Tracking gesture to detect gender

Andrea Stevenson Won, Le Yu, Joris H. Janssen and Jeremy N. Bailenson

Department of Communication, Stanford University
{aswon@stanford.edu, billyue@stanford.edu, joris.h.janssen@philips.com, bailenso@stanford.edu}

Abstract

Nonverbal behavior is a very important part of human interactions; and how this behavior is tracked and rendered is also key to establishing social presence. Tracking nonverbal behavior is useful not only for rendering signals via avatar, but also for providing clues about interactants. In this paper we describe a novel method of determining identity (i.e., gender) using machine learning with input taken from the Microsoft Kinect. Twelve men and twelve women performed a number of gestures in front of the Kinect. A logistic regression used ten posture and gesture features (e.g., angle between shoulders and neck) to predict gender. When presented with a person it has never seen before, the model was 83% percent accurate in predicting whether the person was a man or a woman, even from very short (i.e., ten seconds) exposures to the test participants. We discuss the usefulness of the current research tool for presence, as well as point out practical applications.

Keywords: **presence, co-presence, social presence, Kinect, gesture, gender detection, machine learning, computer vision, nonverbal communication**

1. Tracking gesture to detect gender

Some elements of nonverbal behavior that have been studied in virtual environments include eye gaze (Garau et al., 2003; Bailenson, Blascovich, Beall, & Loomis, 2001), posture (Vinayagamoorthy, Steed & Slater, 2008), touch (IJsselsteijn, 2006), and purposefully communicative gestures such as waving or clapping (Kane, McCall & Collins, 2012). Any technology that can potentially port nonverbal information unobtrusively may be used to augment social presence. "The illusion [of presence] will be more complete if the medium is perceptually and psychologically immersive....and if we encounter people or entities within such a medium, even if there is no possibility of true social interaction with them, we are encouraged to respond to social cues they provide" (Lombard et al., 2000). As Biocca (1997) states "The body transmits information to other bodies through a kind of affective contagion." How that information is mediated is key to the experience of social presence.

Efforts to incorporate nonverbal behavior into virtual interactions must take into account both technical questions and awareness of how gesture informs interactions in real life. Newer video games, for example the Microsoft Kinect and Nintendo Wii, allow for high immersive tracking that captures subtle body movements and gestures of the players. In the current paper we discuss and examine how these high immersion tracking games, in particular the Microsoft Kinect which uses active computer vision to automatically track body movement, can increase the information available to inform a sense of presence by automatically detecting behavior. In addition, the ability of these devices to be linked over the internet allows co-presence to occur cheaply and easily in the home environment.

In reviewing the relevant literature, we first discuss previous work that examines the relationship of nonverbal behavior to social presence. Next we choose a test case—gender—and examine classic methods from psychology and communication that detect gender using behaviors such as gesture and gait. Finally, we will briefly explain the principles of computer vision and machine learning as the tools we use to examine how the digital footprint, collected via the Kinect, can reveal cues about gender.

1.1. Examining and automatically detecting nonverbal behavior

The theoretical framework developed by Walther (1996) provides a way to understand the implications of online interactions in which there are fewer social cues available than there are face to face. Absence of most conventional kinds of nonverbal communication has been notable in online interactions, although substitutions, such as emoticons for facial expressions, have evolved. Despite very limited channels for nonverbal interactions, Yee and colleagues (2007) showed that nonverbal social norms via interpersonal distance carried over to online

virtual environments. Now, video game interfaces such as the Kinect, which allow more complex gestural information to be communicated through networked computers from one physically distant participant to another, hold the promise of adding more cues that resemble true nonverbal behavior. In essence, the Kinect may provide enough behavioral cues to allow a rich nonverbal and verbal social interaction via avatars, but still maintain the benefits of Walther's hyperpersonal nature of computer-mediated communication due to the ability to tailor one's representations.

1.2. Using Kinect and Other Automated Systems to Detect Human Gestures

A number of studies have used automated systems to analyze body language or gesture. Some interesting recent studies have focused on states of mind. Michael, Dilsizian, Metaxas, and Burgoon (2010), tracked hand and head movements relative to the body, head position in three-dimensional space, and facial expressions to detect deception without requiring sensors to be placed on the participant. In another study, a virtual reality-based system using head-mounted displays tracked the head and eye movements of 20 children to compare the attention performance of children with attention deficit hyperactivity disorder (ADHD) to a control group of children (Parsons, Bowerly, Buckwalter, & Rizzo, 2007). The children previously given a diagnosis of ADHD "displayed more overall hyperactivity as measured by five out of six measures of total body movement" (p. 336). Thus, we see that important information that could increase interpersonal communication and thus presence can be picked up through automated systems.

