This is an Accepted Manuscript of an article published by Taylor & Francis, available online:

Radu-Daniel Vatavu. (2017). Beyond Features for Recognition: Human-Readable Measures to Understand Users' Whole-Body Gesture Performance. *International Journal of HumanComputer Interaction*. Taylor & Francis, 1-18.

http://dx.doi.org/10.1080/10447318.2017.1278897

Beyond Features for Recognition: Human-Readable Measures to Understand Users' Whole-Body Gesture Performance

Radu-Daniel Vatavu

MintViz Lab | MANSiD Research Center
University Stefan cel Mare of Suceava
13 Universitatii, 720229 Suceava, Romania

Abstract

Understanding users' whole-body gesture performance quantitatively requires numerical gesture descriptors or features. However, the vast majority of gesture features that have been proposed in the literature were specifically designed for machines to recognize gestures accurately, which makes those features exclusively *machine-readable*. The complexity of such features makes it difficult for user interface designers, non-experts in machine learning, to understand and use them effectively (see, for instance, the Hu moment statistics or the Histogram of Gradients features), which reduces considerably designers' available options to describe users' whole-body gesture performance with legible and easily interpretable numerical measures. To address this problem, we introduce in this work a set of 17 measures that user interface practitioners can readily employ to characterize users' whole-body gesture performance with *human-readable* concepts, such as area, volume, or quantity. Our measures describe (1) spatial characteristics of body movement, (2) kinematic performance, and (3) body posture appearance for whole-body gestures. We evaluate our measures on a public dataset composed of 5654 gestures collected from 30 participants, for which we report several gesture findings, *e.g.*, participants performed body gestures in an average volume of space of 1.0 m^3, with an average amount of hands movement of 14.6 m, and a maximum body posture diffusion of 5.8 m. We show the relationship between our gesture measures and recognition rates delivered by a template-based Nearest-Neighbor whole-body gesture classifier implementing the Dynamic Time Warping dissimilarity function. We also release BOGART, the Body Gesture Analysis Toolkit, that automatically computes our measures. This work will empower researchers and practitioners with new numerical tools to reach a better understanding of *how users perform whole-body gestures* and, thus, to use this knowledge to inform improved designs of whole-body gesture user interfaces.

Keywords: whole-body gestures; gesture measures; gesture analysis; gesture recognition; Dynamic Time Warping; classification; Nearest-Neighbor classifier; human movement; human motion; experiments; toolkit.

1. INTRODUCTION

Processing human movement represents a preliminary task for gesture analysis and gesture user interface design and development, tackled by researchers so far with various gesture features, tools, and techniques (Aslan et al., 2013; Piana et al., 2013; Vatavu, 2013a; Wang and Suter, 2006). Recently, inexpensive, off-the-shelf motion capture sensors, such as the Microsoft Kinect sensor, have enabled researchers and practitioners to capture whole-body gesture movement at decent levels of sensing resolution and, consequently, to leverage the expressiveness of body movement for intuitive and natural human-computer interaction (Microsoft, 2014; Fothergill et al., 2012; Vatavu, 2012a, 2015). However, the existing work has focused almost exclusively on *how to recognize body gestures* in the attempt to provide the community with reliable and robust techniques to classify gestures accurately; see Poppe (2010) for a survey. These recognition techniques, however, employ gesture representations and gesture features that are difficult for user interface designers to understand and operate intuitively. For example, Bobick and Davis (2001) employed Hu moment statistics (Hu, 1962) to implement a human action classifier, which received wide adoption. However, the authors acknowledge that *"one disadvantage is that the Hu moments are difficult to reason about intuitively"* (pp. 261-262), which ulti-

Email address: vatavu@eed.usv.ro (Radu-Daniel Vatavu)
URL: http://www.eed.usv.ro/~vatavu (Radu-Daniel Vatavu)

mately affects the capacity of the practitioner to understand, debug, and fix potential problems that may arise during system usage. This situation can be seen over and over for the gesture features currently in use in the community (Aggarwal and Cai, 1999; Moeslund et al., 2006; Poppe, 2010; Turaga et al., 2008) that are exclusively *machine-readable*; see "principal invariants" and "gradient tensor features" (Ali and Shah, 2010), "Hessian matrices" and "eigenvalues" (Gorelick et al., 2006), or "Fourier features" (Weinland et al., 2006), to name only a few. On the contrary, the Human-Computer Interaction community has always valued and promoted the adoption of pattern recognition and machine learning concepts and techniques that are easy to understand, even for non-experts; see, for instance, the $-family of gesture recognizers (Wobbrock et al., 2007; Anthony and Wobbrock, 2010; Vatavu et al., 2012).

Adversely, there has been considerably less attention in the community for designing *human-readable* gesture measures to understand *how users actually perform whole-body gestures*, although such an understanding would greatly benefit gesture user interface design. For example, knowing how much users vary their body movements in space can inform about the effort to perform such movements (Nielsen et al., 2004; Rekik et al., 2014; Vatavu et al., 2011; Vatavu, 2013a). Or, knowing the level of precision at which users prefer to perform body gestures can inform the design of sensors' resolution accuracy levels for cost-effective gesture acquisition (Vatavu, 2013b). Tackling such aspects is important because of the inherent variation of users' gestures, which are *"spontaneous creations of individual speakers, unique and personal"* (McNeill, 1992, p. 1). In this work, we focus on the practitioner's rather on the machine's perspective by proposing a set of *human-readable gesture measures* that rely on commonly-understood concepts of space and time for user interface practitioners to evaluate and understand users' whole-body gesture performance effectively.

The contributions of this work are as follows:

1. We introduce a set of measures to characterize users' whole-body gesture performance on three distinct execution levels: spatial, kinematic, and body appearance.
2. The majority of our measures (*e.g.*, quantity of movement, difference in gesture movement, ratio of movements, etc.) are customizable, allowing researchers to particularize them to suit their specific investigation goals about gestures, making our set of whole-body gesture measures considerably larger through custom expansion.
3. We compute our measures on a public dataset composed of 5654 whole-body gestures collected from 30 participants (Fothergill et al., 2012), for which we report several gesture discoveries, such as the average volume of all gesture executions was 1.0 m^3. We discuss our gesture performance measures in connection with recognition rates delivered by a template-based gesture classifier implementing the Dynamic Time Warping dissimilarity function.
4. We deliver the community with a software toolkit, BOGART, the BOdy Gestures Analysis Toolkit, to automate computation of our measures in order to encourage further development in this line of work. BOGART is available as a compiled .NET library and C# source code files.

It is our hope that our set of measures will benefit researchers to reach a better understanding of human gesture movement in general, and to develop better gesture technology and user interfaces for whole-body gestures, in particular.

2. RELATED WORK

We are interested in this section in previous work that employed features to represent and describe whole-body gestures and movement; see (Aggarwal and Cai, 1999; Moeslund et al., 2006; Poppe, 2007, 2010; Rosenhahn et al., 2008; Turaga et al., 2008) for extensive surveys. Please note that both "human movement" and "human motion" have been employed in the technical literature discussing recognition techniques and, for the purpose of this work, they are synonymous. In this section, we review gesture representation techniques and features developed in the Computer Vision and Pattern Recognition communities to recognize whole-body movements and human activities. We also review work about describing and interpreting human movement using non-technical descriptors by connecting to the Performing Arts, and we discuss current methodologies employed in Human-Computer Interaction to elicit, analyze, and describe users' gestures, such as the gesture elicitation methodology (Vatavu and Wobbrock, 2015; Wobbrock et al., 2005, 2009). We start with a discussion on the terminology employed across the disciplines interested in the study of human movement and gestures.

2.1. The study of human movement across disciplines

Because human movement has been studied by researchers from various disciplines, it is important to correlate different terminology and concepts to which we refer in this section. For example, human movement has been addressed at various levels of abstraction in the Computer Vision community. Generally, human motion analysis, action detection, action recognition, and gesture recognition are seen as distinct research goals in this community, although they share the same image representations of human movement (Poppe, 2007, 2010). For instance, Moeslund et al. (2006) employed an hierarchy of concepts for movement, such as *action primitives*, *actions*, and *activities* to group research on activity representation and recognition in their survey of the state-of-the-art in techniques for human motion capture and analysis. For example, an action primitive is an atomic entity from which actions are composed, *e.g.*, performing a forehand shot in a tennis match. The corresponding action in this case would be returning the ball, which might be composed of different action primitives, such as running toward the ball and hitting it with the forehand; ultimately, the high-level activity is playing tennis, which assumes different actions performed over time (Moeslund et al., 2006, p. 110). Bobick (1997) referred to *movements* as the most atomic primitives of human action (which do not require contextual knowledge to be recognized), *activities* as sequences of movements (for which knowledge comes from understanding the statistics of the sequence), and *actions* as being large-scale events that typically include interactions with the environment and causal

relationships. Furthermore, Turaga et al. (2008) distinguished between *actions* and *activities* in their survey of techniques for recognizing human activities. In their view, actions are simple patterns of motion performed by a single person, while activities involve coordination of the actions of several people.

In Psychology, gestures have been analyzed in the context of understanding the way people think and express themselves using language. For example, McNeill (1992) looked at gestures as *"movements of the hands and arms that we see when people talk"* that *"reveal the idiosyncratic imagery of thought"* (p. 1), and Kendon (2000) worked with a definition of gestures as *"coordinated movements that achieve some end"* (p. 47) to address the inter-relationships between gesture and language.

In this work, we understand by whole-body gestures any movement performed at the scale of the body that bears meaning for the purpose of interacting with a computer. This interpretation is in line with other researchers, see for instance the overview on gesture-based interaction of Buxton (2011). Our definition of whole-body gestures is in direct correspondence with the *action primitives* of Moeslund et al. (2006) and also with the *movements* of Bobick (1997) under the assumption of a uniquely-associated identifier for gesture movement in a given application, *i.e.*, the gesture's class. However, the measures that we introduce may also be applied to characterize and evaluate human movement at higher levels of abstraction, such as *actions* and *activities* as defined in (Bobick, 1997; Moeslund et al., 2006). Our definition follows a principle illustrated by Kurtenbach and Hulteen (1990) to discriminate between gestures and generic movement: *"A gesture is a motion of the body that contains information. Waving goodbye is a gesture. Pressing a key on a keyboard is not a gesture because the motion of a finger on it's way to hitting a key is neither observed nor significant. All that matters is which key was pressed."*

2.2. Numerical features to represent and describe human movement

Many features have been introduced to represent human action and gesture movement for the purpose of efficient recognition. Frequently employed features rely on edges, body silhouettes and contours, motion, and color extracted from images and videos. Many of these features are well described in the existing surveys on the state-of-the-art in the field. For example, Poppe (2007) overviewed vision-based techniques for human motion analysis with focus on reliable detection of the configurations of body parts over time. A follow-up survey (Poppe, 2010) addressed techniques for recognizing body poses and full-body movements in images and video. Aggarwal and Cai (1999) discussed techniques for interpreting human motion including tracking and recognition. Moeslund et al. (2006) conducted a survey of vision-based methods to capture and analyze human movement, which they structured using a taxonomy discussing model initialization for human capture, segmenting and tracking of humans in images, pose estimation, and human action recognition. Turaga et al. (2008) discussed techniques for recognizing activities in which more people are involved. Jaimes and Sebe (2007) examined techniques applied to the design of multimodal interaction with computing systems using body movement, gestures, and gaze.

