G3Di: A Gaming Interaction Dataset with a Real Time Detection and Evaluation Framework

Victoria Bloom, Vasileios Argyriou, Dimitrios Makris

Kingston University, London, UK
<Victoria.Bloom,Vasileios.Argyriou,D.Makris>@kingston.ac.uk

Abstract. This paper presents a new, realistic and challenging human interaction dataset for multiplayer gaming, containing synchronised colour, depth and skeleton data. In contrast to existing datasets where the interactions are scripted, G3Di was captured using a novel gamesourcing method so the movements are more realistic. Our detection framework decomposes interactions into the actions of each person to infer the interaction in real time. This modular approach is applicable to a virtual environment where the interaction between people occurs through a computer interface. We also propose an evaluation metric for real time applications, which assesses both the accuracy and latency of the interactions. Experimental results indicate higher complexity of the new dataset in comparison to existing gaming datasets.

Keywords: human interaction recognition, multimodal dataset, multiplayer gaming, interaction evaluation metric

1 Introduction

Recognising human interaction is a very active research area in the field of computer vision and is key to a range of domains including security, entertainment and robotics. The goal of interaction recognition is to automatically detect human interactions in a sequence of observations. Conceptually, a two person interaction is composed of a pair of two single actions, an action and a counter action [14]. In traditional human interaction, people interact directly with each other like in a real boxing match, illustrated in Fig. 1. Recent technological developments, such as low cost depth sensors, has enabled a new form of virtual interaction, for example a full body boxing game shown in Fig. 2.

This new generation of games use the human body as the controller and have increased the appeal of gaming to family members of all ages. Multiplayer sports games encourage people to interact with other players across the globe or friends and family in the same living room. The interactions can be collaborative or competitive depending on the specific sport and game mode. Boxing is naturally a competitive sport but team sports can be played either collaboratively with friends on the same team or competitively with friends on the opposing team. For example, one can play table tennis alongside a friend in a doubles match or against a friend in a singles match. The players can act simultaneously or

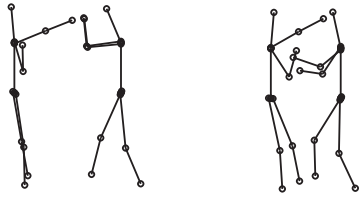

Fig. 1. Boxing interactions: A block (left) occurs when one person punches and the other person defends at the same time, whereas an attack (right) occurs if one person punches and the other fails to defend.

after a short delay depending on the sport. For example, in boxing the actions are concurrent but other sports such as table tennis have a delay between one person acting and the other reacting.

Past research has typically focused on recognising interactions from colour sequences but the recent release of low cost depth sensors combined with a real time pose estimation algorithm has seen the rapid growth of research on depth and skeleton data. Each modality has advantages and disadvantages: colour and depth data contain contextual information but are both dependent on the camera view and the persons' appearance. Depth and skeleton data are more robust than colour data when there are a lot of illumination changes and can even work in total darkness. Skeleton data is both invariant to the camera location and subject appearance, but lacks contextual information and does not work well when the player is not standing or sitting upright. Fusing colour and depth has overcome some limitations of the individual modalities but most current algorithms consider the depth and colour channels independently [2]. The new multimodal dataset presented in this paper with synchronised colour, depth and skeleton data can provide the opportunity to develop algorithms with improved fusion of the different modalities towards producing more robust algorithms that have a wider range of applications.

The contributions of this paper are a realistic and challenging human interaction 3D dataset with a real time detection and evaluation framework. G3Di is a novel multiplayer gaming dataset containing synchronised colour, depth and skeleton data. In contrast to existing datasets the movements are much more complex and realistic. Our new interaction framework recognises individual actions as they occur to infer the interaction in real time. We also propose an evaluation metric for real time applications which assesses both the accuracy and latency of the interactions.

2 Related work

Human interactions are composed of actions therefore we review existing datasets, recognition algorithms and evaluation metrics at both the action and interaction levels of the activity hierarchy.

2.1 Datasets

Historically, human activity datasets were recorded with visible light cameras and consist only of colour data [9] [15] [16] . For a comprehensive review of these also see Aggarwal and Ryoo [1]. The major problem with colour data of human motion is that there is a considerable loss of information [2]. After the recent release of low cost depth sensors there has been a growth of 3D datasets that provide skeleton data with some also providing colour and/or depth data [4] [6] [7] [11] [12] [18] [20].