The Microsoft Kinect is an extremely popular video game interface (Gamesbeat, 2011) that uses an infrared emitter and sensor to capture body movements by isolating the X,Y, and Z coordinates of 20 nodes roughly representing joints in the the body. Since this technology has become available, several studies have already used it as a tool to detect, interpret or represent human motion. Most significant to this paper, as part of an effort to develop a less expensive alternative to a "smart home" for monitoring vulnerable inhabitants, a system using Kinect was designed to identify common daily activities such as opening pill bottles and drinking a glass of water (Sung, Ponce, Selman, & Saxena, 2011). Although the system was better at recognizing activities when it had been trained on a previous recording of a specific participant, it was also able to generalize from that training and recognize activities done by a new participant unknown to the system. One of the most important qualities of this kind of data, along with data obtained from online interactions, is the unobtrusive nature of the data gathering. The user wears no markers, no special clothing, and does not even need to be aware the cameras are in the room.

1.3. The Kinect and Computer Vision

The Kinect provides a unique opportunity to study gestures while minimizing the effect of being hampered by reminders of observation. Previous successful video game interfaces have sought to involve naturalistic body movement without using computer vision, for example by using a base with sensors (Dance Dance Revolution, Konami Corporation) or a handheld wand (the Nintendo Wii). For some time, the ability to track body movements in 3D space has been in use by high end virtual reality research labs and the military, and in entertainment contexts such as computer graphics in film (see Blascovich & Bailenson, chapter 3, for a review). However, as previously utilized, 3D tracking required the users to enter sterile lab setups and wear markers on his or her bodies, an expensive and invasive technology, without much application in daily life. This is now changing, as the Kinect is an inexpensive, portable and unobtrusive device about 12 x 6 x 5 inches in size and weighing slightly more than 3 pounds. Its size allows it to sit on a desk or be mounted on a wall. Each camera can capture movement of up to four people at once, from a range of 4 to 12 feet, even in low light conditions. Also significantly, the Xbox console can be networked so that players' avatars can interact online. The gross movements that are detectable by Kinect are noninvasive and, unlike other face tracking methods, can still be useful from a distance and at low resolution of recording. By assessing a basic characteristic like gender, we can examine the ability of Kinect to assist a theoretically driven research question about gender cues as well as advance practical applications. While Kinect is an extremely popular video game interface, it is also a powerful new kind of medium which tracks and can aid in rendering human nonverbal communication.

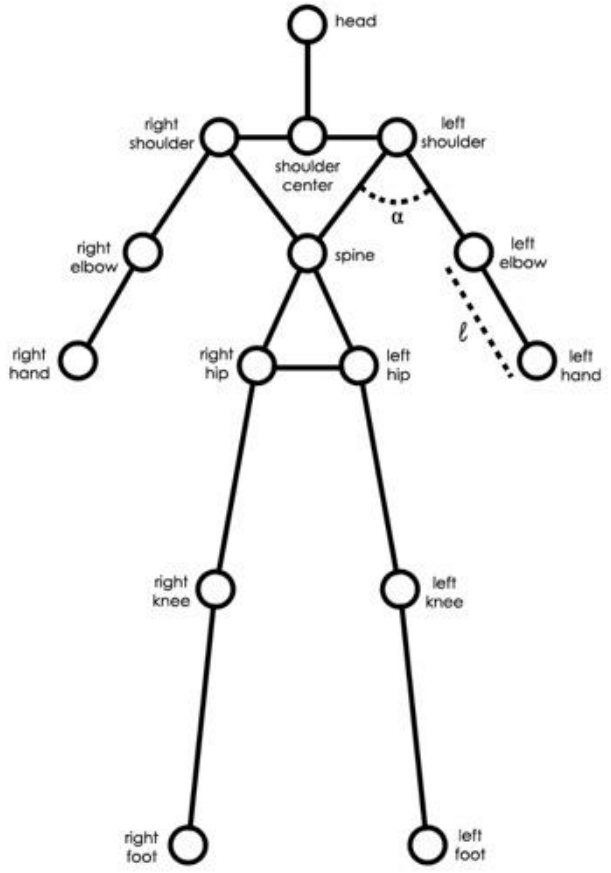

Figure 1. An illustration of Kinect data output in the form of a wireframe. The wireframe consists of 15 nodes with and *x, y,* and *z* value in physical space. Angle *α* describes an example of the angles that we extracted as features. Distance *l* describes an example distance between two nodes that we extracted as features.

