Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2009, Article ID 362651, 17 pages
doi:10.1155/2009/362651

Research Article

A Dynamic Bayesian Approach to Computational Laban Shape Quality Analysis

Dilip Swaminathan, Harvey Thornburg, Jessica Mumford, Stjepan Rajko, Jodi James, Todd Ingalls, Ellen Campana, Gang Qian, Pavithra Sampath, and Bo Peng

Arts, Media, and Engineering Program, Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287, USA

Correspondence should be addressed to Dilip Swaminathan, dilip@asu.edu

Received 11 December 2007; Revised 7 September 2008; Accepted 13 January 2009

Recommended by Daniel Ashbrook

Laban movement analysis (LMA) is a systematic framework for describing all forms of human movement and has been widely applied across animation, biomedicine, dance, and kinesiology. LMA (especially Effort/Shape) emphasizes how internal feelings and intentions govern the patterning of movement throughout the whole body. As we argue, a complex understanding of intention via LMA is necessary for human-computer interaction to become *embodied* in ways that resemble interaction in the physical world. We thus introduce a novel, flexible Bayesian fusion approach for identifying LMA Shape qualities from raw motion capture data in real time. The method uses a dynamic Bayesian network (DBN) to fuse movement features across the body and across time and as we discuss can be readily adapted for low-cost video. It has delivered excellent performance in preliminary studies comprising improvisatory movements. Our approach has been incorporated in *Response*, a mixed-reality environment where users interact via natural, full-body human movement and enhance their bodily-kinesthetic awareness through immersive sound and light feedback, with applications to kinesiology training, Parkinson's patient rehabilitation, interactive dance, and many other areas.

1. Introduction

Recently, much attention has been given to making human-computer interaction (HCI) more "natural;" that is, more similar to everyday human interaction situated in the physical world [1]. In particular, there is increased emphasis on *multimodal* interaction; that is, responding via multisensory feedback to speech, facial expression, bodily gesture, pen movement, and so forth [2–4]. However, few of these developments (if any) address *embodiment*, an essential attribute of everyday human interaction [5–10]. Embodiment directly challenges the traditional (Cartesian dualist) paradigm which posits that humans interact with the environment via separate and sequential stages known as *perception* (forming a mental model of the environment based on sensory input), *cognition* (planning bodily actions based on this mental model), and *action* (executing these actions). The Cartesian view considers mind and body as separate, interfacing only through the "mental model" constructed during perception. By contrast, embodiment posits that perception, cognition, and action are *situated* in the context of everyday activity and are in fact closely intertwined. That is, mind and body continuously interact and cannot be separated in the context of any given activity.

Consider, for instance, the situation where one is thirsty and reaches for a glass of water. The Cartesian paradigm suggests that one initially perceives and constructs a mental model of the glass, and subsequently plans actions based on this model: (1) reach out, (2) open the hand, (3) grasp the glass, and (4) bring it up to the mouth. Only after the model is constructed and the corresponding actions are planned does one actually pick up the glass. However, embodied cognition suggests a much more integrated role for cognition. Motivated by an overall *activity schema* (grasp the glass), one (a) perceives the relation between glass and hand, (b) plans to bring the hand closer to the glass and changes the hand configuration to fit the circumference of the glass, and (c) executes this plan, which in turn alters the perceived relationship between glass and hand. The role of cognition is reduced from planning complex action

sequences to planning simple adjustments that bring the perceived activity closer to the desired goal or schema.

An analogy can be made to the difference between *closed loop* and *open loop* control systems as shown in Figure 1. The goal of a *controller* (such as a thermostat) is to supply the necessary input(s) to achieve the desired output(s) (Figure 1(a)), for example, provide the appropriate heating and cooling inputs to maintain a room at 75°F. This task can be greatly simplified when the error between actual and desired outputs is used to control the input to the heating/cooling system as shown in the *closed loop* configuration of Figure 1(c). If it is too hot the system will turn on the air conditioner; if it is too cool the system will activate the furnace. This rule is much simpler than guessing the sequence of heating/cooling cycles that will keep the temperature at 75°. Moreover, it is well known that feedback is *necessary* to minimize the total squared output tracking error for a fixed energy input, according to the solution of the linear quadratic regulator (LQR) problem [11]. That is, feedback obtains not only a simpler controller but also one that is *optimal* in terms of a generally accepted measure of performance. Analogously, cognition in an embodied interaction framework (as shown in Figure 1(d)) is likely not only to be less complex, but also more effective than cognition in a framework where mind-body separation is enforced. For instance, if the environment undergoes a sudden change (such as the glass tipping over as one tries to grasp it), one can make the necessary adjustments without having to revise the entire action plan. Furthermore, recent neurological evidence has also emerged to support the theory that human motor control largely does follow a servomechanical (i.e., closed-loop) configuration [12, 13].

Unfortunately, traditional HCI (by this we mean the mouse-keyboard-screen or desktop computing paradigm) is quite limited in terms of the range of user actions the interface can understand. These limitations can affect embodied interaction as follows. Instead of users working out their intentions to act in the process of perceiving and acting, they must translate these intentions into emblematic actions—mouse clicks and key presses - prior to acting upon them. This translation process involves cognitive planning in a sense that is divorced from perception and action, breaking the embodied interaction loop. Moreover, a number of researchers have focused on the dynamic, emergent nature of *interaction context* (i.e., the shared conceptual framework that enables user and system to meaningfully interpret each other's actions; cf. [8, 14–16]) as a fundamental consequence of embodied interaction [5, 8, 10]. However, if the user is forced to communicate through specific, emblematic actions, context is by definition fixed by the system, as the user must focus on conforming his/her actions to what he/she knows the system can understand.

Hence, to foster embodied interaction, we need interfaces that can develop a complex, meaningful understanding of intention—both kinesthetic and emotional—as it emerges through natural human movement. It has been well understood in the movement science literature that intention in human movement has a full-body basis; that is, intention is rooted not in the movements of individual limbs and joints, but in the way these movements are patterned and connected across the whole body [17–20]. In the past two decades, a number of mixed-reality systems have been developed which incorporate full-body movement sensing technologies. These systems have been widely applied in diverse areas including exploratory data analysis [21], rehabilitation [22–24], arts and culture [25, 26], gaming [27–29], and education [22, 30, 31]. Movement sensing embedded in these systems largely consists of the following types: 1) recognition of specific, emblematic gestures or static poses [22, 26, 32–34], or 2) extraction of low-level kinematic features (body positions and joint angles) [27, 28, 35]. Unfortunately, these sensing methodologies fall short of supporting embodied interaction unless augmented with a higher-level analysis of intention. Systems that respond only to emblematic gestures or poses retain the aforementioned problems associated with translation, cognitive planning, and system-centered context. Systems that focus only on low-level kinematic features (a system that uses the left elbow height to control the pitch of a sound, the right shoulder joint angle to control its amplitude, etc.) still fail to account for how movement is patterned and connected across the body. Consequently, we must design interfaces based on full-body movement sensing which address the role of intention in the patterning and connection of full-body movement.

To this end, we adopt the framework of Laban movement analysis (LMA), a system developed by Rudolf Laban in the 1920s for understanding and codifying all forms of human movement in an intention-based framework [19, 20]. LMA has not only served as a notational system for expressive movement in dance, it has been widely applied over the past 80 years in kinesiology, developmental psychology, CGI animation, and many other areas. While LMA has some limitations especially in its ability to describe the specific neurological processes underlying motor control, it is nevertheless still finding new applications even in clinical areas such as improving function and expression in patients with Parkinson's disease [36], to better understand social interactions in children with autism [37], to investigate the neuronal development of reach-to-grasp behaviors [38], and to design animated characters with more expressive and believable movement characteristics [39, 40]. LMA is broadly divided among the following categories: *Body*, *Space*, *Effort*, and *Shape*. Analysis of *Body* primarily involves determining body part usage and phrasing (unitary, simultaneous, successive, sequential), but also looks at how the body patterns itself in movement (head-tail, upper-lower, body-half, etc.). *Space* organizes and clarifies the body and its actions by establishing a clear pathway or goal for movement. It concentrates on the size, approach to and use of one's *kinesphere*, or personal space as well as defines a clear spatial matrix around the body. *Effort* primarily addresses the expressive content or style of one's movement. Effort qualities manifest themselves through *space* (not to be confused with the Space category), *time*, *flow*, and *weight* and usually appear in combinations called *states* or *drives*. *Shape*, in general, elicits the form, or forming of the body. One subcomponent of *Shape*, Shape qualities, concerns itself with how the body changes its shape

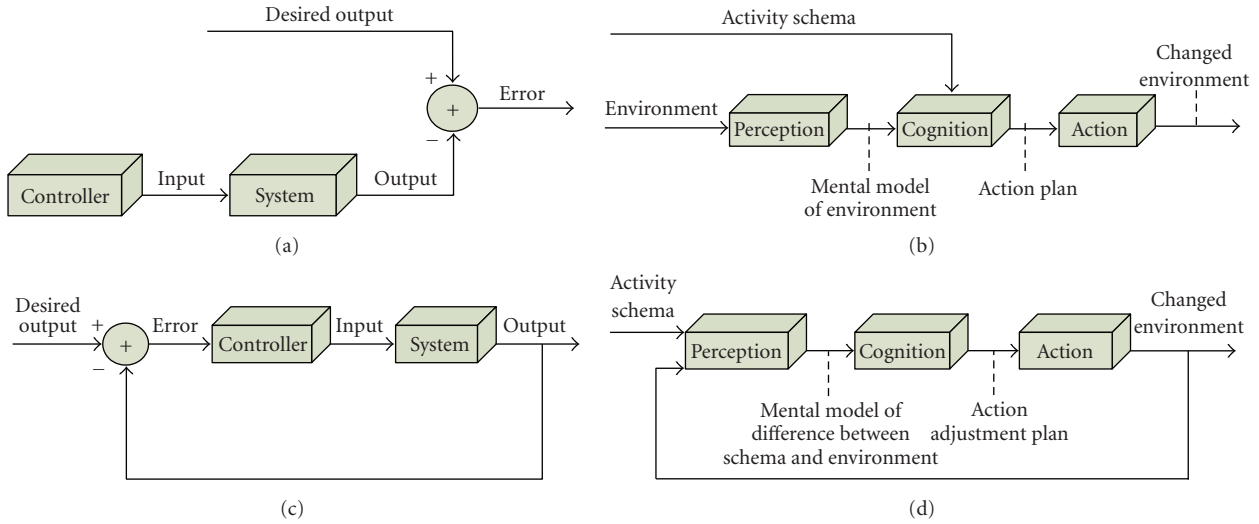

Figure 1: Analogy between closed loop control and embodied interaction.

