A Mixed Reality System for Teaching STEM Content using Embodied Learning and Whole-Body Metaphors

Remo Pillat[*]
Synthetic Reality Lab
University of Central Florida

Arjun Nagendran[†]
Institute for Simulation and Training
University of Central Florida

Robb Lindgren[‡]
Media & Learning Lab
University of Central Florida

Abstract

This paper describes the development of an MR environment that can be used in teaching STEM (Science, Technology, Engineering, and Mathematics) topics. Specifically we seek to create a space for facilitating whole-body metaphors where learners use the physical movement and positioning of their entire bodies to enact their understanding of complex concepts.

A rigorous technical approach comprised of virtual elements, real users, spatial audio, and an integrated sensor network is presented that fulfills the requirements of an embodied learning environment. An algorithm that uses homography-based multi-projector blending is used to create a large, seamless projection on the floor that affords a human-scale interaction environment. To further improve the immersive quality, projectors are strategically overlapped to minimize user shadows on the projected surface. A hybrid sensor solution using a Kinect and a laser scanner is developed that tracks users' physical movements and extracts relevant game parameters such as position and velocity. Requiring no pre-training or props, this tracking setup is adaptable and shows high performance over a wide range of users, from children to adults. An exhibit employing this MR system was field-tested at the Museum of Science and Industry in Tampa, FL.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Sensor fusion

Keywords: mixed reality, embodied learning, whole-body interaction, whole-body metaphors, educational game, sensor fusion

1 Introduction

STEM education has become a top priority in the United States [National Science Board 2010]. Researchers and educational practitioners continue to seek innovative ways to deliver important and challenging science and technology content while maintaining high levels of student engagement and an interest in pursuing STEM careers. It has been widely recognized that video game and simulation technologies have a strong potential to improve education in the sciences [Honey and Hilton 2011], but the specific interface paradigms that effectively augment student understanding - and the precise technology specifications for implementing these paradigms - are still being developed.

Mixed Reality (MR) systems designed for educational purposes have, in particular, received attention for their promise in recent years. Researchers have described the learning affordances of these systems and several applications in both formal and informal STEM education have been developed [Birchfield and Johnson-Glenberg 2010], [Chang et al. 2010], [Hughes et al. 2005], [Kirkley and Kirkley 2002], [Pan et al. 2006]. Many of these papers cite the heightened engagement that students experience when interacting physically with a novel and immersive digital environment. We focus here on a paradigm of MR interactions that we refer to as body-based metaphors.

Body-based metaphors are a type of embodied learning where an individual or a group of individuals use their bodies to enact concepts they are attempting to understand. Previous research in philosophy and education has described how body movement can serve as a starting point for new learning [Gallagher 2005], [Goldin-Meadow et al. 2009] and studies have shown that metaphor can be an effective instrument for conveying difficult science concepts [Cameron 2002], [Christidou et al. 1997]. With MR, conceptual metaphors and body movement can be effectively combined by permitting a learner to take on the role of a system component. The MR environment provides real-time feedback and visualizations that support the metaphor and help the learner to generate correct intuitions. The study of [Lindgren and Moshell 2011] indicated a strong educational potential for MR experiences implemented based on the design requirements for metaphor-based learning.

To successfully realize MR embodied learning experiences there are a number of requirements that have to be met. The first three of those are taken from the design principles for an embodied learning environment described in [Johnson-Glenberg et al. 2011]: an **intuitive interface** that allows for direct manipulation, **user immersion**, and **human scale**. We have adapted these three principles and added two additional requirements that we believe are necessary in particular to support our paradigm of body-based metaphors. The experience should facilitate the **establishing of identity** and provide **rich channels of feedback**. Based on these requirements we now describe an MR system that was designed to address them.

Although the benefits of embodied learning and body-based metaphors are well grounded in the research literature, less attention has been given to developing the technical design principles that would effectively enable these kinds of interactions. These principles should not be based on the constraints of any given technology, but we will give specific recommendations on how they could be implemented given the current state-of-the-art.

2 Prior Work

Several mixed reality systems are presented in the literature that were designed for large-scale embodied learning experiences. The SMALLab interaction space [Birchfield et al. 2006] is based on

[*] e-mail:rpillat@cs.ucf.edu
[†] e-mail:arjun@cs.ucf.edu
[‡] e-mail:robb.lindgren@ucf.edu

