–Independent Work Report Fall, 2015–

USING KINECT TO LEARN HOW TO BALLROOM DANCE

Author: Michael Li (ml13@princeton.edu)

Advisor: Professor Thomas Funkhouser

0. ABSTRACT

This paper details the design, development, and evaluation of a Kinect-powered application to facilitate the instruction of ballroom dance. The application uses the Kinect camera's skeletal tracking capabilities to teach and evaluate users through various ballroom dance positions and concepts, taking the form of a number of training modules that end with a game-like assessment portion. This application aims to fill a hole in ballroom dance instruction, providing the ease of access of self-study materials alongside the quality of instruction of live coaching.

1. INTRODUCTION

Ballroom dance is an often greatly-appreciated, but under-participated-in art form. Many people enjoy watching the graceful progression of a waltz or tango and the energy and sensuality of a well performed cha cha or rumba, but few people know how to perform these dances themselves. Often times, this is not due to lack of desire to learn, but lack of availability of viable and useful teaching materials. Therefore, the goal of my project was to use computer vision to provide people with an easily-accessible, yet intelligent, automated ballroom coach. I intended to do this by creating an application that uses footage captured by a Kinect camera to provide coaching and assistance to people while they practice.

The idea to use technology to aid the acquisition of ballroom dance arose from my own experiences on the Princeton Ballroom Dance team, and a resulting awareness of how both (a) limited the current methods of acquisition are, and (b) how automate-able the teaching process can be. As with many other forms of dance, ballroom dance has an idea of correctness based on the accuracy of the poses struck by the dancer. However, ballroom dance, especially the form that is performed by competitive dancers, is somewhat unique in this regard, in that it is a very prescriptive form of dance. For example, in competitive ballroom, there is actually a syllabus of clearly defined, universally allowed steps, each with their own very clear definitions of right and wrong. This sets it apart from things like contemporary dance or hip hop, where the choreography is much more up to interpretation and the individual's sense of style.

The rigidity of ballroom dance means that in the vast majority of cases, external instruction from a coach is essential. This type of instruction is extremely valuable to beginner dancers in order to provide assistance in developing the muscle memory for how a certain position or step is supposed to feel, because the accuracy of a pose is very difficult to determine by oneself when first starting out. However, this imposes a limitation insofar as progress is often restricted to practice times when there are either more experienced dancers available, or during lessons with the coach. All other practice times, at least in the beginning, are much more hit or miss.

At the same time, this lack of need for interpretation means that ballroom is perfect for something like a Kinect camera to watch and provide feedback for. Since each step has a well defined

Figure 1. Illustration of how positions in ballroom can be scored using a computer. The position above is called frame, and the left arm (shown on the right) should be held upward at a 90˚ angle from the elbow. Deviations from this right angle can objectively be said as incorrect.

idea of ground truth, it then simply becomes a matter of identifying which model to compare against, and then some degree of subjective accuracy relative to that model can easily be obtained and then scored (Figure 1).

From these two factors followed the key idea of this project, which is to use computer vision and the intelligence of a computer to provide instruction that is both as accurate and useful as a live ballroom coach, but that can be accessed according to one's own schedule. This was achieved by using the Kinect camera to obtain key information about a user's position, and then I developed a scoring and feedback algorithm to provide the user with advice and feedback on how to improve.

2. RELATED WORK

Unlike the projects of many of my classmates, there does not exist a large body of existing work in computer science that is working toward the same goal as this specific project, which is the automated instruction of ballroom dance. However, there does exist plenty of related work in ballroom instruction in general, and there are also a number of applications that attempt to combine dance and computer vision in some way. In the following section, we will tackle these two separate categories of related work individually.

2.1 Ballroom Instruction

The main existing methods of ballroom instruction can be summarized as either (1) learning from a live ballroom coach, or (2) self-study from resources like a textbook, or videos, such as instructional DVDs. Each of these methods has their pros and cons, but none of them provide the optimal learning experience on their own.