Human movement can be described at the level of the entire body seen as a single region of interest or at the level of body parts, which results in a set of local measurements, *e.g.*, a set of 3-D points tracking various parts of the body over time, as provided by standard motion capture equipment. Poppe (2010) discussed both *global* and *local* representations of human action in his survey of human action recognition techniques. For example, frequently-employed global representations use the body silhouette's area and contour or the optical flow of motion (Bobick and Davis, 2001; Chen et al., 2006; Howe, 2004). In turn, local representations compute local descriptors, such as space-time points (Laptev, 2005) or SURF features (Willems et al., 2008). In the following, we discuss previous work that employed representations of human movement for the purpose of recognition. In doing so, we focus our discussion on the features that prior work has introduced as we highlight their strong *machine-readable* dimension.

One way to describe body pose is to extract its silhouette, from which other features can be computed, such as statistical moments (Ahad et al., 2008; Bobick, 1997; Gorelick et al., 2006; Hu, 1962). Bobick and Davis (2001) introduced temporal templates to summarize the spatio-temporal motion properties of human movement in terms of location (*i.e.*, where motion has occurred) and recency (*i.e.*, how long since motion has occurred) at each pixel of the image. For instance, motion energy images (MEIs) are binary images that encode the presence of motion accumulated over time at each pixel's location, *e.g.*, pixel (x, y) has value 1 if motion was detected at that pixel's location anytime in the last τ frames. Motion history images (MHIs) are more nuanced versions of MEIs as they employ gray-level intensities to encode the recency of motion with brighter colors showing more recently detected motion, *i.e.*, pixel (x, y) has value $\tau - t$ $(t \leq \tau)$ if motion was last detected at that pixel's location t frames ago. The authors used the seven statistical moments of Hu (1962) extracted from the MEI and MHI images to represent human movement, which they matched against stored templates. However, even though motion images deliver good synthesized visual descriptions of action, Hu moments are difficult to interpret by humans; in the authors' own words, *"One disadvantage is that the Hu moments are difficult to reason about intuitively"* (Bobick and Davis, 2001, pp. 261-262).

The temporal templates of Bobick and Davis (2001) received great popularity among researchers interested in recognizing human movement, and follow-up work introduced modified versions of MEIs and MHIs to represent the spatio-temporal aspects of movement in the form of a single image template. For instance, Bradski and Davis (2002) extended the value range of pixels from MHIs to more than 256 levels of gray by storing the actual floating-point timestamps of motion detected at each pixel's location; the result was the timed motion-history image (tMHI). Davis (2001) extended the MHI to an hierarchy of motion images that compute local motion flow across different directions of motion. Xiang and Gong (2006) proposed pixel change history images (PCHs), which are parametrized

versions of MHIs that allow the practitioner to control how the recency of motion is updated at each pixel's location. Another feature representation technique conceptually similar to the temporal templates of Bobick and Davis (2001) was proposed by Masoud and Papanikolopoulos (2003), which employed multiple feature images to represent human movement over time. The feature image at time t is computed by subtracting from the current frame a weighted-average image of the action occurring up to time t. Motion averaging is controlled by a decay rate, $\alpha \in [0..1]$, which affects the type of motion captured in the feature image, which may be motion relative to the background ($\alpha = 0$), motion relative to the previous frame ($\alpha = 1$), or some other temporal change in the scene ($\alpha \in (0..1)$). Movement is depicted by feature images as a fading trail corresponding to the parts of the body engaged in motion (Masoud and Papanikolopoulos, 2003, p. 732). Weinland et al. (2006) extended 2-D motion-history images to 3-D motion-history volumes (MHVs) that encode in a free-viewpoint manner the human movement captured from multiple video cameras. The authors employed Fourier-based features to represent movement, which were found better suited for recognition than the Hu moments (Bobick and Davis, 2001). Polana and Nelson (1997) defined a periodicity measure working on the spectrum of Fourier-transformed signal of the movement to detect periodic motion. However, Fourier features (Weinland et al., 2006, p. 257) are impossible to decipher by people not trained in interpreting the frequency spectra of multi-dimensional signals.

Body silhouettes are the basis for computing many other body descriptors. For example, Chen et al. (2006) introduced the star skeleton to represent body poses. The contour of the human body is detected in each frame and the "star" is formed by joining the contour's centroid with its most extreme points. Human movement is then described as a sequence of star skeletons. A Hidden Markov Model was employed by Chen et al. (2006) to recognize body actions with 98% reported accuracy. Howe (2004) described body contours with turning angles and employed the Chamfer distance to classify human movement. Gorelick et al. (2007) represented human actions as 3-D shapes induced by the silhouette of the body moving in space and time. Each point in their representation is characterized by the mean time required for a particle to perform a random walk to the boundary of the shape, computable as a solution of a Poisson equation (Gorelick et al., 2006) at each point (x, y, t). Using that representation, the authors derived local and global features that describe human movement, *e.g.*, space-time saliency, space-time orientations, and statistical moments weighted by the characteristic function of the space-time shape, which were then used for detection, recognition, and clustering of human actions. Although suited for interpretation by a machine, these features are off-limits for non-experts, as they involve advanced concepts such as Hessian matrices, eigenvalues, and weighted moments computed for multi-dimensional signals (Gorelick et al., 2006, pp. 2248-2259).

When the body silhouette cannot be reliably computed from the image, motion information, such as optical flow, can be alternatively used to describe human movement. For example, Efros et al. (2003) computed optical flow to represent and recognize actions of people in low resolution video. Ahad et al. (2008) computed motion-energy and motion-history images using optical flow rather than subtracting consecutive frames, which they found to improve classification accuracy of actions affected by occlusion. Dalal and Triggs (2005) encoded the shape of the human body using HOG descriptors (Histogram of Gradients). To compute HOGs, the image is divided into cells, and a 1-D histogram is computed for each cell using edges' orientations at each pixel. The resulted descriptors characterize local aspects of the body and contribute to the overall description of the human movement when performing some action captured by a sequence of frames. Batra et al. (2008) defined space-time shapelets to describe human movement as a histogram over a dictionary of local edge-structures computed over time. Maes et al. (2013) employed spatio-temporal representations and template-based matching for coupled action-perception processes. Ali and Shah (2010) introduced a set of kinematic features to describe the dynamics of human motion, such as divergence, vorticity, and gradient tensor features computed from optical flow. Although suited for interpretation by a machine for recognition purposes (up to 95% classification accuracy was reported), these features are impossible to be interpreted by non-experts, as they rely on advanced concepts, such as derivatives of vector-valued functions and principal invariants of gradient tensor matrices (Ali and Shah, 2010).

2.3. Nonnumerical description of human movement

Description of human movement has been examined by many other disciplines, such as Kinesiology, Physiotherapy, and the Performing Arts, for which researchers and practitioners have devised specific methods to record, store, and analyze the movement of the human body. For instance, Abbie (1974) discussed motivations behind inventing systems of movement notation and provided a review of such systems relevant for Physiotherapy. The author noted that *"The invention of many notation systems suggests that words have been found inadequate to describe movement accurately, and the more complex or unusual a movement, the less adequate and accurate become the words."* (p. 61). Sevdalis and Keller (2011) examined relationships between research on dance and embodied cognition with the purpose of understanding human action and social cognition. In their review of the topic, the authors focused on motor experience and expertise, learning and memory, action, intention and emotion understanding and audio-visual synchrony.

The Performing Arts have seen many systems to record and analyze human movement. Such systems were designed from as early as the 16th and 17th centuries, such as the Beauchamp-Feuillet notation used to record steps for Baroque dance (Waite and Appleby, 2003). Dance notation, or written dance, employs graphical symbols and figures, numerical systems, and letters and words to present educated readers with a visual, symbolic description of human movement during dance. For instance, one of the first methods to record dance in written form comes from Thoinot Arbeau's "Orchésographie" dating from the 16th century, referenced in Hutchinson (1991) (p. 2), which used codes and word abbreviations to depict dance steps; for example, "R" means "reverencia", "re" stands for

"represa", etc. Written descriptions accompanied by names and figure illustrations were used to accompany the musical notation. Names were placed next to corresponding musical notes to create the connection between music and the steps to be performed. Visual systems, such as the "Sténochoréographie" of Arthur Saint-Léon from 1852, referenced in (Hutchinson, 1991), employed stick figures to depict positions of arms, legs, torso, and head during dance, which were placed under the music score to synchronize timings. While discussing systems to represent human movement in dance, Hutchinson (1991) noted that *"Every few years a new system appears"* and that *"Most fall back on one or other of the devices already tried, and most favor one form of dance. As modern technology develops, the emphasis is upon mathematical systems which can be adapted to the computer. It is essential, however, that the human aspect is not lost."* (p. 4)

One of the best known and employed systems is Labanotation or Kinetography Laban invented by Rudolf Laban to represent and store human movement (von Laban and Lange, 1975). Labanotation employs three types of descriptors: motifs, effort-shape, and structural descriptions. Motif Writing describes the theme of the movement, its motivation, aim, and intention. The Effort-Shape description characterizes the quality and the expression of human movement, where effort relates to *energy* and shape describes the *form* that the movement takes during dance. Structural description expresses human movement in terms of the parts of the body involved in movement production, space, time, and dynamics, which are abstracted as visual symbols. These symbols take the form of vertical lines and staff representing the body divided into its right and left sides. The shape of the symbols indicates direction, while their size indicates timing. Secondary symbols (*e.g.*, pins, bows, hooks) illustrate variations in style. The richness and flexibility of the Labanotation system in representing human movement determined its adoption beyond choreography in other fields, such as anthropology (Grim-Feinberg and Santos, 2015; Wulff, 2001), physiotherapy (Abbie, 1974), sports (Dania et al., 2013), and gesture and movement representation for interactive computer applications (Kordts et al., 2015; Laiyang and Junjun, 2014; Hachimura and Ohno, 1987; Loke and Robertson, 2009). The system has even been automated by software that converts human movement into the visual abstractions of Labanotation (Guo et al., 2014; Choensawat et al., 2016).