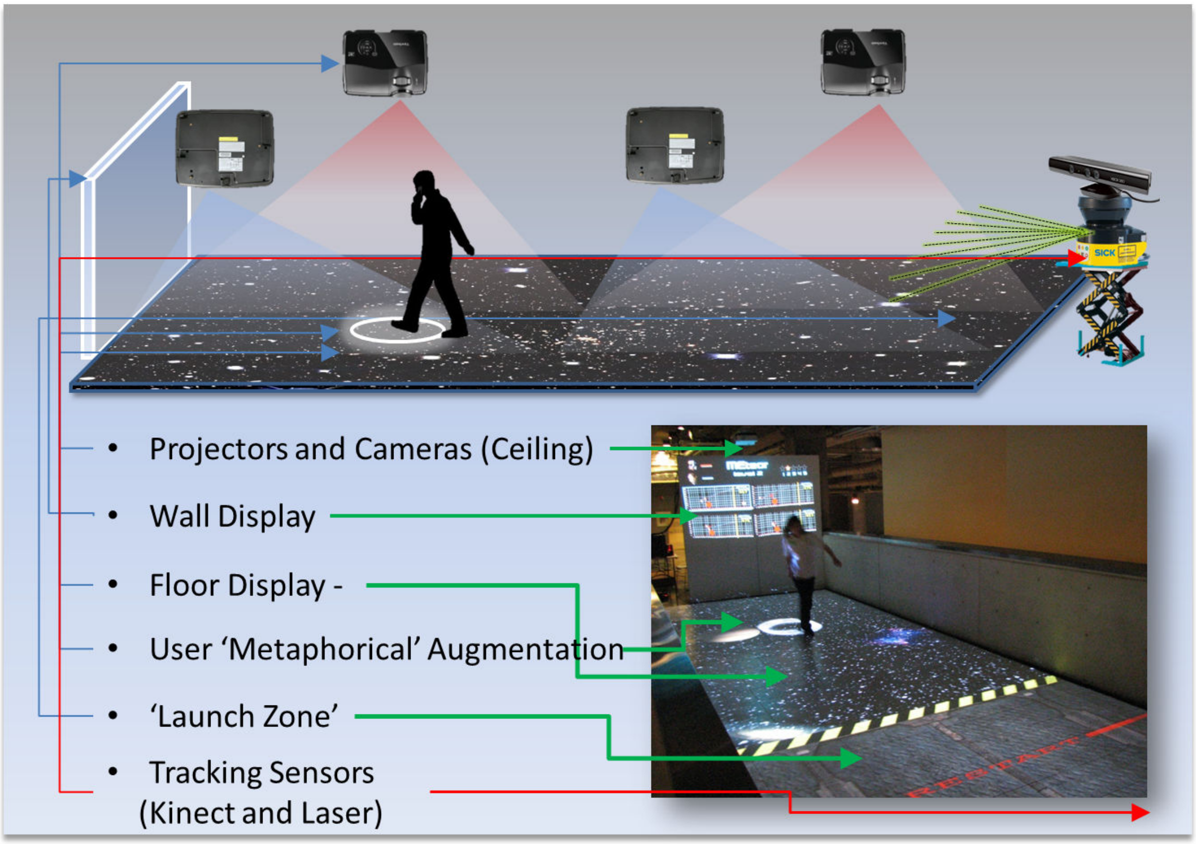


Figure 1: *General system architecture showing interaction in the immersive space during a user study at the Museum of Science and Industry, in Tampa.*

a 4.5 m x 4.5 m interaction area. Visual information is provided through one overhead projector, so that user shadows will obstruct some of the imagery. Users' positions are found through optical cameras using color-based segmentation and tracking. Using optical cameras as tracking devices poses questions of robustness under a variety of lighting conditions and user appearances. In later iterations of the system [Johnson-Glenberg et al. 2011], the users carried a trackable object to resolve their positions.

Age Invaders [Khoo et al. 2008] was designed as a game-based mixed reality interaction space. The tracking system is dependent on RFID tags in the users' shoes, thus limiting interaction to participants wearing the props. The floor is illuminated by large low-resolution LED screens underneath a walkable plastic surface. This circumvents the problem of shadows, but no high-resolution imagery can be displayed and the interaction area is limited to about 3 m x 1.5 m.

An interesting approach to shadow minimization in multi-projector overhead systems is presented in [Nagase et al. 2011]. Shadows were removed by supplementing the existing image through an additional projector. That effectively mitigates the shadow but limits overall projection brightness to the magnitude that any one projector can achieve. In addition, the algorithm requires prior geometrical information and continuous tracking updates about the shadow-casting object.

CAVE [Cruz-Neira et al. 1993] and derivative immersive systems [Sajadi and Majumder 2012] are popular for medium-scale interactive spaces. Usually, these systems are limited in size and are used for 3D visualization purposes rather than interactive, educational experiences. Although larger CAVE environment can reach the size of our interaction space, they are hard to transport and the cost of installation makes these systems not easily available.

Commercial systems from Snibbe Interactive [Snibbe and Raffle 2009] and GestureTek [GestureTek 2012] allow the building of large-scale mixed-reality environments. Both companies make use of the users' shadows as innovative interaction element, but the approach seriously hinders visualization of objects in the users' vicinity.

An early precursor of the proposed system is presented in [Lindgren and Moshell 2011]. The interaction area was limited to about 3 m x 2 m and there was no overlap between projectors. In addition, the tracking implementation employed a single head-mounted prop, thus limiting the movements of the users.

3 System Overview

The system proposed herein is shown in Figure 1. Multiple projectors are used to illuminate a large area in an overlapping configuration that eliminates shadows while a combination of distance measurement sensors is used to facilitate interaction in the space. In specific, four projectors illuminate a large rectangular floor area measuring 9 m x 3 m (30′ x 10′) that can be used to deliver scenario-specific content, while a fifth projector illuminates a wall to display statistics and performance-related metrics to the user. The large scale of the installation allows human scale movements and encourages active engagements by the users (see requirement for human scale). Two webcams are mounted between the projectors to assist in the alignment process.

A SICK LMS-111 laser scanner positioned at the far-end of the floor area is used to track a user in the immersive space. In addition, a Microsoft Kinect is rigidly mounted above the laser scanner and provides accurate user and limb tracking within a section of the interactive area.

Multiple speakers are mounted above the interaction surface and provide a rich auditory feedback.

4 Multi-Projector Blending Minimizing Shadows

An intrinsic feature of every overhead projection is the occurrence of shadowing due to partial occlusion by the user. This has implications for the user's feeling of immersion, one of the initial requirements, as any visual artifacts potentially incur a break in presence for the subject.