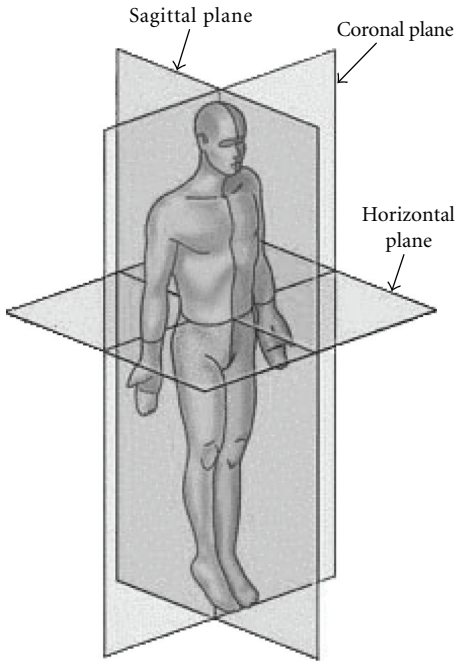

Figure 2: Body-centered coordinate system showing the different body planes.

in a particular direction. Figure 2, adapted from [41], shows a body-centered coordinate system with horizontal, sagittal, and coronal planes. *Rising*/*sinking* fall perpendicular to the horizontal plane; *retreating*/*advancing* fall perpendicular to the coronal plane, and *enclosing*/*spreading* describe motion about the sagittal plane as well as reveal the general folding and unfolding of the body. All movement are comprised of one, two or three Shape qualities depending on the complexity of the movement itself; however one quality usually dominates and can be said to characterize the movement.

1.1. System Goals. Currently, we have focused most of our efforts on Shape quality (SQ) extraction. While it may not seem so to a human observer, doing computational SQ analysis is quite difficult because there is no single, consistent way one can express a particular quality. One may advance, for instance, by walking toward something, pointing, or by craning one's neck forward in a slight and subtle way. Nevertheless, different SQs do imply different *tendencies* in the movement of individual joints and limbs within the context established by the body-centered coordinate system shown in Figure 2. For instance, if someone is *rising*, it is more likely that their torso will rise than sink. Similarly, SQs may imply nonlocal tendencies, such as an upward shift of the body's enter of mass with respect to the horizontal plane. We hence treat SQ inference as a large-scale *information fusion* problem, in which many different tendencies combine to give a holistic picture of the overall movement. Our method is *extensible*; if new sources of information enter, they can be readily incorporated to improve the accuracy of our SQ inference, without having to redesign the method or collect large amounts of data. New information sources can include *new sensing modalities*, for instance, hand-held tangible interface objects [42] or pressure-sensitive floors [43], as well as *higher-level contextual information* such as a person's tendency to emphasize one axis (e.g., *rising/sinking*) in the context of a particular activity. Similarly, if an information source no longer becomes reliable (due to a sensor fault), the system can just ignore the features that depend on this information. Performance will be slightly lessened since there is less information available, but the result with our method will not be catastrophic.

To date, there has been little overall work on computational SQ analysis let alone the kind of extensible, fault-tolerant method we propose. Probably the most extensive, detailed modeling of SQ can be found in the EMOTE system [44], and subsequent efforts [45, 46]. EMOTE introduces computational models for both Shape and Effort, but for movement synthesis (for character animation) rather than

analysis. It remains unclear how EMOTE can be adapted for analysis. Neural network models have been applied for Effort analysis [45–47], and it may be possible to redevelop these models for Shape, although we are presently unaware of such an attempt. However, these neural network-based approaches can only make "hard decisions" regarding the presence or absence of a particular movement quality. This is inadequate for embodied interaction frameworks where continuous changes in the nature of the movement must be coupled to continuous changes in the media feedback. We solve this issue by adopting a Bayesian approach yielding at each time instant, a posterior distribution over all qualities that indicates for each quality the degree of certainty or strength that the quality is present in the movement. Hence the fact of a quality becoming more certain can be easily detected as the posterior concentrates more and more over that quality. Also, the methods proposed in [45–47] seem completely "data-driven," and therefore cannot be readily extended to incorporate new contextual information or sensing modalities without a costly retraining process involving new data sources.

Our method utilizes a dynamic Bayesian network (DBN) to jointly decide the dominant SQ based on raw marker location data from a motion capture system. We output a posterior distribution over dominant SQ/motion segment hypotheses given all sense-data observations from the motion capture system. If information sources are correctly modeled via appropriate conditional probability distributions, marginalizing and then maximizing this posterior with respect to the dominant SQ hypothesis will yield error-optimal decisions [48–50]. However, the raw SQ posterior reveals much about the salience or ambiguity of the qualities expressed in the movement, which would be lost if the system simply makes a decision. That is, if one perfectly isolates a particular SQ in one's movement, the posterior will concentrate completely on that SQ. On the other hand, if one's movement is more ambiguous with respect to SQ, this ambiguity will be reflected in a posterior that is spread over multiple SQs. The *Response* environment, a mixed-reality system aimed at fostering bodily-kinesthetic awareness through multisensory (audio/visual feedback) which incorporates our SQ inference engine and makes extensive use of the dominant SQ posterior, as the concentration of this posterior implicitly reflects a degree of dominance [51].

The remainder of this article is organized as follows. Section 2.1 gives an overview and block diagram of our proposed SQ extraction method encompassing feature selection, probabilistic modeling of feature dynamics, feature fusion via whole-body context to infer the dominant SQ and a description of the *Response* environment. Section 2.2 discusses feature selection and computations, Section 2.3 discusses temporal dynamics modeling of individual features, Section 2.4 presents the full-body fusion model, and Section 2.5 describes the computation of dominant SQ posteriors from raw feature data using this model. Section 3 presents a preliminary study involving a range of movement examples, from highly stylized movements to movements which are more complex and unstructured. The performance of our SQ inference (when the dominant SQ posterior is thresholded according to the error-optimal *maximum a posteriori* (MAP) rule) is quite promising, and has been successfully embedded in the *Response* environment [51] as previously discussed.

2. Proposed Method

2.1. System Overview. The overall method including marker-based optical motion capture, probabilistic motion analysis and multimodal feedback provided by the *Response* environment for interaction is diagrammed in Figure 3. Raw data observations consist of 3D position data from 34 labeled *markers*, which are soft, IR-reflective spheres attached at various positions to one's body via Velcro straps (Figures 4 and 5). Marker positions and labelings are updated every 10 milliseconds using an eight-camera IR motion capture system supported by custom software (EvART) developed by Motion Analysis Corporation [52]. In practice the system sometimes has difficulty tracking all of the markers, so occasionally markers will be reported as missing or the labeling of one marker will be switched with that of another. From this marker set we first compute the body-centered coordinate system consisting of the navel origin and the orientations of the horizontal, coronal, and sagittal planes (Figure 6). Next, we compute a set of features, called *subindicators*, which describe the movements of individual body sections as well as global body movement characteristics with respect to this coordinate system. Subindicator features are designed so that consistent positive/negative changes are highly indicative of one pair of SQs at least for that particular body section (e.g., the right arm is rising/sinking, the torso is advancing/retreating, etc.) Finally, we apply a novel dynamic Bayesian network (DBN) which models (a) the segmental continuity of subindicator features given subindicator (individual body-section) SQ hypotheses, and (b) the temporal dynamics of subindicator SQ hypotheses given the dominant SQ hypothesis. The full-body fusion arises implicitly in the latter part of the DBN, as described in Section 2.3. The output of the computational SQ analysis is a posterior probability distribution of the dominant SQ which drives the interactions provided by the *Response* environment. *Response* leverages the system's capacity for embodied interaction in the following sense: rather than attempting to create very complex movement-feedback mappings, these mappings develop organically through certain natural affinities between feedback and movement.

The *Response* environment consists of two submodules, which we call *pulsar* and *glisson*. The pulsar submodule uses SQ analysis and a measure of overall activity [51] to alter parameters of a bank of pulsar synthesis generators [53]. Pulsar synthesis is a method of sound synthesis based upon the generation of trains of sonic particles. We map the posterior probability of the current SQ hypothesis to various parameters. *Advancing/retreating* and *spreading/enclosing* control the range of the fundamental frequency of overlapping pulsarets, with *advancing/retreating* controlling the lower bound and *spreading/enclosing* the higher bound of a uniform random frequency generator. *Rising/sinking* affect the duty cycle of the pulsar generators, causing wide modulations in formant

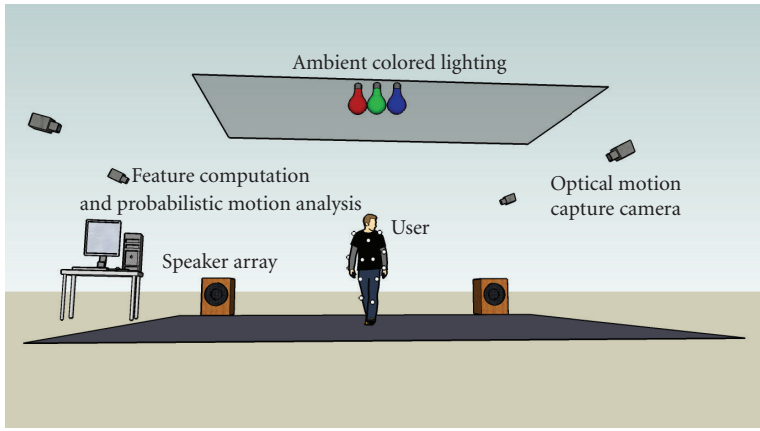

Figure 3: System overview schematic diagram.