Priel (1974) developed a system for the numerical description of human body postures in the form of "posturegrams" that report the position and angle of each body joint relative to a reference level. The Ovako Working Posture Analyzing System (OWAS) was introduced to identify and evaluate poor working postures (Karhu et al., 1977, 1981). OWAS consists of two stages: (1) an observational part, during which working postures are discovered and (2) a set of criteria for redesigning working methods and places to increase the comfort of human body posture during work.

2.4. Methodologies to analyze and describe human movement

In this section, we consider another approach for evaluating gesture performance by reviewing studies that collected and analyzed users' gesture preferences for interacting with a computer system. Wobbrock et al. (2005) introduced a methodology for computing agreement between users from which proposals were elicited for abstract symbols. The methodology was applied to reveal users' preferences for multi-touch interaction and to compile a set of gestures reflective of users' behavior for multi-touch surfaces (Wobbrock et al., 2009). Vatavu and Wobbrock (2015) and Vatavu and Wobbrock (2016) developed the methodology to include more measures, such as disagreement and coagreement rates, and statistical inference tests.

The agreement rate methodology was applied to study whole-body gestures as well. For instance, Vatavu (2012b) found an average agreement rate of .415 (from a maximum of 1.000) when participants were asked to propose gestures to control standard functions of the TV set. Follow-up studies revealed more about users' conceptual models of interacting with whole-body gestures (Vatavu, 2013a) and high-resolution finger and hand postures and movements (Vatavu and Zaiti, 2014; Zaiti et al., 2015). For example, users prefer to associate symmetric gestures to dichotomous tasks, use two hands to increase the expressiveness of their gestures, employ cultural signs and gestures, and perform gesture movements with reference to specific body parts, such as covering the ears for lowering down the TV audio volume (Vatavu, 2013a). Connell et al. (2013) looked at children's whole body gestures and reported a potential effect of contextual cues on spatial interaction and navigation tasks, an influence of age on the proposed gestures, and some preference for egocentric (toward the body and body-centered) gestures. Silpasuwanchai and Ren (2014) investigated body gestures for video games and reported an average agreement of .370. The authors discussed the opportunity of transferring gesture commands between different body parts, such as between hands and legs. Morris (2012) elicited users' gesture and speech preferences for controlling a web browser in the living room and reported an influence of participants' previous WIMP experience on their proposals for commands (*i.e.*, the "legacy bias") and identified common conventions employed by participants, such as referring links with numbers, *e.g.*, "go to link 2". Nebeling et al. (2014) replicated the study and extended the methodology toward obtaining reproducible user-defined gesture sets.

Stern et al. (2008) looked into hand gestures that would be intuitive for users to execute and defined the intuitiveness of a gesture command as the strength of *"the cognitive association between a command or intent, and its physical gestural expression."* The authors introduced a measure that describes the intuitiveness of the association between a gesture command and a function as the number of participants in consensus with that association. In a similar manner, Vatavu (2013a) employed the confidence value of a referent as the maximum percent of participants in agreement for that referent and Morris (2012) used the max-consensus measure for the percent of participants suggesting the most popular proposal and the consensus-distinct ratio for the percent of distinct proposals for a given referent.

Beyond elicitation studies, researchers have explored creative ways for people to effect commands using whole-body gestures. For instance, Holz and Wilson (2011) introduced "data miming," a technique to understand gestures used by peo-

ple to describe concrete physical objects. To inform the development of their technique, the authors conducted a study during which they observed participants describing objects using hand gestures. Results showed that participants made use of simultaneous and symmetric hand movements, different postures of the hand, and spatial gestures to describe size and shape. Gustafson et al. (2010) conducted several studies to understand users' ability to employ imaginary interfaces by using their hands and visuo-spatial memory. For example, participants drew characters and sketches in mid-air with good within-stroke alignment in the absence of visual feedback, but the alignment of multiple strokes was more challenging. Also, participants' accuracy of pointing to imaginary targets became worse as the target was farther away from the reference hand. Overall, experimental findings showed that users' short-term memory can replace visual feedback provided by conventional interfaces for mid-air gesture interaction. Follow-up studies exploited people's ability to point, gesture, and perform whole-body movements for a variety of application scenarios, even in the absence of visual feedback (Baudisch et al., 2014; Dezfuli et al., 2012; Gustafson et al., 2011).

2.5. Summary

The literature of human action recognition has proposed many features that work well to classify whole-body gestures accurately. However, these features are either too complex or uninformative to be used by practitioners to describe users' gesture performance during the design of gesture sets or gesture user interfaces. Clearly, gradient tensor features (Ali and Shah, 2010), Hessian matrices (Gorelick et al., 2006), or Fourier features (Weinland et al., 2006) are off-limits for most user interface designers without advanced training in machine learning, but who nevertheless wish to describe users' gesture performance numerically. Other systems, such as (von Laban and Lange, 1975) are purely descriptive and do not provide numerical quantification of human movement. In this work, we are interested in *human-readable measures* that can be employed to characterize users' whole-body gesture performance to help researchers and practitioners reach a clear and accurate understanding of how users actually perform gestures, a methodological aspect of gesture analysis left unattained by previous work.

3. MEASURES FOR WHOLE-BODY GESTURES

In this section, we introduce a set of seventeen (17) measures for researchers and practitioners to evaluate users' whole-body gesture performance by leveraging commonly-understood concepts for describing numerical quantities, such as area, volume, time, etc., and simple arithmetic operations, such as differences and ratios. We define our measures with respect to the following three dimensions along which whole-body gesture movement unfolds relevant information:

1. The *spatial* characteristics of body movement capture aspects related to the area, volume, and amplitude of gesture movement performed by the whole body or by individual body parts inside a 3-D space. Spatial gesture measures provide answers to questions that start with *where* (*e.g.*, where was movement produced in the room?) and with *how much* (*e.g.*, how much movement was produced with the dominant hand?).
2. The *kinematic* dimension of gesture performance characterizes the temporal aspects of whole-body gesture production, such as the absolute and relative speed of individual body parts. Consequently, kinematic measures are able to provide answers to questions that start with *when* and *for how long* body movement took place.
3. The *body appearance* dimension describes the particularities of gesture movement that determine expressiveness and differentiate body movements by their meaning. For example, the amount of distinct body postures employed during the articulation of a body gesture is a measure of expressiveness. Appearance measures answer questions that start with *how* (*e.g.*, how was the gesture actually produced in terms of the body postures adopted by the user?).

Most of the measures that we introduce are general and, consequently, are customizable to capture various aspects of gesture performance, leading to even more measures. We highlight these measures and the opportunities they provide next in the paper. By using the three dimensions mentioned above, our set of whole-body gesture measures is able to characterize the *spatio-temporal-appearance* aspects of users' whole-body gesture performance for a multitude of experimental situations and scientific investigation goals. For the rest of the paper, we consider a whole-body gesture defined as a set of n body postures P_i unfolding in time t_i, $i = 1..n$, with each body posture composed of J joints that are being tracked and reported by a motion sensor, such as the Microsoft Kinect sensor or the Vicon Motion Capture system, for instance:

$$\left\{(P_i, t_i) \,|\, P_i = \left\{p^i_j \,|\, p^i_j = \left(x^i_j, y^i_j, z^i_j\right), j = 1..J\right\}, i = 1..n\right\} \quad (1)$$

For example, in the case of the Microsoft Kinect sensor, the number of joints tracked in each video frame is J=20 (for SDK versions up to v1.8) and J=30 joints for Kinect for Windows SDK v2.0; see (Microsoft, 2016a). Other systems, such as the Vicon Motion Capture system, allow a configurable number of joints to be tracked in 3-D using retroreflective markers attached to the human body.

3.1. Spatial measures

Spatial measures describe the overall amplitude and amount of users' whole-body movements in 3-D space as well as of the individual movements of specific body parts that may be of particular importance to the researcher, *e.g.*, hands, arms, or head. Spatial measures are used to characterize general gesture movement in space, *e.g.*, How much are users moving when performing gesture g_1?; or What is the difference in the amount of movement between the dominant and non-dominant hand during the performance of gesture g_2? Spatial measures can also serve to inform the design of user interfaces employing whole-body gesture commands, for instance by determining the exact volume of space around the body needed by users to perform

gestures safely, in line with current recommendations and operating guidelines accompanying motion-capture sensors, such as the Microsoft Kinect operating guidelines (Microsoft, 2014).

Our 10 spatial measures for characterizing and analyzing users' whole-body gesture performance are as follows:

1. Gesture Volume (GV) represents the volume of the 3-D space in which the whole-body gesture is performed:

$$\mathrm{GV} = \prod_{\delta \in \{x,y,z\}} \left(\max_{i,j} \left\{ \delta_j^i \right\} - \min_{i,j} \left\{ \delta_j^i \right\} \right) \tag{2}$$

where δ represents each of the x, y, and z dimensions of gesture movement, i enumerates all body postures P_i, and j enumerates all the joints tracked by the sensor. Gesture volume values may be reported in sensor or screen units (*e.g.*, pixels^3 or voxels) or in physical units, such as cubic meters, as reported in the Evaluation section of this paper.

2. Gesture Area (GA) represents the 2-D area *in front* of the sensor in which the whole-body gesture is performed:

$$\mathrm{GA} = \prod_{\delta \in \{x,y\}} \left(\max_{i,j} \left\{ \delta_j^i \right\} - \min_{i,j} \left\{ \delta_j^i \right\} \right) \tag{3}$$

where δ represents each of the x and y dimensions, i enumerates all body postures P_i, and j all the tracked joints. This measure represents a particularization of Gesture Volume, valuable for a large majority of whole-body gestures that take place *in front* of a display and that consist of movement mostly performed along the x and y axes. While Gesture Volume can serve to inform the design of full 3-D body-centered user interfaces, such as the SpaceSensor of Hong and Woo (2006), Gesture Area focuses on interactions that are performed in a plane in front of the user's body to control content on a remote display; see (Microsoft, 2016b). In this work, we report GA values in physical units, such as meters squared.

3. Quantity of Movement ($\mathrm{Q_M}$), defined as the total amount of movement performed by the user and computed as the sum of Euclidean distances between the corresponding joints of time-consecutive body frames:

$$\mathrm{Q_M} = \frac{1}{\lambda} \sum_{i=2}^{n} \sum_{j=1}^{J} \left\| p_j^i - p_j^{i-1} \right\| \tag{4}$$

where $\left\| p_j^i - p_j^{i-1} \right\|$ represents the Euclidean distance between two 3-D points denoting the same joint j, $j = 1..J$, belonging to time-consecutive body postures P_{i-1} and P_i, $i = 2..n$. The factor λ acts as a normalization factor that controls how the quantity of movement is being interpreted in connection with the number of joints that are tracked by the sensor. For instance, λ may take the value 1, case in which $\mathrm{Q_M}$ reports the cumulative movement produced by all the body joints. Or, λ can be equal to the number of joints J, case in which $\mathrm{Q_M}$ reports the average quantity of movement per joint. Alternatively, λ may take any other value that the practitioner finds suitable for normalizing the quantity of human movement with respect to the number of tracked joints for a specific experiment or application scenario. In this work, we use $\lambda = 1$ and report $\mathrm{Q_M}$ values in meters.

