Utilizing Depth Sensors for Analyzing Multimodal Presentations: Hardware, Software and Toolkits

Chee Wee Leong, Lei Chen, Gary Feng, Chong Min Lee and Matthew Mulholland
Research & Development
Educational Testing Service
Princeton, New Jersey, USA
{cleong, lchen, gfeng, clee001, mmulholland}@ets.org

ABSTRACT

Body language plays an important role in learning processes and communication. For example, communication research produced evidence that mathematical knowledge can be embodied in gestures made by teachers and students. Likewise, body postures and gestures are also utilized by speakers in oral presentations to convey ideas and important messages. Consequently, capturing and analyzing non-verbal behaviors is an important aspect in multimodal learning analytics (MLA) research. With regard to sensing capabilities, the introduction of depth sensors such as the Microsoft Kinect has greatly facilitated research and development in this area. However, the rapid advancement in hardware and software capabilities is not always in sync with the expanding set of features reported in the literature. For example, though Anvil [21] is a widely used state-of-the-art annotation and visualization toolkit for motion traces, its motion recording component based on OpenNI is outdated. As part of our research in developing multimodal educational assessments, we began an effort to develop and standardize algorithms for purposes of multimodal feature extraction and creating automated scoring models. This paper provides an overview of relevant work in multimodal research on educational tasks, and proceeds to summarize our work using multimodal sensors in developing assessments of communication skills, with attention on the use of depth sensors. Specifically, we focus on the task of public speaking assessment using Microsoft Kinect. Additionally, we introduce an open-source Python package for computing expressive body language features from Kinect motion data, which we hope will benefit the MLA research community.

Categories and Subject Descriptors

H.5.1 [Information Systems]: Information Interfaces and Presentation—Multimedia Information Systems

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICMI '15 November 09 - 13, 2015, Seattle, WA, USA

Copyright is held by the owner/author. Publication rights licensed to ACM. Copyright 2015 ACM 978-1-4503-3912-4/15/11...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2830605.

Keywords

Depth sensors; Kinect; Multimodal Learning Analytics

1. INTRODUCTION

Effective communication between humans for a specific cause requires adaptive measures to convey the intent and message clearly to the other party. Kinesics are often used along with words to infer the intent of the speaker [4]. In the context of classrooms or study groups, teachers tend to teach effectively by utilizing body gestures, voice modulation, visual aids, etc., in addition to use of the blackboard. Likewise, students learn better not just by listening, but also by providing feedback through nodding, vocal acknowledgment, facial expressions, all of which are channels of feedback potentially helpful for improving pedagogic processes [7]. In fact, cognitive and communication research produced evidence that mathematical knowledge can be embodied in gestures made by teachers and students [1], along with other recent studies that lend weight to similar hypotheses that proper use of body movements and gestures can lead to instructional scaffolding to achieve better learning outcomes [23, 3].

While the above classroom scenarios are reflective of a conducive learning environment that leverages on multimodal cues, constructing automated systems for fostering such scenarios is still difficult with the current level of multimodal learning analytics research. Some of the main blocking issues relates to spatial constraints (e.g. how to mount sensors conveniently without being intrusive), background noise, lighting issues, classroom artifacts (e.g. partially occluded body due to desks and chairs), and even psychological effects (e.g. different behavioral patterns induced by the thought that one is being tracked). Nevertheless, the central goal in multimodal learning analytics relates to the effective deployment of algorithms and systems to collect and analyze similar communication cues, and build automated systems with predictive models that have good correlation with human ratings on the target learning or assessment task. Intuitively, such a system comprises the following components (1) A network of input sensors for capturing data from multiple modalities, including voice, video, haptics, etc. (2) A communication protocol to process data streams from the sensors for the purposes of time and events synchronization. (3) A set of algorithms implemented for real-time feature tracking and machine learning outcomes prediction. (4) A set of output devices rendering the learner outcomes textually, aurally, or visually or e.g. interactive avatars, audio feedback, etc. (5) A data storage protocol and devices im-
implementing the protocol for holding persistent sensor data, feature values and learner predictions along with other metadata for offline analysis and model training.

With respect to multimodal learning analytics, there is significant ongoing work in the literature. [38] documented the study of how expertise in building stable, well-engineered structures can be identified through meaningful utterances and coordinated hand movements, which are extracted from time-series, process-oriented video and voice data. In another line of work, [15] employed automated facial expression tracking to predict time instances when students become disengaged or frustrated during interactions with tutoring systems. [30] proposed a multi-camera tracking framework for detecting engagement/disenagement of students and teaching performance in classroom. When combining this framework with other input such as after-class questionnaires and in-depth interviews, a viable system for predicting the quality of a class can be deployed. In addition, previous MLA challenges also yielded important contributions where there are efforts to extract multimodal cues from audio, video and motion traces to automatically score presentations of students [10, 13].