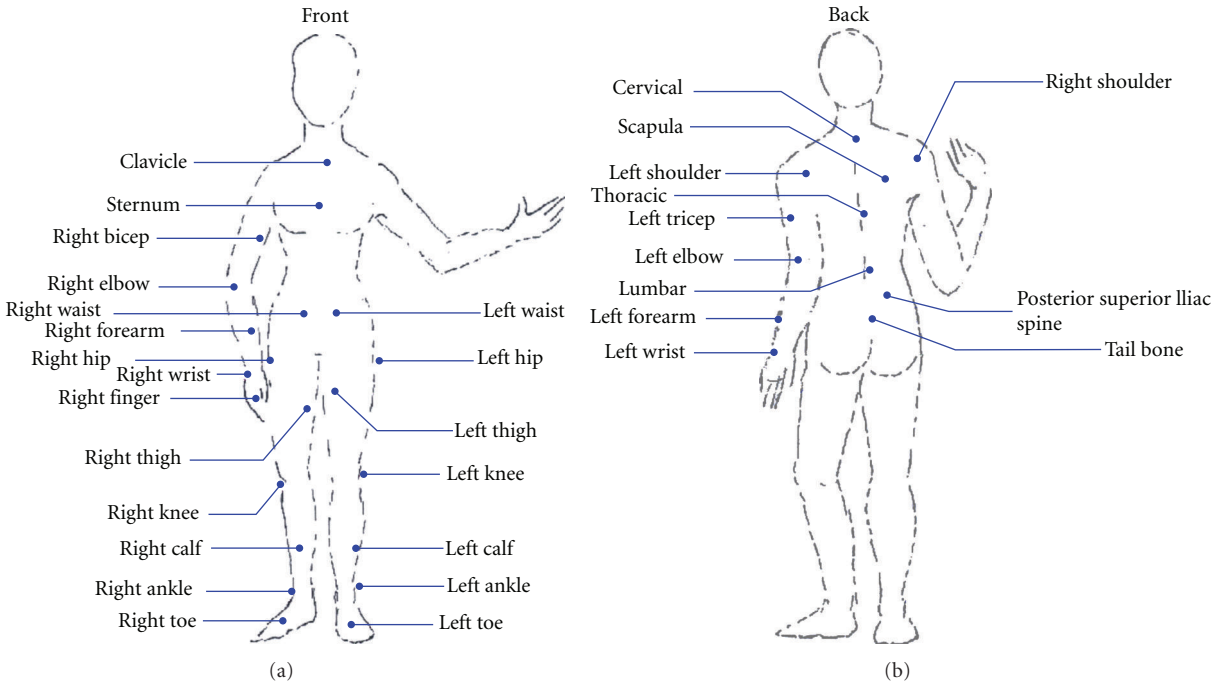

Figure 4: Marker set schematic showing front and back sides of the body.

frequencies. Activity level is also mapped directly to overall amplitude of the bank of pulsar generators and the intensity of specific color of lights (blue). Additionally, the sound feedback is spatialized so as to be located in surrounding speakers where the activity is being sensed. The glisson submodule uses a bank of glisson particle generators [53]. Glissons are short grains of sound that have a frequency trajectory (or glissando) with very short time frames of the grain. Depending on grain length, the affect can be anywhere from clicks to chirps to large masses of shifting sounds. In this case the glissons are shorter (20–90 milliseconds). The trajectory of the glissons is mapped to the *rising/sinking* probability of the SQ analysis. *Rising* movement causes individual glissons to rise in frequency and *sinking* has the opposite affect. *Advancing* increases the durations of the glissons while retreating lowers them. In the submodule white light is mapped to the activity of the user. These two submodules, experienced in alternation, encourage the participants to focus on the experience of movement, rather than on how the system works, and to explore new creative possibilities through their movement. We now proceed to describe in detail the computation of the body-centered coordinate system and the subindicator features in the next section.

2.2. Feature Extraction. Our goal is to extract features from raw marker position data for which *changes* in these features are highly indicative of the SQs (rising/sinking, advancing/retreating, enclosing/spreading). As a first step we obtain the body-centered coordinate system, specifically the

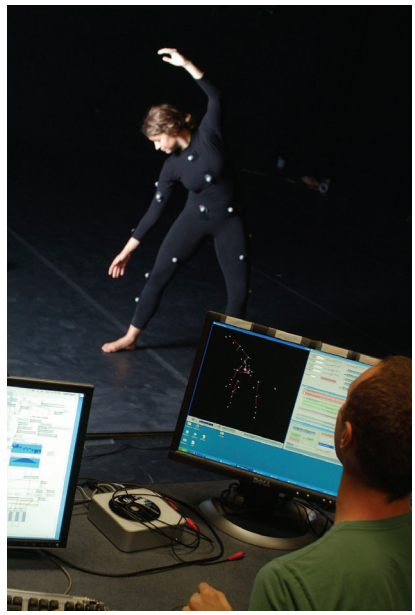

Figure 5: Image of a user interacting in the *Response* environment, which incorporates our SQ analysis framework.

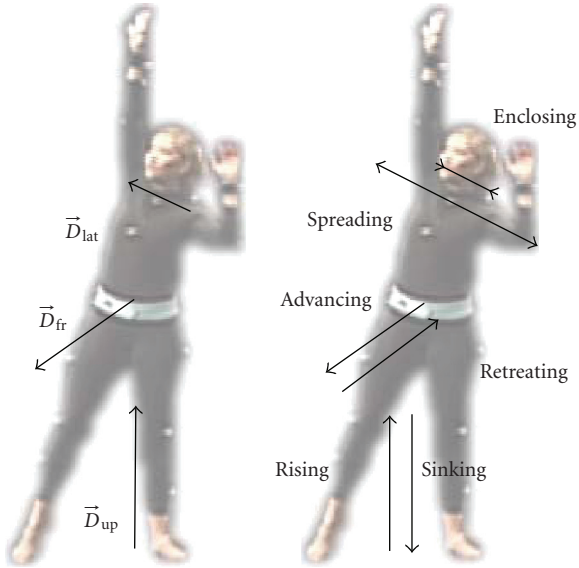

Figure 6: Directional vectors used in feature calculation, and Shape qualities corresponding to each vector. *Spreading/enclosing* is expected to occur along the lateral ($\vec{D}_{\text{lat}}$) direction, *advancing/retreating* along the front ($\vec{D}_{\text{fr}}$) direction, and *rising/sinking* over the up ($\vec{D}_{\text{up}}$) direction.

orientations of the horizontal, coronal, and sagittal planes shown in Figure 2. Let us define three vectors: $\vec{D}_{\text{up}}$ points upward, perpendicular to the horizontal plane; $\vec{D}_{\text{lat}}$ points left to right, perpendicular to the sagittal plane; $\vec{D}_{\text{fr}}$, the *front direction*, points back to front, perpendicular to the coronal plane. These directional vectors are illustrated in Figure 6. The *up direction* ($\vec{D}_{\text{up}}$) is provided by the motion capture system, and points straight up (perpendicular to the floor). We choose the y axis of the motion capture coordinate system as the up directon, that is, $\vec{D}_{\text{up}} = (0, 1, 0)$.

The *lateral direction* ($\vec{D}_{\text{lat}}$) is defined with the tail in the left shoulder, and points in the direction of the right shoulder. If M_{ls} is the marker position of the left shoulder, and M_{rs} is the marker position of the right shoulder, then $\vec{D}_{\text{lat}} = (M_{\text{rs}} - M_{\text{ls}})/\|M_{\text{rs}} - M_{\text{ls}}\|$.

Finally, the *front direction* ($\vec{D}_{\text{fr}}$) is determined by the person's attitude toward the surrounding space, and is calculated from his/her movement. The *front direction* is necessary in determining the extent to which the person is *advancing* or *retreating*. Specifically, movement in the front direction is interpreted as *advancing*, and movement against it is interpreted as *retreating*. Rather than using a front direction that is fixed, we allow the person to establish (and change) the front direction through his/her movement.

In simple circumstances, for example, if the person's entire body is facing a specific direction for a substantial length of time, the front direction can be determined from the facing direction of the pelvis. In more complicated situations, the front direction can stay the same even though the pelvis is rotating. An example is the advancing of a discus thrower. In this case, the athlete's attitude toward the space is such that he or she is *advancing* toward a specific point in space, where the discus will be launched forward. Even though the entire body is spinning, the front direction stays the same.

The front direction is first initialized to the facing direction of the pelvis in the horizontal plane, $\vec{D}_{\text{pel}}$ (which is a unit vector calculated from the positions of the markers at the left waist, right waist, and the cervical). From that point on, we calculate the front direction as a weighted mean of the previous front direction and the current facing direction of the pelvis.

In particular, let M_p and M'_p be the positions of the pelvis at the current and previous frame, respectively (these are approximated by taking the mean of markers placed at the left and right sides of the waist). The horizontal pelvis movement across these two frames is then given by $\Delta M_p = (M_p - M'_p) \cdot (1, 0, 1)$. Then, we compute $\vec{D}_{\text{fr}} = c \cdot \vec{D'}_{\text{fr}} + (1 - c) \cdot (\Delta M_p / |\Delta M_p|)$, where $\vec{D}'_{\text{fr}}$ is the previous front direction and c is given by

$$c = \begin{cases} \min\{1, s_{\text{fore}}(\Delta M_p \cdot \vec{D}_{\text{pel}})\}, & \Delta M_p \cdot \vec{D}_{\text{pel}} \geq 0, \\ \min\{1, -s_{\text{back}}(\Delta M_p \cdot \vec{D}_{\text{pel}})\}, & \Delta M_p \cdot \vec{D}_{\text{pel}} < 0. \end{cases} \quad (1)$$

The constants s_{fore} and s_{back} specify how much movement of the pelvis either forward or backward (with respect to itself) influences the front direction. In our experiments, we used $s_{\text{fore}} = 4\, s_{\text{back}}$, that is, forward pelvis motion was considered 4 times as indicative of the front direction than backward pelvis motion. The exact values depend on the frame rate.

Using these coordinate vectors, we obtain the following features.

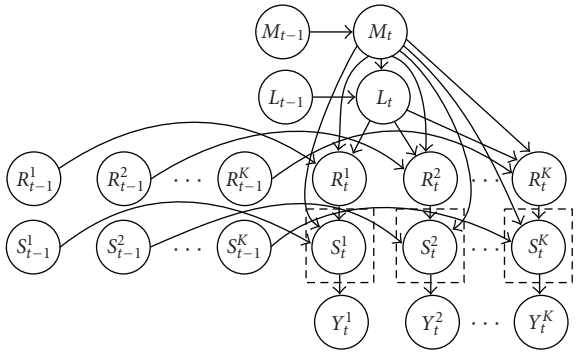

Figure 7: Single time slice of the DAG corresponding to the overall dominant Shape quality inference model.

(i) *Mean marker height* describes the global body position along $\vec{D}_{\text{up}}$. Specifically, we compute the mean of all marker positions and project this mean onto $\vec{D}_{\text{up}}$. A positive change in mean marker height indicates *rising*, while a negative change indicates *sinking*.

(ii) *Right/left elbow height* is the projection of the corresponding elbow marker onto $\vec{D}_{\text{up}}$. Changes in either feature indicate rising/sinking, since people often do so with their upper body which includes the arms. Up/down arm movements are usually coordinated with similar movements of the elbow.

(iii) *Accumulated frontward shift* is the running sum of the change in mean marker position as projected onto $\vec{D}_{\text{fr}}$. A positive change in accumulated frontward shift indicates *advancing*, while a negative change indicates *retreating*.

(iv) *Lateral marker variance* is the magnitude variance of all marker positions as projected onto $\vec{D}_{\text{lat}}$ (perpendicular to $\vec{D}_{\text{fr}}$). That is, we first project all marker positions onto $\vec{D}_{\text{lat}}$, compute their covariance matrix, and take the trace of this matrix. A positive change in lateral marker variance indicates *enclosing/spreading*, while a negative change indicates *advancing/retreating*.

