming the pose estimation problem into a per-pixel classification problem. 3D proposals of body joints are then generated, based on the classification results. Later, in [6], a more advanced approach is proposed, based on a voting scheme to directly infer the positions of joints.

Recent prominent techniques have focused on solving the skeleton tracking problem in regular monocular color cameras instead of depth sensors. In the work presented in [7], the authors propose a deep convolutional neural network based pipeline that detects candidate body parts in the scene, as well as evaluated relations between them. It then proceeds to solve a heavily relaxed graph matching problem and extract connected skeletons for all persons detected in the scene.

A. Issues with skeleton data from depth sensors

Studies such as the one by Cosgun *et al.* [8] have investigated the accuracy of conventional depth sensing techniques with respect to the accuracy of extracted skeletons. In this study, it is claimed that average joints errors have been observed to usually be more than 5 cm. Such fluctuations in the position of a body part can be critical in accurate sensing of the human position and motion, especially in healthcare applications. Another study [9] places that error at about 10cm, especially in general, non-constrained postures. Moreover, depth-based skeletal tracking struggles with occluding body parts and objects in the scene.

Another issue, found both in depth-based and RGB-based skeletal tracking, is the sensitivity of the result with respect to initialization, especially when dealing with unconstrained scenes [10]. In other words, the way at which we begin the attempt of detecting skeleton joints affects the final result. Repetition of an action, or background processing of the sequence may leverage this issue and obtain more refined results.

978-1-5386-4359-4/18/$31.00 ©2018 IEEE
Authorized licensed use limited to: Panteion University. Downloaded on December 03,2024 at 19:49:02 UTC from IEEE Xplore. Restrictions apply.

B. The current state of the THETIS dataset

Our work in this paper focuses on the THETIS dataset [11]. It is comprised by 31 amateurs and 24 experienced players performing tennis shots. Data capturing was done using a Kinect sensor placed in front of the subjects. The action classes in the dataset are: backhand with two hands, backhand, backhand slice, backhand volley, forehand flat, forehand open stands, forehand slice, forehand volley, service flat, service kick, service slice and smash.

Although THETIS includes RGB and depth videos (scaled to grayscale), original (raw) depth information can be obtained using the unconstrained ONI files. So skeleton tracking needs to be handled from scratch. Another major issue, though, is that, in some cases, other subjects interfere in the scene, as data capturing took place mostly in a gym. Samples from the THETIS dataset can be seen in figure 1.

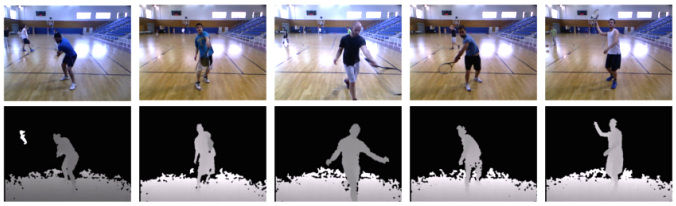

Fig. 1. RGB and depth samples from the THETIS database for the backhand, flat service, forehand flat, slice service and smash action classes.

III. Denoising the THETIS dataset

When dealing with a pre-existing dataset, assumptions as to what needs to be detected in the scene can be made, considering the specifics of the problem. In the THETIS dataset, this refers to the detection of the skeletal joints of the tennis player in the foreground of the scene. Joints are extracted in a per-frame fashion, using the method by Shotton *et al.* [6]. As it has already been discussed, there exist cases of individuals interfering during the execution of the action by the subject. However, we can expect that the tennis player will eventually be the subject detected in the vast majority of the frames of the sequence. We call this subject the *prominent subject*. Other persons detected in the sequence are subsequently rejected.

The complete pipeline is presented in algorithm 1. Due to the fact that a re-run of the sequence or another repetition of the action will, in most cases, give results of varying precision (as discussed previously), we leverage many iterations of the same sequence to obtain the final, refined annotations. In every iteration of the algorithm, we calculate the value of a factor to which the current iteration will impact the final result. This factor depicts the number of times the prominent subject has been detected in a specific frame. Starting from 0, we increment the factor in every iteration in which the prominent subject has been detected in this frame.