The decision to employ overhead projection is usually made for cost-conscious reasons. One other alternative for large floor interaction surfaces is the use of walkable platforms that incorporate under-floor projectors. The necessities of platform stability and floor transparency increase the price point of this kind of system considerably.

Since the projectors in our system are ceiling-mounted, shadows cast by the users could pose a problem in this mixed reality environment. Shadows can potentially degrade or completely hide the projected imagery, thus decreasing the level of immersion. Our solution ensures that each floor area is covered by multiple projectors that are mounted in complementary positions and angles. Essentially, an image is projected by one projector onto the shadow of the user that is cast on the floor by another projector. This allows the user to still clearly observe all elements of the floor-projection, albeit at reduced brightness in the shadowed areas.

No attempt is made to actively remove the shadows, because any such technique would have to reduce the projection brightness in other areas. Since we expect that most of the users' movements occur along the long axis of the interaction area, their shadows will appear perpendicular to their motion vector. The darkened area will thus only be perceived through peripheral vision, minimizing the impact on the feeling of immersion. During the deployment of the sample implementation described in Section 7 our design assumptions were confirmed, as none of the users expressed a dissatisfaction with the appearance of the areas with reduced brightness.

Because multiple projectors cover the same area, an increased surface brightness can be achieved. This allows the use of lower-cost projectors while maintaining acceptable brightness levels.

4.1 Projector Alignment

Since the projectors have a high percentage of overlap, an accurate alignment of the projected imagery is important to further the goal of user immersion. The procedure presented here is based on the work of [Chen et al. 2002]. Similar methods can also be found in [Raskar et al. 1999; Raskar et al. 2002] and [Cotting et al. 2011].

The following assumptions can be made about the system. The projection surface is reasonably flat and can thus be considered an idealized plane. All the cameras are assumed to be internally calibrated, so they can be treated as pinhole cameras. The projectors are considered to be following the pinhole model as well. Note that the external position of the cameras is irrelevant for the alignment process, as long as each camera's viewing frustum covers about half of the space.

Under these conditions, each transformation between points in camera-, projector- and screen-space can be described by a planar homography **H** [Hartley and Zisserman 2003]:

$$\mathbf{x}' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \mathbf{Hx} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \qquad (1)$$

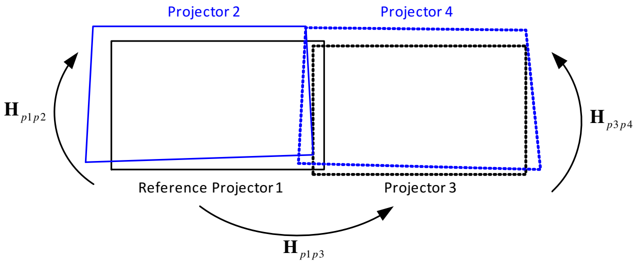


Figure 2: *A four projector setup with mutually overlapping projector pairs. Also shown are the conceptual homographies* $\mathbf{H}_{p1p2}$*,* $\mathbf{H}_{p1p3}$*, and* $\mathbf{H}_{p3p4}$*.*

There are two types of homographies that are significant: ones that transform points from projector pixel space to camera space and others that directly translate between projector pairs. The former will be denoted by $H_{p\mathbf{i}c}$, where **i** is replaced with the currently considered projector. These homographies can be directly estimated from image correspondences. The latter type of homography will be derived from the projector-camera ones and is written as $H_{p\mathbf{i}p\mathbf{j}}$ where **i** and **j** are two distinct projectors.

The basic idea of our algorithm is that several horizontal and vertical calibration line patterns will be displayed by each projector separately. Webcams mounted in the ceiling record an image of the projected patterns on the floor. The lines can be extracted from these frames through dynamic thresholding. A connected component analysis is used to reject noisy outliers and extract the dominant lines in the image. Finding a linear least-squares approximation of these components yields lines whose horizontal and vertical intersection points establish correspondences between projectors and the camera. Based on at least 4 point correspondences, a homography can be calculated. We use the Gold Standard estimation method described in [Hartley and Zisserman 2003] that combines a linear estimation step with a RANSAC outlier rejection and a final non-linear refinement. To improve the robustness of the estimation, we typically use in the order of 100 - 200 correspondences.

One projector-camera homography is calculated for each projector, but the ultimate goal of the calibration procedure is to find a mapping between the different projector coordinates. As an example take the homographies $\mathbf{H}_{p1c}$ and $\mathbf{H}_{p2c}$ for projectors 1 and 2. The derived homography that relates projector 1 and projector 2 can be found by simple algebraic concatenation:

$$\mathbf{H}_{p1p2} = \mathbf{H}_{p2c}^{-1}\,\mathbf{H}_{p1c} \qquad (2)$$

Similarly, homographies between arbitrary projector pairs can be calculated as long as a transformation between them exists in the transitive homography tree [Chen et al. 2002].

In our application, we use the described method to calibrate a 4-projector system. The approximate projection areas and specific homographies are indicated in Figure 2.

The whole alignment procedure is automated and is completed in less than one minute. With the calculated homographies, the scene can be rendered once and post-processed by a specialized vertex shader for each projector output. Since the homography transformation is a linear operation, it can be implemented with minimal overhead on the GPU, thus leaving enough computing resources for the rest of the application.