To reduce the effect of noise in the marker positions, as well as marker mislabeling and occlusion, each feature is partially denoised using a second-order Savitzky-Golay filter [54] over a window of 0.2 seconds (20 frames at 100 fps). For each frame t, we denote the feature vector as $Y_t^{1:5}$ where the individual feature correspondences are as follows: Y_t^1—mean marker height, Y_t^2—right elbow height, Y_t^3—left elbow height, Y_t^4—accumulated frontward shift, Y_t^5—lateral marker variance. Given these feature vectors, we specify how we model the feature dynamics and how it is influenced by the dominant SQ probabilistically in the following section.

2.3. Probabilistic Model. Let the dominant SQ hypothesis at frame t be L_t, and

$$L_t \in \{Ri, Si, Ad, Re, Sp, En, Ne\} \tag{2}$$

corresponding, respectively, to *rising*, *sinking*, *advancing*, *retreating*, *spreading*, *enclosing*, and *neutral*.

We model the influence of L_t on raw feature observations Y_t^i, $i \in 1:5$ using a DBN for which a single time slice of the corresponding directed acyclic graph (DAG) is shown in Figure 7. We describe intermediate variables as follows.

(i) $M_t \in \{0, 1\}$ provides an overall segmentation of the full-body gesture. Where $M_t = 1$, the user begins a new gesture; where $M_t = 0$, the user is continuing to perform the same gesture. It is possible, but not necessary, that the dominant SQ changes when $M_t = 1$. For instance, a person can be rising with his/her torso and head, and during this motion decide also to lift up his/her left arm. When the arm first begins to lift $M_t = 1$; however, the dominant SQ does not change.

(ii) The *subindicator* R_t^i, for $i \in 1:5$, encodes the extent to which the *inherent* feature corresponding to Y_t^i is increasing, decreasing, or neutral. (The inherent feature is hidden; each Y_t^i is at best a noisy observation of the feature). We encode $R_t^i \in \{-1, 0, 1\}$, where $R_t^i = 1$ corresponds to increasing, $R_t^i = -1$ to decreasing, and $R_t^i = 0$ to neutral (neither increasing nor decreasing to any significant degree.) For instance, $R_t^2 = 1$ indicates that the right elbow is rising. The dominant SQ L_t induces tendencies on each of the subindicators. From the definition of M_t, we prohibit R_t^i from changing unless $M_t = 1$.

(iii) S_t^i, the *subindicator state*, is a vector containing the inherent feature X_t^i, of which Y_t^i equals X_t^i corrupted by noise, plus additional, auxiliary variables necessary to describe the influence of M_t and R_t^i on X_t^i as a first-order Markov dependence, $P(S_t^i \mid S_{t-1}^i, R_t^i, M_t)$. Further details are given in Section 2.4.

To summarize, the joint distribution corresponding to the DAG in Figure 7 admits the following factorization:

$$\begin{aligned} &P(M_{1:T}, L_{1:T}, R_{1:T}^{1:K}, S_{1:T}^{1:K}, Y_{1:T}^{1:K}) \\ &\quad = P(M_1)P(L_1 \mid M_1) \\ &\qquad \times \prod_{i=1}^{K} P(R_1^i \mid L_1, M_1)P(S_1^i \mid R_1^i, M_1)P(Y_1^i \mid S_1^i) \\ &\qquad \times \prod_{t=2}^{T} P(M_t \mid M_{t-1})P(L_t \mid L_{t-1}, M_t) \\ &\qquad \times \prod_{i=1}^{K} P(R_t^i \mid R_{t-1}^i, L_t, M_t)P(S_t^i \mid S_{t-1}^i, R_t^i, M_t)P(Y_t^i \mid S_t^i). \end{aligned} \tag{3}$$

In the following section, we give explicit descriptions of the dependences in (3).

2.4. Distributional Specifications. We first describe the inherent subindicator feature dynamics as encoded via $P(S_t^i \mid S_{t-1}^i, R_t^i, M_t)$, coupled with the observation dependence $P(Y_t^i \mid S_t^i)$. As previously discussed, S_t contains X_t^i, the

inherent subindicator feature, for which Y_t^i is a "noisy" version:

$$Y_t^i \sim \mathcal{N}(X_t^i, \sigma_{Y^i}^2). \tag{4}$$

However, S_t^i contains additional information necessary to model the influence of R_t^i and M_t on its dynamics using a first-order Markov dependence. That is,

$$S_t^i = \text{vec}\{V_{0,t}^i, V_t^i, X_t^i\}, \tag{5}$$

where

(i) V_t^i is the inherent feature velocity; that is, rate of change in X_t^i;

(ii) $V_{0,t}^i > 0$ is a constant, *nominal feature speed* associated with the current gesture. Gestures can be slow or fast; during the current gesture, V_t^i varies smoothly ($V_t^i \approx V_{t-1}^i$) while $V_t^i \approx V_{0,t}^i$ if $R_t^i = 1$, $V_t^i \approx -V_{0,t}^i$ if $R_t^i = -1$, and $V_t^i \approx 0$ if $R_t^i = 0$. The nominal speed itself can vary, albeit slowly, throughout the gesture.

The full dependence, $P(S_t^i \mid S_{t-1}^i, R_t^i, M_t)$, factors according to the expanded, single-feature DAG as shown in Figure 8; that is,

$$P(S_t^i \mid S_{t-1}^i, R_t^i, M_t) = P(V_{0,t}^i \mid V_{0,t-1}^i, M_t) P(V_t^i \mid V_{t-1}^i, V_{0,t}^i, R_t^i) \times P(X_t^i \mid X_{t-1}^i, V_t^i), \tag{6}$$

where $P(X_t^i \mid X_{t-1}^i, V_t^i)$ concentrates deterministically on $X_t^i = X_{t-1}^i + V_t^i$. In specifying $P(V_t^i \mid V_{t-1}^i, V_{0,t}^i, R_t^i)$, we must simultaneously satisfy competing modeling assumptions regarding the proximity of V_t^i to V_{t-1}^i as well as to a suitable function of $V_{0,t}^i$. These assumptions can be resolved in the form of a conditional Ornstein-Uhlenbeck (OU) process:

$$P(V_t^i \mid V_{t-1}^i, V_{0,t}^i, R_t^i) = \mathcal{N}(\alpha V_{t-1}^i + (1-\alpha)\delta_t^i, \beta\sigma_{V^i}^2)\ . \tag{7}$$

In (7) $\beta \triangleq (1-\alpha)/(1+\alpha)$, and

$$\delta_t^i \triangleq \begin{cases} V_{0,t}^i, & R_t^i = 1, \\ 0, & R_t^i = 0, \\ -V_{0,t}^i, & R_t^i = -1. \end{cases} \tag{8}$$

Here α controls the degree which $V_t^i \approx V_{t-1}^i$ and σ_{V^i}, the variance of the process about δ_t^i, controls the assumption $V_t^i \approx \delta_t^i$. Since the OU process is mean-reverting [55], its use in modeling the trajectory V_t^i helps greatly in ensuring that small, rapid fluctuations in the subindicator features due to involuntary motions are registered as neutral, $R_t^i = 0$, rather than as rapid oscillations in the subindicators themselves. For example, someone performing wave-like motion using their arms is probably neither *rising* nor *sinking*, at least as far as intention is concerned. In this way, the OU process modeling goes a long way toward modeling the user's intention, as consistent with the overall LMA philosophy.

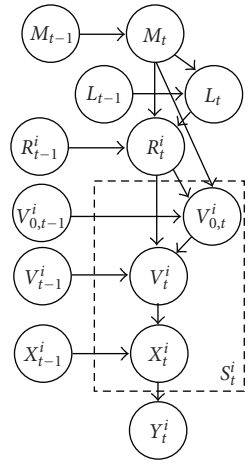

Figure 8: Expanded DAG for single feature displaying the factorization of $P(S_t^i \mid S_{t-1}^i, R_t^i, M_t)$ over the components of S_t^i.

The nominal feature speed $V_{0,t}^i$ is always positive and, if the corresponding subindicator is *active* (i.e., $R_t^i \neq 0$), is expected to drift slowly during a gesture, and reset upon the onset of a new gesture. Furthermore, concerning an analogy between gesture speed and musical tempo, we expect the drift in $V_{0,t}^i$ to be proportional to its value. Similar generative models for tempo variation are well known [50, 56, 57], among others. When the subindicator is *inactive* ($R_t^i = 0$), we note the notion of feature speed becomes meaningless and furthermore, via (6)–(8), $V_{0,t}^i$ does not directly influence other observations or states. Hence we model $V_{0,t}^i$ as always resetting to anticipate the onset of a new gesture. In summary, we model $P(V_{0,t}^i \mid V_{0,t-1}^i, M_t, R_t^i)$ as follows:

$$\log V_{0,t}^i \sim \begin{cases} \mathcal{N}(\log V_{0,t-1}^i, \sigma_{V_0^i}^2), & M_t = 0,\ R_t^i \neq 0, \\ \mathcal{N}(\log V_{00}^i, \epsilon^{-1}), & M_t = 1 \text{ or } R_t^i = 0, \end{cases} \tag{9}$$

where $\epsilon \ll 1$.

To obtain the remaining distributions in (3), we specify that the dominant Shape quality and each subindicator change only the onset of a new gesture; that is, if $M_t = 0$, then $L_t = L_{t-1}$ and $R_t^i = R_{t-1}^i$ with probability 1. When $M_t = 1$, L_t may change, but does not have to. A new gesture need not be caused by a change in dominant SQ. Let us first consider the modeling of $P(L_t \mid L_{t-1}, M_t = 1)$ in more detail. We model this dependence as a mixture of two distributions; one encoding the tendency that L_t remains consistent, the other, $P_0(L_t)$ specifying the stationary distribution after change:

$$P(L_t \mid L_{t-1}, M_t = 1) = \rho_L \delta_{\{L_t = L_{t-1}\}} + (1-\rho_L) P_0(L_t). \tag{10}$$

Indeed, as long as $\rho_L < 1$, a stationary distribution for L_t exists and equals $P_0(L_t)$. Lacking additional information, we model $P_0(L_t)$ as uniform.