However, most 3D datasets are restricted to activities performed by a single human subject which subsequently limits the development of 3D recognition algorithms to a single person [2] [4] [7] [11] [18]. The problem with the existing 3D gaming datasets, MSRAction3D [11], MSRC-12 [7] and G3D [4] is that they are single player, whereas commercial games are often multiplayer.

Another major limitation of the existing gaming datasets is that the scenarios were scripted so the movements are not realistic. In scripted datasets, the participants were instructed beforehand on how and when to perform the actions. Furthermore, in the MSR Action3D and MSRC-12 datasets each sequence contains only a single action class and the transition between repetitions often includes the neutral position. The G3D dataset contains more realistic gaming scenarios as there are multiple action classes in a sequence, but as there is a delay between actions the subject often returns to the neutral position when changing action. In fast paced competitive games like boxing, players do not return to the neutral position between actions which creates complex action transitions.

The key features of the gaming 3D datasets are illustrated in Table 1. The table was ordered on the number of data sources provided which increases by row. G3D is the only existing gaming dataset to contain all three modalities (colour, depth and skeleton data). The SBU [20] and K3HI [10] traditional person to person interaction datasets contain all three modalities but in both of these datasets the people were captured from a side view and partially clipped, which created noisy and unreliable skeleton data.

To overcome the limitations of the current datasets we propose a new multiplayer gaming dataset, G3Di with synchronised colour, depth and skeleton. The people were captured from the front view and interacted indirectly with each other through a computer interface. Sports games introduced the element of competition between the players so the actions captured were more realistic and challenging in comparison to scripted actions.

2.2 Recognition Algorithms

In human activity recognition there is a vast wealth of research on interaction recognition and traditionally approaches were appearance based as low level features could be quickly extracted from colour sequences. Recent work [2] [11] [12] suggests that human activity recognition accuracy can be improved by using features from 3D data. Pose based features from skeleton data are a very effective

Table 1. Comparison of 3D gaming datasets.

Dataset	Classes	Subjects	Data sources	Instruction Modality	Scenario
MSRC-12 [7]	12	30	Skeleton	Scripted	Actions
MSRAction3D [11]	20	10	Depth+ Skeleton	Scripted	Actions
G3D [4]	20	10	Colour+ Depth+ Skeleton	Scripted	Actions
G3Di	15	12	Colour+ Depth+ Skeleton	Game-sourced	Actions+ Interactions

representation for human motion [3] [7] [10] [19] [20] so we focus on pose based approaches.

Due to the development of a real time pose estimation algorithm [17] from depth streams many recent activity recognition algorithms are based on skeletal joint information. In a recent review of human activity recognition from 3D data [2], the authors concluded that most current approaches only deal with a single human subject. Subsequently, the features are based on joints from a single skeleton such as the pairwise joint location difference feature [3] [7] [19].

These pose based features were extended to multiple skeletons by Yun et al. [20] to model human interactions. Their experiments showed that the distance between all pairs of joints was the optimum set of joint features for real time interaction. This feature measures the pairwise joint distance in each skeleton, as well as between the two skeletons. This feature set was specifically designed for person to person interaction where the distance between the joints of the people aids the classification. For example, the distance between two people can easily be used to differentiate between approaching and departing. However, this feature set is not so relevant in virtual human interaction where there is no physical interaction between the people.

Further research by Hu et al. [10] with pose based features from multiple skeletons discovered that an interaction can be represented by a positive and negative action. Their results showed that the positive action on its own was discriminative enough to classify the interactions in their dataset, so the interaction recognition was simplified to positive action recognition. This works for simple scenarios where there is only one outcome from an action, such as the punching in their dataset where the first person punches and the second person falls away from the hit. However, in more complex scenarios there are more than one possible reactions from a punch, for example, a hit as just described or a block where the second person defends themselves by raising their hands in front of their face. If the skeletal information from the second person is ignored it will be very difficult to differentiate between these two interactions.

To overcome these limitations, our framework decomposes the interaction into the actions of both people and infers the interaction from the action pair in

real time. The pose based features are only extracted within a skeleton and not between skeletons so that our approach can be applied to virtual human interaction. Moreover, decoupling the action and counter action reduces the number of instances required for training and allows new interactions to be detected that are not in the training data.