In terms of quality of instruction, learning from a live ballroom coach is the obvious winner. The combined knowledge of the instructor, with their ability to give real-time feedback on how well a student is doing means that live instruction possesses the highest theoretical rate of improvement. However, there are many practical considerations and limitations that significantly hamper this theoretical rate. The first of these considerations is availability. Believe it or not, ballroom coaches exist in relatively limited supply, which both means that scheduling a private lesson can be very difficult, and that most coaches charge exorbitant rates, making the entry cost ballroom very high. An alternative to private lessons is group lessons, such as those provided by groups like our own Princeton Ballroom Dance team. But this comes with its own issues, namely lack of individual attention, greatly watered down rate of progression, as well as its own slew of scheduling difficulties.

This brings us to the more accessible method, which is self-study. While self-study can be done at one's own availability, it is nigh impossible to reach the same level of accuracy and clarity as guided lessons. Instructional videos provide helpful visuals, but from very limited angles, and are usually geared towards the leader's steps, which means that followers are often left out of luck. It

is also very hard to simultaneously watch an instructional video and figure out what you are actually supposed to do, which means that learning from these videos is a very inefficient, as much time is wasted re-watching the same clips, and extraordinarily awkward experience, often involving attempting to do the steps with computer in hand. Textbooks provide the same limitations as instructional videos, but with the added benefit of slightly more explicit explanation of what to do, and additional limitation of less effective visuals.

Figure 2. Current applications of the Kinect camera that relate to dance

It is painfully clear that none of the existing resources for learning ballroom dance check all the boxes.

2.2 The Intersection of Dance and Computer Vision

There are a number of existing applications that use computer vision to judge peoples' dancing. In fact, many of the more successful titles for the Kinect are dancing games, such as *Just Dance*, which is now in its 7th flagship iteration, *Michael Jackson: The Experience*, which is actually a themed spin-off of *Just Dance*, and *Dance Central* (Figure 2). All of these games operate around the central premise of presenting the player with some choreography set to a number of predetermined songs, and using pose detection to score how well the user executes said choreography. These games are usually extremely popular both because of the intrinsic appeal of music and

dance, and because of the engaging nature of a game. In many ways, this is very similar to my project, in that is uses computer vision to grade a users dancing and provide real-time feedback, in this case, in the form of a score.

However, the problem with these games is that insofar as an instructional resource, they are pretty useless [1]. While the number of moves asked of the player are numerous and varied, they are also very arbitrary and are not explicitly taught, without any explanation regarding how to perform the moves more effectively. The scoring formulae used by these games are also very hard to understand and interpret, especially since they are based in complex machine learning algorithms [2]. In the heat of the game, one notices little more than whether their last move was “Excellent!” or simply “Good”, much less understanding what those assessments actually mean. A final score of S vs C gives somewhat of an indication of skill, but aside from running the same level over and over again with trial and error, gives little in the way of teaching the player how to improve.

3. KEY IDEA

The key idea of my project that solves goal of ballroom instruction better than any of the existing alternatives, is that by bringing computer vision into the mix, it is able to combine the engagement of a video game, with the instructional value of a professional ballroom coach, with the accessibility of a self-study resource, to create an all purpose ballroom instruction suite.