Once the semantic features of visual data have been captured through computer vision, the data must be interpreted to provide meaningful information. This is done through the use of algorithms that interpret the input to pull out the desired information. In the case of the Microsoft Kinect, a skeleton (Figure 1) with 20 nodes roughly representing joints is extrapolated from the data, and the movements of the nodes are then tracked in three dimensions, providing the position and movement of the body parts of the person being tracked.

1.4. Gender Identification Through Movement and Gesture

Although gender is linked to sexual characteristics that are rooted in biology, it is also an important social construct that influences many avenues of interpersonal communication (Birdwhistell, 1970). There are an enormous number of cues available to human observers in real life to answer the important question of whether another human is to be considered male or female, including facial recognition, voice tone, speech patterns, and gesture. Of these cues, gesture is considered to be extremely salient, as stated by Yang (2010) "[These] gender-specific gestures are so clearly distinguished that the people of the same cultural group who observe them with their eyes and without access to linguistic utterances can quickly tell, with the least difficulty or doubt, that they are typical of males or females" (p. 366).

However, beyond these culturally specific differences, there are differences in gesture and posture which can make it possible to distinguish men from women cross-culturally (Yu, Tan, Huang, Jia & Wu, 2009). These motions can be broken down into basic elements, as became apparent through various experiments using point light displays, starting in the 1970's (Johansson, 1973). These displays provide extremely limited visual information, consisting only of light disks (affixed to the joints of real or simulated walkers) moving against a dark background, showing gross movements in a single plane. Despite the minimal nature of this information, human observers looking through this dark background (i.e., they can only see the lights shining through, not the human behind the background) have been able to identify actions, gender (Cutting, Proffitt, & Kozlowski, 1978), emotions (Atkinson, Dittrich, Gemmell & Young, 2004), and even states of mind (Michalak, Troje, & Heidenreich, 2011) based only on these displays. The data generated by Kinect is similar to these point light displays in that they consist of coordinates that indicate joint position; however, a) the Kinect also provides accurate information on the Z axis and b) the Kinect does not require that the user wear an obtrusive setup with lights on it which likely interferes with normal behavior.

1.5. Machine learning

To investigate how well gender can be recognized from Kinect data, machine learning was used. In a book

chapter written for communication scholars, Ahn and colleagues (2010) summarized the machine learning procedures and reviewed previous research in communication that has successfully implemented the technique. During machine learning, a computer takes raw data and utilizes a train-and-test paradigm to uncover patterns or relationships in the data (for a recent example using machine learning and nonverbal behavior, see Michael, Dilsizian, Metaxas, & Burgoon, 2010). As such, machine learning is the process of training machines to predict values or classes based on input data (see Witten & Frank, 2011, and Bishop, 2006, for detailed manuals). The power of machine learning is rooted in the fact that it can quickly process a great volume of data and learn complex interrelationships that are captured in the data. As this is a bottom-up approach, machine learning algorithms can come up with relationships between variables that have not been identified before.

A popular statistical technique which can be described as a form of machine learning algorithm is linear regression. With linear regression, input predictors which have a continuous value are combined linearly to predict a continuous outcome value. For instance, the size of a house can be used to predict the price of that house. Consider a set of house prices and sizes for one hundred houses. When feeding this data to a linear regression algorithm to fit the regression line, the resulting regression line might look something like Figure 2. Subsequently, for a new house from which only the size is known, the house's price can be predicted by using the size as input to the linear regression. There might also be cases in which one does not want to predict continuous values but only discrete classes or values. For instance, for the case of house prices, we might only know if a house is cheap or expensive. Linear regression is not well suited for this problem as it might make prediction errors when the input values for two different discrete classes are close together (see Figure 3 for an example). Instead, another algorithm, called logistic regression, can be used to create a predictor that is able to separate two different classes. An example logistic regression line is presented in Figure 3. From Figure 3, it can be seen that all the data points for houses have only one of two possible values on the house price axis. The logistic regression line works like a threshold model. Based on the example data points it predicts that when a house is smaller than the threshold of 45 square meters, it will be cheap, and that when it is bigger than the threshold 45 square meters it will be expensive. Finally, note that in these examples we used only one predictor, but these examples can be extended to use more predictors. These predictors are often referred to as features.