2.3 Evaluation metrics

A common performance measure used for activity recognition is classification accuracy which is applied to the entire sequence. For example, an interaction label is predicted for each frame in the sequence and a majority decision over all frames is taken to decide the interaction label for the complete sequence. However, this approach can only be applied to simple sequences containing the same action class which is not the case for the new dataset.

To overcome the limitations of sequence based evaluation, frame based evaluation metrics have been developed [6] [15]. Escalera et al. [6] introduced a Jaccard index that can evaluate sequences with multiple action/interaction classes with respect to time. Ryoo and Aggarwal [15] proposed spatial and temporal bounding boxes to evaluate sequences with multiple interactions with respect to both space and time. Both approaches are evaluated based on the overlap between the system detection and the ground truth labels. These application metrics include temporal constraints but do not evaluate the latency of the detection.

Low latency detection is critical for real world applications such as surveillance and gaming. Nowozin et al. [13] proposed a latency aware performance metric for online human action recognition. They introduced 'action points' as temporal anchors for the detection and evaluation of single person actions in real time. According to [13], an action label is correct if it is detected within a specific time window around the ground truth action point. We propose an interaction evaluation metric for real time applications which assesses both the accuracy and latency of the interactions by exploiting the generality of the action point metric.

3 G3Di dataset

A new multimodal interaction dataset has been captured, for real time multiplayer gaming and is publicly available[1]. G3Di contains synchronised colour, depth and skeleton data. The dataset was captured using a novel gamesourcing approach where the users were recorded whilst playing computer games.

Our recording environment as illustrated in Fig. 3 allowed us to capture realistic gaming actions. The inherent competitive nature of the games resulted in the players putting more effort into their movements. The setup shows two players as the current generation of depth sensors are limited to full skeleton tracking of two people. However, the same setup could be used for up to six

[1] G3Di can be downloaded from `http://dipersec.kingston.ac.uk/G3D/`

Fig. 2. Synchronised colour, depth and skeleton data from a boxing game.

players when the next generation of depth sensors are released. The recording environment contains two overlapping depth sensors: one for playing full body games on a standard games console and the other to capture the colour, depth and skeleton data. The disadvantage of using two sensors with overlapping fields of view is that considerable noise is introduced to the depth data and consequently the skeleton data, due to infrared interference. Specifically, the depth sensor we used the Kinect, derives depth by projecting a structured light code onto the scene and comparing the reflected pattern with the stored pattern. To overcome this problem a motor was attached to one depth sensor to vibrate it and therefore reduce the interference between them as observed in experiments by Butler et al. [5].

Due to the formats selected, it is possible to view all the recorded data and metadata without any special software tools. The three streams were recorded at 30fps in a mirrored view. The depth and colour images were stored as 640x480 PNG files and the skeleton data in XML files. Each skeleton contains the player's position and pose: the pose comprises of 20 joints and the joint positions are given in X, Y and Z coordinates in meters. These positions are also mapped into the depth and colour coordinate spaces. The skeleton data includes a joint tracking state, displayed in Fig. 2 as tracked (green), inferred (yellow) and not tracked (red). The joint tracking state provides the confidence of the data for each joint. If the data is tracked, the confidence in the data is very high. Whereas, if the data is inferred by calculating it from other tracked joints, the confidence in the data is very low. This is important information for developers of multimodal algorithms fusing data between the skeleton data and other modalities.

To the best of our knowledge this is the first dataset comprised of virtual interactions, meaning that two players interact with each other through a computer interface. This dataset contains 12 people split into 6 pairs. Each pair performed 15 gaming actions, for five sports games: boxing (right punch, left punch, defend), volleyball (serve, overhand hit, underhand hit, or jump hit), football (kick, block and save), table tennis (serve and hit), sprint (run) and hurdles (run and jump). Most sequences contain multiple action classes in a controlled indoor environment with a fixed camera, a typical setup for gesture based gaming. The people played the game in a training mode to become familiar with the movements before they were recorded. The actual game was recorded

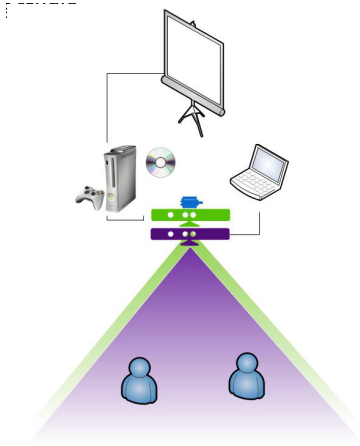

Fig. 3. Recording environment with 2 depth cameras for simultaneous gameplay and recording.

and particular sections where several different actions were performed multiple times by each player were selected for the dataset.