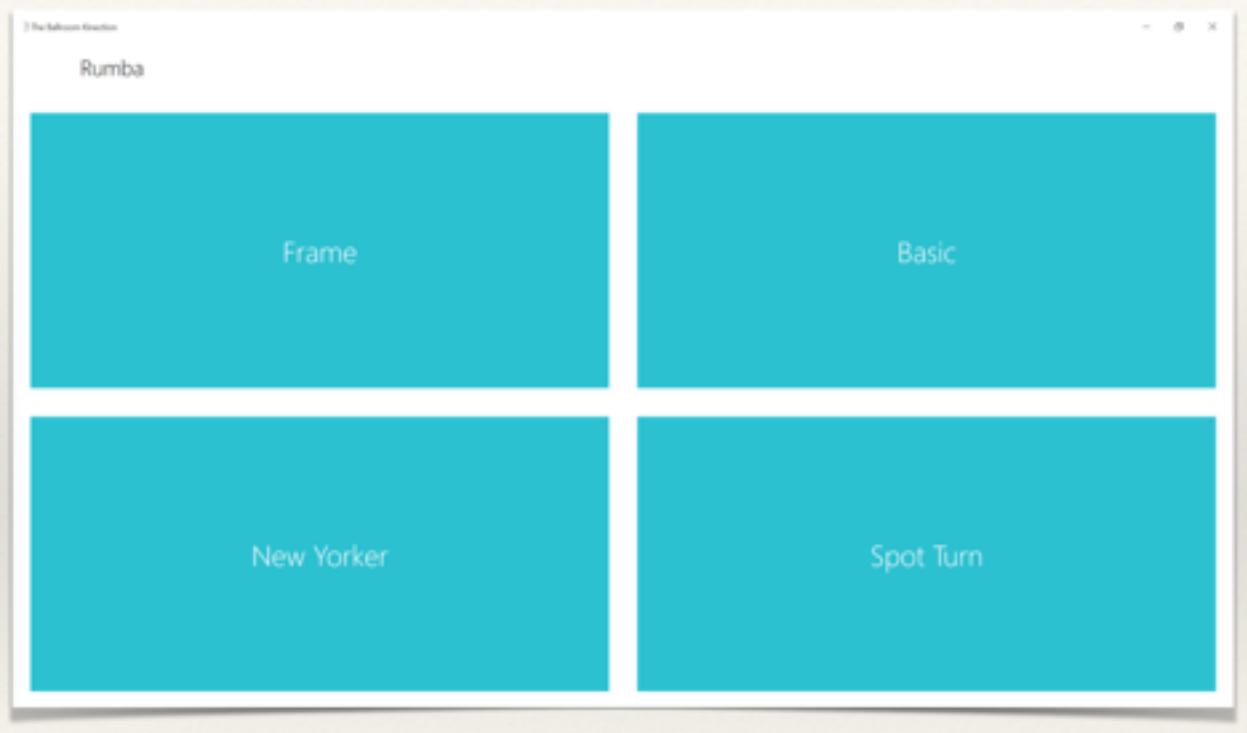

Figure 3: The main screen of my application. From it, the user can choose one of four training modules to practice.

4. OVERVIEW OF THE SYSTEM

The end result of my project was a coach/game hybrid program, that takes a user through a number of training modules to teach a number of key steps in ballroom dance (Figure 3), and then takes the form of a video game to test and reinforce these principles. In order to limit the scope of the project to something attainable within a semester, I decided to only teach one dance–rumba, and to limit the number of steps/positions to four–frame, basic, New Yorker, and spot turn. Rumba was decided on as the dance to teach because it does not require the dancer to travel significantly around the floor, which means it was suitable for a the relatively small space I expected most likely available to my users, and that the Kinect camera would have a good view of the user throughout the steps. Rumba was also chosen because the steps are essentially identical for

both leaders and followers. This meant I would only need to program each step once, instead of having handle each role separately.

Frame, basic, New Yorker, and spot turn were chosen as the four steps to teach, because they form the basis of any beginner rumba routine, yet at the same time, provide sufficient variety to prevent the experience of using my program from being repetitive or boring. These are also the steps that our coach starts with when teaching new members to the team, so they felt appropriate for my program to teach.

As four steps are taught by the program, the program itself is also split into four main training modules. Each explains to the user the importance of the step being used and things to pay attention to in execution of the step. For example, for frame, the program articulates that frame is the

