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.

Copyright © 2009 Dilip Swaminathan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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.

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 center 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
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
Shape qualities corresponding to each vector. Spreading shown in Figure 2. Let us define three vectors:

- The **lateral direction** ($\vec{D}_{lat}$) 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 $\vec{D}_{lat} = (M_{rs} - M_{ls})/||M_{rs} - M_{ls}||$.

- The **up direction** ($\vec{D}_{up}$) 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 direction, that is, $\vec{D}_{up} = (0, 1, 0)$.

Finally, the **front direction** ($\vec{D}_{fr}$) 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, $\vec{D}_{pel}$ (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 $\vec{D}_{fr} = c \cdot \vec{D}_{fr}' + (1 - c) \cdot (\Delta M_p/||\Delta M_p||)$, where $\vec{D}_{fr}'$ is the previous front direction and $c$ is given by

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

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 $s_{fore} = 4s_{back}$, 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 $D_{up}$. Specifically, we compute the mean of all marker positions and project this mean onto $D_{up}$. 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 $D_{up}$. 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 $D_{fr}$. 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 $D_{lat}$ (perpendicular to $D_{fr}$). That is, we first project all marker positions onto $D_{lat}$, 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

$$L_t \in \{R_i, S_i, Ad, Re, Sp, En, Ne\}$$

The DAG in Figure 7 admits the following factorization:

$$P(M_{1:T}, L_{1:T}, R_{1:T}^{1:5}, S_{1:T}^{1:5}, Y_{1:T}^{1:5}) = P(M_1)P(L_1 | M_1)$$

$$\times \prod_{i=1}^{K} P(R_i^1 | L_1, M_1)P(S_i^1 | R_i^1, M_i)P(Y_i^1 | S_i^1)$$

$$\times \prod_{t=2}^{T} P(M_t | M_{t-1})P(L_t | L_{t-1}, M_t)$$

$$\times \prod_{i=1}^{K} P(R_i^1 | R_i^{1:T-1}, L_t, M_t)P(S_i^1 | S_{i-1}^1, R_i^{1:T}, M_t)P(Y_i^1 | S_i^1).$$

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

$$P(M_{1:T}, L_{1:T}, R_{1:T}^{1:5}, S_{1:T}^{1:5}, Y_{1:T}^{1:5}) = P(M_1)P(L_1 | M_1)$$

$$\times \prod_{i=1}^{K} P(R_i^1 | L_1, M_1)P(S_i^1 | R_i^1, M_i)P(Y_i^1 | S_i^1)$$

$$\times \prod_{t=2}^{T} P(M_t | M_{t-1})P(L_t | L_{t-1}, M_t)$$

$$\times \prod_{i=1}^{K} P(R_i^1 | R_i^{1:T-1}, L_t, M_t)P(S_i^1 | S_{i-1}^1, R_i^{1:T}, M_t)P(Y_i^1 | S_i^1).$$

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 $P(S_i^1 | S_{i-1}^1, R_i^{1:T}, M_t)$, coupled with the observation dependence $P(Y_i^1 | S_i^1)$. As previously discussed, $S_i$ contains $X_i^1$, the
“inherent” subindicator feature, for which \( Y_i \) is a “noisy” version:

\[
Y_i \sim N(X_i, \sigma^2_{Vi}).
\]  

(4)

However, \( S_i \) contains additional information necessary to model the influence of \( R_i \) and \( M_i \) on its dynamics using a first-order Markov dependence. That is,

\[
S_i = \text{vec}\{V_i, V_i^i, X_i\},
\]

(5)

where

(i) \( V_i \) is the inherent feature velocity; that is, rate of change in \( X_i \);

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

The full dependence, \( P(S_i | S_{i-1}, R_i, M_i) \), factors according to the expanded, single-feature DAG as shown in Figure 8; that is,

\[
P(S_i | S_{i-1}, R_i, M_i) = P(V_i | V_{i-1,0}, M_i)P(V_i | V_{i-1,0}, R_i)P(X_i | X_{i-1,0}, V_i),
\]

(6)

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

\[
P(V_i | V_{i-1,0}, R_i) = \mathcal{N}(\alpha V_{i-1,0} + (1 - \alpha) \delta_i (R_i), \beta \sigma^2_{Vi}).
\]

(7)

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

\[
\delta_i \equiv \begin{cases} V_{i,0}, & R_i = 1, \\ 0, & R_i = 0, \\ -V_{i,0}, & R_i = -1. \end{cases}
\]

(8)