4 Interaction Detection Framework

Our novel framework detects individual actions from multiple people, to infer the interaction between them. This modular approach is applicable for virtual interaction and enables interaction between people that are not in the same physical location. The three key stages of the interaction framework are: training, testing and evaluation, as illustrated in Fig. 4. The training phase is performed offline for each action and uses the training data to learn action models. The testing phase is executed for each frame in real time to provide online interaction recognition. Actions from different people are detected independently. At each frame, these detections are combined to infer the current interaction.

An existing approach for online action recognition is to represent each action by a reference point [3] [7] . An 'action point' is defined as a single time instance that an action is clear and can be uniquely identified for all instances of that action [13]. For example, the action point of a punch is defined as 'the time at which the arm is maximally extended'. An action point has no temporal duration which accurately represents some actions, for example a punch. However, this is not the case for all actions, such as the defend, which is defined as 'the time when two hands are positioned in front of the face' as in reality the hands remain in front of the face for a significant period of time.

To overcome the limitation of action points we propose action segments. In contrast to an action point, an action segment has temporal duration. The duration of the action segment is important for training action classifiers with

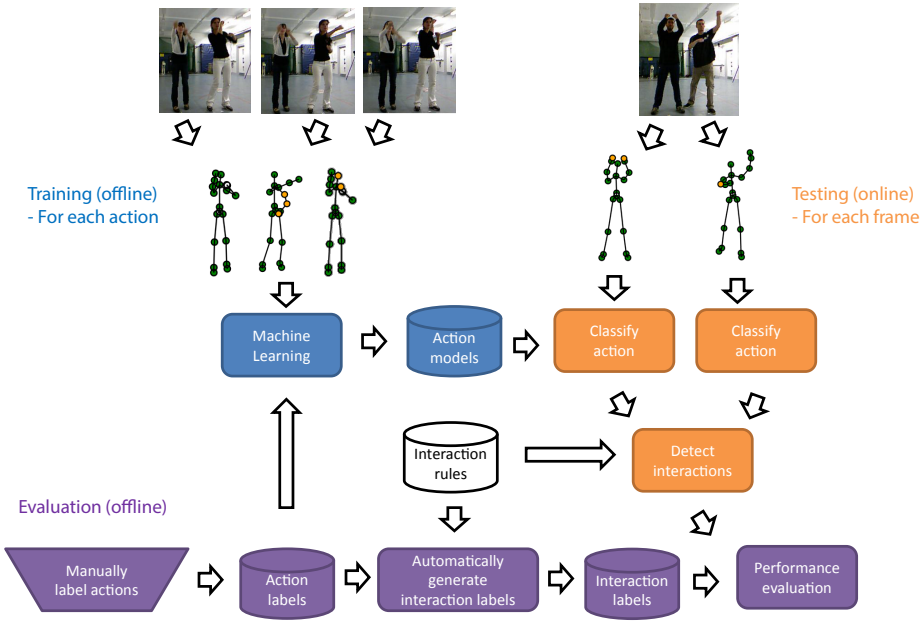

Fig. 4. The interaction framework comprising of three key stages: training, testing and evaluation.

consistent samples and should improve detection accuracy. It is also critical for recognising multiple interactions when one subject performs one long action and the other multiple short actions. An example is where one subject defends whilst the other subject punches him multiple times. These should all be detected as blocking interactions, but without considering the duration of the defend action, only the first would be detected as a block and the subsequent punches incorrectly as attacks.

4.1 Training

The training phase transpires offline and uses the data for each action to learn a model. The ground truth labels are used to select the frames and the subject within each frame that is performing the specific action. For each subject instance a feature vector is constructed and concatenated with all subject instances for the same action. The framework is generic so features from the colour, depth, skeleton or a fusion of these features can be used to train any classifier to learn action models.