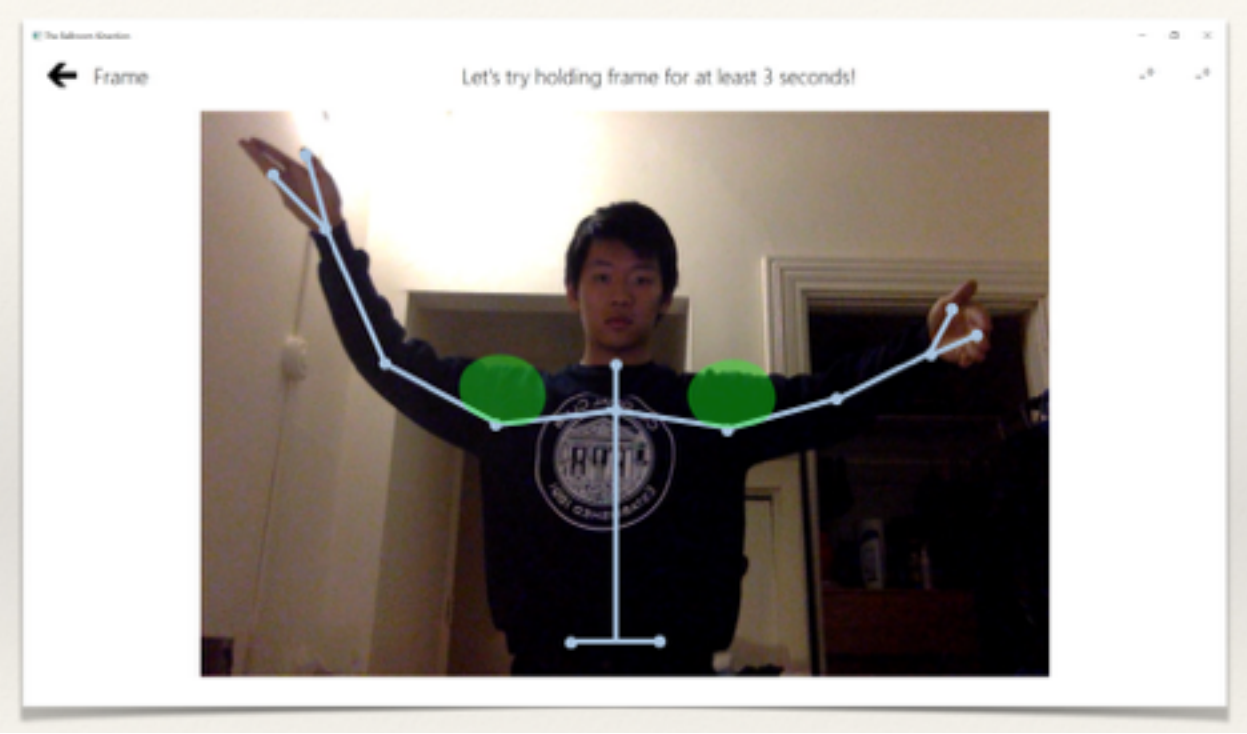

Figure 4: An example of an exercise for frame

basic arm position that one should maintain throughout the routine, and that it consists of keeping the shoulders down, spine straight, and elbows up.

Each module then takes the user through a number of exercises to try to establish the aforementioned muscle memory of each step. For frame, the program asks the user to assume frame, holding it for at least five seconds each time. Throughout the exercise, the program also gives real-time instruction on the screen on what the user may be doing incorrectly (Figure 4). After the user decides they have had enough chances to practice, there is a graded run of the exercise, where the user is timed and asked to do the exercise three times in quick succession, after which a score is calculated based on how quickly they are able to accomplish the objective. The scoring algorithm for this was as follows: Each exercise calculates how long it should theoretically take for an optimal performance. It then adds a slight buffer to this (equal to one-half of the total time) as a grace period. Any additional time taken follows a exponential deduction proportional to the excess of the expected time taken. The grace period is added in order to account for things like transition time and time needed to read the instructions in order to make perfect scores possible for humans, and then the score deductions reflect that there is still room for improvement. In other words, the final score for a run taking T seconds on a challenge with expected completion time E is:

$$S = 100 \times \min\left(1,\ e^{(1.5E-T)/E}\right).$$

For example, holding frame for five seconds three times takes at optimal, 15 seconds. An attempt taking 25 seconds would be given 84.6%.

I will now go more in-depth into the pose recognition and scoring part of my project, which is arguably the most interesting part. The pipeline for this can broadly be split into three main stages:

4.1 Using the Kinect Camera to capture information about the user

My program utilizes two of the five data streams that the Kinect camera exposes: the color data stream and the body data stream. The color data stream works as a normal camera would, and returns the RGB values of all the pixels within the Kinect camera's range of sight. This, although not directly related to the scoring portion of my program, acts as the backbone for the real-time feedback my program provides the user. As was seen back in Figure 4, while the user goes through the various training modules, they are presented with a color image of what they are doing so it is easier to visualize what their interpretation of the position looks like and what they can do to improve it. Guidelines are also overlaid on the color image to provide further feedback, the drawing of which will be discussed in *Section 4.3*.