Ultimately, in the i-th iteration, the coordinates of joint J in a particular frame are updated using equation 1. J^i_{final} is the updated (final) values of coordinate vector J in the i-th iteration, J^{i-1}_{final} are the final values calculated by the previous iteration and J^i_{new} are the latest values in the current iteration, before updating. Essentially, what is achieved is an averaging scheme, tailored to the needs of a skeletal joint dataset. The algorithm finishes when the contribution of the latest iteration becomes negligible, i.e. $\frac{1}{\boldsymbol{factor}} < threshold$, where $threshold$ is a predefined value.

$$\boldsymbol{J^i_{final}} = \frac{\boldsymbol{factor} - 1}{\boldsymbol{factor}} \boldsymbol{J^{i-1}_{final}} + \frac{1}{\boldsymbol{factor}} \boldsymbol{J^i_{new}} \quad (1)$$

Algorithm 1 Pseudocode for the kinect-based skeletal information refinement pipeline

```
initialize factor vector
threshold ← a
repeat
  repeat
    detect all subjects
    for all detected subjects do
      increment number of occurrences
    end for
  until end of video sequence
  determine prominent subject
  for all video sequence frames do
    if prominent subject present in frame then
      recalculate prominent subject's joint positions in
      frame, based on equation 1
      factor++
    end if
  end for
until 1/factor < threshold
```

An alternative to this pipeline would be another averaging filter such as median filtering. This filter is a non-linear digital filtering technique, mainly focused on smoothing out individual noise and spikes. However, as this is not particularly the case in skeleton tracking, where noise is mostly uniformly distributed, this may not have given the desired result. In figure 2, we can see a set of refined skeleton joints from a sample of the "backhand" action of THETIS. Finally, the complete set of refined skeletal annotations can be found in the THETIS website[1].

IV. Experimental results

In order to verify the robustness of the produced annotations, we used the refined skeletal data as input for two different action analysis scenarios with wide applicability in the world of human-computer interaction and exertion games. The first scenario is that of generic activity recognition and is well represented by studies such as the ones presented in [4] and [12]. The second is the qualitative assessment of actions in context. A typical study in the field is the one in [13], which focuses on characterizing the level of expertise of tennis players.

[1] http://thetis.image.ece.ntua.gr/

Authorized licensed use limited to: Panteion University. Downloaded on December 03,2024 at 19:49:02 UTC from IEEE Xplore. Restrictions apply.

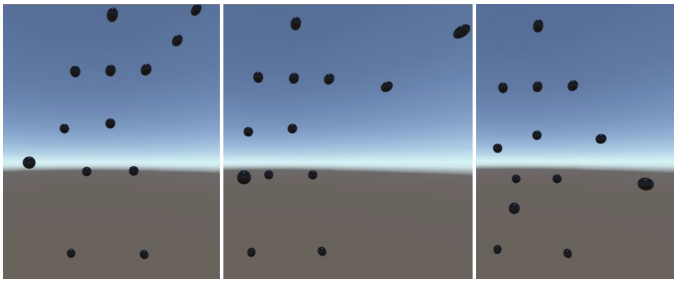

Fig. 2. Samples of extracted and refined skeleton joints of a "backhand" action from the THETIS dataset, as shown in a Unity-based application window. Viewpoint (perspective) distortions are visible.

A. Action recognition using Activity Feature Vectors

In the recent work by Cippitelli *et al.* [12], an action recognition pipeline was implemented, based on skeleton data, as the authors deem them as the most compact and representative form of the human presence in a scene. The algorithm is comprised by the following steps:

- Extraction of feature vectors representing human postures, using raw skeletal data. Essentially, a set of spatial (distance) features is calculated and normalized, based on the distance of the joints from the center of the torso.
- Selection of the most important postures for every activity class, using a clustering algorithm and picking the cluster centers as representative feature vectors.
- Computation of the final Activity Feature Vectors representing the activity as a whole. Each activity posture is characterized by the clustering of the previous step and the final Feature Vector contains the most important postures, by the order at which they appear in the activity sequence.

This method produces relatively short feature vectors, so dimensionality reduction is not necessary in most cases. Finally, a multiclass SVM is utilized to perform the classification.