4.2 Projector Synchronization

When the number of projectors increases, the question arises how a graphics input can be provided to all of them simultaneously. In most cases, a synchronization mechanism has to be implemented to ensure image coherence in frame rate and appearance. Different commercial solutions exist (e.g. Mersive, Coolux) that allow a multi-computer rendering synchronization.

In this application, it was decided to drive all projectors through one computer to minimize the hardware requirements. For the presented 4-projector solution, this was achieved by subdividing the rendered image into quadrants. Each quadrant represents the input to one projector and the respective rendering step accounts for the necessary transformation to ensure projector alignment (see section 4.1). This single rendered image is then split into 4 external Dido LT $^{\text{TM}}$ video wall processors by Aurora Multimedia. Each module crops one quadrant and resizes it to the attached projector's native resolution. At a minimal loss of resolution, one graphics output can thus feed imagery to 4 projectors simultaneously.

5 Hybrid User Tracking

Traditionally, tracking in mixed reality spaces is accomplished using props such as retro-reflective markers, and viewed by an array of infrared cameras. The use of such props in a fully immersive mixed reality experience can limit the range of motion and place restrictions on the movement of users engaged in the experience. Non-intrusive technology is therefore employed for user tracking in the proposed system: a laser scanner and a Microsoft Kinect. Although the two sensors have different characteristics, they complement each other in the returned modalities.

In addition to not requiring any props, our proposed system allows user tracking in the whole interaction area and requires no pre-training for different users. This aides the goal of an intuitive user interface and fulfills another one of our requirements. Depending on the requirements of the simulation, the calculation of other modalities like user velocity or movement vector might be required. This section will detail how a high fidelity in user position and velocity estimation can be achieved by combining multiple sensors and fusing their measurements.

The used SICK LMS-111 laser scanner has a 270° field of view, and a maximum ranging distance of nearly 15 m. This ensures full coverage of our interaction area with arbitrary sensor placement. The 2D laser has a scanning frequency of 50 Hz at an angular resolution of 0.5° and returns a point cloud of data points in its field of view.

The Microsoft Kinect returns a 3D depth and aligned color image. Its maximum depth range is 4.5 meters, although noise levels increase proportionally with depth as well. Because of its limited 58° horizontal opening angle, its effective tracking area in our system is limited to approximately 2.5 m x 3 m. One of the strengths of the Kinect is the richness of its returned information. Besides the raw depth image, users in the sensor's field-of-view are segmented reliably and abstract skeleton information (positions and orientations of several predetermined body points, e.g. head, shoulder, wrist) is extracted at a high rate. Data from both the Kinect and the laser sensor were analyzed with respect to the band of noise they produced when estimating velocities. In essence, this is a representation of how smooth the velocity estimation curve from each of the sensors is.

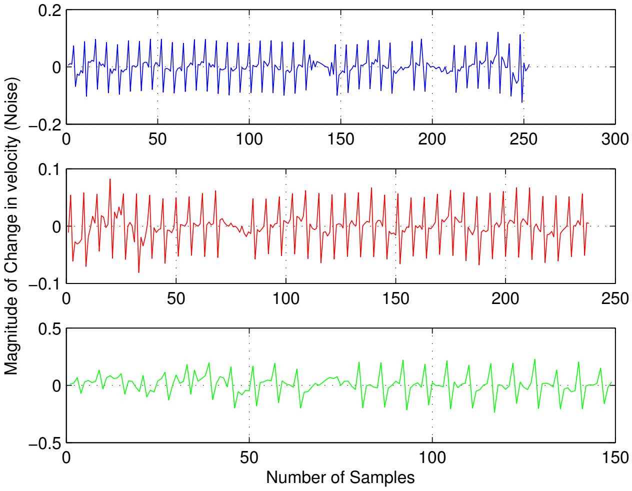


Figure 3: *Noise profile of the velocities from the laser scanner reveals a band whose magnitude increases with velocity.*

It was found that the Kinect performs particularly well when estimating velocity (both magnitude and direction), but is limited in its range. The laser scanner, allows estimates of velocity to be obtained over the entire area, but has a reduced accuracy owing to the processing of point-cloud data. From Figure 3 and 4, it is evident that the noise profile of the Kinect has a much narrower band (magnitude of approximately 0.2) compared to that of the laser scanner (magnitude (jitter) ranging from 0.2 to 0.6). The Kinect is therefore better suited to get accurate velocity estimates (both magnitude and direction) while the laser scanner is used for position estimates in the area.

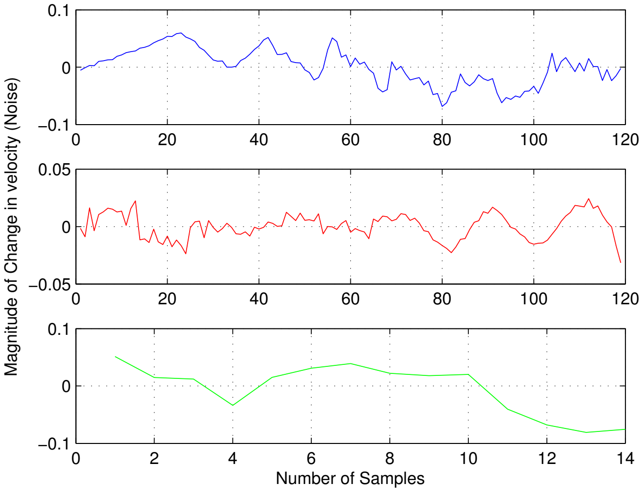


Figure 4: *The noise profile of the velocities from the Kinect shows a lot less variation in magnitude (jitter).*