In this paper, we are particularly interested in the use of depth sensors for multimodal learning analytics. Specifically, we focus on the application of a low-cost depth sensor, Microsoft Kinect, to evaluate a target assessment task, public speaking, which is a core competency skill for college and workforce success. The main contribution of our work is to provide a cookbook style of discussion, where we elaborate from beginning to end how we leverage state-of-the-art sensors, toolkits and algorithms to build a preliminary presentation assessment system with good correlation with human annotations on the same task.

The paper proceeds in the following manner. First, we highlight and explain the emerging importance of depth sensors, how they operate in principle, and why they should be utilized as part of the sensing network. This is especially relevant given that traditional input sensors which are restricted to video and audio tend to limit the performance of multimodal systems [17]. With the advent of 3D depth sensors, an orthogonal form of information can be used together with information from other sensors in the same set of input features. In particular, as our depth sensor of choice, we focus on the Microsoft Kinect. Second, we provide an overview of existing integrative multimodal systems that can be leveraged to perform recording and synchronization of data streams in the data collection process, which is elaborated next with technical details via two case studies. Finally, we highlight a set of high-level features extracted from Kinect that are shown to have promising predictive power for multimodal presentation assessments. Extraction of these features from other datasets is possible through our Python scripts to be released as open-source software, which we hope to benefit the MLA research community.

2. DEPTH SENSORS

A 3D depth sensor is a device that assigns distance from an anchor point in a 3D space to a 2D scene. Specifically, each pixel in the 2D scene is attached with relative depth information from the anchor point. Once calibrated, the actual physical units of distance (e.g. inches or feet) can be mapped. The role of 3D sensors has been significantly amplified over the years, graduating from its origination in research labs and increasingly finding its way into our everyday lives [39, 20, 8, 37]. This can be attributed to the continuous innovation in sensors, improvement in sensing accuracy, decreasing costs of sensor components, availability of APIs for sensors and related software libraries for developing useful applications. More importantly, since depth sensors have the ability to track body movements as the scene changes, it allows human body language understanding, a research problem proven formidable challenging with video cameras [17]. It is this enabling technology that gives rise to a new way of human computer interaction.

A brief section on the inner workings of optical depth sensors is provided here in order to create a bridge from their hardware obscurity to their practical usefulness. There are generally three principles by which these devices operate, namely structured light, time-of-flight (ToF), and stereo vision. Structured light sensors use a projector to produce a laser, typically Infrared (IR), that passes through a diffraction grating to produce a set of IR dots which are projected onto the scene. The projected IR pattern is random, but all the IR dots are distinct from one another, which allows differentiation between dots. Since the relative geometry between the IR projector and IR sensor is known during calibration, including the angles of the diffraction, triangulation can be used to render each dot in the image taken by the IR sensor in 3D, thus allowing depth information to be obtained. Unlike structured light, a Time-of-Flight (ToF) depth sensor also employs laser-inducing light to illuminate the scene, but uses the actual reflected light to compute depth. Each light pulse is modulated i.e. “information” is being attached to the light by varying its amplitude or frequency, hence making it identifiable from emission to detection. Taking into account the travel time of the light pulse (hence, “time-of-flight”), the distance from the depth sensor to a single spot in the scene can be computed. To create a 3D point cloud of the entire scene, a ToF sensor must enable the time taken for millions of light pulses to return to the sensor, typically achieved in picoseconds. In stereo vision sensor systems, two or more cameras mounted in parallel but separated by a short distance are utilized to collectively compute depth information. Specifically, the focal length and optical axes of each camera are parameters used to project a specific, shared 2D point on the images into 3D Cartesian space. The 3D projections from each camera are necessarily different from one another due to the difference in the geometric setup of each camera, hence the discrepancies can be plugged into a formula to compute the depth distance from the cameras to the point.

Apart from their different working principles, depth sensors can be categorized by their range of sensing and tracking ability. There are also several aspects of depth sensors that we believe are important for researching learning analytics. First, while there are depth sensors operating on similar principles that focus on scene understanding and reconstruction, such as the 3D structure sensor or Intel’s R200 3D sensor2, we believe understanding human body movements to be key to valuable multimodal learning applications. These user-centric depth sensors are designed and calibrated to work optimally for motion capture and gesture tracking, ranging from the entire body skeleton to tracking

---

1http://structure.io/
2https://software.intel.com/en-us/articles/realsense-r200-camera
Table 1: Technical specification summary of commonly used depth sensors