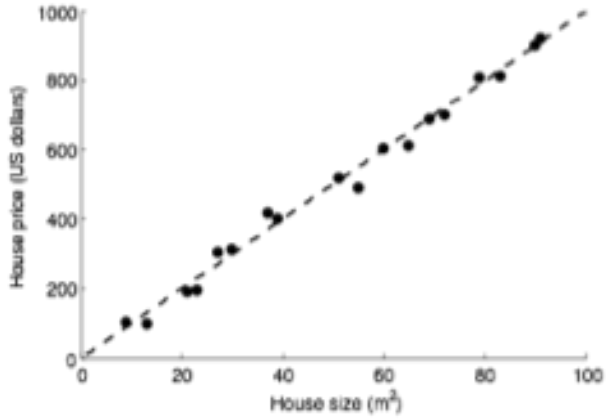

Figure 2. A linear regression line describing the relationship between house size and house prices. The black dots depict individual data points of houses on which the regression line is based. The horizontal axis depicts the size of the house (in m2), and the vertical axis depicts the price of the house (in US dollars).

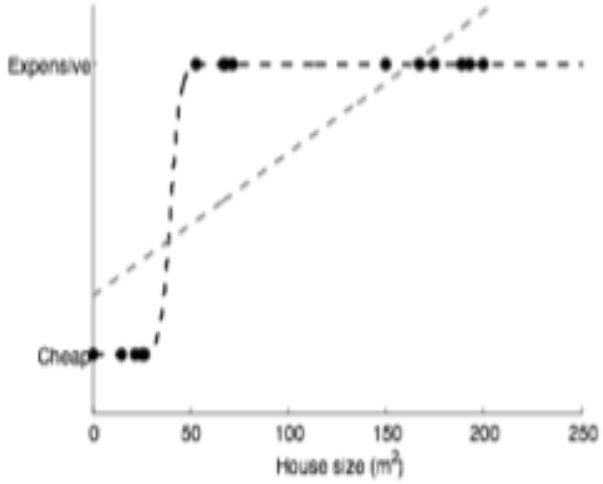

Figure 3. A logistic regression line describing the relationship between house size and whether or not a house is expensive. The black dots depict individual data points of houses on which the regression line is based. The threshold between a cheap and an expensive house is around 50 m2. The horizontal axis depicts the size of the house (in m2), and the vertical axis depicts whether the house is cheap or expensive. The logistic regression shows that the higher the house size is, the more likely the house is to be expensive. The gray line depicts a linear regression model based on the same data points. This would make prediction errors for some of the data points of the expensive houses.

1.6. Current Study

In previously cited work, automatic detection and interpretation of human motion,using a combination of computer vision and machine learning, including the use of Kinect, has been successfully attempted. Other studies have shown (as in point-light displays) that it is possible to detect gender, along with other actions, and even states of mind, given very limited visual input such as the movement of the major joints of the body in two dimensions.

In the following study, we tested this inexpensive and widely available method of tracking human activity to see if it can be used to detect gender using a combination of gestural and postural information obtained from a brief series of poses. We hypothesized that we will be able to predict gender at a level greater than chance. We also investigated what the relevant contributions are of static features that roughly map onto "body shape" (e.g., height, distance between shoulders) compared with movement features that roughly map onto "gesture" (e.g., changes in angle between nodes over time). Getting a preliminary sense of the types of body configurations and movements which differentiate via gender will add to previous research on gender differences.

2. Methods

2.1. Participants and design

Participants were 12 male and 12 female students aged 18 to 39 years (median: 22 years) from an American West-coast university. Gesture was manipulated as a within-participant factor, so each participant performed all twelve different gestures.

2.2. Materials

Participants were positioned in the center of a room of 6.0 meters by 5.5 meters, at a tape mark. An XBOX Kinect camera was attached to the wall directly in front of the participant, at a height of 1.35 meters. The distance between the participant and camera was 2.6 meters. The experimenter was located in an adjacent control room and was able to talk to the participant and see the participant through a window and vice versa.