4. Generalized Quantity of Movement ($\mathcal{Q}_\mathrm{M}$) represents the generalized version of the $\mathrm{Q_M}$ measure, for which Euclidean distances are weighted (w_j, $j = 1..J$) to denote their relative importance in the overall sum:

$$\mathcal{Q}_\mathrm{M} = \frac{1}{\lambda} \sum_{i=2}^{n} \sum_{j=1}^{J} w_j \cdot \left\| p_j^i - p_j^{i-1} \right\| \tag{5}$$

The generalized quantity of movement is useful when the practitioner needs to isolate specific parts of the whole-body gesture movement that may be relevant for answering specific research questions during gesture analysis, such as how the hands are moving. In this example, weights corresponding to hand joints will be set to 1.0, while the rest of the weights to 0.0. The outcome measure is valuable and informative on its own to consider as a distinct measure by itself, as follows:

5. Quantity of Hands Movement ($\mathrm{Q_{Hands}}$) represents the amount of movement performed by the user's hands.

The movement of other body parts can be extracted in the same way by setting appropriate weight values for body joints. Another example is assigning weights of different magnitudes to different body parts, such as 1.0 for the dominant hand, and a lower value, such as 0.5, for the non-dominant hand for evaluating a gesture that was specifically designed as unimanual.

6. Difference of Movement ($\mathrm{D_M}$) captures the difference between the amount of movement produced by two different body parts. We define $\mathrm{D_M}$ as the difference between values reported by the generalized quantity of gesture movement measure computed with different sets of weights:

$$\mathrm{D_M} = \mathcal{Q}_\mathrm{M}(\mathrm{weights}_1) - \mathcal{Q}_\mathrm{M}(\mathrm{weights}_2) \tag{6}$$

There are many interesting combinations of weights corresponding to specific body parts for which movement may be compared, which we ultimately leave to the researcher to explore, according to the particular goal of their investigations. In this work, we evaluate the difference between the movements produced by the right and left hands, as follows:

7. Difference of Hands Movements ($\mathrm{D_{Hands}}$):

$$\mathrm{D_{Hands}} = \mathcal{Q}_\mathrm{M}(\mathrm{weights}_r) - \mathcal{Q}_\mathrm{M}(\mathrm{weights}_l) \tag{7}$$

where the $\mathrm{weights}_r$ array has weights equal to 1.0 for the joints of the right hand and 0.0 otherwise, and $\mathrm{weights}_l$ is defined similarly for the left hand.

$\mathrm{D_M}$ can be further customized by the researcher to include more body parts, resulting in measures that compute sums and differences of more than two values. Note how the computation of the $\mathrm{D_M}$ measure resembles that of the Haar-like features of Papageorgiou et al. (1998), widely used in the Computer Vision community, which we adopted and adapted in this work to evaluate whole-body gestures.

8. Ratio of Movement ($\mathrm{R_M}$) computes the relative amount of movement performed by one body part with respect to another. We define $\mathrm{R_M}$ as the ratio between two values reported by the

generalized quantity of gesture movement measure computed with two different sets of weights:

$$R_M = \frac{Q_M(\text{weights}_1)}{Q_M(\text{weights}_2)} \quad (8)$$

Again, there are many combinations of body parts for which relative movement may be compared, which we leave to the researcher's choice. In this work, we evaluate the ratio of hands movement relative to the entire movement of the body:

9. Ratio of Hands-to-Body Movement ($R_{Hands:Body}$):

$$R_{Hands:Body} = \frac{Q_M(\text{weights}_{hands})}{Q_M} \quad (9)$$

as well as the ratio between the movement of hands and legs:

10. Ratio of Hands-to-Legs Movement ($R_{Hands:Legs}$):

$$R_{Hands:Legs} = \frac{Q_M(\text{weights}_{hands})}{Q_M(\text{weights}_{legs})} \quad (10)$$

with weights arrays properly defined with 1.0 and 0.0 values. As we report in the Evaluation section, hand movements represent an important part of whole-body gestures. Also, knowing the ratio of movements of body limbs may be used for fatigue analysis; see the Evaluation section for discussion. Note that all R_M measures report dimensionless quantities.

3.2. Kinematic measures

Kinematic measures characterize users' whole-body gesture performance in the temporal domain by considering the timestamps t_i associated to each body posture constituting the gesture; see eq. 1. Kinematic measures are employed to understand how long it takes users to perform gestures, to monitor how performance improves with practice (*e.g.*, do users get equally faster with practice for all gesture types?), and to design whole-body gestures that are performed efficiently (*e.g.*, fast) by users. We present below performance time and speed, and point the reader to variations of these measures.

11. Performance Time (T) represents the total duration of the gesture performance, reported in seconds:

$$T = t_n - t_1 \quad (11)$$

where t_n and t_1 represent the timestamps associated to the last (P_n) and first (P_1) body postures of the gesture (eq. 1).

12. Average Gesture Speed (S) represents the magnitude of the average velocity at which body movement is performed. We compute gesture speed as the ratio between the quantity of gesture movement and performance time:

$$S = \frac{Q_M}{T} \quad (12)$$

for which we use $\lambda = J$.

By employing the generalized quantity of gesture movement measure at the numerator of eq. 12, we can compute the average speed of specific body parts, such as:

13. Average Hands Speed (S_{Hands}). We report on this measure in the Evaluation section of this paper.

The researcher interested in relative differences between the average speed of different body parts (*e.g.*, the difference in the speed at which the dominant hand moves compared to the speed of the non-dominant hand), may compute differences in speed using different weighting schemes, as shown previously for D_M and R_M. For space concerns, we omit actual definitions here, but we point the reader to the opportunity of exploring such options as well as to the variety of kinematic data that can be obtained with various weighting schemes, differences and ratios of the speed of movements of the various body parts.

3.3. Appearance-based measures

Appearance-based measures characterize the body postures that compose the whole-body gesture; see eq. 1. They are useful to understand how gestures decompose into simple units of movement, *i.e.*, body postures, and how units compare to each other. The importance of characterizing gesture appearance has been remarked by other researchers. For example, Bobick and Davis (2001) considered that *"The most primitive level, however, is movement - motion whose execution is consistent and easily characterized by a definite space-time trajectory in some feature space. Such consistency of execution implies that for a given viewing condition there is consistency of appearance. Put simply, movements can be described by their appearance."*

In this section, we define posture variation, diffusion, density, and body posture rate:

14. Body Posture Variation (BPV) computes the average deviation of body postures from the centroid posture of the whole-body gesture:

$$BPV = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\left\|p_j^i - \overline{p_j}\right\| \quad (13)$$

where n is the number of body postures (eq. 1), and $\overline{p_j}$ ($j = 1..J$) represent the joints of the average posture $\overline{P}$ of the gesture. The average posture is computed by averaging the x, y, and z coordinates of all the joints of all the body postures in the gesture. BPV is a measure of dispersion of the individual units that make up the movement from the centroid posture. Large values of this measure inform the practitioner that users adopted different body postures while they performed the gesture. We report body posture variation in physical units, such as meters.

15. Body Posture Diffusion (BPD) represents the maximum difference (or dissimilarity) between the body postures constituting the whole-body gesture:

$$BPD = \max_{1\le i<k\le n}\left\{\sum_{j=1}^{J}\left\|p_j^i - p_j^k\right\|\right\} \quad (14)$$

While BPV reports the amount of spread of individual body postures around a representative average, BPD reports their maximum dissimilarity with respect to each other. We report BPD values in physical units as well.

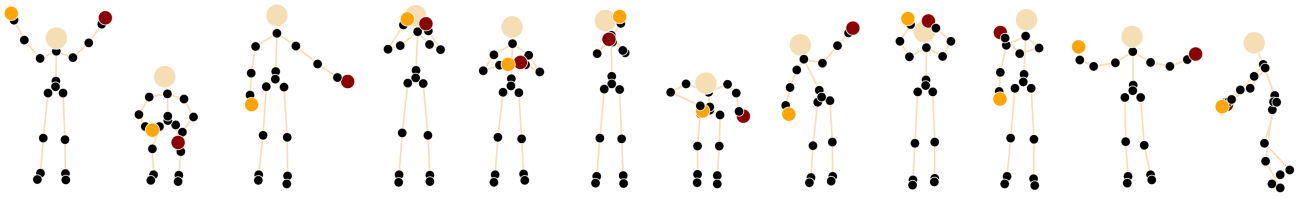

Figure 1: Body postures extracted from a public dataset (Fothergill et al., 2012) to illustrate each of the 12 whole-body gestures evaluated in this work. From left to right: "start music", "crouch/hide", "next menu", "put on goggles", "wind up music", "shoot pistol", "end music (bow)", "throw", "protest", "change weapon", "move up tempo", and "kick". In total, we evaluated 5,654 body gestures collected from 30 participants, consisting in a total number of 581,246 body frames. Note: the right hand is shown in dark red.

16. Body Posture Density (BPρ) represents the variation of body postures over the volume of the space in which body movement was produced and recorded:

$$\text{BP}\rho = \frac{\text{BPV}}{\text{GV}} \tag{15}$$

We report BPρ values in physical units: meters of variation in body posture per one cubic meter of space.

17. Body Posture Rate (BPR) represents the variation of body postures over the time duration during which body movement was produced and recorded:

$$\text{BPR} = \frac{\text{BPV}}{\text{T}} \tag{16}$$

We report BPR values in physical units (m/s).

Body Posture Density and Rate are measures that integrate information about users' body appearance during gesture production with spatial and kinematic description of the whole-body gesture. The next section shows how these measures are able to reveal more findings about users' whole-body gesture performance than the measures from which they were derived.

4. EVALUATION: SHOWCASING MEASURES APPLICATION AND REVEALING PRACTICAL ASPECTS OF WHOLE-BODY GESTURE PERFORMANCE

In this section, we show the usefulness of our set of measures to characterize users' whole-body gesture performance by computing and evaluating the measures for the gesture types of the public dataset of Fothergill et al. (2012). The dataset contains 5,654 records of 12 distinct body gestures (see Figure 1) that were collected from 30 participants using the Microsoft Kinect sensor, with a total number of 581,246 body frames[1]. Please note that it is not our goal to comprehensively describe users' performance for this dataset or for these specific gesture types, but rather we use this section to showcase the convenience of our measures and their capability to reveal practical aspects of whole-body gesture performance.