<table>
<thead>
<tr>
<th></th>
<th>Microsoft Kinect v1</th>
<th>Microsoft Kinect v2</th>
<th>Intel Realsense (F200)</th>
<th>LEAP Motion Controller</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>Structured Light</td>
<td>Time-of-Flight</td>
<td>Time-of-Flight</td>
<td>Stereo Cameras</td>
</tr>
<tr>
<td><strong>RGB camera</strong></td>
<td>640x480 @30fps</td>
<td>1920x1080 @30fps</td>
<td>1920x1080 @30fps</td>
<td>undisclosed</td>
</tr>
<tr>
<td><strong>Depth sensors</strong></td>
<td>640x480 @30fps</td>
<td>512x424 @30fps</td>
<td>640x480 @60fps</td>
<td>undisclosed</td>
</tr>
<tr>
<td><strong>Microphone</strong></td>
<td>Quad-array microphone</td>
<td>Quad-array microphone</td>
<td>Dual array microphone</td>
<td>-</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>0.4 to 4.5 m</td>
<td>0.5 to 4.5 m</td>
<td>0.2 to 1.2 m</td>
<td>0.03 to 0.6 m</td>
</tr>
<tr>
<td><strong>Skeletons tracked</strong></td>
<td>2</td>
<td>6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Horizontal Field of View (FOV)</strong></td>
<td>57°</td>
<td>70°</td>
<td>70°</td>
<td>150°</td>
</tr>
<tr>
<td><strong>Vertical FOV</strong></td>
<td>43°</td>
<td>60°</td>
<td>43°</td>
<td>120°</td>
</tr>
<tr>
<td><strong>USB</strong></td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Gestures tracking</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Body joints</strong></td>
<td>20</td>
<td>25</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>SDK</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Portability</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

of hands and fingers, and hence are more appropriate for our study. Second, in view of versatile learning and assessment scenarios, we consider the portability of a sensor, measured as its ability to be attached to a tablet or laptop, to be a valuable characteristic. This allows for an extension of learning environments beyond classrooms to other situations that require mobility, where feature tracking and analytics can take place on demand and on the fly. Third, the availability of a Software Development Kit (SDK) to create customized solutions is central to meeting the diverse needs of learning tasks and assessments. Such middleware allows the transformation of lower-level data streams to high-level features, many of which can be more interpretable. Hence, it yields transparency to scoring models. Table 1 shows selected 3D depth sensors noted for their relevance to user-centric sensing and tracking capabilities. It is by no means exhaustive, but rather represents sensors that have been used with groundings in research studies [12, 6, 16]. The table also summarizes important characteristics of each sensor, which collectively creates a usability scenario that can be matched against the needs of different learning tasks or assessments. For instance, in our public speaking assessment task, Microsoft Kinect is the depth sensor of choice, since it allows the capture of holistic body movements. A full, in-depth treatment of the inner workings of Kinect and how body movements can be inferred is described in [39]. Other scenarios, such as when students attending online courses via laptops require intervention at times of doubt, are probably better served by Intel Realsense. Further, it is possible to use the LEAP motion controller, a highly precise hand gesture tracker, for conducting tutorials in virtual worlds, such as mixing chemicals from test tubes in chemistry lessons.

3. INTEGRATED MOTION CAPTURE SYSTEMS

Advances in low-cost consumer-grade depth sensors such as the Kinect have highlighted the need for standardization of these natural interaction devices. OpenNI\(^3\) is an organization established to shoulder this task. Additionally, it provides an open-source framework for the development of drivers and software for accessing and utilizing Natural User Interface (NUI) devices. Compliant devices include the Kinect and the original PrimeSense sensor, and a growing list of other depth sensors. The Kinect itself has also provided its own SDK\(^4\) with sample code supporting gesture and voice recognition and facial tracking, among other things. To date, five versions of SDK have been released: SDK versions 1.5, 1.6, 1.7 and 1.8, supporting Kinect for Windows v1, while SDK version 2.0 supports Kinect for Windows v2. There have also been a multitude of projects in online repositories in github, kinecthacks.com and codeplex.com that address ways to capture depth sensor data. However, to our knowledge, many of these programs are created as ad hoc solutions for isolated problems e.g. [25], and did not undergo stress testing. In the context of our evaluative task, it is also imperative that the data collection system is easy to calibrate and deploy, allows for repetitive use while being fault-tolerant, and permits the interoperability of depth sensors with other sensing equipment to record synchronized data streams. The last characteristic is important to create a rich feature set spanning across multiple modalities. While it is true that the SDK of many depth sensors also captures and allows extraction of non-depth data such as video and speech, synchronous recording by other higher fidelity devices is sometimes required. For example, for higher speech recognition rates, a lapel microphone is preferred to the built-in microphone array from Kinect. We may also find the need to record data using a non-traditional, non-depth sensor, such as a haptic sensor or sociometric badges for fu-

\(^3\)http://openni.ru/index.html
ture exploitations. Here, we introduce two multimodal data collection systems that we found to be reliable for repetitive use. An attractive notion is that they are available in free, fully-functional version, and are actively supported. Both are based on the Windows operating system.