In this work, we compare two approaches for training our action models, the first based on action points and the second based on action segments. Existing action recognition algorithms are based on action points, they use n frames before and after the action point to train their classifiers [3] [7] . This is appropriate for short actions but this new dataset confirms that under realistic

gaming conditions some actions have a long duration. Therefore, we recommend that action classifiers should be trained with all the frames that are part of the action segment in addition to a few n frames before and after the segment. Our results in section 5.2 show that adopting action segments instead of action points improves both the action and interaction accuracy.

To assess the complexity of our realistic dataset we use an established online action recognition algorithm with published results for multiple existing gaming datasets [3]. Specifically, we use Adaptive Boosting (AdaBoost) [8] with the same parameters and skeleton based features as reported in [3].

4.2 Testing

Testing sequences are processed online for real time detection. Each frame is divided into different people, which are classified into individual actions. These classifications are then combined to infer the current interaction.

Action recognition For each subject, a feature vector is created containing the same features as used for training the action models. The action model with the highest response depicts the action label for the current subject to provide real time action classification. The action model responses for the subject are summed over a sliding window of w frames to smooth the results and increase accuracy. This temporal filtering prevents broken actions caused by individual bad frames and therefore reduces the number of false positives. After smoothing, the highest action result determines the action label for the current subject. A change in action label is the detected action point, which is also the start of the action segment. To incorporate the duration of the action we also record the end of the action.

Interaction recognition To detect interactions for multiple people we must identify the interaction rules between people. These rules are application specific and include the valid combinations of actions together with timing constraints. These rules can be tailored by the application designer to include any necessary additional constraints.

The interactions for the G3Di dataset are depicted in Table 2, for compactness just two scenarios are shown. The action a and counter action ca, are checked at each frame together with a timing constraint f to detect interactions in real time. The timing constraint depends on the scenario, for example all the interactions in boxing are instant ($f = 0$), the action and counter action co-occur. However, other scenarios such as table tennis have a delay between the action and counter action ($f > 0$).

In this work we evaluate two approaches for detecting interactions, the first based on modelling actions as a single point in time and the second based on actions with temporal duration. In both cases, the detected interaction points can be compared with the ground truth interaction points to obtain a single F1 score. In the first case action points t, are used to represent the actions and

Table 2. Gaming interactions for the boxing and table tennis scenarios in G3Di.

Sport	Action	Counter Action	Interaction
Boxing	Right Punch	Defend	Block
	Left Punch	Defend	Block
	Right Punch	Other	Attack
	Left Punch	Other	Attack
	Right Punch	Right Punch	Attack
	Right Punch	Left Punch	Attack
	Left Punch	Left Punch	Attack
Table Tennis	Serve	Forehand hit	Rally
	Serve	Backhand hit	Rally
	Serve	Other	Miss
	Forehand hit	Forehand hit	Rally
	Forehand hit	Backhand hit	Rally
	Forehand hit	Other	Miss
	Backhand hit	Backhand hit	Rally
	Backhand hit	Other	Miss

interactions are detected if the action and counter action occur either at the same time or after a fixed delay, as described by Equation (1).

$$\phi\left(a_t, ca_t\right) = \begin{cases} 1 & ca_t - a_t \;=\; f \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In the second case action segments are used to represent the actions and interactions are detected if the action and counter segments overlap either at the same point in time or after a fixed delay, as described by Equation (2).

$$\psi\left(a_s, a_e, ca_s, ca_e\right) = \begin{cases} 1 & \text{if } \left(a_s + f \leq\; ca_e\right) \;\&\; \left(ca_s \leq\; a_e + f\right) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Where s and e represent the start and end of the action segment respectively and $s \leq\; e$.

4.3 Interaction Evaluation Framework

To evaluate the performance of both action and interaction recognition algorithms on this new dataset, action and interaction online metrics and ground truth annotation are required. For action recognition, an existing evaluation metric is the action point metric [13]. Action points enable latency aware evaluation of online action recognition systems. This metric will be used to assess the timing of the action points with respect to the ground truth.