[1] `http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/`

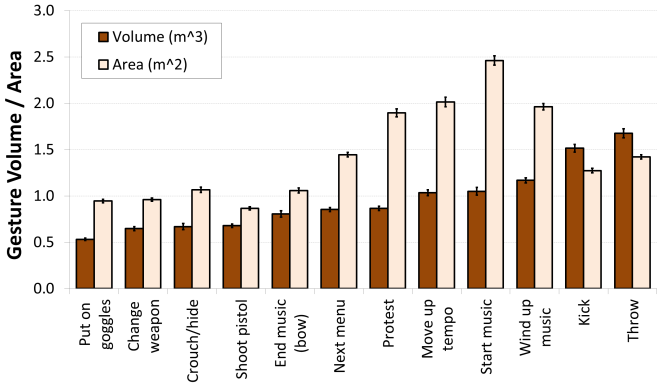

Figure 2: Average Gesture Volume and Gesture Area values reported for the whole-body gestures dataset of Fothergill et al. (2012). Notes: error bars show 95% CIs; gestures are shown in ascending order of their volume.

4.1. Spatial measures

We found a significant effect of gesture type on both Gesture Volume ($\chi^2(11)=1614.070$, $p<.001$) and Gesture Area ($\chi^2(11)=1924.533$, $p<.001$), with "throw" showing the largest volume (1.68 m^3) and "start music" the largest area (2.46 m^2); see Figure 2. At the opposite end of the scale were "put on goggles" (with a volume of 0.53 m^3) and "shoot pistol" (area of 0.87 m^2). The values reported by these two measures confirm intuition as "throw" requires the hand to move from back to front and, likely, the body leaning to the front as well, which gives large depth to the overall movement. However, users normally stand still during "put on goggles," with only the hands moving toward the eyes, which explains low volume for this gesture type. Similarly, there is considerably more arm movement along the x and y axes for "start music," during which users raised and stretched their arms, see Fothergill et al. (2012) (p. 1740), than for "shoot pistol" that only requires movements of hands in front of the body. In average, all the gestures were contained within a volume of 1.0 m^3 (SD=0.35 m^3) with an average area of 1.5 m^2 (SD=0.52 m^2). Figure 2 shows average volume and area values side by side to facilitate easy interpretation of the depth dimension of each gesture type. Note how these two measures are able to characterize the amplitude of users' gesture movements precisely, despite their simple definitions. A correlation analysis showed that GV and GA deliver distinct information (Pearson's $r_{(N=12)}=.386$, $p>.05$, *n.s.*).

We also found a significant effect of gesture type on the

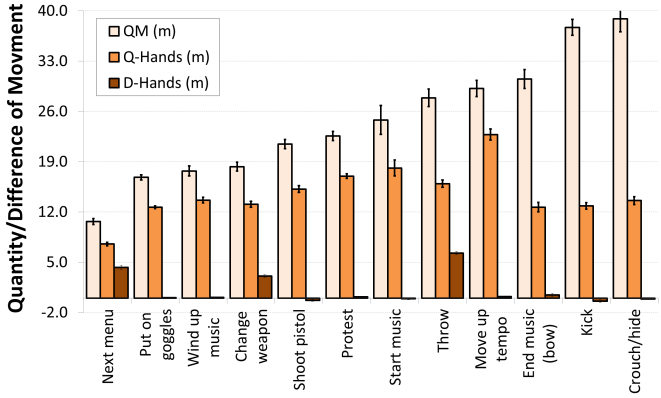

Figure 3: Average Quantity and Difference of Movement values reported for the whole-body gestures dataset of Fothergill et al. (2012). Notes: error bars show 95% CIs; gestures are shown in ascending order of Q_M.

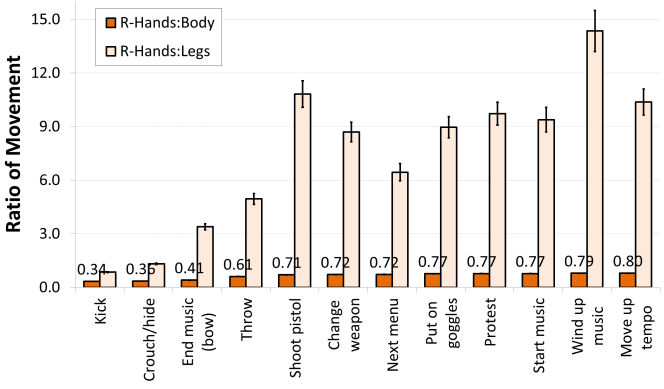

Figure 4: Average Ratio of Movement reported for the whole-body gestures dataset of Fothergill et al. (2012). Notes: error bars show 95% CIs; gestures are shown in ascending order of the hands-to-body ratio of movement.

quantity of gesture movement Q_M ($\chi^2(11)=1483.394$, $p<.01$); see Figure 3. The largest amount of movement was generated when performing the "crouch" and "kick" gestures: 38.9 m and 37.7 m, respectively, for which a post-hoc Wilcoxon signed-rank test did not detect any significant difference ($Z=-1.146$, $p>.05$, *n.s.*). The "next menu" gesture presented the smallest amount of body movement (10.7 m), being performed with one hand only. We also found a significant effect of gesture type on the quantity of hands movement Q_{Hands} ($\chi^2(11)=1144.586$, $p<.01$), with an average movement of 14.6 m per gesture (SD=3.7 m). When compared to the amount of body movement overall (average 24.7 m, SD=8.5 m), we found that hands accounted for more than 60% of all body movement. Analysis of individual gesture types showed large hand movements occurring for "throw", "protest", and "move up tempo". Gesture type also affected significantly the difference in the amount of movement performed by the two hands D_{Hands} ($\chi^2(11)=997.273$, $p<.01$), with the right hand traveling with 1.1 m more on average than the left hand (SD=2.2 m). Out of the 12 gesture types, only 3 stand out in Figure 3 with large differences in terms of hand movements, *i.e.*, "throw", "change weapon", and "next menu", showing participants' preferences for using the right hand to perform these gestures. All the remaining D_{Hands} values are below 0.46 m, which shows equal bimanual involvement, which might have been either explicit (*i.e.*, intended by users during performance) or implicit (*i.e.*, hands move together with the body because of inertia mechanisms); a finding that the practitioner can now use as a start point for new investigations in this direction, *e.g.*, how much does body inertia affect the appearance of movement?

When analyzing the ratios of movement, we found that hands account for 65% of the entire body movement of the gestures in the dataset. Gesture type had a significant effect on the hands-to-body movement ratio $R_{Hands:Body}$ ($\chi^2(11)=1620.288$, $p<.01$). Some gestures used hands extensively, such as "wind up music" (ratio value of 0.79), "put on goggles" (0.77) or "shoot pistol" (0.71), while others required hands less, such as "crouch/hide" (0.36) or "kick" (0.34); see Figure 4. These results show the importance of hand movements overall in the production of whole-body gestures, function of gesture type. We also found that "throw," a gesture defined by a remarkable hand movement, had a value of 0.61, which is generally high, yet lower than other gesture types and explainable by the general body movement accompanying the hand, *i.e.*, the body leans backward and forward during the throw. By looking at the ratio of the movement of hands relative to the movement of legs, we can characterize gesture performance at even finer levels of detail; see Figure 4. Overall, gesture type had a significant effect over this measure as well ($\chi^2(11)=1195.233$, $p<.01$). "Shoot pistol" presented a large value (10.8), showing high relevancy of hands over legs movement during this gesture, as expected. On the other hand, "crouch/hide" (1.3) showed an almost equal amount of movement of hands and legs, while the "kick" ratio was subunitary (0.8), showing more leg than hand movements. The practitioner can now use these results to inform further analysis of the fatigue involved by whole-body gesture production. For instance, gestures that require all limbs to move will likely be perceived differently in terms of difficulty than gestures involving only few limbs. Also, different ratios of limb movements may be perceived differently by users in terms of the relative difficulty of producing body movements. For some interaction contexts (*e.g.*, when there is little space available for comfortable or safe movement) or for some user categories (*e.g.*, age groups or motor impairments), gestures with various ratios of limb movements may be preferred and their performance evaluated with our set of measures.

4.2. *Kinematic measures*

We found a significant effect of gesture type on Performance Time ($\chi^2(11)=729.154$, $p<.01$) as well as on users' Average Speed ($\chi^2(11)=1460.858$, $p<.01$). Gestures took on average 3.5 seconds to execute (SD=0.6), at an average speed of 0.36 m/s (SD=0.12); see Figure 5. Some gestures were fast, such as "kick" (0.6 m/s) or "crouch/hide" (0.5 m/s), suggesting action types for which quick response is needed from users (otherwise, negative consequences may occur, such as loosing points in a video game). Other gesture types were performed more slowly when judged at the scale of the whole body, such

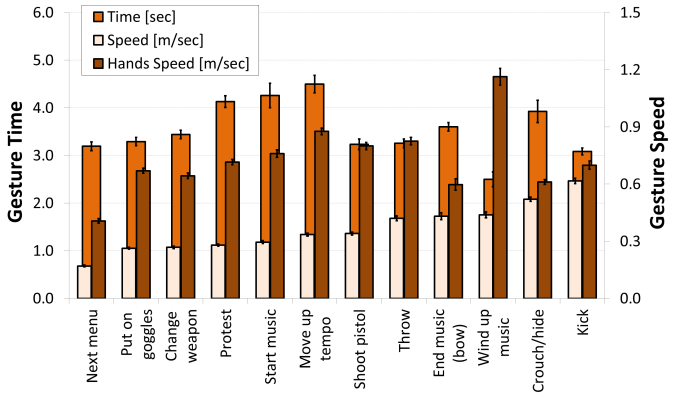

Figure 5: Average Performance Time and Speed reported for the whole-body gestures dataset of Fothergill et al. (2012). Notes: error bars show 95% CIs; gestures are shown in ascending order of their average speed.

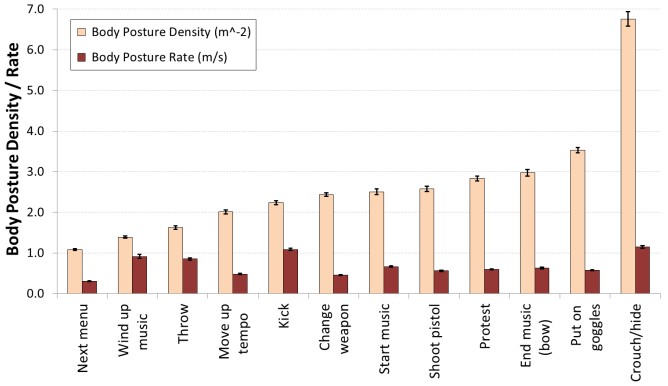

Figure 7: Average Body Posture Density and Rate values reported for the set of whole-body gestures of Fothergill et al. (2012). Notes: error bars show 95% CIs; gestures are shown in ascending order of their density.

as "put on goggles" (0.26 m/s) or "change weapon" (0.27 m/s), suggesting more carefulness from users during the performance of more precise movements. Speed analysis focused on hands revealed even more findings. For example, hands were two times faster than the overall speed of the whole body, with an average of 0.73 m/s (SD=0.18). Even though "put on goggles" and "change weapon" were evaluated as slow executions when judging speed at the scale of the body (all joints considered), they were fast in terms of hands movement (0.67 and 0.64 m/s, respectively), a finding that we obtained by employing weights and the generalized quantity of movement Q_M in eq. 12.