3.1 Brekel

Brekel\(^5\) is a proprietary, integrated motion capture system that records video, speech and skeletal data streams in real-time using SDK from Kinect SDK and OpenNI and NITE. It allows for quick recognition and tracking of up to 2 persons at a time. Writing of synchronous data streams is direct to disk in FilmBox (FBX), BioVision Hierarchical (BVH) and TXT formats. In total, 58 joints covering all body parts are tracked, with major tracking points such as hips, left/right arms, left/right hands, left/right legs. While file outputs are possible, there is also direct streaming to 3D applications e.g. Autodesk MotionBuilder for creating virtual reality software. The basic version, Brekel Kinect, which supports Kinect v1, is free for research use; other Brekel packages provide enhanced tracking capabilities e.g. Brekel Pro Body v2 tracks humans using Kinect v2, while facial feature and head pose tracking is possible with Brekel Pro Face v2. The experiments reported in this paper are performed using Brekel Pro Body Kinect tracking software, which is similar to Pro Body 2 but operates with Kinect v1.

3.2 Multisense

Multisense\(^6\) is a open-source framework for integrating data streams from multiple sensors for the purpose of audio and video sensing, as well as non-verbal behavior processing. Developed by USC ICT initially as part of the Virtual Human Architecture\(^7\) (VHT), it is a modular platform for generating probabilistic models for human behavior perception and understanding. Multisense is capable of tracking, in real-time or offline mode, facial expressions, body posture, acoustic and linguistic patterns, and higher levels of behavioral descriptors, such as attention span for example. This tracking is enabled via multiple input modalities including video, audio, depth streams from devices such as Kinect, webcam and high-fidelity microphones embedded into the framework. The data streams are synchronized but processed by the appropriate trackers independently e.g. CLM-Z for facial and head tracking [2], skeleton tracking by Microsoft Kinect SDK and FAAST [34] for skeleton action encoding. In total, 24 joints covering all body parts are tracked, with major tracking points such as head, neck, torso, and waist, as well as left/right shoulders, left/right knees, etc. The richness of the multimodal feature set created can lead to inference of higher-level information such as mental and emotional states. Architecturally, Multisense operates on multithreading processes allowing for the different trackers to run in parallel and in real-time. This functionality is built and improved on top of the Social Signal Interpretation (SSI) framework [36] from which MultiSense borrows basic functional elements. A compiled, basic version of Multisense is freely available at the VHT website by request. This version includes the AVRecorder application which allows for multithreaded, synchronous recording of Kinect v1, a webcam, and external microphones.

4. CASE STUDIES

4.1 Task

Oral communication skills are critical to success in both learning and career development. For example, it has been consistently rated as one of the most valued workforce skills in large scale surveys [22], and its importance is also reflected in the newly developed national K-12 education standards [28]. Consequently, there is a pressing need for obtaining valid, reliable, and cost-efficient ways to evaluate oral communication skills, such as on public speaking tasks.

Public speaking involves a range of constructs, including but not limited to, content, organization, language use, vocal quality, and the effective use of non-verbal cues. Evidence for these qualities is distributed across multiple modalities, from linguistic (e.g., coherence of the message and word choice), vocal (e.g., intonation and disfluencies), facial expressions, to hand and body gestures. A competent public speaker coordinates all aspects of these modalities to achieve an engaging and effective performance. It is not surprising that most existing rubrics for public speaking skills [29, 24, 32] evaluate both verbal and nonverbal aspects of communication. Clearly, low-cost commodity depth sensors, such as Microsoft Kinect, have the potential of providing a cheap solution for tracking and understanding body postures of presenters.

From 2013 to 2014, we have worked on creating a multimodal public speaking performance corpus. Two types of public speaking tasks were used in our data collection, namely informative and impromptu presentations. In an informative speech, presenters were given a pre-prepared slides deck and up to 20 minutes to prepare their presentation delivery. In an impromptu speech, presenters were required to advocate an idea that was not actually favorable to them. No visual aids were provided for this type of speech. The experience of the presenters in public speaking varied widely from beginners to experienced members of the Toastmasters club.