The XBOX Kinect camera identifies a skeletal wireframe at 30 frames per second for each person in front of the camera. The wireframe model consists of 20 nodes, each positioned in three dimensional space (i.e., outputting an x, y, and z value for each node). Additionally, the Kinect camera outputs whether a node was tracked or inferred. Inferring a note happens when it could not be accurately tracked and was therefore inferred from the position of other nodes.

We selected twelve different gestures and postures that the participants had to perform. Gestures and postures were selected so that they contained a variety of different movements that together engaged the entire body. The different gestures and postures were all given a name and put in a fixed order. A list of the gestures and postures can be found in Table 1.

2.3. Procedure

When participants arrived they were led to the recording room and instructed to stand on the tape mark. Next, participants were told they would perform twelve different postures or gestures, each for ten seconds. The

Table 1. A list of all the gestures that participants had to perform during the experiment, each for ten seconds. The third column depicts the classification accuracy for predicting gender when using only that gesture.

Number	Gesture	Classification accuracy
1	Head tilt from side to side	83%
2	Arm wave above head from side to side	71%
3	Idle with hands on hips	67%
4	Idle with arms crossed	63%
5	Hula hoop motion (without actual hoop)	75%
6	Marching with knees to the side	58%
7	Lasso movements with right arm	75%
8	Twist (dance)	88%
9	Snake-like arm waves with arms stretched side to side	67%
10	Pointing forward with left arm	75%
11	Pointing forward with right arm	63%
12	Hand language signaling other to go on	79%

experimenter would then go to the control room and start the recording.

For each gesture recording, the experimenter first told the participant the name of the gesture and asked the participant to perform it to make sure it was understood correctly. Most of the gestures were correctly understood the first time. In case a participant did not correctly perform the gesture, the experimenter demonstrated the posture to the participant. The experimenter then asked the participant to perform the gesture until a stop signal was given. Next, the experimenter marked the start of the recording in the data and started a timer. After ten seconds, the experimenter again marked the data and told the participant to stop. This procedure was repeated for all twelve gestures. The entire procedure took less than five minutes per participant.

2.4. Machine learning analysis

We used the Kinect data to create a machine that can automatically detect the gender of the participant performing the gestures. The process we used for this consisted of the following steps that will be discussed in more detail in the following paragraphs: feature extraction, feature selection, training, and testing.

To capture the most important parts of the data and reduce the noise in the measurements we calculated specific features from 15 of the 20 nodes that constituted the output of the Kinect. (Instead of using both the hand and the wrist node, we used only the wrist nodes, and for the feet, only the ankle node- see Figure 1.) First, we extracted features that described the *static* length of different body parts. For instance, features included the distance from right shoulder to left shoulder, the distance from the left elbow to the left shoulder, and the distance from the head to the neck. In total, we extracted fifteen static (body shape) features.

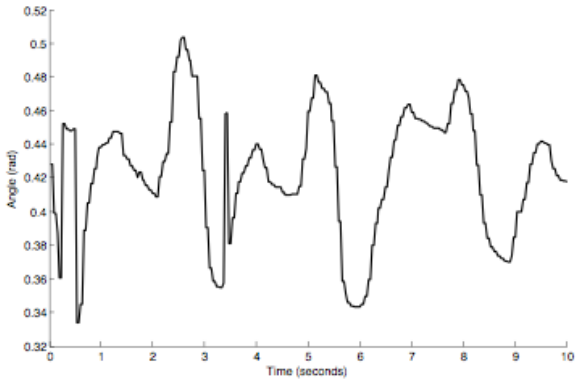

Figure 4. An example of a trace describing the change of angle between right shoulder and right arm during the arm wave gestures.

Second, we extracted *movement* features. For this, we calculated the angles between two different bones of the skeleton (see Figure 1). Thus, for each frame (of the 30 fps) that we recorded, we extracted 18 angles (e.g., the angle from the spine-to-neck bone to the neck-to-left-shoulder bone). Then, for each angle we got a trace as depicted in Figure 4. Subsequently, from this trace, we took the mean, standard deviation, and skewness as features. This resulted in 54 movement features (i.e., features that can change over time).