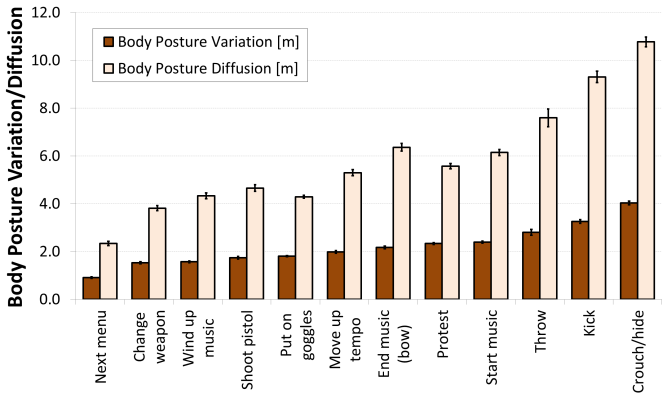

Figure 6: Average Body Posture Variation and Diffusion values reported for the set of whole-body gestures of Fothergill et al. (2012). Notes: error bars show 95% CIs; gestures are shown in ascending order of BPV.

4.3. Appearance-based measures

We found a significant effect of gesture type on both Body Posture Variation ($\chi^2(11)=1694.525$, $p<.01$) and Body Posture Diffusion ($\chi^2(11)=1738.918$, $p<.01$). Body postures varied from their centroids with 2.2 m on average (SD=0.8 m), with an average Posture Diffusion of 5.8 m (SD=2.4 m); see Figure 6. Some gestures were more dynamic in appearance, as indicated by larger BPV values. For example, "crouch/hide" and "kick" are movements composed of body postures that are more different from their centroid than are the postures of "next menu" or "shoot pistol" (4.0 and 3.3 m versus 0.9 and 1.7 m); see Figure 6. When analyzing Body Posture Diffusion, we found that individual body postures may vary greatly for some gesture types (*e.g.*, 10.8 m for "crouch/hide," all joints considered), suggesting more joint movements required to perform those gesture types and, consequently, more ways for users to perform them. The practitioner can use these results for further investigations on what caused such variations that, ultimately, may impact negatively users' adoption of such gesture types.

The Body Posture Density (BPρ) and Body Posture Rate (BPR) measures revealed more findings about the appearance of the body during gesture production with respect to the volume of space and the duration of the gesture. For instance, we found that Density varied between 1.1 m per cubic meter for the "next menu" gesture and 6.8 m per cubic meter for "crouch/hide". On average, the variation in how body postures changed with respect to the volume of space in which those variations occurred was of 2.7 m for each cubic meter of volume encompassing movement; see Figure 7. Gestures with equal variation in body posture, such as "change weapon", "wind up music", "shoot pistol", and "put on goggles" (ranks 2, 3, 4, and 5 in the ordered list of gestures shown in Figure 6) are better differentiable by their Density values (ranks 2, 6, 8, and 11 in Figure 7), even when they are performed in similar volumes of space (see Figure 2). A Friedman test showed a statistically significant effect of gesture type on BPρ ($\chi^2(11)=1738.918$, $p<.01$). Body postures varied in their appearance at a rate of 0.69 m/s, with the slowest rate obtained for "next menu" (0.30 m/s) and the highest for "crouch/hide" (1.15 m/s); see Figure 7. Gesture type had a significant effect on BPR ($\chi^2(11)=1738.918$, $p<.01$). Just like Density, Rate is able to differentiate between gestures with similar variation in body posture: "change weapon", "wind up music", "shoot pistol", and "put on goggles" are now ranked second, fourth, fifth, and tenth by their rate; see Figure 7. Furthermore, a correlation analysis showed that BPρ and BPR deliver distinct information (Pearson's $r_{(N=12)}=.466$, $p>.05$, *n.s.*).

4.4. Relationship to gesture recognition performance

The gesture measures that we propose in this paper were designed to help the practitioner develop a better understanding of how users actually perform whole-body gestures in terms of their spatial, kinematic, and appearance characteristics. The previous sections showed how evaluating these measures on actual gestures reveals interesting aspects of human gesture performance, such as computing the lower and upper margins for the volume of space in which movement takes place, the overall quantity of movement produced, or the relationships between the movement of body parts, such as hands and legs, and the overall movement of the whole body. However, it is also interesting to understand how our measures relate to gesture recognition performance. To this end, we conducted a recognition experiment for the gesture dataset of Fothergill et al. (2012) that we employed in the previous sections to evaluate our gesture performance measures. For this experiment, we implemented the Nearest-Neighbor (1-NN) classifier (Webb, 2002) and the Dynamic Time Warping (DTW) function (Myers and Rabiner, 1981; Bodiroža et al., 2013; Ferguson et al., 2014; Jiang et al., 2015; Lou et al., 2017) to evaluate the dissimilarity between whole-body gestures following Vatavu (2012a) (p. 87):

$$\text{candidate C} \in \text{Class}(T_j) \text{ if } T_j = \underset{T_i \in \mathcal{T}}{\arg\min} \{\text{DTW}(C, T_i)\} \quad (17)$$

where T_i are gesture templates from the training set $\mathcal{T}$ and the DTW function between gesture candidate C and template T is iteratively computed using a cost matrix ζ that optimally aligns points C_i to T_j to minimize the overall Euclidean distance between the two gestures, as follows:

$$\begin{aligned} \text{DTW}(C, T) &= \zeta_{|C|,|T|}, \text{ where} \quad (18)\\ \zeta_{1,1} &= \|C_1 - T_1\| \\ \zeta_{1,j} &= \zeta_{1,j-1} + \left\|C_1 - T_j\right\| \\ \zeta_{i,1} &= \zeta_{i-1,1} + \|C_i - T_1\| \\ \zeta_{i,j} &= \min\left\{\zeta_{i-1,j-1}, \zeta_{i-1,j}, \zeta_{i,j-1}\right\} + \left\|C_i - T_j\right\| \end{aligned}$$

where $|C|$ and $|T|$ represent the number of points of the candidate and template gestures. Prior to classification, gestures were normalized as follows: (1) all gestures were uniformly resampled in the time domain into a fixed number of $n = 32$ body postures, (2) they were scaled down to the unit box, and (3) translated to the origin so that the centroid of all sequences of movement was zero. We applied these normalization steps by following previous practices of gesture preprocessing from the literature (Anthony and Wobbrock, 2010; Vatavu et al., 2012; Wobbrock et al., 2007). Recognition rates were computed for each gesture type in a *user-independent* training scenario by varying the number of participants employed for training, as follows. P participants were randomly selected for training from the available 30 participants of the dataset of Fothergill et al. (2012) and their gesture samples were added to the training set. Another participant, different from the first P, was randomly selected for testing and his/her gestures were submitted to classification. This procedure was repeated for 100 times for each value of P, where P varied from 1 to 2, 4, 8, and 16 training participants. This procedure follows the standard practice of evaluating gesture recognizers; see (Anthony and Wobbrock, 2010; Vatavu et al., 2012; Wobbrock et al., 2007).

Figure 8, left illustrates the user-independent recognition rates obtained for the whole-body gesture dataset of (Fothergill et al., 2012), function of the number of participants employed for training. On average, recognition rate was 90.4% (SD=29.5%) for all gesture types and all training conditions, computed from a total number of 94,500 classification trials. Recognition rate increased from 83.5% (SD=37.2%) when using training data from one participant only (P=1) to 95.2% (SD=21.4%) with training data collected from P=16 participants (representing +11.7% more accuracy on average). A Cochran's Q test conducted on classification results interpreted as success rates confirmed a statistically significant effect of the number of training participants P on recognition rate ($\chi^2(4)$=147.416, p<.001). Post-hoc Wilcoxon signed-rank tests (Bonferroni corrected at p=.05/4=.0125) showed significant differences between training conditions with 1 and 2 participants, 2 and 4, and 8 and 16 training participants. Figure 8, right shows the recognition rates obtained for each gesture type, which varied from a minimum of 75.8% (SD=42.8%) for "move up tempo" to 99.8% (SD=4.8%) for "crouch/hide". A Cochran's Q test revealed a statistically significant effect of gesture type on recognition rate ($\chi^2(11)$=500.726, p<.001). Overall, eight out of the twelve gesture types were recognized with over 90% accuracy, out of which four gesture types ("crouch/hide", "kick", "put on goggles", and "end music (bow)" were classified with over 95% accuracy.

To understand the relationship between our gesture performance measures and the recognition rates obtained with the DTW 1-NN gesture classifier, we computed Pearson's r correlation coefficients; see Table 1 for coefficients listed in decreasing order of their absolute magnitude. We found that the Quantity of Hands Movement (Q_{Hands}) and the two ratio measures involving hands movement ($R_{Hands:Body}$ and $R_{Hands:Legs}$) had the highest correlations with the average recognition rate ($r_{(N=12)}$=−.701, −.678, and −.616, all p<.05), but also with recognition rates obtained for specific training conditions, from P=2 to P=16 (at p<.05 and p<.01). These results reveal the importance of hands' movements to discriminate between the various whole-body gesture types of the dataset that we evaluated. The results can be confidently generalized to other whole-body gesture types, knowing the importance of hand movements as principal motor implementers of the imagery of thought (McNeill, 1992) and hand gestures providing the imagistic content for speech in the context of a unitary human language system (Xiong and Quek, 2006). All the three measures correlated negatively with recognition rate: more hand movement (measured individually with Q_{Hands} or in relation to other body parts using $R_{Hands:Body}$ and $R_{Hands:Legs}$) made gestures less recognizable. This result is explained by our preprocessing steps (*e.g.*, resampling all gestures into the same number of body postures) and the specific functionality of the DTW algorithm that strives to optimize alignments between two time series. We also found that Gesture Area (GA) also correlated highly with the average recognition rate ($r_{(N=12)}$=−.576, p<.05), but Gesture Volume (GV) did not ($r_{(N=12)}$=−.106, p>.05, *n.s.*). This result shows

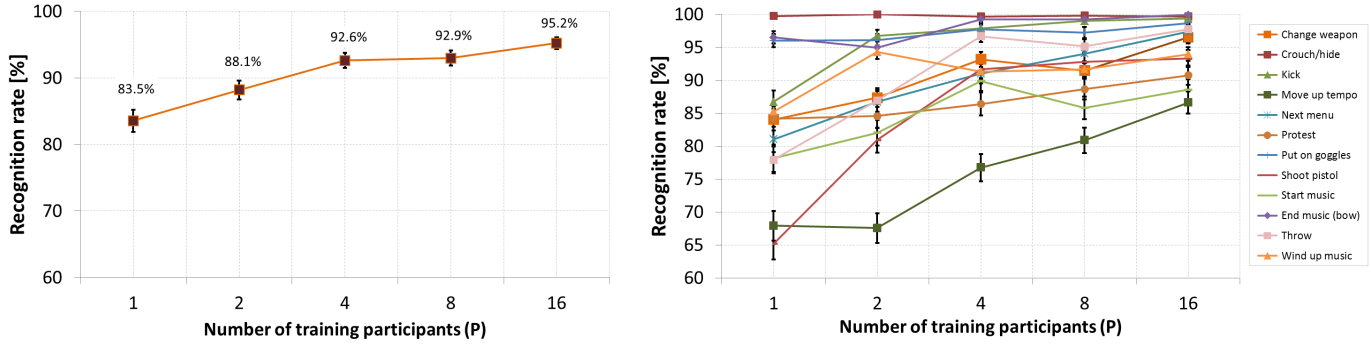

Figure 8: User-independent recognition rates for the set of whole-body gestures of Fothergill et al. (2012): average recognition rates across all gesture types and training conditions (left) and recognition rates per gesture type (right). Notes: error bars show 95% CIs.