4.2 Data Collection

The selection of the hardware and software used in data collection are based on the following technical requirements:

- **complete motion recording**: in contrast to event-based motion recording, which usually lasts several seconds, a complete presentation session may last up to 5 minutes.
- **full body tracking**: full body tracking of presenters must be recorded for body language understanding and interpretation.
- **complicated background**: PowerPoint presentations were used in the informative presentation task, and therefore interfered as background “noise” lighting, affecting brightness and contrast during video recording.
- **high-resolution face videos**: facial expressions shown in presentations are important cues for detecting psychological states e.g. nervousness, which is useful for rating performance.
- **data streams synchronization**: multiple data streams (motion, video, and audio) may be recorded independently by different devices, hence they must be precisely synchronized.
the ground to restrict movements of presenters to within a view constraints of Kinect v1, markers had to be placed on motion recording. Due to horizontal and vertical field of on a laptop with modest configuration to support complete technical requirements. Brekel Pro Body operated reliably in the data collection. This system satisfied most of our total of 17 participants from our organization participated high-resolution videos/audios recorded by the camcorder. A traces, videos/audios recorded by Brekel software, as well as by the end of 2013. The corpus contains Kinect motion traces were recorded independently by the camcorder and Kinect respectively, manual, imprecise synchronization had to be performed via the clapping action. Attempting to focus on faces of presenters using digital zooming in the camcorder would fail from time to time due to excessive pacing of the presenters.

One major issue was with lighting in the background. Due to illumination from The SMART Board used for displaying presentation slides during the informative talk task, a “lens flare” effect was generated, thereby reducing contrast on the face and subsequently hindering with facial expression analysis.

To address these issues found in 2013, in our 2014 data collection, we made several important improvements to our data collection system. Firstly, to address the challenge of precisely synchronizing multiple data streams, we utilized the MultiSense software from USC. The consumer-grade camcorder previously used was replaced with a HD webcam (Logitech C615). Similar to our 2013 setup, Kinect v1 was used for recording body traces for the presentations. Data streams from the different sensors i.e., Kinect v1 and webcam were time synchronized using the VReeper application of MultiSense, which allows for multithreaded, synchronous recording of the input sensors. USB 3.0 connections are made between the Kinect, webcam and the recording laptop. Given the large real-time data throughput requirements, a high-end mobile workstation with Intel Core i7-4800MQ quad-core processor with 8 threads, 16 GB RAM, and a 512 GB SSD hard drive was used in 2014. More details on the setup of sensors and equipment as well as the recording environment are illustrated in Figure 2.

Using the above-mentioned data collection system, we successfully created our first multimodal presentation corpus by the end of 2013. The corpus contains Kinect motion traces, videos/audios recorded by Brekel software, as well as high-resolution videos/audios recorded by the camcorder. A total of 17 participants from our organization participated in the data collection. This system satisfied most of our technical requirements. Brekel Pro Body operated reliably on a laptop with modest configuration to support complete motion recording. Due to horizontal and vertical field of view constraints of Kinect v1, markers had to be placed on the ground to restrict movements of presenters to within a trapezium-like area. This prevented motion traces recording from being dropped during presentations.

However, there were other requirements that were not met fully. Since the high-resolution videos/audios and the motion traces were recorded independently by the camcorder and Kinect respectively, manual, imprecise synchronization had to be performed via the clapping action. Attempting to focus on faces of presenters using digital zooming in the camcorder would fail from time to time due to excessive pacing of the presenters.

Secondly, to address the “lens flare” effect caused by the SMART Board, we utilized an Autocue/QTV 3 head soft-box lighting kit to provide contrast adjustment for presenters. Two softbox heads, each with 85W fluorescent lamps, are placed 10 feet away from each end of the SMART Board.
respectively, projecting at an angle of 45° inwards toward the presenter. Using this new lighting toolkit, we conducted a pretest/posttest on facial features extraction based on CLM-Z [2] head/face tracking application, and concluded that the additional lighting significantly decreases the percentage of corrupted video frames tracked from 19.2% to 5.24%, as well lowering the mean and standard deviation of tracking uncertainty over those frames. Figure 3 shows the effect of using a lighting softbox.

![Figure 3: Applying two Autocue/QTV 3 lighting softboxes significantly improves contrast on the faces of the presenters: the left/right pictures reveal the effects before/after the softboxes are applied.](image)

In summary, there are several technical requirements for collecting multimodal data from public speaking sessions in order to allow for a meaningful analysis and interpretation of these data for learning or assessment purposes. In our initial recording setup in 2013, we managed to address a subset of these requirements. In 2014, our improved setup has enabled us to meet all of the requirements. While the delivery of each speaking task represents a scenario in the lab, there are already significant challenges encountered, such as the inadequately field of view of Kinect v1 that led to the requirement of a marked out area for restricting presenters’ movements, or the diminished contrast on presenters’ faces due to the “lens flare” effect, which is still somewhat problematic for ideal facial expression analysis. Migrating the tasks delivery to a real-world classroom would inevitably bring more challenges. Nevertheless, we hope that the advancement of sensors and tracking technology would allow some of these problems to be solved. For example, Microsoft Kinect v2 boasts a larger effective FOV and should lift some of the movement restrictions. It is a prospect we are interested in pursuing as part of our future work. For the remaining sections in this paper, we focus on feature extraction and machine learning experiments on the 2013 corpus.