In our experimental setup, we use the leave-one-person-out cross-validation protocol described thoroughly in [4]. Using this scenario, the pipeline based on the aforementioned technique and the refined skeleton dataset achieves an average accuracy of 93.65%. Table I showcases class-specific performance.

Average accuracy of the proposed pipeline is then compared with the other published techniques tested on the skeletal subset of THETIS [11] [4]. As we can see in table II, the pipeline using the novel technique in [12] on the refined skeletal data achieves higher average performance. Most notably, the method outperforms the novel approach presented in [4], which is based on a 3D Trace transform of the spatio-temporal volume of action sequences. However, this method has shown better performance in other modalities and is therefore more extensively applicable.

B. Assessing the level of expertise of tennis players

Although generic classification of human actions is a widely investigated field, qualitative assessment of actions has remained in relative obscurity. In the context of serious games focusing at player tutoring, rehabilitation and other activity-related tasks, the ability to make assumptions based on the quality of a movement is crucial. For that reason, in the work presented in [13], simple variance-based shape descriptors are utilized to classify between amateurs and experienced players in the THETIS dataset. The level of experience is self-declared, as experienced players either have been regularly practicing tennis or have attended tennis courses. Amateurs, on the other hand, have seldom or never played tennis before. For the purpose of capturing the dataset, they executed tennis actions with the help of a tutor.

TABLE I
CLASS-SPECIFIC RESULTS OF THE ALGORITHM PRESENTED IN [12], WHEN USED ON THE REFINED THETIS SKELETON DATASET

Action class	Accuracy (%)
Backhand	0,9549
Backhand with two hands	0,9775
Backhand slice	0,9275
Backhand volley	0,9388
Forehand flat	0,9308
Forehand open	0,9678
Forehand slice	0,9147
Service flat	0,9275
Service kick	0,9275
Service slice	0,9163
Smash	0,9195
Volley	0,9356

TABLE II
AVERAGE ACCURACY COMPARISON BETWEEN FOUR PUBLISHED PIPELINES AND THE ACTIVITY FEATURE VECTORS BASED PIPELINE ON THE REFINED THETIS DATASET

Method	Accuracy (%)
STIPs with HOG/HOF descriptors [14]	54.40
Dense Trajectories: Trajectory [15]	46.84
Dense Trajectories: MBH [15]	46.84
Dence Trajectories: Combination [15]	53.08
3D CTT - Selective STIP [16] based VTFs [4]	86.06
Activity Feature Vectors [12] on refined data	**93.65**

The pipeline in [13] calculates the variance of the points in a frame from the mean point (average of all points) and constructs a variance vector characterizing each action sequence. These vectors are then used as input in a simple K-NN classification scheme. However, due to the difference in length between sequences, Dynamic Time Warping [17] is utilized in two different classification scenarios:

- As a distance metric for the K-NN algorithm, as it calculates the minimum distance between two sequences
- By time-aligning the input sequences and then using Euclidean distance as a metric.

Classification was performed in an activity class specific manner. Both scenarios showed promising class-specific results, given the simplicity of the calculated features and the complexity of the task at hand.

Similar to [4], this method uses Selective Spatiotemporal Interest Points [16] as input. STIPs in general, as originally proposed by Laptev *et al.* [14] [18] represent, in the context

Authorized licensed use limited to: Panteion University. Downloaded on December 03,2024 at 19:49:02 UTC from IEEE Xplore. Restrictions apply.

TABLE III
PER-CLASS ACCURACY OF THE SHAPE DESCRIPTOR-BASED PIPELINE [13], WITH SELECTIVE STIPS [16] AND OUR REFINED SKELETAL JOINTS AS INPUT

Action class	Selective STIPs		Skeletal Joints	
	Not aligned, DTW metric	Time-aligned, Euclidean distance	Not aligned, DTW metric	Time-aligned, Euclidean distance
Backhand	74.55	64.85	**79.59**	77.55
Backhand with two hands	**70.91**	66.06	61.11	62.96
Backhand slice	64.85	**69.09**	68.52	55.56
Backhand volley	**69.09**	63.64	60.42	64.58
Forehand flat	**69.09**	58.18	53.57	53.57
Forehand open stands	**68.48**	68.48	65.45	67.27
Forehand slice	66.06	61.21	**70.91**	54.55
Forehand volley	66.06	60.61	**71.70**	62.26
Service flat	63.64	63.03	**63.83**	61.70
Service kick	**72.73**	70.91	62.22	64.44
Service slice	**67.88**	66.67	61.54	59.62
Smash	**64.85**	63.03	58.49	62.26

of human motion, points on the human body that express movement. Skeleton joints can be considered a very concise but representative subset of these points and therefore can be treated the same way.