Likewise, we model $P(R_t^i \mid R_{t-1}^i, L_t, M_t = 1) = P_0(R_t^i \mid L_t)$, where $P_0(R_t^i \mid L_t)$ is the corresponding stationary

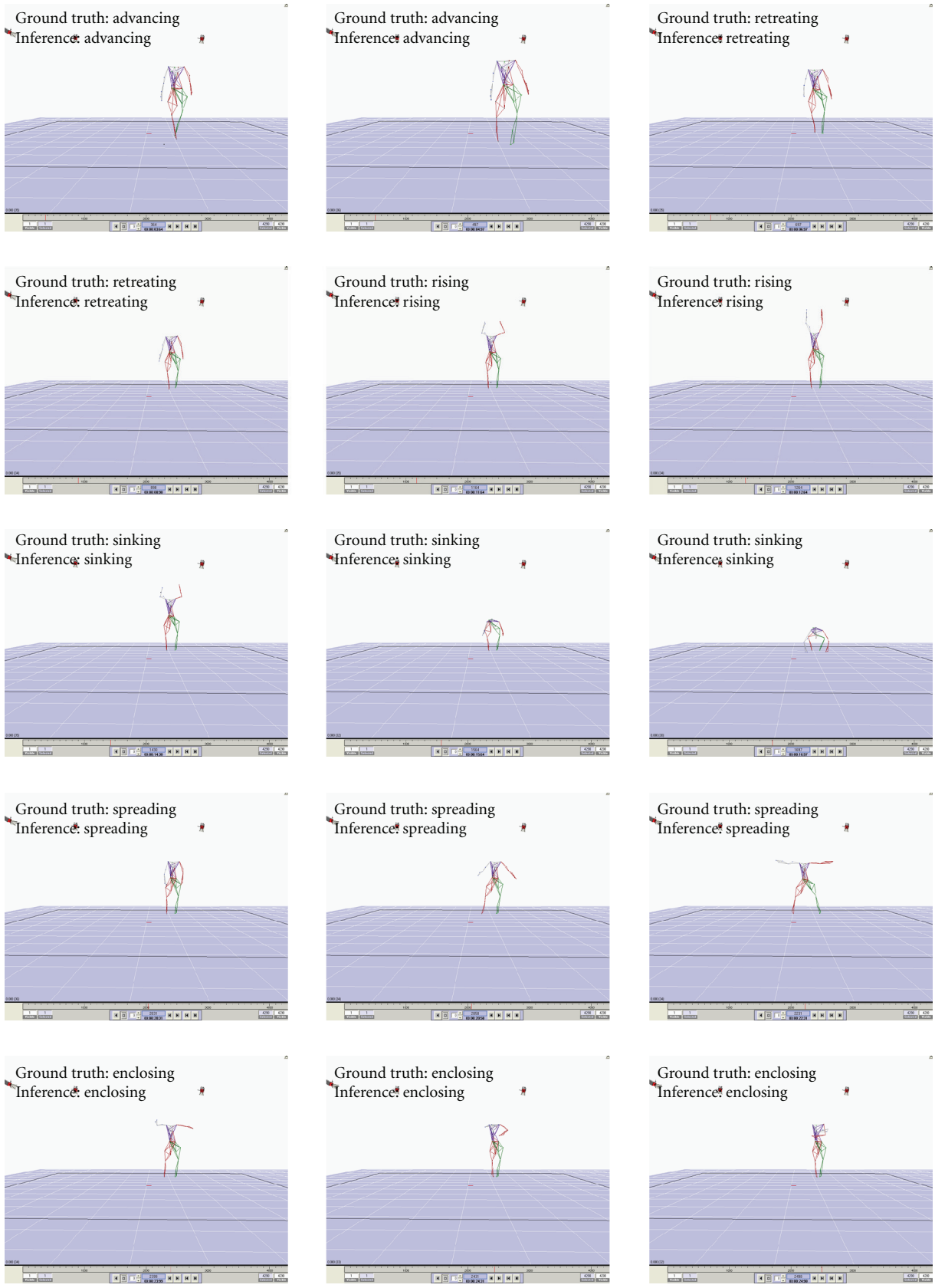

Figure 9: Image sequence of "Menu 1" (dancer 1) movement data showing the ground truth and inference results of the dominant Shape quality expressed.

Table 1: Probabilistic constraints of subindicator states given the dominant Shape quality for specifying $P_0(R_t^i \mid L_t)$.

Feature	$L_t = Ri$	$L_t = Si$	$L_t = Ad$
R_t^1, R_t^2, R_t^3	$p^- \ll 1,\ p^+ \gg p^-$	$p^+ \ll 1,\ p^- \gg p^+$	$p^- \ll 1,\ p^+ \ll 1$
R_t^4	$p^- \ll 1,\ p^+ \ll 1$	$p^+ \ll 1,\ p^- \ll 1$	$p^- \ll 1,\ p^+ \gg p^-$
R_t^5	$p^- \ll 1,\ p^+ \ll 1$	$p^+ \ll 1,\ p^- \ll 1$	$p^- \ll 1,\ p^+ \ll 1$
Feature	$L_t = Re$	$L_t = Sp$	$L_t = En$
R_t^1, R_t^2, R_t^3	$p^- \ll 1,\ p^+ \ll 1$	$p^- \ll 1,\ p^+ \ll 1$	$p^- \ll 1,\ p^+ \ll 1$
R_t^4	$p^+ \ll 1,\ p^- \gg p^+$	$p^- \ll 1,\ p^+ \ll 1$	$p^- \ll 1,\ p^+ \ll 1$
R_t^5	$p^+ \ll 1,\ p^- \ll 1$	$p^- \ll 1,\ p^+ \gg p^-$	$p^+ \ll 1,\ p^- \gg p^+$
Feature	$L_t = Ne$		
R_t^1, R_t^2, R_t^3	$p^- \ll 1,\ p^+ \ll 1$		
R_t^4	$p^- \ll 1,\ p^+ \ll 1$		
R_t^5	$p^- \ll 1,\ p^+ \ll 1$		

Table 2: Design of specifications for $P_0(R_t^i \mid L_t)$.

Feature	$L_t = Ri$	$L_t = Si$	$L_t = Ad$
R_t^1, R_t^2, R_t^3	$p^-0.15,\ p^+ = 0.75$	$p^+ = 0.15,\ p^- = 0.75$	$p^- = 0.1,\ p^+ = 0.1$
R_t^4	$p^- = 0.1,\ p^+ = 0.1$	$p^- = 0.1,\ p^+ = 0.1$	$p^-0.01,\ p^+ = 0.98$
R_t^5	$p^- = 0.1,\ p^+ = 0.1$	$p^+ = 0.1,\ p^- = 0.1$	$p^- = 0.1,\ p^+ = 0.1$
Feature	$L_t = Re$	$L_t = Sp$	$L_t = En$
R_t^1, R_t^2, R_t^3	$p^- = 0.1,\ p^+ = 0.1$	$p^- = 0.1,\ p^+ = 0.1$	$p^- = 0.1,\ p^+ = 0.1$
R_t^4	$p^+ = 0.01,\ p^- = 0.98$	$p^- = 0.1,\ p^+ = 0.1$	$p^- = 0.1,\ p^+ = 0.1$
R_t^5	$p^- = 0.1,\ p^+ = 0.1$	$p^- = 0.01,\ p^+ = 0.98$	$p^+ = 0.01,\ p^- = 0.98$
Feature	$L_t = Ne$		
R_t^1, R_t^2, R_t^3	$p^- = 0.01,\ p^+ = 0.01$		
R_t^4	$p^- = 0.01,\ p^+ = 0.01$		
R_t^5	$p^- = 0.01,\ p^+ = 0.01$		

distribution for R_t^i assuming L_t is constant. Essentially, $P_0(R_t^i \mid L_t)$ specifies how the subindicator features are influenced by the presence or the absence of a dominant SQ; that is, this distribution encodes the *full-body context* discussed in Section 1. For example, suppose $L_t = Ri$; that is, the dominant Shape quality is *rising*. Now, we do not expect the three associated subindicators; namely, R_t^1, R_t^2, and R_t^3 to always be positive, as this would mean whenever a person rises, he will always lift his arms. Rather, we expect merely that (a) it is unlikely that either R_t^1, R_t^2, or R_t^3 will be negative; and (b) it is much more likely that each will be positive than negative. Regarding the subindicators generally associated with other qualities; R_t^4, R_t^5, it will be improbable that either is positive or negative. A full set of constraints on $P_0(R_t^i \mid L_t)$ is shown in Table 1, where p^+ is shorthand for $P(R_t^i = 1 \mid L_t)$, and p^- represents $P(R_t^i = -1 \mid L_t)$. The complete specification of $P_0(R_t^i \mid L_t)$ is given via Table 2.

Finally, regarding $P(M_t \mid M_{t-1})$, we currently encode only the expectation that boundary events are sparse; that is, M_t is modeled as Poisson [58] with $P(M_t = 1) = p$, effectively severing the dependence of M_t on M_{t-1}. However, much human movement exhibits a rich temporal structure, for instance, rhythmic dance movements set to music. Hence we can use $P(M_t \mid M_{t-1})$ to encode this temporal structure, perhaps by also augmenting M_t to include additional states which encode the elapsed duration since the most recent boundary event. For instance, the temporal expectancy framework of [59] can be directly applied in this setting, and we plan to incorporate it in future work.

2.5. Inference Methodology. To decide the dominant Shape quality at time t, given observations $Y_{1:t}^{1:K}$, we first compute the filtered posterior $P(L_t \mid Y_{1:t})$ and choose L_t which maximizes this posterior. It is well known that this choice of L_t yields the minimum-error decision [48]. However, some hidden variables, for instance, M_t, L_t, and $R_t^{(1:K)}$, are discrete, and others, for instance, $V_{0,t}$ and V_t are continuous with first-order Markov dependences which depend on the discrete layer. The overall dynamic Bayesian network is in the form of a nonlinear, non-Gaussian switching state space model. Exact filtering in switching state-space models is exponential-time [60] and thus cannot be implemented in real time. Assuming conditional, linear Gaussian dependences at the continuous layer which we still do not have, a number of approximate filtering strategies: interacting multiple model (IMM) [61], second-order generalized pseudo-Bayes (GPB2) [61], and/or Rao-Blackwellized particle filter (RBPF) [62] become tractable. In our present model there are a large