5. OPEN-SOURCE TOOLKIT FOR GENERATING MULTIMODAL FEATURES

We hypothesize several high-level features that can be transformed from low-level sensors data that are potentially useful in a machine learning framework for our public speaking tasks assessment. To our knowledge, there are no publicly available implementations of these features, hence we hope to benefit the MLA community through the release of an open source toolkit that facilitates their extraction for evaluation of public speaking tasks, as well as other similar pedagogic tasks related to speaking. In this section, we first elaborate on details for motivating these features, and provide details on the open-source toolkit that generates them.

5.1 Motion-based Body Features

Note that, though we have a multitude of multimodal features as outputs from the sensing network, we focus on the use of the depth sensor and its efficacy in our work. While we are aware that Microsoft and OpenNI SDKs provide gestures recognition ability, they are mostly limited to a set of closed-form gestures e.g. hand swiping to the left, etc, for improving interaction with NUI devices. Such gestures have limited interpretability for evaluating public speaking, and also do not occur frequently during speaking presentations (feature sparseness problem). In contrast, we aspire to extract higher-level features that are more generalizable and applicable to measuring the constructs of speaking proficiency.

Various methods have been proposed to compute expressive features related to body language in several research areas, e.g., affective computing, virtual agents, and multimodal dialogic systems. For example, [26] systematically summarized the methods used for analyzing expressive body language performances. They categorized the extractable features into the three layers: (a) low-level, such as hands velocity, (b) medium-level, such as hands symmetry, and (c) high-level, such as gestures bearing communicational meanings. Glowinski et al. [14] presented a framework for finding a minimal representation set of affective gestures. They first processed head and hand motion data through an array of expressive feature extraction modules for measuring energy, spatial extent, smoothness, symmetry, and head posture. Next, they computed statistics such as the maximum, mean, and standard deviation (SD) from these measurements. Finally, the minimal set of representational features was obtained using PCA-based dimension reduction.

Following [14, 26], we extracted a number of visual features related to spatial, temporal and dynamic aspects from the Kinect motion data, a frame-by-frame 3D-coordinated (XYZ) recording of each body part as stored in the BVH file. Specifically, we focus on the following body parts: hip, spine, left/right forearms, and left/right hands, due to these body parts’ dynamism during presentation.

- **Spatial**: for each motion data frame, (a) the distance between the hands and the body (hands-spine); (b) the distance between the arms and the body (arms-spine), and (c) the distance between the two hands (hands);

- **Temporal**: the first order derivatives of the above spatial measurements; the corresponding features’ names contain a special token “1dv”. For example, hands-spine-1dv indicates the first order derivative of hands-spine.

- **Power**: second order derivatives of the above spatial measurements; the corresponding features’ names contain a special token “2dv”. For example, hands-spine-2dv indicates the second order derivative of hands-spine.

Note that the temporal feature relates to the rate of change of the spatial feature e.g. the speed at which both hands are moving away from each other. Similarly, power relates to the rate of change of the temporal feature e.g. the rate at which the speed is changing when both hands are moving away from each other. In additional to the spatial features, both temporal and power features approximate “smoothness” of
the movement of body parts. Furthermore, [26] also suggested several derived features to measure full-body motions, such as Kinetic Energy (KE), Posture, and Symmetry.

• Kinetic Energy: We postulate the KE as a measure of the power/energy exhibited by the presenter. It measures the amount of overall movement of the body, and is somewhat correlated to the speaker being “physically active” during the presentation. KE is computed based on the velocities of upper-body segments and their corresponding percentage of mass, as described in the formula below:

\[ KE = 0.5 \times \sum_i m_i v_i^2 \]

where \( m_i \) and \( v_i \) refer to the normalized mass and velocity of body part \( i \) respectively. In our work, we focus exclusively on body parts that are in the upper-body region – namely, hips, spine, shoulders, arms, forearms and hands – noted for their importance in a related study in Kinesiology [27], from which we also observed the normalized mass of these body parts.\(^7\)

• Posture: The posture of a presenter can be approximated using the concept of a Bounding Volume (BV) described in [26], whereby using the hip as the center, we construct an imaginary, pseudo-cuboid whose length, width and height are formed by the furthest distance spanning any two arbitrary body parts in the \( X \), \( Y \) and \( Z \) dimensions respectively. The BV of a presenter at any frame can be computed simply by taking the volume of this imaginary cuboid. Intuitively, BV provides an approximation of the degree of body “openness” shown by the presenter.