After all features were extracted, we selected the features which were most useful. First, we created three groups of features. The first group contained all features. The second group only contained movement features (i.e., features that can change over time) and the third group only contained static features (i.e., features that cannot change over time). Finally, for each group, $\chi 2$ ranking was applied, a standard technique to reduce the number of features in a machine learning input set (Witten & Frank, 2011). This ranks features based on $\chi 2$ scores that indicate each feature's ability to predict the correct class (in our case the gender of the participant). After the ranking was made, we selected the top ten features for classification. This process resulted in three groups of ten features.

For each individual feature group we developed machine learning algorithms using training and testing. To do this in an ecologically valid way, we did a separate training and testing session for each of the 24 persons in our dataset. For each person, we used data from the 23 other persons to train the machine learning algorithm, and tested how well this algorithm was able to detect the gender of the person left out of the training dataset. This

Table 2. The number of data points for all participants that were classified as men or women based on the top ten *overall* (both movement and static) features. The table indicates that men were misclassified slightly more often than women.

		Predicted gender		
		Men	Women	Total
Actual gender	Men	134	10	144
	Women	38	106	144
	Total	172	116	

Table 3. The number of data points for all participants that were classified as men or women based on the top ten *static* features. The table indicates that men were misclassified slightly more often and women were misclassified equally often.

		Predicted gender		
		Men	Women	Total
Actual gender	Men	111	33	144
	Women	50	94	144
	Total	161	127	

Table 4. The number of data points for all participants that were classified as men or women based on the top ten *movement* features. The table indicates that men were misclassified slightly more often and women were misclassified equally often.

		Predicted gender		
		Men	Women	Total
Actual gender	Men	115	29	144
	Women	38	106	144
	Total	153	135	

was repeated for all 24 persons in our dataset, and their results were averaged to an overall recognition accuracy. This method is called "leave-one-person-out cross validation" and was done to make sure that validation was done as if the machine learning algorithm was tested on new people (i.e., a person it had not received training data from before).

3. Results

3.1. Overall classification

Classification using the top ten of all features resulted in 83% recognition accuracy of gender, for persons that were not in the training dataset (i.e., using leave-one-person-out cross validation). This was significantly higher than chance level of 50% ($Z = 3.23$; $p < .001$). Men were misclassified as women slightly more often than women were misclassified as men, as seen in **Tables 2, 3 and 4**. A ranked list of the selected features is depicted in Table 5, which showed that almost all the selected features were static features and that only one was a movement feature.

Table 5. A ranked list of the most useful *overall* (movement and static) features for classifying gender.

Rank	Feature	Static or movement features
1	Distance from left knee to left hip	Static
2	Distance from left foot to left knee	Static
3	Distance from right shoulder to left shoulder	Static
4	Distance from neck to tailbone	Static
5	Standard deviation of the angle from right hip to spine	Movement
6	Distance from right elbow to right shoulder	Static
7	Height of person	Static
8	Distance from head to neck	Static
9	Distance from right hand to right elbow	Static
10	Distance from left hip to right hip	Static

3.2. Static versus movement features

As there was such an apparent discrepancy between the number of movement and static features that were selected for classification of gender (Table 5), we also explicitly compared classification with these two different feature types. Movement features were the features that were based on variability over time, whereas static

Table 6. A ranked list of the most useful *static* features for classifying gender.

Rank	Feature
1	Distance from left knee to left hip
2	Distance from right shoulder to left shoulder
3	Distance from left foot to left knee
4	Distance from right elbow to right shoulder
5	Distance from neck to tailbone
6	Distance from right foot to right knee
7	Distance from left hand to left elbow
8	Distance from head to neck
9	Distance from left hip to right hip
10	Distance from left shoulder to left elbow

Table 7. A ranked list of the most useful *movement* features for classifying gender.