5.1 User Extraction

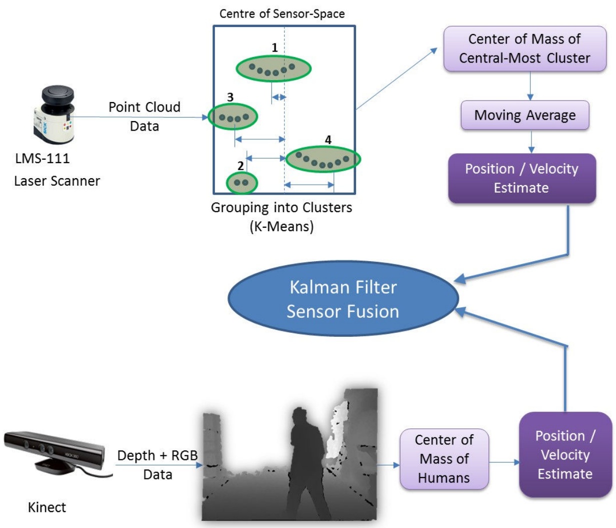

Centre of Sensor-Space
1
3
4
2
Point Cloud Data
LMS-111 Laser Scanner
Grouping into Clusters (K-Means)
Center of Mass of Central-Most Cluster
Moving Average
Position / Velocity Estimate
Kalman Filter Sensor Fusion
Depth + RGB Data
Kinect
Center of Mass of Humans
Position / Velocity Estimate

Figure 5: *Overview of processing steps for user position and velocity tracking.*

When the system is first launched, a series of calibration algorithms are used to ensure that all projector images are aligned during every rendered frame and the user's movements are accurately mapped in the simulation. This is a one-step process and does not need to be repeated unless the physical setup is disturbed.

The K-Means clustering algorithm is used to group the point cloud data from the laser scanner, following which the cluster that is most central to the laser scanner is passed onto the next processing block. A moving average filter is then implemented on the mean of the point cloud subset that was passed into this processing block. Combined with a tracking solution, this has the potential to be used for multiple users in the future. The scanner itself is adjusted to be slightly below shoulder height of a user, since this 2D plane returns the least scattered point cloud. A calibration routine is used to map the raw polar coordinates of the laser data to image-space pixel coordinates, so that the user's position directly corresponds to the position of the virtual object that she controls.

The Kinect simplifies the extraction of user positions, as humans are automatically segmented by the device. Their center of mass is extracted and a moving average filter analogous to the laser scanner is employed to calculate velocity information.

A graphical overview of the user extraction steps for the two range sensors is shown in Figure 5. This figure also shows a sensor fusion module that will be detailed in the following section.

5.2 Sensor Fusion

Owing to the differences in range and noise characteristics of the Kinect and the laser scanner, using either of them in an independent configuration results in performance drawbacks. A better solution is to combine the information obtained from both the sensors to derive better measurements for use in the mixed reality environment. The nature of the data obtained from the laser sensor (point cloud) means that reliably computing the center of gravity of a user depends on the number of data points acquired. As a person gets further away from the sensor, the number of obtained data points decreases naturally, owing to the angular resolution (0.5°) of the laser scan. In addition, if a person rotates in the play area so that his shoulders are perpendicular to the scanner, the number of acquired data points drops. This causes subtle variations in estimated position that may not be very noticeable, but critically affect the accuracy of the derived velocity vector.

The Kinect on the other hand is free from post-processing steps, since it directly returns the center of mass from the skeleton of the user. However, it is incapable of tracking the user outside its $4.5\,m$ range, with noise increasing with depth. This leaves us with a minimized area to work with for the immersive experience when solely relying on the Kinect.

To overcome these problems, we follow an approach based on Kalman filtering to 'fuse' the information gathered from both the sensors and improve our position and velocity estimates over the entire area. Under the assumption that all state variables are perturbed by zero-mean normal-distributed noise, the Kalman filter is an optimal recursive estimator for the state variables of linear dynamical systems:

$$\mathbf{x_k} = \mathbf{A}\mathbf{x_{k-1}} + \mathbf{w_k}$$
$$\mathbf{z_k} = \mathbf{H}\mathbf{x_k} + \mathbf{v_k}$$

The state vector $\mathbf{x} = [p_x, p_y, v_x, v_y]^\intercal$ consists of the user's 2D position (p_x, p_y) and his velocity vector (v_x, v_y). The state transition matrix A relates the state estimate from the last time step to a new (a priori) estimate. For our problem, the matrix A is the identity matrix with dt replacing elements $(1,3)$ and $(2,4)$ (row-first) of a 4x4 identity matrix.

Here dt denotes the time (in seconds) since the last filter update and will change dynamically due to the potentially different sensor update rates.

The process noise can be described by $\mathbf{w} \sim N(0, \mathbf{Q})$. Here, $\mathbf{w}$ is assumed to be white, zero-mean Gaussian noise with variance matrix $\mathbf{Q} = [0.01, 0.01, 0.05, 0.05]\ I$. The high confidence in the process model that is reflected in $\mathbf{Q}$ is justified because at the high update rate of the sensors the user's movement can be approximated by the linearization implicit in $\mathbf{A}$.

In its measurement stage, the Kalman filter will incorporate readings from a sensor and use the comparison to the a priori state estimate to calculate a better a posteriori estimate. In our case all state variables are directly observable (see Section 5.1 and 7) so the measurement matrix $\mathbf{H}$ is simply the identity matrix.