For interaction evaluation the existing frame based metrics [6] [15] include temporal constraints but do not evaluate the latency of the detection. To overcome these limitations we propose a new interaction point based evaluation metric that can evaluate both the accuracy and latency of the interactions. The

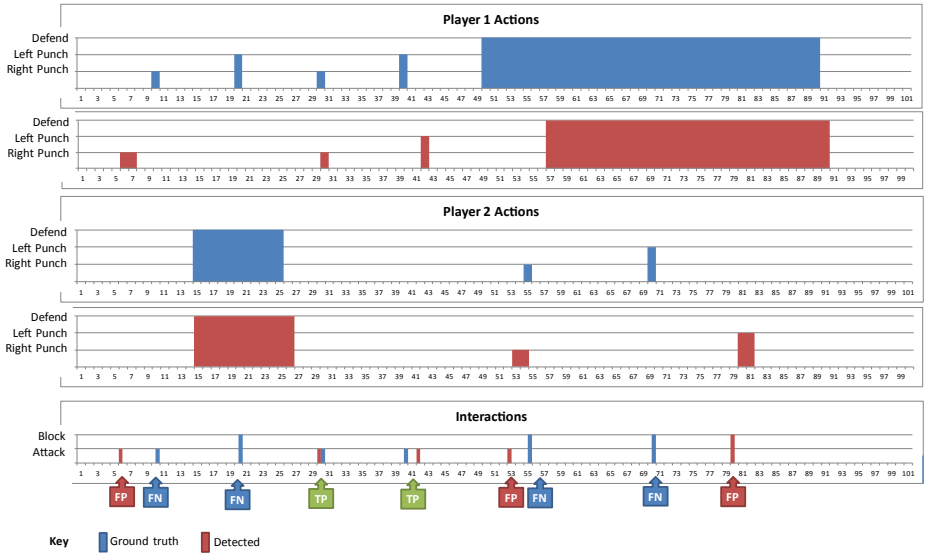

Fig. 5. An example timeline for a boxing game, showing the true positives (TP), false positives (FP) and false negatives (FN). A TP, is a correct interaction identified within Δ frames of the ground truth. A FN, is an undetected interaction on the ground truth. A FP, is an incorrect interaction detected.

interaction points can be evaluated in a similar manner to action points, to obtain a single F1 score for an easy comparison of different interaction algorithms.

Dataset annotation The ground truth for the action dataset was conventionally annotated by manually labelling each action point and each action segment. The interaction ground truth could have also been manually labelled but it was more efficient to automatically construct the interaction labels from the action ground truth labels. The ground truth interactions are automatically labelled based on the set of rules that govern the interactions for a particular game (as described in Section 4.2).

Interaction evaluation metric To evaluate the timing accuracy of an interaction we adapt the existing action point metric [13] to assess the timing of the interaction points with respect to the ground truth. The interaction points are assessed for detection and timeliness and an F1 score is generated. The acceptable latency of the interaction is application specific and can be adjusted with the Δ parameter. To clarify the assessment of interaction points a dummy timeline for a boxing game has been created (Fig. 5), showing the ground truth and the detected points for actions and interactions. The precision and recall are measured for each interaction and both of these measures are combined to

calculate a single interaction F1-score (F1). To measure accuracy for multiple interactions, the mean interaction F1-score is calculated over all interactions.

5 Real Time Results

The interaction framework proposed in Section 4 can be used with any classifier. To obtain our results we used a multiclass implementation of Gentle AdaBoost. Following [3] [7] we use a 'leave one person out' protocol. As there are 12 people, this process is repeated 12 times with different subsets to obtain the average performance.

5.1 Action Recognition Results

To evaluate the complexity of the actions in the new dataset G3Di in comparison with actions in existing gaming datasets we recreate experiments previously performed on those datasets. For a fair comparison, we use the same classifier, pose based features, parameters and action point evaluation metric as published in [3]. Specifically, a vector of 297 features is extracted for each skeleton in each frame which is a concatenation of 57 position difference features, 60 position velocity features, 20 position velocity magnitude features, 80 joint angle features and 80 angle velocity features. The latency parameter Δ was fixed at $330ms$ for consistency with previously published results [3] [7].

The previously published action recognition F1 results for the G3D [3] and MSRC-12 [7] datasets and the new result for the G3Di dataset are shown in Table 3. The F1 for the new dataset is the lowest, indicating that G3Di is more challenging, especially as the actions in the G3Di boxing scenario are a subset of those actions found in the G3D fighting scenario.