In this experimental scenario, the refined skeleton dataset is used as input for the original algorithm in [13], instead of STIPs. Again, two classification scenarios are followed, and the results can be found in table III. It can be noted that, using the refined dataset, the algorithm demonstrates similar performance with the original STIP-based implementation. In some activity classes, the joint-based pipeline performs better. The reasons why certain classes show better results than others, as well as what part the difference between STIPs and skeletal joints has to play, form legitimate investigation directions for future studies.

V. CONCLUSION

In this work, we have noted the importance of accurate human skeleton information extraction in human-computer interaction related applications, such as exertion games. We have seen the issues that plaque average skeletal joint extraction techniques and we extended a published dataset with problematic raw data, by demonstrating a simple 3D skeleton data fairing pipeline. Experiments on two current human action related scenarios show that the refined dataset, when used with state-of-the-art algorithms, can form a basis for accurate experimentation on action-related applications.

ACKNOWLEDGMENT

This work was supported by the EC-funded project ECoWeB - Emotional Competence and Well-Being (grant agreement No 754657).

REFERENCES

[1] F. Mueller, R. A. Khot, K. Gerling, and R. Mandryk, "Exertion games," *Found. Trends Hum.-Comput. Interact.*, vol. 10, no. 1, pp. 1–86.

[2] R. Cowie, E. Douglas-Cowie, K. Karpouzis, G. Caridakis, M. Wallace, and S. Kollias, "Recognition of emotional states in natural human-computer interaction," in *Multimodal User Interfaces*. Springer, Berlin, Heidelberg, pp. 119–153.

[3] F. Deboeverie, S. Roegiers, G. Allebosch, P. Veelaert, and W. Philips, "Human gesture classification by brute-force machine learning for exergaming in physiotherapy," in *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, Sept 2016, pp. 1–7.

[4] G. Goudelis, G. Tsatiris, K. Karpouzis, and S. Kollias, "3d cylindrical trace transform based feature extraction for effective human action classification," in *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, Aug 2017, pp. 96–103.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. IEEE Computer Society, 2011, pp. 1297–1304.

[6] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, Dec 2013.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] A. Cosgun, M. Bunger, and H. Christensen, "Accuracy analysis of skeleton trackers for safety in hri," in *Proceedings of the Workshop on Safety and Comfort of Humanoid Coworker and Assistant (HUMANOIDS), Atlanta, GA, USA*, 2013, pp. 15–17.

[9] Š. Obdržálek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel, "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2012, pp. 1188–1193.

[10] T. Schubert, A. Gkogkidis, T. Ball, and W. Burgard, "Automatic initialization for skeleton tracking in optical motion capture," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 734–739.

[11] S. Gourgari, G. Goudelis, K. Karpouzis, and S. Kollias, "Thetis: Three dimensional tennis shots a human action dataset," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 676–681.

[12] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using skeleton data from rgbd sensors," *Intell. Neuroscience*, vol. 2016, pp. 21–, Mar. 2016.

[13] G. Tsatiris, K. Karpouzis, and S. Kollias, "Variance-based shape descriptors for determining the level of expertise of tennis players," in *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, Sept 2017, pp. 169–172.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[15] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *CVPR 2011*, June 2011, pp. 3169–3176.

[16] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzlez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396 – 410, 2012.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb 1978.

[18] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, no. 2-3, pp. 107–123, Sep. 2005.

Authorized licensed use limited to: Panteion University. Downloaded on December 03,2024 at 19:49:02 UTC from IEEE Xplore. Restrictions apply.