The other main data stream my program uses is the body data stream, and this provides all the necessary joint and skeletal information that the program processes in the scoring of the user in their various attempts at the ballroom positions. All data streams are broken down into frames, and each frame of the body data stream gives a variety of preprocessed information regarding the users in the camera's line of sight. The first thing the body data stream does is identify the individual humans it sees, and ranks them in order of prominence. This is useful because it allows

me to only take in data for the primary user and provide feedback for them, which alleviates clutter and confusion.

The camera also gives XYZ-coordinates of a number of predetermined joints and body parts. These include things like the head, should joint, and elbow joint. This data is crucial to my program as it performs matching and scoring of the poses based on joint angles. Joint angles were chosen as the input to process as it meant that bodies of all different sizes would be able to use the same ground truth values, instead of needing to recalculate based on the lengths of individuals' wingspan and height. Joint angles were also used because they were simpler for the program itself to handle, and were also a more elegant metric through which to provide the user with feedback, especially since the locations of three separate joints can essentially be summarized by the one angle.

After reading in the information from the body frames, we do some basic trigonometry to calculate the aforementioned joint angles. The angles are calculated using the X and Y values, using a simplifying assumption that all depth (or Z) values were the same. We are somewhat able to get away with this assumption because in the poses that we care about, the target joint angles are all for joints that should fall in relatively similar depth planes. However, this does end up giving rise to some false negatives and positives, so an improvement would be to properly calculate these joint angles taking all three dimensions into consideration.

4.2 Calculate and Quantify the Angle Deltas Between the Ground Truth Models and the User

After the target joint angles have been calculated for the user, they are compared to the ground truth models stored in the program that were obtained by tracking my ballroom coach. For each target angle, I established a band of acceptable values, and then the program identifies whether the user's model falls within these bands, or if they are either too high or too low, returning to the feedback algorithm essentially a three-value enumeration off which to base feedback. A band of acceptable values was used, because although the positions in ballroom are well-defined, there is still some margin of variation depending on an individual's body and personal comfort, plus any given angle is very difficult to maintain to the nearest degree for any reasonable amount of time.

4.3 Generate and Deliver to the User Real-Time Feedback

The next part of the pipeline takes the three-value enumerations generated by the scoring algorithm and translates them into human-understandable feedback. This part mainly consisted of identifying if an angle was too large, too small, or just right, and then based on the individual body parts involved in that angle, provide customized feedback for how to fix the error. For example, for frame, if the left elbow-shoulder-sternum joint, illustrated back in Figure 4, is too small, that means that the elbow is too high and the angle is being pinched. The program takes this assessment and then translates it into a suggestion to let the user relax their left arm slightly. This feedback is then displayed in the top center of the screen, above the rendering of the user, to allow them to dynamically adjust their position. Arcs for the target joint angles are also rendered and colored to reflect correctness: green for correct, red for too much activation (in the case of

frame, too small of an angle caused by lifting the elbow too high), and yellow for too little (too large of an angle caused by a drooping arm).

An additional piece of feedback is in rendering of the skeleton, which is overlaid on the user's body. This is to make it clear exactly which body parts the program is tracking, and to try to reduce the mystery in where the angles are coming from. This particular decision was in response to the existing Kinect dance games, which are extremely enigmatic in their scoring of players, partially due to the sophisticated machine learning algorithms employed by most of these games. With the skeletons, it is incredibly transparent what the camera is tracking, and therefore also very easy to understand which body parts need to be moved.

For clarity and to create a good user experience, the skeletons are rendered intelligently, only drawing the bones and joints that are completely visible in the frame. Bones and joints will be added and removed as body parts move in and out of the line of sight. This is originally an issue because invisible joints do not have a defined position, and so they default to the origin, which meant that as soon as a joint moved out of the screen, a number of large and obstructive lines would be rendered from the top left corner to various places in the frame.