Table 3. Gaming dataset action results highlighting the complexity of the new dataset.

	G3D dataset (Fighting)	MSRC-12 dataset (FPS)	G3Di dataset (Boxing)
Action F1	0.896	0.643	0.426

5.2 Interaction Recognition Results

To demonstrate our interaction detection framework on the new dataset G3Di we initally use action points for detection and the same experimental setup as described in the previous section to get a baseline result for the G3Di boxing dataset. We then incorporate action segments into our interaction detection framework and repeat the same experiments. We performed quantitative evaluation using the action point and interaction point metrics and qualitative evaluation by visually analysing our failure cases. The quantitative results are

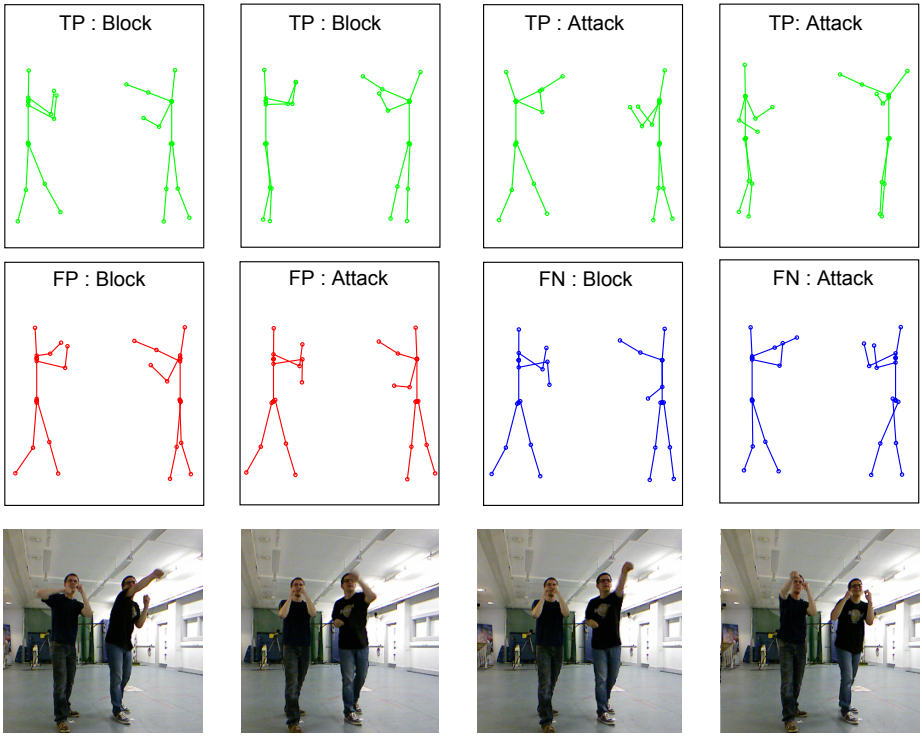

Fig. 6. Examples of real time interaction detection. Correct detections TP, are displayed in green and incorrect detections are shown in red for FP and blue for FN. The action detected by the system is displayed above each pair of skeletons. The colour images below each failure cases highlight the major sources of error, transition error between actions and noisy joints.

shown in Table 4, which highlight an increase in both the action recognition and interaction performance by 13%. This confirms that the duration is important for detecting actions and subsequently interactions in real time.

The qualitative results are displayed in Fig. 6, including examples of correct and incorrect detected interactions. The colour images below each failure cases highlight the major sources of error, which are transition error between actions

Table 4. Gaming dataset interaction detection results.

	Baseline method (action points)	Proposed method (action segments)
Action F1	0.426	0.561
Interaction F1	0.448	0.578

and noisy joints. The transition errors occur when a player moves quickly from one action class directly to another without passing through a neutral state. As Fig. 6 illustrates, when a player transition from a block to a punch it is difficult to infer from the skeleton data alone the current action. The colour images provide additional information to help differentiate the current action and suggest a fusion of colour and skeleton may improve detection in these cases. The transition errors support our claim that this new gaming dataset is more challenging than existing gaming datasets with simple transitions. Some of the failure cases were also related to noisy joints in our skeleton data but it is important to have some noise in a realistic gaming dataset as in a real home environment there may be noise caused by direct sunlight. The noisy joints endorse our inclusion of joint confidence and additional modalities (depth and colour) in our new dataset.