Measures of whole-body gesture performance	**Recognition rate**†					
	average (all P)	**P=1**	**P=2**	**P=4**	**P=8**	**P=16**
Quantity of Hands Movement (Q_{Hands})	−.701*	−.505	−.692*	−.663*	−.736**	−.784**
Ratio of Hands to Body Movement ($R_{Hands:Body}$)	−.678*	−.531	−.613*	−.680*	−.748**	−.692*
Ratio of Hands to Legs Movement ($R_{Hands:Legs}$)	−.616*	−.503	−.502	−.622*	−.682*	−.678*
Gesture Area (GA)	−.576*	−.337	−.473	−.638*	−.733**	−.763**
Body Posture Rate (BPR)	.512	.415	.593*	.503	.475	.362
Body Posture Density (BPρ)	.487	.584*	.444	.408	.403	.305
Performance Time (T)	−.450	−.145	−.536	−.497	−.537	−.563
Average Speed (S)	.435	.302	.479	.442	.474	.379
Body Posture Diffusion (BPD)	.410	.366	.396	.426	.410	.302
Body Posture Variation (BPV)	.387	.375	.375	.392	.364	.252
Hands Speed (S_{Hands})	−.348	−.337	−.200	−.323	−.393	−.447
Quantity of Movement (Q_M)	.243	.239	.206	.250	.265	.172
Gesture Volume (GV)	−.106	−.251	−.059	−.011	−.040	−.030
Difference of Hands Movement (D_{Hands})	.000	−.149	−.091	.107	.077	.255

† Recognition rates are reported function of the number of training participants P from which gesture samples were collected; see Figure 8.
* Correlation is significant at $p = .05$.
** Correlation is significant at $p = .01$.

Table 1: Pearson correlation coefficients ($N = 12$) computed between our set of whole-body gesture performance measures and gesture recognition rates, function of the number of training participants P. Note: gesture measures are listed in descending order of the absolute magnitude of the correlation coefficient.

that gestures were performed with preponderance in the plane facing the sensor and that the z dimension (movements forward and backward in front of the sensor) had little importance for discriminating between gesture types. This result rewards our intuition to consider Gesture Area as a distinct spatial measure next to Gesture Volume (see the spatial measures section), as a large majority of whole-body gestures are performed *in front* of a display with movement mostly taking place along the x and y axes; see for instance the standard gesture types implemented for the XBox 360 console (Microsoft, 2016b). Nevertheless, other gesture types, such as walking toward the display, will need the depth information to be discriminated from other gestures, and Gesture Volume will most likely catch the differences in depth between various gesture types.

Other gesture measures correlated moderately (although not statistically significant) with recognition rate, such as Body Posture Rate (.512), Body Posture Density (.487), Performance Time (−.450) and Average Speed (.435). It is interesting to note that only spatial measures correlated significantly with recognition rate (at either $p<.01$ or $p<.05$), a result that is explained by the specifics of our gesture classifier: the DTW function employs exclusively the spatial characteristics of gesture points to compute the dissimilarity between gestures; see eq. 19. However, it is likely that significant correlations will be observed between our kinematic and appearance-based measures and recognition rates obtained with other gesture classifiers, such as statistical classifiers that rely on gesture features computed from timestamps or the appearance of the body, such as the features from (Bobick and Davis, 2001; Masoud and Papanikolopoulos, 2003; Weinland et al., 2006; Chen et al., 2006; Howe, 2004). While our goal in this section was simply to showcase the relationships between our measures of whole-body gesture performance and gesture recognition results, future work will likely reveal more interesting findings in this direction. Toward this goal, we offer practitioners a large palette of whole-body gesture measures that they can select from and even particularize for their specific evaluation scenarios or according to the specifics of their gesture classifiers.

4.5. Summary

We showcased in this section how to apply our set of measures in practice to characterize users' whole-body gesture

movement on a public dataset (Fothergill et al., 2012), for which our results can be easily reproduced. We evaluated human movement and gestures using human-readable concepts, such as area, volume, and quantity, for which we reported several gesture findings. While it was not our goal to comprehensively describe users' performance in this dataset or for these specific gesture types, we used the opportunity provided by this section to showcase the convenience of our measures and their capability to reveal practical aspects of whole-body gesture performance. We also showed that spatial measures correlated highly with the recognition rates delivered by a gesture classifier implementing a dissimilarity function based on the spatial characteristics of human movement. Researchers and practitioners can employ our set of measures in a similar manner to examine their users' whole-body gestures and they can even adapt our measures to suit their particular analysis needs; *e.g.*, quantity of movement, difference in gesture movement, ratio of movements, etc. represent customizable measures that allow for particularization to suit specific scientific investigation goals about whole-body gestures.

5. CONCLUSION

We introduced in this paper a set of measures to evaluate users' whole-body gesture performance. We showed how our measures can be employed to characterize whole-body gestures by reporting and analyzing their values on a large gesture dataset. To assist with computing the measures, we release in the community BOGArT, the Body Gestures Analysis Toolkit, freely available to download from `http://www.eed.usv.ro/~vatavu`. BOGArT computes all our measures and exports results into `.csv` files, supported by mostly all data visualization and statistical analysis software. It is important to see this work as a first step toward designing *human-readable* gesture measures. Obviously, our set of measures is not exhaustive, and future work and practice with the measures will expand it further. For this reason, we also provide the source code for BOGArT (C#, .NET Framework 4.5, project built in Visual Studio Community 2015) to foster further development and exploration of new whole-body gesture measures in the community. We believe that the contributions of this work will empower researchers and practitioners with new numerical tools to understand users' gesture performance better and, consequently, to inform improved designs of whole-body gesture interfaces.

ACKNOWLEDGMENTS

The author would like to thank the anonymous reviewers for their valuable comments and suggestions that were incorporated in the final version of this manuscript. This research was supported by the project *"Computational psychology of human movement to understand gestures and body kinesics"* (PSYKINESICS), financed by UEFISCDI, Romania. Research was conducted in the Machine Intelligence and Information Visualization Lab (MintViz) of the MANSiD Research Center. The infrastructure was provided by the University of Suceava and was partially supported from the project *"Integrated center for research, development and innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for fabrication and control"*, No. 671/09.04.2015, Sectorial Operational Program for Increase of the Economic Competitiveness, co-funded from the European Regional Development Fund.

References

Abbie, M. (1974). Movement notation. *Australian Journal of Physiotherapy*, 20(2):61–69.

Aggarwal, J. K. and Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440.

Ahad, M., Ogata, T., Tan, J., Kim, H., and Ishikawa, S. (2008). Motion recognition approach to solve overwriting in complex actions. In *Proceedings of FG '08, the 8th IEEE International Conference on Face and Gestures*, pages 1–6.

Ali, S. and Shah, M. (2010). Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):288–303.

Anthony, L. and Wobbrock, J. O. (2010). A lightweight multistroke recognizer for user interface prototypes. In *Proceedings of Graphics Interface 2010*, GI '10, pages 245–252, Toronto, Ont., Canada, Canada. Canadian Information Processing Society.

Aslan, I., Primessnig, F., Murer, M., Moser, C., and Tscheligi, M. (2013). Inspirations from honey bees: Exploring movement measures for dynamic whole body gestures. In *Proceedings of ITS'13, the ACM International Conference on Interactive Tabletops and Surfaces*, pages 421–424.

Batra, D., Chen, T., and Sukthankar, R. (2008). Space-time shapelets for action recognition. In *Proceedings of the 2008 IEEE Workshop on Motion and Video Computing*, WMVC '08, pages 1–6, Washington, DC, USA. IEEE Computer Society.

Baudisch, P., Pohl, H., Reinicke, S., Wittmers, E., Lühne, P., Knaust, M., Köhler, S., Schmidt, P., and Holz, C. (2014). Imaginary reality basketball: A ball game without a ball. In *Proceedings of CHI EA '14*, pages 575–578. ACM.

Bobick, A. (1997). Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358):1257–1265.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267.

Bodiroža, S., Doisy, G., and Hafner, V. V. (2013). Position-invariant, real-time gesture recognition based on dynamic time warping. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, pages 87–88, Piscataway, NJ, USA. IEEE Press.

Bradski, G. R. and Davis, J. W. (2002). Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184.

Buxton, B. (2011). Gesture based interaction (chapter 14). http://www.billbuxton.com/input14.Gesture.pdf.

Chen, H.-S., Chen, H.-T., Chen, Y.-W., and Lee, S.-Y. (2006). Human action recognition using star skeleton. In *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks*, VSSN '06, pages 171–178. ACM.

Choensawat, W., Nakamura, M., and Hachimura, K. (2016). *Dance Notations and Robot Motion*, chapter Applications for Recording and Generating Human Body Motion with Labanotation, pages 391–416. Springer International Publishing.

Connell, S., Kuo, P.-Y., Liu, L., and Piper, A. M. (2013). A Wizard-of-Oz elicitation study examining child-defined gestures with a whole-body interface. In *Proceedings of the 12th International Conference on Interaction Design and Children*, IDC '13, pages 277–280. ACM.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893.

Dania, A., Tyrovola, V., Koutsouba, M., and Hatziharistos, D. (2013). Labankido: The evaluation of a multimedia tool designed for the teaching of

basic skills and concepts in dance education. *The International Journal of Sport and Society*, 3(3):137–148.

Davis, J. (2001). Hierarchical motion history images for recognizing human motion. In *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video*, pages 39–46.

Dezfuli, N., Khalilbeigi, M., Huber, J., Müller, F., and Mühlhäuser, M. (2012). PalmRC: Imaginary palm-based remote control for eyes-free television interaction. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroiTV '12, pages 27–34. ACM.