Rank	Feature
1	Mean angle between the left shoulder and neck via the center shoulder
2	Mean angle between the right hip and spine via the tailbone
3	Mean angle between the left shoulder and spine via the center shoulder
4	Mean angle between the shoulders and spine
5	Mean angle of left hip and right knee via the right hip
6	Standard deviation of the angle between the shoulders and spine
7	Standard deviation of the angle between the shoulders and tailbone
8	Standard deviation of the angle between the right hip and the right foot via the right knee
9	Standard deviation of the angle between the right hip and spine via the tailbone
10	Mean angle from spine to head via the center shoulder

features described body properties that did not change over time. Classification based on only the static features resulted in 77% recognition accuracy, which is significantly higher than chance ($Z = 2.65$; $p < .005$). The top ten features selected for that are depicted in Table 6. Classification based on only the movement features resulted in 71% recognition accuracy, which is significantly higher than chance ($Z = 2.06$; $p < .02$). The top ten features selected for that are presented in Table 7. However, combining the top movement feature with the top static features increased the accuracy to 83%.

3.3. Classification per gesture

Finally, we also looked at classification per gesture. Although this resulted in lower overall classification accuracies because there was less data to train on, it was our goal to compare the different gestures for differentiating between men and women. The classification accuracies between men and women for each individual gesture are depicted in Table 1. In this study, the movements named head tilt and the twist were most successful in distinguishing between men and women. Hence, this suggests that men and women differed more when doing these gestures than when doing the other gestures. This resonates with previous work that demonstrates that lateral sway in the shoulders and hips was an important cue used in distinguishing men from women, since both these movements involved considerable lateral movement.

4. Discussion

Our machine learning algorithms in combination with Kinect computer vision performed significantly above chance in determining gender from a brief series of gestures. We expect that it will be a useful method for tracking nonverbal behaviors crucial to facilitating presence in virtual social interactions in the future.

4.1. Possible uses

The ability to sense human body language unobtrusively and in real time may also be useful in designing better user interfaces, as an assist to social signal processing (Vinciarelli, Pantic, Bourlard, 2009) and to provide input to drive robot/avatar behavior (Tomari, Kobayashi, & Kuno, 2011) both in real life and in virtual or augmented reality (Vera, Gimeno, Coma, Fernández, 2011). Since Kinect is inexpensive, portable, and usable in a home environment, it shows potential to be an interface that can allow for more immersive, more interactive social presence experiences in the home.

Another interesting subject for future research would be to explore the finding that in this study, Kinect was actually more successful when using static features than when using movement features. Since our participant population was small and consisted of a convenience sample of students and university affiliates of a similar age, it is reasonable to assume a fairly culturally homogenous group. Thus, according to Yang, we might predict greater success from movement features, at least when humans are assessing the data. Other work suggests that movement heavily influences gender judgments made by humans. In Mather and Murdoch's 1994 paper, they state that gender discrimination by humans in point-light displays is strongly dependent on the dynamic cue of lateral body sway in shoulders and hips. Birdwhistell's work (1970) points out that although humans do exhibit sexual dimorphism, much of gender identification as done by humans is through nonverbal communication, including movement. This opens up the opportunity to use Kinect to look at the social construct of gender, as in agents, as distinct from secondary sexual characteristics, since Kinect can consider anatomically based static features separately from movement features.

References

Ahn, S. J., Bailenson, J.N., Fox, J., & Jabon, M. E. (2010). Using Automated Facial Expression Analysis for Emotion and Behavior Prediction. In Doeveling, K., von Scheve, C., & Konjin, E. A. (Eds.) *Handbook of Emotions and Mass Media* (349-369) London/New York: Routledge.

Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception,* 33(6), 717-746.

Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators and Virtual Environments* 10(6), 583-598.

Bailenson, J. N., Blascovich, J. (2011). Infinite Reality-Avatars, Eternal Life, New Worlds, and the Dawn of the Virtual Revolution. New York: William Morrow

Birdwhistell, R.L.(1970) Kinesics and Context. Philadelphia: University of Pennsylvania Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer. Bush, R. (2009) IARPA BROAD AGENCY ANNOUNCEMENT. Retrieved 11/01/11 from: http:// www.iarpa.gov/Reynard_BAA_Amend1.pdf

Biocca, F., Harms, C., & Gregg, J. (2001). The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. Paper presented at the International Workshop on Presence, Philadelphia, PA. Available at: http://astro.temple.edu/~lombard/P2001/Biocca2.pdf.

Biocca, F. (1997). The cyborg's dilemma: progressive embodiment in virtual environments. Journal of Computer-Mediated Communication, 3 (2).

Cutting, J. E., Proffitt, D. R., & Kozlowski, L. T. (1978). A biomechanical invariant for gait perception. *Journal of experimental psychology: Human perception and performance,* 4(3), 357-72.

Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., and Sasse, A. M. (2003). The impact of avatar realism and eye gaze control on the perceived quality of communication in a shared immersive virtual environment. *Proceedings of SIGCHI,* 529–536.

GamesBeat- Interpreting Innovation (January 27, 2011) *www.venturebeat.com* Microsoft's Kinectified game business grows 55 percent — fastest-selling consumer electronics device in history. Retrieved 11/01/11 from: http://venturebeat.com/2011/01/27/ microsofts-kinectified-game-business-grows-55-percent/

IJsselsteijn, W. (2003) Staying in Touch. Social Presence and Connectedness through Synchronous and Asynchronous Communication Media. *Proceedings of HCI International Conference on Human-Computer Interaction*, Lawrence Erlbaum Associates, New Jersey 924-928.

IJsselsteijn, W. (2006) Mediated social touch: a review of current research and future directions, *Virtual Reality*

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis' attention. *Perception and Psychophysics,* 14(2), 201–211.

Kane, H. S., McCall, C., Collins, N. L., (2012) Mere presence is not enough: Responsive support in a virtual world *Journal of Experimental Social Psychology*, 48(1), 37-44 DOI: 10.1016/j.jesp.2011.07.001

Mather, G., Murdoch, L. (1994) Gender discrimination in biological motion displays based on dynamic cues. *Proceedings: Biological Sciences,* 258(1353) 273-279.

Michael, N., Dilsizian, M., Metaxas, D., & Burgoon, J. K. (2010). Motion profiles for deception detection using visual cues. *Computer Vision (Lecture Notes in Computer Science), 6316,* 462-475.

Michalak, J., Troje, N. F., & Heidenreich, T. (2011). The effects of mindfulness-based cognitive therapy on depressive gait patterns. *Journal of Cognitive and Behavioral Psychotherapies,* 11, 13–27.

Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., & Jay, F. (2005). Detecting deception through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems, 20(5),* 36-42 .

Parsons, T. D., Bowerly, T., Buckwalter, J. G., & Rizzo, A. A. (2007). A controlled clinical comparison of attention performance in children with ADHD in a virtual reality classroom compared to standard neuropsychological methods. *Child Neuropsychology, 13(4),* 363-381.

Schroeder, R. (2002). Social interaction in virtual environments: Key issues, common themes, and a framework for research. In R. Schroeder (Ed.), The social life of avatars: Presence and interaction in shared virtual environments, (pp. 1–18). London: Springer-Verlag.

Shen, C., & Williams, D. (2011). Unpacking time online: Connecting internet and massively multiplayer online game use with psychosocial well-being. *Communication Research,* 38 (1), 123-149.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2011). Human activity detection from RGBD images. *AAAI 2011 Workshop: Plan, Activity, and Intent Recognition.*

Tomari, R. (2011). Multi-view head detection and tracking with long range capability for social navigation planning. Advances in Visual Computing. Retrieved from http:// www.springerlink.com/index/13128L1H47706416.pdf

Vera, L., Gimeno, J., Coma, I., & Fernández, M. (2011). Augmented mirror: Interactive augmented reality system based on Kinect. *International Federation For Information Processing,* 483-486.

Vinayagamoorthy, V., Steed A., Slater M. (2008) The impact of a character posture model on the communication of affect in an immersive virtual environment *IEEE Transactions on Visualization And Computer Graphics*, *14 (5),* 965-981.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing, 27(12),* 1743-1759.

Walther, J. B. (1996). Computer-mediated communication. Communication research, 23(1), 3. Sage Publications. Retrieved from http://crx.sagepub.com/content/23/1/3.short

Witten, I. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufman, Burlington MA

Yang, P. (2010). Nonverbal gender differences: Examining gestures of university-educated Mandarin Chinese speakers. *Text & Talk - An Interdisciplinary Journal of Language, Discourse & Communication Studies, 30(3),* 333-357.

Yee, N., Bailenson, J.N., Urbanek, M., Chang, F., & Merget, D. (2007). The unbearable likeness of being digital; The persistence of nonverbal social norms in online virtual environments. *Cyberpsychology and Behavior,* 10, 115-121.

Yu, S., Tan, T., Huang, K., Jia, K., & Wu, X. (2009). A study on gait-based gender classification. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society, 18(8),* 1905-1910.