6 Conclusions

A novel realistic and challenging human interaction dataset, G3Di for real time multiplayer gaming is introduced. It overcomes the limitations of existing 3D gaming datasets that only contain a single player with simple action sequences. Our interaction framework recognises individual actions with low latency for real time interaction detection. The incorporation of the action duration in our framework improved both the action and interaction performance. We also proposed an interaction evaluation metric for real time applications which assesses both the accuracy and latency of the interactions. Experimental results indicate higher complexity of the new dataset in comparison to the existing gaming datasets, highlighting the importance of this dataset for designing algorithms suitable for realistic interactive applications. Our future work is to develop an algorithm that fuses features from the depth or colour with the skeleton features to improve the performance of our interaction detection framework. Additionally, we will incorporate person to person features in the interaction framework to recognise traditional human interactions.

7 Acknowledgments

We would like to thank the staff, students and interns of Kingston University for participating in the gamesourcing. We would also like to thank Kevin Bottero and Nicolas Ferrand from Ecole Nationale Suprieure d'Ingnieurs de CAEN (ENSICAEN) for their assistance in collecting and annotating the new dataset.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys 43(3), 16:1–16:43 (2011)
2. Aggarwal, J., Xia, L.: Human Activity Recognition From 3D Data: A Review. Pattern Recognition Letters (2014)

3. Bloom, V., Argyriou, V., Makris, D.: Dynamic Feature Selection for Online Action Recognition. In: Salah, A., Hung, H., Aran, O., Gunes, H. (eds.) Human Behavior Understanding, pp. 64–76. No. 8212 in LNCS, Springer International Publishing (2013)
4. Bloom, V., Makris, D., Argyriou, V.: G3D: A gaming action dataset and real time action recognition evaluation framework. Workshop on Computer Vision for Computer Games, IEEE Conference on Computer Vision and Pattern Recognition (CVPR Workshops) pp. 7–12 (2012)
5. Butler, A., Izadi, S., Hilliges, O., Molyneaux, D., Hodges, S., Kim, D.: Shake'n'Sense : Reducing Interference for Overlapping Structured Light Depth Cameras. Human Factors in Computing Systems pp. 1933–1936 (2012)
6. Escalera, S., Baró, X., Gonzàlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce, V., Escalante, H.J., Shotton, J., Guyon, I.: ChaLearn Looking at People Challenge 2014: Dataset and Results. ECCV workshop (2014)
7. Fothergill, S., Mentis, H., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems pp. 1737–1746 (2012)
8. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm (1996)
9. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29(12), 2247–2253 (2007)
10. Hu, T., Zhu, X., Guo, W., Su, K.: Efficient Interaction Recognition through Positive Action Representation. Mathematical Problems in Engineering 2013, 1–11 (2013)
11. Li, W., Way, O.M.: Action Recognition Based on A Bag of 3D Points. IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPR Workshops) pp. 9–14 (2010)
12. Ni, B., Moulin, P.: RGBD-HuDaAct: A color-depth video database for human daily activity recognition. IEEE International Conference on Computer Vision Workshops (ICCV Workshops) pp. 1147–1153 (2011)
13. Nowozin, S., Shotton, J.: Action Points: A Representation for Low-latency Online Human Action Recognition. Tech. Rep. MSR-TR-2012-68, Microsoft Research Cambridge (2012)
14. Park, S., Aggarwal, J.: Event semantics in two-person interactions. IEEE Conference on International Conference on Pattern Recognition (ICPR) 4, 227–230 (2004)
15. Ryoo, M.S., Aggarwal, J.K.: UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA) (2010)
16. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. IEEE Conference on International Conference on Pattern Recognition (ICPR) 4, 32–36 (2004)
17. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1297–1304 (2011)
18. Wang, J.: Mining actionlet ensemble for action recognition with depth cameras. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1290–1297 (2012)
19. Yao, A., Gall, J., Fanelli, G., Gool, L.V.: Does Human Action Recognition Benefit from Pose Estimation? Proceedings of the British Machine Vision Conference pp. 67.1–67.11 (2011)

20. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D., Brook, S.: Two-person Interaction Detection Using Body-Pose Features and Multiple Instance Learning. IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPR Workshops) pp. 28–35 (2012)