• Symmetry: Following [26], we compute the symmetry index (SI) of a presenter’s two-handed gestures. SIs measure the degree of symmetry exhibited by both hands along a specified spatial dimension. For example, we compute \( X \) dimension’s SI as:

\[ SI_X = \frac{||X_P - X_L|| - ||X_P - X_R||}{||X_L - X_R||} \]

where \( X_L \) and \( X_R \) are the \( X \)-dimension coordinates of the left and right hands, and \( X_P \) is the \( X \) coordinate of the pivoting body part i.e. hips. A value of 0 for \( SI_X \) suggests perfect symmetry while a value of 1 suggests perfect dissymmetry. \( SI \) for the \( Y \) and \( Z \) dimensions can be computed accordingly. Symmetry-based features were shown to be relevant for measuring affective and behavioral status, and is also concerned with detecting the underlying attitude of a speaker e.g. relaxed or tensed [14]. From symmetry measurements of the individual axes, we also derived the following averages as additional measurements: \( SI_{XY} = (SI_X + SI_Y)/2 \) and \( SI_{XYZ} = (SI_X + SI_Y + SI_Z)/3 \).

Finally, for each frame-wise measurement described above \((f)\), a set of statistics were computed to serve as features for the entire response, including:

\[ \text{mean}: \text{mean value, which is } mean(f) \]
\[ \text{mmrt}: \text{ratio of the max value to the mean, which is computed as } max(f)/mean(f) \]
\[ \text{mean-log}: \text{mean of log-scaled value, which is computed as } mean(log(f)) \]
\[ \text{mmrt-log}: \text{ratio of max log-scaled value to the mean of log-scaled value, which is computed as } max(log(f))/mean(log(f)) \]
\[ \text{SD}: \text{standard deviation, which is computed as } SD(f) \]
\[ \text{SD-log}: \text{standard deviation of log-scaled values, which is computed as } SD(log(f)) \]

As a way to efficiently generate these high-level features, we implemented the Visualization and Analysis for Multimodal Presentation (VAMP)\(^8\) toolkit, which is a toolkit consisting of a set of scripts in Python, chosen for its fast prototyping speed for research. Currently, only processing of Kinect-based skeletal data is supported. In the future, we hope to extend support to other depth sensors, while also providing continuous updates to the library of high-level features found to have high relevance for the automated scoring of public presentation tasks.

The toolkit receives as inputs a set of synchronized video, audio and skeletal data, and transforms them into feature contours for visualization. Figure 4 shows the operating flowchart of our toolkit. The core function emphasized here is the application of a recursive parser to extract the frame-level features elaborated above using the skeletal data from Kinect. Note that there are other functions that extract multimodal cues from the other modalities as well, and these would also be available as part of the toolkit. Of those, Praat \([5]\) can be used to determine the prosody of the speech given during the presentation, while FAVE \([31]\) can be used to align the phones and words if an audio transcription is available. Video-based facial emotion extraction is an optional function which relies on the outputs generated by FACET\(^9\), a state-of-the-art computer-based automated emotion recognition toolkit that requires separate licensing. A more elaborate explanation of the motivation for including these multimodal cues can be found in [11]. Finally, all feature contours from the different modalities are normalized to be range-bound for optimal display in Anvil. Standardization of codecs formatting for video and audio is performed to comply with visualization specifications. Note that, in addition to the purposes of visualization, the time-series data from each modality can also be saved to persistent storage for the purposes of further analysis and model training.

As mentioned, we used Anvil\(^10\) as our visualization module. Originally developed as a gesture research tool, Anvil is now widely used as a video annotation toolkit in Human-Computer Interaction, Linguistics, Ethology, Anthropology, Psychotherapy, Embodied Agents, Computer Animation and many other fields. It provides annotations that are accurately aligned with the various input modalities e.g. video, audio and skeletal. Additionally, the annotations are colored and hierarchically defined, and are fully customizable.

\(^7\)When a body part listed in our BVH motion tracking result is missing from [27], we selected the normalized mass weight of the closest body part as its replacement e.g. pelvis weight is used for hips, abdomen weight is used for spine.

\(^8\)https://github.com/EducationalTestingService/VAMP

\(^9\)http://emotient.com/

\(^10\)http://www.anvil-software.org/
using the XML markup language. By visualizing the time-series data in one place, researchers can perform observation and garner intuition and evidence about the potential usefulness of certain features or feature types for building assessment or learning models. Furthermore, interesting event markers can be annotated for further discussion and analysis.