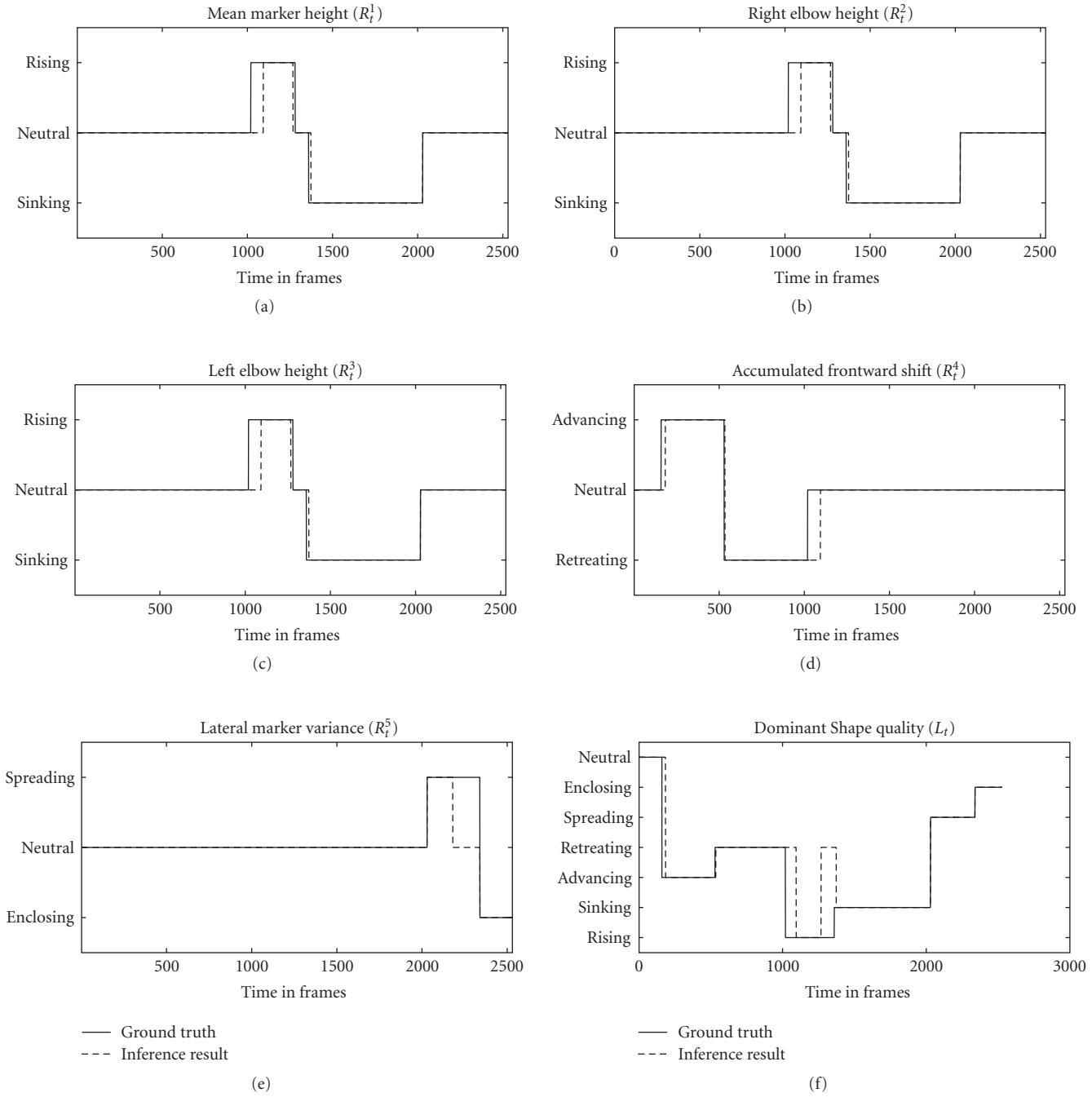

FIGURE 10: Segmentation performance on "Menu 1" (dancer 1) movement data showing the fusion of different features to infer the dominant Shape quality.

number of discrete states ($\#L_t \times \#M_t \times \#\prod_{i=1}^{5}[\#R_t^i] = 3402$) and thus only the RBPF, with 1000 particles over discrete states, functions appropriately for real-time inference. The nonlinearity and non-Gaussianity of the continuous state dynamics are handled with the RBPF framework by replacing the Kalman time update with the appropriate version of the unscented transform [63]. As the results in Section 3 show, our algorithm yields quite acceptable performance at 100 fps, with some latency due to the real-time nature of the decision process.

3. Experimental Results and Discussion

In order to test the capabilities of our dominant SQ inference we tested its performance on data collected from three dancers utilizing improvisation. The main reason to use trained dancers and focus on dance movements is that dancers' enhanced movement expertise and experience with choreography makes it much easier for certified Laban movement analysts to obtain the ground truth. In the context of dance, improvisation can be described as *free* movement

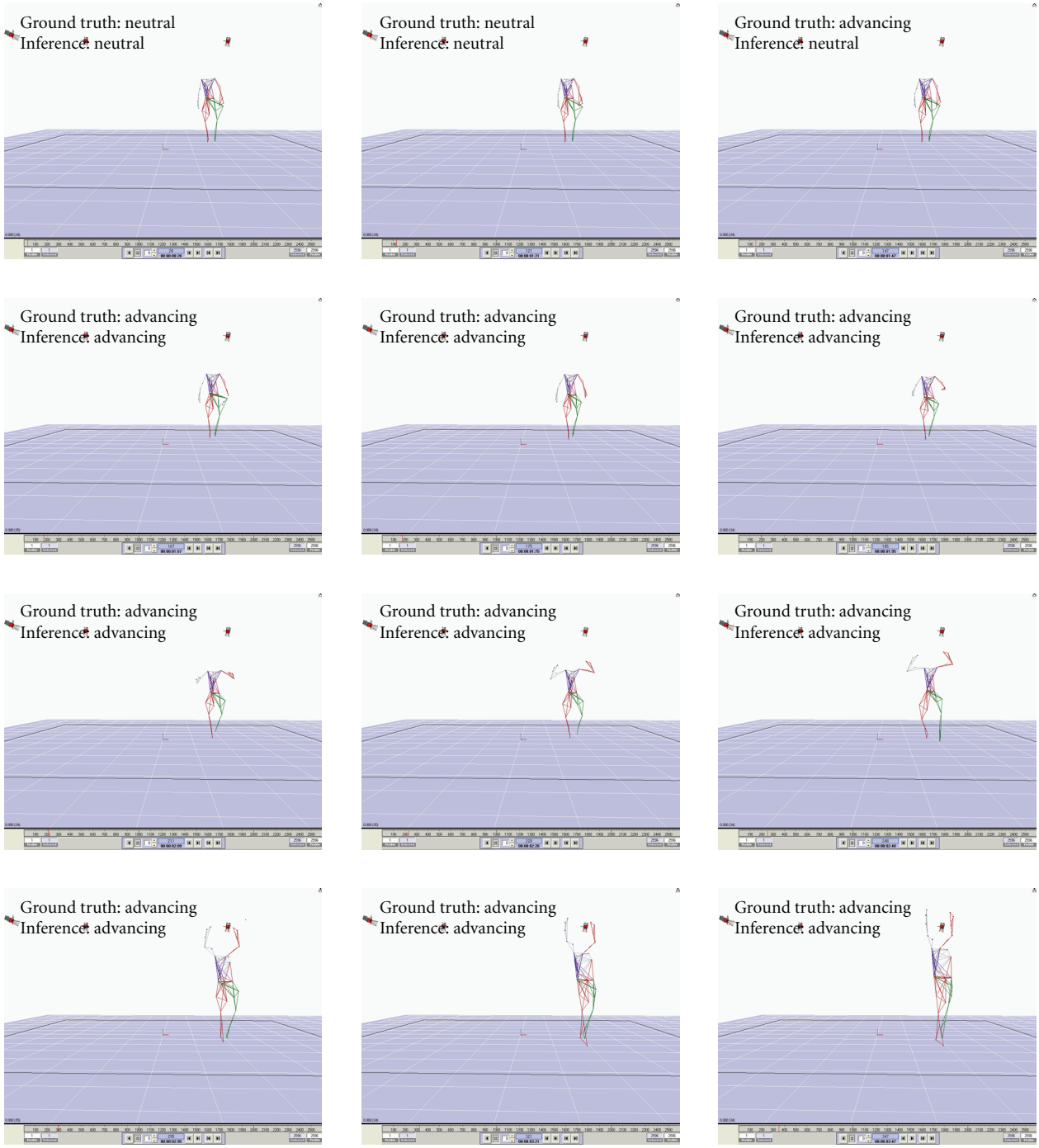

Figure 11: Image sequence of "Menu 3" (dancer 1) movement data showing the ground truth and inference results of the dominant Shape quality expressed between frames 1–350.

that is spontaneously created in the moment but often within certain guidelines. For the purposes of data collection and validation of our analyses, trained dancers performed a series of improvisatory movements following a set sequence. We call these sequences *improvisational menus*. In our case these menus consist of sequences of dominant SQs. For example, a menu might be *(rising → spreading → retreating → sinking)*, wherein the menu outlines the overall sequence of SQs, but gives no indication as to how or for what duration they should occur. This allows the dancer to explore how differently she can perform the same set of SQs through multiple repetitions of the menu. For our experimental analysis, each dancer performed four improvisational menus, of which two were simple menus (menus 1 and 2) and two were complex (menus 3 and 4). During the simple menus, the dancer attempted to perform movements expressing the individual

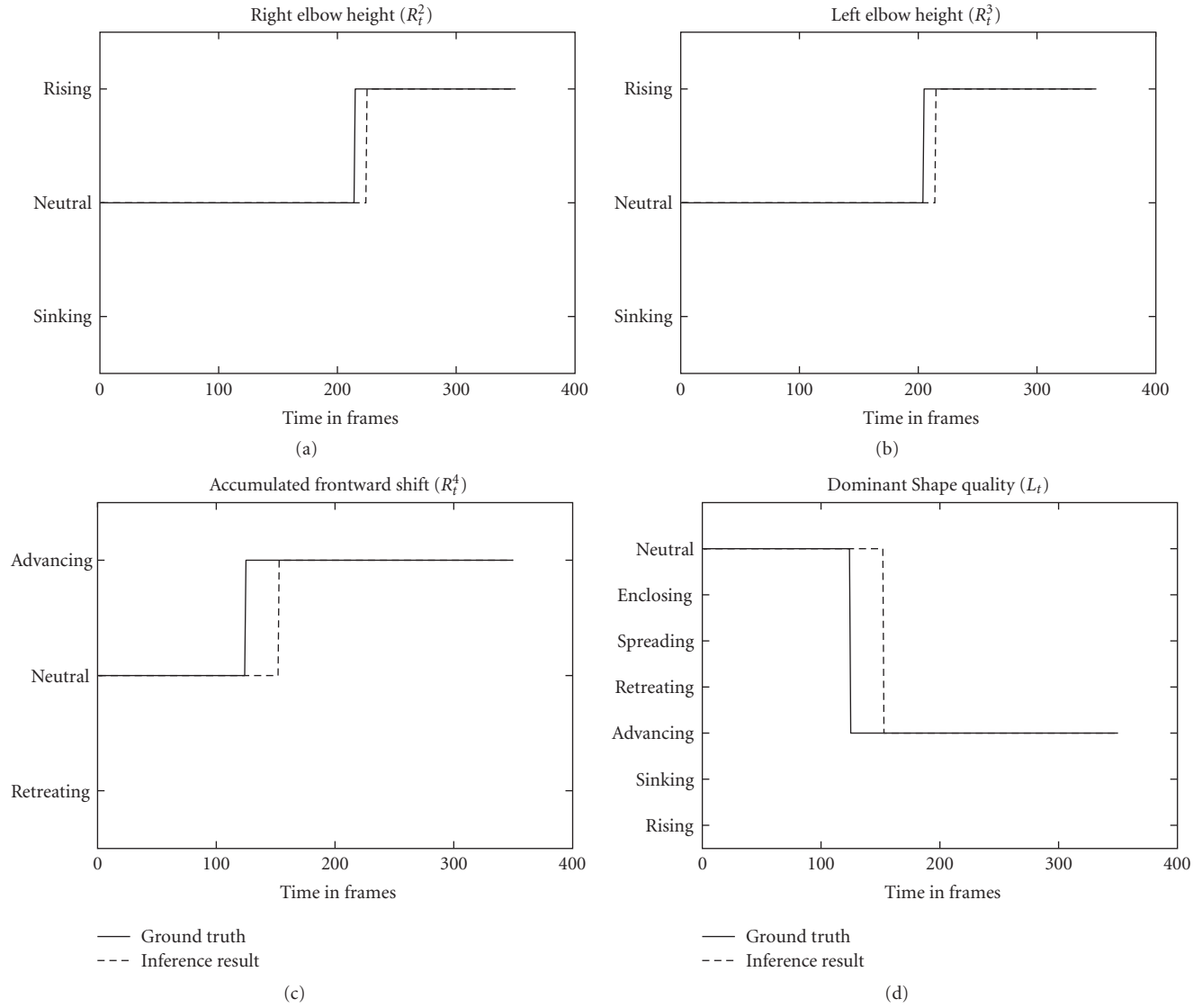

Figure 12: Dominant Shape quality segmentation performance on frames 1–350 of "Menu 3" (dancer 1) movement data shows that even though the person is advancing with the arms rising, our method correctly infers the dominant Shape quality as advancing.