Efros, A., Berg, A., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, pages 726–733.

Ferguson, S., Schubert, E., and Stevens, C. J. (2014). Dynamic dance warping: Using dynamic time warping to compare dance movement performed under different conditions. In *Proceedings of the 2014 International Workshop on Movement and Computing*, MOCO '14, pages 94:94–94:99, New York, NY, USA. ACM.

Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of CHI '12, the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Gorelick, L., Galun, M., Sharon, E., Basri, R., and Brandt, A. (2006). Shape representation and classification using the poisson equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1991–2005.

Grim-Feinberg, K. and Santos, M. F. A. W. (2015). Labanotation and the study of human movement in anthropology. In *Proceedings of the Congress on Research in Dance Conference*, pages 59–67.

Guo, H., Miao, Z., Zhu, F., Zhang, G., and Li, S. (2014). *Automatic Labanotation Generation Based on Human Motion Capture Data*, volume 483, pages 426–435. Springer Berlin Heidelberg.

Gustafson, S., Bierwirth, D., and Baudisch, P. (2010). Imaginary interfaces: Spatial interaction with empty hands and without visual feedback. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 3–12. ACM.

Gustafson, S., Holz, C., and Baudisch, P. (2011). Imaginary phone: Learning imaginary interfaces by transferring spatial memory from a familiar device. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 283–292. ACM.

Hachimura, K. and Ohno, Y. (1987). A system for the representation of human body movement from dance scores. *Pattern Recognition Letters*, 5(1):1–9.

Holz, C. and Wilson, A. (2011). Data miming: Inferring spatial object descriptions from human gesture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 811–820. ACM.

Hong, D. and Woo, W. (2006). A 3D vision-based ambient user interface. *International Journal of HumanComputer Interaction*, 20(3):271–284.

Howe, N. R. (2004). Silhouette lookup for automatic pose tracking. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, CVPRW '04, pages 15–22. IEEE Computer Society.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187.

Hutchinson, A. (1991). *Labanotation or Kinetography Laban: The System of Analyzing and Recording Movement*. Routledge/Theatre Arts Book, New York.

Jaimes, A. and Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1-2):116–134.

Jiang, F., Zhang, S., Wu, S., Gao, Y., and Zhao, D. (2015). Multi-layered gesture recognition with kinect. *J. Mach. Learn. Res.*, 16(1):227–254.

Karhu, O., Härkönen, R., Sorvali, P., and Vepsäläinen, P. (1981). Observing working postures in industry: Examples of OWAS application. *Applied Ergonomics*, 12(1):13–17.

Karhu, O., Kansi, P., and Kuorinka, I. (1977). Correcting working postures in industry: A practical method for analysis. *Applied Ergonomics*, 8(4):199–201.

Kendon, A. (2000). Language and gesture: Unity or duality. In *D. McNeill (Ed.) Language and Gesture: Window into Thought and Action*, pages 47–63. Cambridge University Press, Cambridge.

Kordts, B., Altakrouri, B., and Schrader, A. (2015). Capturing and analysing movement using depth sensors and labanotation. In *Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS '15, pages 132–141, New York, NY, USA. ACM.

Kurtenbach, G. and Hulteen, E. (1990). Gestures in human-computer communications. In *B. Laurel (Ed.), The Art of Human Computer Interface Design*, pages 309–317. Addison-Wesley.

Laiyang, L. and Junjun, G. (2014). Fingers' movement analysis based on labanotation. In *Proceedings of the IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA)*, pages 1307–1311.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123.

Loke, L. and Robertson, T. (2009). Design representations of moving bodies for interactive, motion-sensing spaces. *International Journal of Human-Computer Studies*, 67(4):394–410.

Lou, Y., Wu, W., Vatavu, R.-D., and Tsai, W.-T. (2017). Personalized gesture interactions for cyber-physical smart-home environments. *Science China Information Sciences*, 60(7):072104:1–15.

Maes, P.-J., Amelynck, D., Lesaffre, M., Leman, M., and Arvind, D. (2013). The "Conducting Master": An interactive, real-time gesture monitoring system based on spatiotemporal motion templates. *International Journal of HumanComputer Interaction*, 29(7):471–487.

Masoud, O. and Papanikolopoulos, N. (2003). A method for human action recognition. *Image and Vision Computing*, 21(8):729–743.

McNeill, D. (1992). *Hand and Mind: What gestures reveal about thought*. The University of Chicago Press, Chicago.

Microsoft (2014). Kinect for Windows: Human Interface Guidelines. http://go.microsoft.com/fwlink/?LinkID=247735.

Microsoft (2016a). JointType enumeration. https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx.

Microsoft (2016b). Kinect gestures. https://support.xbox.com/en-US/xbox-360/accessories/body-controller.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126.

Morris, M. R. (2012). Web on the wall: Insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '12, pages 95–104. ACM.

Myers, C. and Rabiner, L. (1981). A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical J.*, 60(7):1389–1409.

Nebeling, M., Huber, A., Ott, D., and Norrie, M. C. (2014). Web on the wall reloaded: Implementation, replication and refinement of user-defined interaction sets. In *Proceedings of the 9th ACM International Conference on Interactive Tabletops and Surfaces*, ITS '14, pages 15–24. ACM.

Nielsen, M., Strring, M., Moeslund, T., and Granum, E. (2004). A procedure for developing intuitive and ergonomic gesture interfaces for hci. In Camurri, A. and Volpe, G., editors, *Gesture-Based Communication in Human-Computer Interaction*, volume 2915 of *Lecture Notes in Computer Science*, pages 409–420. Springer Berlin Heidelberg.

Papageorgiou, C., Oren, M., and Poggio, T. (1998). A general framework for object detection. In *Proceedings of the 6th International Conference on Computer Vision*, pages 555–562.

Piana, S., Stagliano, A., Camurri, A., and Odone, F. (2013). A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. In *IDGEI Int. Workshop*.

Polana, R. and Nelson, R. (1997). Detection and recognition of periodic, non-rigid motion. *International Journal of Computer Vision*, 23(3):261–282.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1-2):4–18.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vision Computing*, 28(6):976–990.

Priel, V. Z. (1974). A numerical definition of posture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 16:576–584.

Rekik, Y., Vatavu, R.-D., and Grisoni, L. (2014). Understanding users' perceived difficulty of multi-touch gesture articulation. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 232–239, New York, NY, USA. ACM.

Rosenhahn, B., Klette, R., and Metaxas, D., editors (2008). *Human Motion: Understanding, Modelling, Capture, and Animation*. Springer.

Sevdalis, V. and Keller, P. E. (2011). Captured by motion: Dance, action understanding, and social cognition. *Brain and Cognition*, 77(2):231–236.

Silpasuwanchai, C. and Ren, X. (2014). Jump and shoot!: Prioritizing primary and alternative body gestures for intense gameplay. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 951–954, New York, NY, USA. ACM.

Stern, H., Wachs, J., and Edan, Y. (2008). Optimal consensus intuitive hand gesture vocabulary design. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, pages 96–103.

Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488.

Vatavu, R. and Wobbrock, J. (2016). Between-subjects elicitation studies: Formalization and tool support. In *Proceedings of CHI'16, the 34th ACM SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. ACM.

Vatavu, R.-D. (2012a). Nomadic gestures: A technique for reusing gesture commands for frequent ambient interactions. *Journal of Ambient Intelligence and Smart Environments*, 4(2):79–93.

Vatavu, R.-D. (2012b). User-defined gestures for free-hand TV control. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroiTV '12, pages 45–48, New York, NY, USA. ACM.

Vatavu, R.-D. (2013a). A comparative study of user-defined handheld vs. free-hand gestures for home entertainment environments. *International Journal of Ambient Intelligence and Smart Environments*, 5(2):187–211.

Vatavu, R.-D. (2013b). The impact of motion dimensionality and bit cardinality on the design of 3D gesture recognizers. *Int. J. Hum.-Comput. Stud.*, 71(4):387–409.

Vatavu, R.-D. (2015). Audience silhouettes: Peripheral awareness of synchronous audience kinesics for social television. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '15, pages 13–22, New York, NY, USA. ACM.

Vatavu, R.-D., Anthony, L., and Wobbrock, J. O. (2012). Gestures as point clouds: A $P recognizer for user interface prototypes. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 273–280. ACM.

Vatavu, R.-D., Vogel, D., Casiez, G., and Grisoni, L. (2011). Estimating the perceived difficulty of pen gestures. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part II*, INTERACT'11, pages 89–106, Berlin, Heidelberg. Springer-Verlag.

Vatavu, R.-D. and Wobbrock, J. O. (2015). Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1325–1334, New York, NY, USA. ACM.

Vatavu, R.-D. and Zaiti, I.-A. (2014). Leap gestures for TV: Insights from an elicitation study. In *Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '14, pages 131–138. ACM.

von Laban, R. and Lange, R. (1975). *Laban's principles of dance and movement notation*. MacDonald & Evans, London.

Waite, P. and Appleby, J. (2003). *Beauchamp - Feuillet Notation: A Guide for Beginner and Intermediate Baroque Dance Students*. Consort de Danse Baroque.

Wang, L. and Suter, D. (2006). Informative shape representations for human action recognition. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 2, pages 1266–1269.

Webb, A. R. (2002). *Statistical Pattern Recognition, 2nd Edition*. Wiley.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257.

Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision*, ECCV '08, pages 650–663. Springer.

Wobbrock, J. O., Aung, H. H., Rothrock, B., and Myers, B. A. (2005). Maximizing the guessability of symbolic input. In *Proceedings of CHI '05 EA*, pages 1869–1872.

Wobbrock, J. O., Morris, M. R., and Wilson, A. D. (2009). User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1083–1092. ACM.

Wobbrock, J. O., Wilson, A. D., and Li, Y. (2007). Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07, pages 159–168. ACM.

Wulff, H. (2001). Dance, anthropology of. *International Encyclopedia of the Social & Behavioral Sciences*, pages 3209–3212.

Xiang, T. and Gong, S. (2006). Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51.

Xiong, Y. and Quek, F. (2006). Hand motion oscillatory gestures and multimodal discourse analysis. *International Journal of HumanComputer Interaction*, 21(3):285–312.

Zaiti, I.-A., Pentiuc, S.-G., and Vatavu, R.-D. (2015). On free-hand TV control: Experimental results on user-elicited gestures with Leap Motion. *Personal and Ubiquitous Computing*, 19(5–6):821–838.

ABOUT THE AUTHOR

Radu-Daniel Vatavu is a Professor of Computer Science at the University of Suceava, where he conducts research in Human-Computer Interaction, Ambient Intelligence, and Entertainment Computing. He received a PhD in Computer Science from University of Lille 1 & University of Suceava (2008) and a HDR in Computer Science (2014).