5.2 Motion-based Body Features Applied to Public Speaking Assessment

The motion-based body features have previously been used in research and have been found to be valuable for representing body language behaviors. Using our 2013 multimodal presentation corpus (containing 56 completed sessions with all data streams), we applied a well-established human rating rubric [33] to fully double-score all presentations. Due to the surprisingly low inter-rater reliability found on our corpus, we ended up using the adjudicated scores (denoted as final score), which are holistic judgments by human raters, determined by a range of presentation performance aspects, e.g., word choice, verbal expression, and also nonverbal communication. The low inter-rater reliability is probably due to the subjective nature of the scoring process itself, hence the adjudicated score is a more reliable metric for measuring presentation performance.

The adjudication process is described as follows: All presentations were double-scored by two raters using a score-scale from 1 to 4. In the cases in which the two scores differed by more than 1 point, a third scorer was called in to adjudicate. The final “operational” score is the average of all scores assigned to a given dimension in the scoring rubric for a given presenter. Consequently, a psychometric analysis of the human scores was published in [19], but is not included here since it is not the focus of this paper.

A correlation study has been done on the extracted Kinect data with final holistic scores that reflect overall presentation performance. Table 2 reports some representative correlations between the features extracted and the non-verbal behavior score, as well as the holistic score. $|r|$ values of more than 0.4 are regarded as significant. We find that the features we proposed and computed by using the toolkit show considerable indicative abilities for predicting overall performance. For example, a large standard deviation in hands-spine displacement and bounding volume is positively correlated to the non-verbal behavioral score. This probably indicates that the appropriate use of hands (e.g. gesturing) and body parts extension (e.g., a form of embracing) are indicative of a non-verbal behavioral style that is appealing to the audience. Despite these promising results, note that this study was based on quite a small data set ($n = 56$), and a larger dataset is needed for further validation.

These Kinect body-based, high-level features were used in a machine learning experiment for building an automatic prediction model to evaluate a presentation based on multimodal input features. In this study, a set of linguistic and speech related features were extracted from the audio streams and their corresponding transcriptions. From the visual channel, these Kinect features were used for representing body language. Furthermore, extra visual features from head orientation, gaze tracking, and facial expressions were considered as well. The visual features (including Kinect features) were found to provide supplementary support for the presentation evaluation task. More on the experimental details can be found in [11].

6. CONCLUSION

Multimodal learning analytics is a growing multi-disciplinary research area that leverages expertise from education, computer vision, natural language processing, linguistics, cogni-
tive science and psychology. Expanding beyond the usage of standard sensing equipment such as video and audio, we proposed the exploitation of depth sensors for extracting construct-relevant features for modeling assessment tasks. Specifically, in this paper, we focused on the use of Microsoft Kinect to obtain motion-based, low-level data and transform them into higher-level features for measuring public speaking skills proficiency. We show how Kinect can be integrated into existing integrative recording platforms for performing data collection via two case studies, and provide technical details in those studies. The preliminary assessment of scoring models that were based on the Kinect’s depth sensing capabilities and built in a machine learning framework provides promising results. While there is relevant work in the literature addressing the use of depth sensors for multimodal learning analytics, few efforts have been expended to standardize algorithms and protocols for performing learning or assessment tasks, probably due to the exploratory nature of this line of work. We aspire to create a standard toolkit that allows data transformation, feature extraction and data visualization on the most commonly used input sensor modalities for research purposes, and we hope that this contribution streamlines efforts in the MLA research community. In the future, we also hope to pursue research in other learning analytics scenarios utilizing other low-cost, commodity depth sensors, such as portable learning assessments or virtual reality applications for learning simulations. Finally, given the absence of sensors in everyday, non-experimental learning scenarios, we hope to pursue another interesting research direction that involves detailed analysis of whether their presence would have effects over the validity of the predictions of the two speaking tasks.

7. REFERENCES


<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>NVB</th>
<th>Holistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD hands-spine-disp</td>
<td>Spatial</td>
<td>0.405</td>
<td>0.376</td>
</tr>
<tr>
<td>mean hands-spine-disp-1dv</td>
<td>Temporal</td>
<td>−0.292</td>
<td>−0.322</td>
</tr>
<tr>
<td>mean arms-spine-disp-2dv</td>
<td>Power</td>
<td>−0.429</td>
<td>−0.635</td>
</tr>
<tr>
<td>lg SD bounding-volume</td>
<td>Posture</td>
<td>0.499</td>
<td>0.354</td>
</tr>
<tr>
<td>mean kinetic-energy</td>
<td>Kinetic Energy</td>
<td>−0.315</td>
<td>−0.426</td>
</tr>
<tr>
<td>symm3d.mean</td>
<td>Symmetry</td>
<td>0.220</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Table 2: Table showing the Pearson correlation of a few representative Kinect-based motion features to two types of human-rated scores. Note: NVB denotes Non-Verbal behavior score while Holistic denotes overall holistic scores.