5. EVALUATION

5.1 METHODS

My project's success by performing a series of test sessions with actual dancers and users. The tests was structured as follows. Each test subject was given a 15-minute block with the program

& Kinect set up. The user was asked to get through as much of the training program as they were able, following the logical progression of frame, basic, New Yorker, and finally, spot turn. As the training program gave context and instructions on the various steps it contained, no external instruction was provided, and the user would rely on the program itself to indicate how to proceed through the modules and how to perform the various steps. Once the user felt sufficiently confident with each taught step, the program continued on to the aforementioned game portion, where the user was presented with a set of challenges related to the training module they had just completed. An example of a challenge would be "Maintain frame for a continuous 5 seconds three times", which would then be scored according to the grading policy mentioned in the overview of the system.

In order to get a trend in the scores, each challenge was asked to be done three times (so for the exercise about, the user would have held frame for five continuous seconds nine times). The trend in these score served as an objective view of the success of the project. For example, if the user's scores demonstrate an upward trend, then that will be considered a success.

Additionally, surveys were given to each participant in order to ask them more subjectively about the success of the project. The exact questions presented are as follows:

- Did you enjoy the training program?
- Did you understand everything that was asked of you?
- Was the feedback useful?
- Do you think you are a better dancer having gone through it?

		Take 1	Take 2	Take 3
Participant 1	**Time**	28	25	26
	Score	69.30	84.65	79.19
Participant 2	**Time**	33	28	32
	Score	49.66	69.30	53.08
Participant 3	**Time**	27	28	34
	Score	74.08	69.30	46.46
Participant 4	**Time**	35	32	24
	Score	43.46	53.08	90.48
Participant 5	**Time**	32	25	24
	Score	53.08	84.65	90.48
Participant 6	**Time**	24	20	21
	Score	90.48	100.00	100.00
Participant 7	**Time**	25	24	22
	Score	84.65	90.48	100.00
Average	**Score**	66.39	78.78	79.96

Figure 5. Scores for the seven participants in the exercise for the frame module

		Take 1	Take 2	Take 3
Participant 1	**Time**	52	57	49
	Score	79.19	67.03	87.52
Participant 2	**Time**	55	57	52
	Score	71.65	67.03	79.19
Participant 3	**Time**	58	53	57
	Score	64.83	76.59	67.03
Participant 4	**Time**	55	54	52
	Score	71.65	74.08	79.19
Participant 5	**Time**	49	48	51
	Score	87.52	90.48	81.87
Participant 6	**Time**	48	44	43
	Score	90.48	100.00	100.00
Participant 7	**Time**	42	45	39
	Score	100.00	100.00	100.00
Average	**Score**	80.76	82.17	84.97

Figure 6. Scores for the seven participants in the exercise for the rumba basic module

- Do you see yourself using the program for your own practice in the future?
- How does the program compare it to other methods, such as a live coach or instructional videos?

5.2 Data Set

A total of seven people were asked to undergo the evaluation process as outlined above. Five subjects were complete novices to ballroom dance, and so they served as an indication of how useful the program is as a step-zero instructional tool. The remaining two test subjects were members of the ballroom team, and so acted as a metric of how accurate the training program was as a trainer. In the results, Participants 1-5 are the novices, while the ballroom team members are Participants 6 and 7. The score sheets are colored to make the distinction clearer.

As 15 minutes was not sufficient time for any of the participants to make it past module two, the module for rumba basic, only the results for these first two will be considered in the following evaluation. Participants' scores are summarized in figures 5 and 6.

The survey results are less conducive to visual summary, and so will be instead be discussed at length in the following section.

5.3 Interpretation of Results

5.3.1 Scores

The scores for the participants did not show any kind of substantial trend in either direction, with scores increasing and decreasing somewhat randomly across the board. That said, a number of observations can still be made from the data.

The first observation is that while the differences in scores may not have proportionally displayed differences in experience, the experienced dancers did noticeably better in both exercises when compared to the novice dancers. This is encouraging because it means that the experienced users had an advantage in completing the challenges, which indicates that proper ballroom technique was being asked of the users. The fact that they did not perform significantly better may likely have been a result of the artificialness of the tasks asked, combined with slight finickiness of the program itself. Therefore, the conclusion we can draw from this is that while the content of the program is more or less proper ballroom technique, the specifics of how the users are scored and the methods of assessment may need some additional work.