The measurement noise $\mathbf{v} \sim N(0, \mathbf{R})$ describes the uncertainty of the obtained measurements and is usually inherent in the used sensors. $\mathbf{v}$ has similar characteristics to $\mathbf{w}$ and its variance $\mathbf{R}$ is modeled as follows:

$$\mathbf{R} = \mathrm{diag}\left(f(p_y), f(p_y), g(p_y), g(p_y)\right) \tag{3}$$

The functions $f(p_y)$ and $g(p_y)$ relate the frontal distance of the user to the expected accuracy of the used sensor. Since these functions will be different dependent on the used sensor, the matrix $\mathbf{R}$ is also modified dependent on which sensor measurement is currently integrated in the Kalman Filter.

Although no comprehensive sensor characterization of the LMS 111 laser scanner exists, the closely related model LMS 200 was thoroughly investigated in [Ye and Borenstein 2002]. For the laser scanner the functions $f(p_y)$ and $g(p_y)$ take the following form:

$$f(p_y) = 0.0036 \times p_y$$
$$g(p_y) = v^{(L)} f(p_y) \tag{4}$$

The Kinect was recently characterized in [Khoshelham and Elberink 2012] and the following relationship between user depth p_y and error variance was found:

$$f(p_y) = 0.00285 \times p_y^2$$
$$g(p_y) = v^{(K)} f(p_y) \tag{5}$$

Both $v^{(K)}$ and $v^{(L)}$ are constants that are application-dependent and should be determined heuristically.

We decided to implement two separate Kalman filters, one for the Kinect's data and one for the laser scanner data. Through our experiments, we found the tracking accuracy of the Kinect to be more than adequate if the user is closer than $4\,m$ to the sensor. Conversely, the laser scanner is the only sensor that can resolve the user's position past $5\,m$ distance.

Denote with $\mathbf{x}^{(\mathbf{L})}$ and $\mathbf{x}^{(\mathbf{K})}$ the current state estimates of the laser scanner and Kinect Kalman filters, respectively.

The combined state estimate can then be composed of a linear combination of these filter outputs:

$$\mathbf{x} = \alpha\, \mathbf{x}^{(\mathbf{L})} + \beta\, \mathbf{x}^{(\mathbf{K})} \tag{6}$$

Here α and β determine the relative weighting of the Kinect versus the laser scanner and the functions ensure a smooth transition when the Kinect reaches its measurement limits. The functions are defined as complementary logistic functions:

$$\alpha(p_y) = \frac{1}{1 + \exp(-s(p_y - 4))}$$
$$\beta(p_y) = -\alpha(p_y) + 1 \tag{7}$$

The Kinect Kalman filter will be heavily favored up to about a range of $4\,m$, whereas the laser scanner will gradually receive a higher weight with increasing measurement distance. The variable s determines how quickly the transition between the Kalman filters occurs.

With this system in place, the user position and velocity can be tracked reliably throughout the whole interaction area. The recursive nature of the Kalman filter allows for the sensor fusion algorithm to run on-line while the system is operating.

6 Results

In this section we present the performance of user tracking with the developed sensor fusion approach.

Section 5 quantified some of the noise characteristics of the Kinect and the laser scanner and developed a sensor fusion solution to handle the integration of both modalities into one tracking framework. In this subsection, we would like to present experimental results that confirm that our sensor fusion algorithm provides better positional as well as velocity estimation compared to what any one sensor can achieve.

For the velocity calculations in the user extraction stage, we chose a moving window size of 12 samples for both sensors. The slope value in Equation (7) is initialized with $s = 10$ to provide a quick transition between the sensor modalities when the Kinect reaches its detection limit. The velocity variance multipliers in Equation (4) and (5) are set to $v^{(L)} = 1000.0$ and $v^{(K)} = 5.0$. The high value for $v^{(L)}$ is owed to the fact that the noise characteristics of the laser-based velocity estimation are inferior to the Kinect.

For experimental validation we recorded the movements of a user in the interaction area. Movement speeds and patterns varied considerable in the data set. Under these conditions, Figure 6 shows some snapshots of the tracking performance for the laser scanner, the Kinect, and our sensor fusion solution. Figure 6(a) displays a spatial plot of the area where the Kinect loses tracking (on the y-axis between 4 and $4.5\,m$). When the user approaches this area, the tracking relies mostly on the Kinect Kalman filter, but when the Kinect performance degrades, the sensor fusion algorithm smoothly integrates the laser scanner state estimation. Notice the continuous "Fused Sensor Data" plot, while the dashed Kinect data shows significant outliers. This behavior leaves the user unaware of the switching of sensor modalities and does not degrade the quality of the interface.

Figures 6(b) and (c) show plots of the estimated velocity magnitudes. Similarly to the position data, the Kinect estimation stops at about $8.5\,s$, but the sensor fusion module achieves a smooth handover to the Kalman filter for the laser scanner. Subsequent velocity estimates are based purely on laser tracking data, but still show superior noise characteristics compared to the unprocessed laser velocity shown with a dash-dotted line.