Table 3: Dancer 1 segmentation results.

Data	% Recall	% Precision	Detection delay
Menu 1	100.0	85.7	0.1566 seconds
Menu 2	100.0	87.5	0.0733 seconds
Menu 3	100.0	70.0	0.2833 seconds
Menu 4	85.7	75.0	0.1916 seconds

Table 4: Dancer 2 segmentation results.

Data	% Recall	% Precision	Detection delay
Menu 1	100.0	85.7	0.0688 seconds
Menu 2	100.0	100.0	0.1433 seconds
Menu 3	83.33	72.72	0.2342 seconds
Menu 4	87.5	77.0	0.2071 seconds

Shape qualities listed on the menu without expressing other, less dominant Shape qualities. For the complex menus, the dancer focused her/his intent on articulating the listed Shape qualities as the most dominant, but allowed for other, less dominant Shape qualities to also be present. Segmentation of the ground truth was done by a certified Laban movement Analyst (Jodi James) watching the movement data offline.

Table 5: Dancer 3 segmentation results.

Data	% Recall	% Precision	Detection delay
Menu 1	100.0	100.0	0.1840 seconds
Menu 2	100.0	85.7	0.1216 seconds
Menu 3	83.33	75.0	0.2383 seconds
Menu 4	100	83.33	0.1250 seconds

Table 6: Confusion matrix.

L_t	*Ri*	*Si*	*Ad*	*Re*	*Sp*	*En*	*Ne*	Total	% Accuracy
Ri	1698	0	0	0	43	0	0	1741	97.5
Si	0	1167	103	0	0	0	9	1279	91.2
Ad	11	148	2634	0	6	0	0	2799	94.1
Re	0	20	9	4527	17	46	0	4619	98.0
Sp	279	0	0	0	1389	0	0	1668	83.2
En	0	160	0	0	0	1446	23	1629	88.7

Tables 3, 4, and 5 show the segmentation performance of our method on all four menus for each of the dancers. "% Recall" computes the percentage of times our method detected a segment of a Shape quality present in the ground truth. "% Precision" computes the ratio of number of segments that were correctly classified to the total number of segments that were detected. 'Detection delay' measures the average delay for our method to correctly detect the onset of a segment, by computing time difference between ground truth and the inference results.

We observe that our method performs excellently on menus 1 and 2 across all the dancers where the movement complexity is fairly simple, with very high average recall (100.0%) and precision (90.76%) rates. In the case of complex menus, namely menus 3 and 4, we observe an overall decrease in performance (89.97% recall and 75.5% precision). Having minimal detection delay is crucial in developing fully embodied multimedia interactions. We observe that our method performed reasonably well in all four menus for all the dancers, having low average detection delays (0.1689 seconds) and even the worst performance was 0.2833 seconds for menu 4 movement performed by dancer 1 which is still acceptable for providing real-time feedback in some situations.

However, there is a noticeable loss of performance on the complex menus. A possible reason for the decrease in precision and recall rates and increase in detection delay is that the dancer becomes more free to incorporate other less dominant SQs in his/her movement. This becomes particularly problematic in the case of *enclosing*/*spreading*. *Rising, sinking, advancing,* and *retreating* all relate to specific spatial directions (forward, backward, up, and down), which in turn helps us determine the dominant SQ comparatively easy. However, *spreading* and *enclosing* have a tendency to be directionally ambiguous because they are often more about folding or unfolding the body rather than moving the body along the horizontal axis. In this case *spreading* and *enclosing* were more difficult to detect because the dancer would usually associate these with other Shape qualities in the vertical or sagittal plane. For example, we observed that our method confuses *spreading* and *rising* with one another because the dancer would usually incorporate some amount of *rising* when she/he is *spreading*. The same affinitive relationship was also true for *enclosing* and *sinking*. The confusion matrix presented in Table 6 supports these hypotheses.

The confusion matrix shows the frame level dominant SQ estimation results comprising of all the movement menus of all the dancers. As discussed earlier, we observe very high estimation accuracy for *rising, sinking, advancing,* and *retreating* and a reduction in accuracy for *spreading* and *enclosing*. We also observe that majority of the errors in estimating *spreading* and *enclosing* were attributed to *rising* and *sinking* respectively. Hence in these circumstances it is particularly hard to identify the correct SQ as the person moving can intend to express a particular SQ but this can be difficult to analyze accurately from an outsider's perspective. Nevertheless, an overall average accuracy of 92.1% indicates that our dominant SQ inference is generally effective.

Figure 9 shows the image sequence and Figure 10 shows the subindicator and dominant SQ segmentation results on menu 1 data of dancer 1. In Figures 11 and 12, a specific example comprising the first 350 frames from menu 3 performed by dancer 1 is detailed. In this example we can see the strength of our fused subindicator approach which analyzes full body movements to infer the dominant SQ. In this particular movement sequence we observe that the dancer starts in a *neutral* state and begins to *advance* (Figure 12(c)) with her whole body while each of her arms begin versus *rising* (Figures 12(a) and 12(b)) at different instances of time. Our model was able to correctly segment the individual features of right elbow height and left elbow height as *rising* (Figures 12(a) and 12(b)) and the frontward marker placement as *advancing* (Figure 12(c)). In spite of the differences in feature level segmentation our model was able to correctly infer the dominant SQ as *advancing* (Figure 12(d)) even though both the arms were *rising*. This fusion of tendencies which sometime compete and other times reinforce each other across the whole body is extremely critical as in everyday human movement there is no prescribed way to express a given SQ.

4. Conclusions and Future Work

In this paper, we have described a novel method for Shape quality (SQ) inference as an integral part of the Laban movement analysis (LMA) framework. Our method performs quite well on preliminary studies using both simple and complex movement sequences, with, on average 94.9% recall, 83.13% precision, and 0.1689 seconds detection delay. As we established in Section 1, the LMA framework is essential toward developing a complex understanding of natural human movement at the level of intention. This understanding, in turn, is essential toward affording human-computer interactions that are embodied, similar to everyday human interactions situated in the physical world.

In embodied interaction, context is not fixed by the system but emerges dynamically through interaction.

Recently, we have begun to embed this real-time SQ analysis in a number of immersive multisensory environments, in which dominant SQ posteriors are tied directly to specific elements or parameters of an audiovisual feedback stream, such as the *Response* environment (Section 2.1) where the user can leverage his/her movement invention and creative play to build a personalized repertoire of creative expression. Additionally, *Response* demonstrates potential far beyond that of a movement-based creative tool. Techniques from this environment, particularly the embedded SQ analysis, can be applied as a training tool in sports for performance improvement or injury prevention, a rehabilitation tool for Parkinson's disease, and so forth. These domains are particularly well-suited to the techniques we have described because they require a general, yet detailed, real-time computational representation of movement, specifically movement that is meaningful to the user. Moreover, as in *Response*, they involve situations where the goal of the system is two-fold: (1) to allow users to focus on their own movements and (2) to encourage/discourage particular types of movements on the part of the user.

One critical challenge for further development is removing the dependence of our method on expensive, nonportable motion capture technology, and developing a video-based system based on a low-cost multiview framework. Recent work [64, 65] has shown much promise in terms of full-body kinematics recovery from video and we are rapidly expanding upon and improving this work. By applying skeleton building techniques, we can extract virtual marker positions and labelings from raw kinematic data by extending techniques presented in [66, 67]. Since obtaining these positions and labelings from 34 markers may still prove a quite challenging problem, we note that the marker set may be very much reduced especially if the body-centric coordinate system can be derived from raw multiview observations. While some issues, particularly the issue of a reduced marker set, remain unresolved, initial efforts toward developing a low-cost, portable multivideo framework appear quite promising.

Acknowledgments

This paper is based upon work partly supported by the National Science Foundation CISE Infrastructure and IGERT Grants nos. 0403428 and 0504647. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation (NSF). The authors would like to thank Tim Trumble for the photograph (Figure 5) used in this paper. They would also like to thank the reviewers for the valuable comments, which helped them improve the quality and presentation of this paper.

References

[1] N. Sebe, M. S. Lew, and T. S. Huang, "The state-of-the-art in human-computer interaction," in *Proceedings of the ECCV Workshop on Computer Vision in Human-Computer Interaction (HCI '04)*, vol. 3058 of *Lecture Notes in Computer Science*, pp. 1–6, Prague, Czech Republic, May 2004.

[2] S. Oviatt, P. Cohen, L. Wu, et al., "Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions," *Human-Computer Interaction*, vol. 15, no. 4, pp. 263–322, 2000.

[3] S. Oviatt, "Multimodal interfaces," in *Handbook of Human-Computer Interaction*, J. Jacko and A. Sears, Eds., pp. 286–304, Lawrence Erlbaum and Associates, Mahwah, NJ, USA, 2003.

[4] S. Oviatt, "Advances in robust multimodal interface design," *IEEE Computer Graphics and Applications*, vol. 23, no. 5, pp. 62–68, 2003.

[5] F. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambridge, Mass, USA, 1991.

[6] P. Dourish, *Where the Action Is: The Foundations of Embodied Interaction*, MIT Press, Cambridge, Mass, USA, 2001.

[7] M. Wilson, "Six views of embodied cognition," *Psychonomic Bulletin and Review*, vol. 9, no. 4, pp. 625–636, 2002.