Here \( \alpha \) controls the degree which \( V_i \approx V_{i-1,0} \) and \( \sigma_{Vi} \), the variance of the process about \( \delta_i \), controls the assumption \( V_i \approx \delta_i \). Since the OU process is mean-reverting [55], its use in modeling the trajectory \( V_i \) helps greatly in ensuring that small, rapid fluctuations in the subindicator features due to involuntary motions are registered as neutral, \( R_i = 0 \), 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 9: Image sequence of “Menu 1” (dancer 1) movement data showing the ground truth and inference results of the dominant Shape quality expressed.
distribution for $R_i^j$ assuming $L_t$ is constant. Essentially, $P_0(R_i^j \mid L_t)$ 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, $R_i^1, R_i^2$, and $R_i^3$ 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 $R_i^1, R_i^2$, or $R_i^3$ 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; $R_i^4, R_i^5, R_i^6$, it will be improbable that either is positive or negative. A full set of constraints on $P_0(R_i^j \mid L_t)$ is shown in Table 1, where $p^+$ is shorthand for $P(R_i^j = 1 \mid L_t)$, and $p^-$ represents $P(R_i^j = -1 \mid L_t)$. The complete specification of $P_0(R_i^j \mid L_t)$ 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 $Y_{1:t}^{L,M}$, 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, $M_t, L_t, and R_i^{(1,K)}$, are discrete, and others, for instance, $V_{1:2}$ 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

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>$L_t = Ri$</th>
<th>$L_t = Si$</th>
<th>$L_t = Ad$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_i^1, R_i^2, R_i^3$</td>
<td>$p^- &lt; 1, p^+ &gt; p^+$</td>
<td>$p^- &lt; 1, p^+ &gt; p^+$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
</tr>
<tr>
<td>$R_i^1$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
</tr>
<tr>
<td>$R_i^2$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
</tr>
<tr>
<td>$R_i^3$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
<td>$p^- &lt; 1, p^+ &lt; 1$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>$L_t = Ri$</th>
<th>$L_t = Si$</th>
<th>$L_t = Ad$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_i^1, R_i^2, R_i^3$</td>
<td>$p^- = 0.15, p^+ = 0.75$</td>
<td>$p^- = 0.15, p^+ = 0.75$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
</tr>
<tr>
<td>$R_i^1$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
</tr>
<tr>
<td>$R_i^2$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
</tr>
<tr>
<td>$R_i^3$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
<td>$p^- = 0.1, p^+ = 0.1$</td>
</tr>
<tr>
<td>$R_i^1, R_i^2, R_i^3$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
</tr>
<tr>
<td>$R_i^1$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
</tr>
<tr>
<td>$R_i^2$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
<td>$p^- = 0.01, p^+ = 0.01$</td>
</tr>
</tbody>
</table>
Figure 10: Segmentation performance on "Menu 1" (dancer 1) movement data showing the fusion of different features to infer the dominant Shape quality.

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.
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 \((\text{rising} \rightarrow \text{spreading} \rightarrow \text{retreating} \rightarrow \text{sinking})\), 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 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.
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 3: Dancer 1 segmentation results.**

<table>
<thead>
<tr>
<th>Data</th>
<th>% Recall</th>
<th>% Precision</th>
<th>Detection delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menu 1</td>
<td>100.0</td>
<td>85.7</td>
<td>0.1566 seconds</td>
</tr>
<tr>
<td>Menu 2</td>
<td>100.0</td>
<td>87.5</td>
<td>0.0733 seconds</td>
</tr>
<tr>
<td>Menu 3</td>
<td>100.0</td>
<td>70.0</td>
<td>0.2833 seconds</td>
</tr>
<tr>
<td>Menu 4</td>
<td>85.7</td>
<td>75.0</td>
<td>0.1916 seconds</td>
</tr>
</tbody>
</table>

**Table 4: Dancer 2 segmentation results.**

<table>
<thead>
<tr>
<th>Data</th>
<th>% Recall</th>
<th>% Precision</th>
<th>Detection delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menu 1</td>
<td>100.0</td>
<td>85.7</td>
<td>0.0688 seconds</td>
</tr>
<tr>
<td>Menu 2</td>
<td>100.0</td>
<td>100.0</td>
<td>0.1433 seconds</td>
</tr>
<tr>
<td>Menu 3</td>
<td>83.33</td>
<td>72.72</td>
<td>0.2342 seconds</td>
</tr>
<tr>
<td>Menu 4</td>
<td>87.5</td>
<td>77.0</td>
<td>0.2071 seconds</td>
</tr>
</tbody>
</table>

**Table 5: Dancer 3 segmentation results.**

<table>
<thead>
<tr>
<th>Data</th>
<th>% Recall</th>
<th>% Precision</th>
<th>Detection delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menu 1</td>
<td>100.0</td>
<td>100.0</td>
<td>0.1840 seconds</td>
</tr>
<tr>
<td>Menu 2</td>
<td>100.0</td>
<td>85.7</td>
<td>0.1216 seconds</td>
</tr>
<tr>
<td>Menu 3</td>
<td>83.33</td>
<td>75.0</td>
<td>0.2383 seconds</td>
</tr>
<tr>
<td>Menu 4</td>
<td>100.0</td>
<td>83.33</td>
<td>0.1250 seconds</td>
</tr>
</tbody>
</table>
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, non-portable 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


Submit your manuscripts at
http://www.hindawi.com