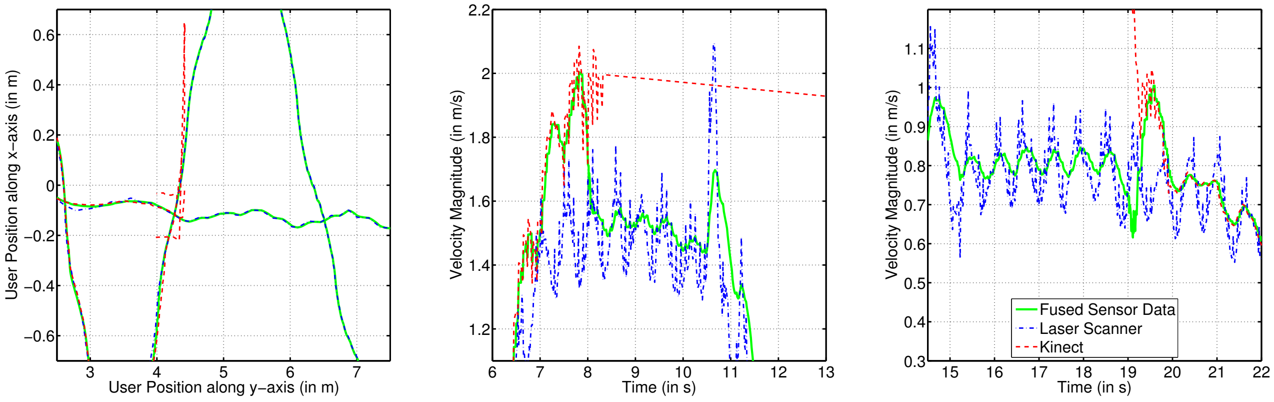
The Kinect is reacquired after $19\,s$ and quickly dominates the velocity estimates.

7 Sample Implementation

We believe that the MR configuration described above has the potential to deliver potent learning experiences for a wide range of content areas. In our first implementation, which we call *MEteor*, we focused on planetary astronomy and a few select principles of physics typically introduced to students in middle school. Rather than having learners experience these principles as external observers - as they are typically taught in school - we created a simulation that allowed the child to become part of the planetary system and gave them an insider's perspective on their operation. Specifically, the participant was assigned the role of an asteroid. The participant's job was to launch the asteroid such that it hit a target or achieved some other objective such as getting into a stable orbit around a planet.

The metaphor of participant-as-asteroid was established at the outset of the game experience with instructions to "become the asteroid". When the user steps onto the asteroid at the center of the interactive platform there is an electrifying visual and audio effect that signals the transformation. For the rest of the simulation game the asteroid either followed the participant's movements (pre-launch) or the participant was expected to predict the movement of the asteroid as it interacted with planets and other celestial objects (post-launch). The metaphor was reinforced with a scoring system that rewarded users who stayed with their asteroid as it curved around a planet or orbited at varying speeds. Both real-time feedback in the form of color indicators and after-action review in the form of graphs displayed via a wall projection were used to improve the users' performance across multiple trials. The overall objective was for the users to develop intuitions about how things move in space by using their bodies to enact these trajectories, and to make connections between these intuitions and the formal tools (e.g., graphs) that are typically used to scientifically represent this movement.

The launched asteroid is imparted with the velocity of the user in the launch zone. To achieve this, a moving average filter is used to compute the instantaneous velocity of the user at every time instant. The launch zone can be thought of as a rectangular constraint in the



(a) (b) (c)

Figure 6: *Estimation performance of the presented sensor fusion system for user position in (a) and velocity magnitude in (b) and (c). The sensor fusion system successfully mediates multiple transitions from Kinect to laser scanner tracking. Compared to the single sensor solutions, it also provides superior noise characteristics for the velocity magnitude estimation. Please refer to Section 6 for more details.*

field of view of the sensor. Velocity components are used together with a filtering algorithm to compute a magnitude and direction for the velocity vector of the asteroid being launched in the play area. This velocity data from the sensor is then transformed via a mapping function into universal coordinates, i.e. the physics engine uses accurate values of masses, velocities and positions of the planetary objects to evolve the system over time. For instance, a simulated Earth orbit takes exactly 365.25 days for one complete revolution around the sun. Iterative computation of gravitational influence between multiple planetary objects is used to determine forces and accelerations within the system; an inverse function then re-projects these computed values into image-space to create a physically realistic simulated experience of the planetary system on the floor.

We successfully completed a round of testing with approximately 120 middle school students in our lab and the MEteor game is currently deployed at the Museum of Science and industry in Tampa, FL where we are able to test learning and immersion in a more authentic informal learning environment.

8 Conclusion and Future Work

In this work, we have described a large-scale mixed reality system for use in educational simulations through a paradigm called 'metaphor-based learning'. The development of the prototype was guided by five design principles for an embodied learning environment that were detailed in Section 1. Subsequently, the system itself was deployed at the Museum of Science and Industry in Tampa to facilitate learning in an informal environment.

Our unique contributions include:

- Incorporating requirements of embodied learning and whole-body metaphors into technical design of a mixed reality system.
- Using a low-cost projection system to provide a large interaction surface while minimizing shadows in the display area.
- Developing a novel sensor fusion technique of a laser scanner and a Kinect to exploit multiple sensing modalities.

Multiple Kinects placed strategically may allow the extraction of skeletal data in the whole interaction area, and coordinate system transforms between the laser and the Kinect(s) could allow us to uniquely track an individual's action in the MR space. We plan to extend this system to support multiple users by augmenting the sensor fusion algorithm.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant DRL-1114621. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agency.

We recognize the entire MEteor team that contributed to the work described in this paper. In particular we would like to thank Charles Hughes and J. Michael Moshell for their contributions to the MR technology and conceptual design of the experience.