[8] P. Dourish, "What we talk about when we talk about context," *Journal of Personal and Ubiquitous Computing*, vol. 8, no. 1, pp. 19–30, 2004.

[9] J. Rambusch and T. Ziemke, "The role of embodiment in situated learning," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pp. 1803–1808, Stresa, Italy, July 2005.

[10] H. Sundaram, "Participating in our multisensory world," Tech. Rep. AME-TR-2005-09, Arts, Media, and Engineering, Arizona State University, Tempe, Ariz, USA, March 2005.

[11] B. D. O. Anderson and J. Moore, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

[12] S. H. Scott, "Optimal feedback control and the neural basis of volitional motor control," *Nature Reviews Neuroscience*, vol. 5, no. 7, pp. 532–546, 2004.

[13] R. D. Seidler, D. C. Noll, and G. Thiers, "Feedforward and feedback processes in motor control," *NeuroImage*, vol. 22, no. 4, pp. 1775–1783, 2004.

[14] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Proceedings of the Workshop on Mobile Computing Systems and Applications (MCSA '94)*, pp. 85–90, Santa Cruz, Calif, USA, December 1994.

[15] A. K. Dey, G. D. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications," *Human-Computer Interaction*, vol. 16, no. 2–4, pp. 97–166, 2001.

[16] A. Mani and H. Sundaram, "Modeling user context with applications to media retrieval," *Multimedia Systems*, vol. 12, no. 4-5, pp. 339–353, 2007.

[17] R. Laban, *Choreutics*, Macdonald and Evans, London, UK, 1966.

[18] R. Laban, *The Language of Movement: A Guidebook to Choreutics*, Plays, Boston, Mass, USA, 1974.

[19] P. Hackney, *Making Connections: Total Body Integration through Bartenieff Fundamentals*, Routledge, London, UK, 1998.

[20] J. Newlove and J. Dalby, *Laban for All*, Nick Hern Books, London, UK, 2003.

[21] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The CAVE: audio visual experience automatic virtual environment," *Communications of the ACM*, vol. 35, no. 6, pp. 64–72, 1992.

[22] N. Parés, P. Masri, G. van Wolferen, and C. Creed, "Achieving dialogue with children with severe autism in an adaptive multisensory interaction: the "MEDIATE" project," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 6, pp. 734–743, 2005.

[23] Y. Chen, H. Huang, W. Xu, et al., "The design of a real-time, multimodal biofeedback system for stroke patient rehabilitation," in *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 763–772, Santa Barbara, Calif, USA, October 2006.

[24] Y. Chen, W. Xu, H. Sundaram, T. Rikakis, and S.-M. Liu, "Media adaptation framework in biofeedback system for stroke patient rehabilitation," in *Proceedings of the 15th ACM International Multimedia Conference and Exhibition (MM '07)*, pp. 47–57, Augsburg, Germany, September 2007.

[25] T. Blaine, "The convergence of alternate controllers and musical interfaces in interactive entertainment," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME '05)*, Vancouver, Canada, May 2005.

[26] J. James, T. Ingalls, G. Qian, et al., "Movement-based interactive dance performance," in *Proceedings of the 14th Annual ACM International Conference on Multimedia (MM '06)*, pp. 470–480, Santa Barbara, Calif, USA, October 2006.

[27] P. Maes, T. Darrell, B. Blumberg, and A. Pentland, "The ALIVE system: wireless, full-body interaction with autonomous agents," *Multimedia Systems*, vol. 5, no. 2, pp. 105–112, 1997.

[28] C. R. Wren, F. Sparacino, A. J. Azarbayejani, et al., "Perceptive spaces for performance and entertainment: untethered interaction using computer vision and audition," *Applied Artificial Intelligence*, vol. 11, no. 4, pp. 267–284, 1997.

[29] D. Birchfield, H. Thornburg, M. C. Megowan-Romanowicz, et al., "Embodiment, multimodality, and composition: convergent themes across HCI and education for mixed-reality learning environments," *Advances in Human-Computer Interaction*, vol. 2008, Article ID 874563, 19 pages, 2008.

[30] D. Birchfield, T. Ciufo, H. Thornburg, and W. Savenye, "Sound and interaction in K-12 mediated education," in *Proceedings of the International Computer Music Conference (ICMC '06)*, New Orleans, La, USA, November 2006.

[31] D. Birchfield, T. Ciufo, and G. Minyard, "*SMALLab*: a mediated platform for education," in *Proceedings of the ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '06)*, Boston, Mass, USA, July-August 2006.

[32] K. Vaananen and K. Bohm, "Gesture driven interaction as a human factor in virtual environments: an approach with neural networks," in *Virtual Reality Systems*, pp. 93–106, Academic Press, New York, NY, USA, 1993.

[33] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Gestural interface to a visual computing environment for molecular biologists," in *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 30–35, Killington, Vt, USA, October 1996.

[34] S. Rajko, G. Qian, T. Ingalls, and J. James, "Real-time gesture recognition with minimal training requirements and on-line learning," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[35] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[36] "Motivating moves," http://www.motivatingmoves.com/index.html.

[37] M. Miyahara, M. Tsujii, M. Hori, K. Nakanishi, H. Kageyama, and T. Sugiyama, "Brief report: motor incoordination in children with Asperger syndrome and learning disabilities," *Journal of Autism and Developmental Disorders*, vol. 27, no. 5, pp. 595–603, 1997.

[38] A. Foroud and I. Whishaw, "On the development of reaching from infancy to pre-adolescence," in *Proceedings of the Society for Neuroscience Conference*, Atlanta, Ga, USA, October 2006.

[39] S. Kalajdziski, V. Trajkovic, S. Celakoski, and D. Davcev, "Multi 3d virtual dancer animation," in *Proceedings of IASTED International Conference on Modeling and Simulation*, pp. 20–25, Marina Del Ray, Calif, USA, May 2002.

[40] L. Torresani, P. Hackney, and C. Bregler, "Learning motion-style synthesis from perceptual observations," in *Proceedings of the 20th Annual Conference Neural and Information Processing Systems (NIPS '06)*, pp. 1393–1400, Vancouver, Canada, December 2006.

[41] "Anatomical terminology: planes of the body," http://training.seer.cancer.gov.

[42] H. Ishii and B. Ullmer, "Tangible bits: towards seamless interfaces between people, bits and atoms," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '97)*, pp. 234–241, Atlanta, Ga, USA, March 1997.

[43] P. Srinivasan, *Design of a large-area pressure sensing floor*, M.S. thesis, Department of Electrical Engineering, Arizona State University, Tempe, Ariz, USA, 2006.

[44] D. Chi, M. Costa, L. Zhao, and N. I. Badler, "The EMOTE model for effort and shape," in *Proceedings of the 27th ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*, pp. 173–182, New Orleans, La, USA, July 2000.

[45] L. Zhao, *Synthesis and acquisition of Laban movement analysis qualitative parameters for communicative gestures*, Ph.D. thesis, CIS, University of Pennsylvania, Philadelphia, Pa, USA, 2001.

[46] L. Zhao and N. I. Badler, "Acquiring and validating motion qualities from live limb gestures," *Graphical Models*, vol. 67, no. 1, pp. 1–16, 2005.

[47] D. Bouchard and N. I. Badler, "Semantic segmentation of motion capture using laban movement analysis," in *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA '07)*, vol. 4722 of *Lecture Notes in Computer Science*, pp. 37–44, Paris, France, September 2007.

[48] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[49] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 2003.

[50] H. Thornburg, *Detection and modeling of transient audio signals with prior information*, Ph.D. thesis, Department of Electrical Engineering, Stanford University, Palo Alto, Calif, USA, 2005.

[51] T. Ingalls, J. James, G. Qian, et al., "Response: an applied example of computational somatics," in *Proceedings of the 4th International Conference on Enactive Interfaces (Enactive '07)*, Gernoble, France, November 2007.

[52] http://www.motionanalysis.com.

[53] C. Roads,*Microsound*, MIT Press, Cambridge, Mass, USA, 2002.

[54] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[55] O. Vasicek, "An equilibrium characterization of the term structure," *Journal of Financial Economics*, vol. 5, no. 2, pp. 177–188, 1977.

[56] A. T. Cemgil, H. J. Kappen, P. Desain, and H. Honing, "On tempo tracking: tempogram representation and Kalman filtering," in *Proceedings of the International Computer Music Conference (ICMC '00)*, pp. 352–355, Berlin, Germany, August 2000.

[57] D. Swaminathan, H. Thornburg, T. Ingalls, J. James, S. Rajko, and K. Afanador, "A new gestural control paradigm for musical expression: real-time conducting analysis via temporal expectancy models," in *Proceedings of the International Computer Music Conference (ICMC '07)*, Copenhagen, Denmark, August 2007.

[58] S. Ross, *Stochastic Processes*, Wiley Interscience, Yorktown Heights, NY, USA, 1995.

[59] H. Thornburg, D. Swaminathan, T. Ingalls, and R. Leistikow, "Joint segmentation and temporal structure inference for partially-observed event sequences," in *Proceedings of the 8th IEEE Workshop on Multimedia Signal Processing (MMSP '06)*, pp. 41–44, Victoria, Canada, October 2006.

[60] V. Pavlovic, B. Frey, and T. Huang, "Variational learning in mixed-state dynamic graphical models," in *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI '99)*, pp. 522–553, Morgan Kaufmann, Stockholm, Sweden, July-August 1999.

[61] Y. Bar-Shalom, *Tracking and Data Association*, Academic Press Professional, San Diego, Calif, USA, 1987.

[62] A. Doucet, N. deFreitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer, Berlin, Germany, 2001.

[63] E. Wan and R. van der Merwe, *Kalman Filtering and Neural Networks*, chapter 7, John Wiley & Sons, New York, NY, USA, 2001.

[64] R. Urtasun, D. J. Fleet, and P. Fua, "3d people tracking with Gaussian process dynamical models," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 1, pp. 238–245, IEEE Computer Society, New York, NY, USA, June 2006.

[65] F. Guo and G. Qian, "3D human motion tracking using manifold learning," in *Proceedings of the 14th IEEE International Conference on Image Processing (ICIP '07)*, vol. 1, pp. 357–360, San Antonio, Tex, USA, September 2006.

[66] S. Rajko and G. Qian, "Autonomous real-time model building for optical motion capture," in *Proceedings of the International Conference on Image Processing (ICIP '05)*, vol. 3, pp. 1284–1287, Genova, Italy, September 2005.

[67] S. Rajko and G. Qian, "Real-time automatic kinematic model building for optical motion capture using a markov random field," in *Proceedings of the 4th IEEE International Workshop on Human-Computer Interaction (HCI '07)*, vol. 4796 of *Lecture Notes in Computer Science*, pp. 69–78, Rio de Janeiro, Brazil, October 2007.