Another observation is that the experienced dancers' scores tended to rise, which means that their ability to complete the tasks improved. One possible interpretation of this is that users' dancing abilities were improving with successive runs. However, a more likely interpretation is that this is a result of better understanding what each task was asking and essentially learning how to play the game, rather than any substantial improvement in dancing ability. To some degree, this is to be expected considering the limited time frame and room for learning provided to the users be-

tween successive trials. This implies that the evaluation method itself may have been intrinsically flawed, and so restructuring of evaluation methods should be done in the future to obtain more meaningful results.

A final observation is that overall, the participants did fairly well on the tasks, with scores averaging from the high 70s to low 80s. While this may be an indication of an overly lenient grading scale, it does seem to imply that the teaching modules were relatively successful in instructing the user on how to perform each step, which bodes well for the overall success of the project.

5.3.2 Surveys

Generally speaking, the survey results tell a similar story to the scores. The participants universally said that they enjoyed their experience with the program, though it was mentioned that it was occasionally a bit boring and repetitive, largely in part due to the similar nature of all four modules.

The participants also mostly noted that the instructions were clear, but that the individual pieces of feedback occasionally did not seem appropriate, and that there were a couple of instances where false positives or negatives were reported by the program. One example was that when a user's arms were completely at rest, the program reported an elbow-shoulder-sternum angle that was very small, and told the user to all their elbow to drop slightly, which is obviously incorrect. This means that either some adjustments to how joint angles are calculated is necessary, or better handling of specific edge cases needs to be built in.

Users also said that, aside from the occasional hiccup, the feedback was very useful, especially the skeleton and the highlighted target angles. Combined with the textual feedback, it was very easy to see what they were doing wrong and how to fix it.

However, the overall conclusions from most participants was that they were not significantly better dancers after using the training program, and that they were not sure they would necessarily choose this over a live coach, at least not as the primary method of instruction. Participants noted a lack of instruction on rhythm and counting, lack a final section outlining a cohesive and comprehensive routine, and somewhat limited variety in types of feedback as reasons this would not be preferable over a live coach for beginners. However, participants did also mention that the added accessibility of the program in comparison to both private and group lessons meant that they would be very happy to use it for guided individual practice, especially if they were in a space that did not have a proper dance mirror.

Overall, it appears that the program has more value as a training resource than a teaching resource, with the distinction being that complete novices are better off opting for an alternative, more hands-on teacher, while slightly more experienced dancers will be able to use this for things like drills. Thus, while some additional work needs to be put into the program to make it fully functional as a standalone instructional resource, it was a relatively successful attempt.

6. CONCLUSION

We have presented a Kinect-based computer program that attempts to fill a gap in ballroom education, providing both the instructional value of a live coach and the low-cost of entry of self-study learning materials. The program uses body and joint data pulled from the Kinect camera to analyze user's dancing compared to ground truth models, and presents real-time feedback and suggestions for improvement, in the format of a series of training modules and assessment challenges. While additional work needs to be put into the content of the program for it to fully reach this goal, it serves as a basis for what could be a very cheap, but effective manner of learning how to ballroom dance.

This additional work consists mainly of a wider range of content for the program to provide. Aside from the obvious improvement of teaching all ten dances in the ballroom syllabus, this would include things like better instruction of basic dance principles, such as as timing, hip rotation, and progression across the floor, a wider variety of taught steps as well as how to incorporate these steps into a cohesive routine, and a more varied set of drills to solidify these teachings.

On the technical side, improvements still need to be made to the manner in which joint angles are calculated to make the system more robust, and a method to analyze the positions of body parts not explicitly tracked by the Kinect SDK is also necessary. For example, in frame, the alignment of the spine is very important, but this is difficult to detect with the Kinect as it only captures the sternum and pelvis locations, instead of the entirety of the spine.

Implementation of these changes will bring my program closer to being the comprehensive ball-room instruction suite that it set out to be.

7. RESOURCES

(1) Clark, Noelene. "'Dance Central 3' vs. 'Just Dance 4': Which One Has the Right stuff?" Hero Complex. Los Angeles Times, 18 Oct. 2012. Web. 15 Dec. 2015.

(2) Raptis, Michalis, and Darko Kirovski. "Real-Time Classification of Dance Gestures from Skeleton Animation." (n.d.): n. pag. *Microsoft Research*. Eurographics, 2011. Web. 15 Dec. 2015