References

Birchfield, D., and Johnson-Glenberg, M. 2010. A Next Gen Interface for Embodied Learning: SMALLab and the Geological Layer Cake. International Journal of Gaming and Computer-Mediated Simulations 2, 1, 49–58.

Birchfield, D., Ciufo, T., and Minyard, G. 2006. SMALLab: A Mediated Platform for Education. In ACM SIGGRAPH - Educators Program, ACM Press, New York, New York, USA, 1–7.

Cameron, L. 2002. Metaphors in the Learning of Science : A Discourse Focus. British Educational Research Journal 28, 5, 673–688.

Chang, C.-W., Lee, J.-H., Wang, C.-Y., and Chen, G.-D. 2010. Improving the authentic learning experience by integrating robots into the mixed-reality environment. Computers & Education 55, 4 (Dec.), 1572–1578.

Chen, H., Sukthankar, R., and Wallace, G. 2002. Scalable alignment of large-format multi-projector displays using camera homography trees. In IEEE Visualization (VIS2002), IEEE, 339–346.

CHRISTIDOU, V., KOULAIDIS, V., AND CHRISTIDIS, T. 1997. Children's Use of Metaphors in Relation to their Mental Models: The Case of the Ozone Layer and its Depletion. Research in Science Education 27, 4, 541–552.

COTTING, D., NEBEL, I., GROSS, M. H., AND FUCHS, H. 2011. Towards a Continuous, Unified Calibration of Projectors and Cameras. Tech. rep., ETH Zuerich, Department of Computer Science.

CRUZ-NEIRA, C., SANDIN, D. J., AND DEFANTI, T. A. 1993. Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. In ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). 135–142.

GALLAGHER, S. 2005. How the Body Shapes the Mind. Oxford University Press, USA.

GESTURETEK, 2012. Gesture Recognition & Computer Vision Control Technology & Motion Sensing Systems for Presentation & Entertainment.

GOLDIN-MEADOW, S., COOK, S. W., AND MITCHELL, Z. A. 2009. Gesturing Gives Children new Ideas about Math. Psychological Science 20, 3 (Mar.), 267–272.

HARTLEY, R., AND ZISSERMAN, A. 2003. Multiple View Geometry, 2nd ed. Cambridge University Press.

HONEY, M. A., AND HILTON, M. L., Eds. 2011. Learning Science Through Computer Games and Simulations. The National Academies Press.

HUGHES, C., STAPLETON, C., HUGHES, D., AND SMITH, E. 2005. Mixed Reality in Education, Entertainment, and Training. IEEE Computer Graphics and Applications 25, 6 (Nov.), 24–30.

JOHNSON-GLENBERG, M., KOZIUPA, T., BIRCHFIELD, D., AND LI, K. 2011. Games for Learning in Embodied Mixed-Reality Environments: Principles and Results. In International Conference on Games + Learning + Society Conference (GLS), 129–137.

KHOO, E. T., CHEOK, A. D., NGUYEN, T. H. D., AND PAN, Z. 2008. Age Invaders: Social and Physical Inter-Generational Mixed Reality Family Entertainment. Virtual Reality 12, 1 (Mar.), 3–16.

KHOSHELHAM, K., AND ELBERINK, S. O. 2012. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. Sensors 12, 2 (Feb.), 1437–1454.

KIRKLEY, S. E., AND KIRKLEY, J. R. 2002. Creating Next Generation Blended Learning Environments Using Mixed Reality , Video Games and Simulations. TechTrends 49, 3, 42–54.

LINDGREN, R., AND MOSHELL, J. M. 2011. Supporting Children's Learning with Body-Based Metaphors in a Mixed Reality Environment. In International Conference on Interaction Design and Children (IDC), ACM Press, New York, New York, USA, 177–180.

NAGASE, M., IWAI, D., AND SATO, K. 2011. Dynamic Defocus and Occlusion Compensation of Projected Imagery by Model-Based Optimal Projector Selection in Multi-Projection Environment. Virtual Reality 15, 2-3 (Aug.), 119–132.

NATIONAL SCIENCE BOARD. 2010. Preparing the Next Generation of STEM Innovators: Identifying and Developing our Nations Human Capital. Tech. rep., National Science Foundation.

PAN, Z., CHEOK, A. D., YANG, H., ZHU, J., AND SHI, J. 2006. Virtual Reality and Mixed Reality for Virtual Learning Environments. Computers & Graphics 30, 1 (Feb.), 20–28.

RASKAR, R., BROWN, M., WELCH, G., TOWLES, H., SCALES, B., AND FUCHS, H. 1999. Multi-projector displays using camera-based registration. In IEEE Visualization (VIS1999), IEEE, 161–522.

RASKAR, R., BAAR, J. V., AND CHAI, J. 2002. A low-cost projector mosaic with fast registration. In Asian Conference on Computer Vision (ACCV), 114–119.

SAJADI, B., AND MAJUMDER, A. 2012. Autocalibration of Multiprojector CAVE-Like Immersive Environments. IEEE Transactions on Visualization and Computer Graphics 18, 3 (Mar.), 381–393.

SNIBBE, S. S., AND RAFFLE, H. S. 2009. Social Immersive Media: Pursuing Best Practices for Multi-User Interactive Camera/Projector Exhibits. In International Conference on Human Factors in Computing Systems (CHI), 1447–1456.

YE, C., AND BORENSTEIN, J. 2002. Characterization of a 2D laser scanner for mobile robot obstacle negotiation. In International Conference on Robotics and Automation (ICRA), IEEE, Washington, DC, 2512–2518.