Rewarding the Original: Explorations in Joint User-Sensor Motion Spaces

John Williamson
jhw@dcs.gla.ac.uk

Roderick Murray-Smith
rod@dcs.gla.ac.uk

School of Computing Science, University of Glasgow, G12 8QQ, Scotland, UK

ABSTRACT

This paper presents a systematic and general technique for establishing a set of motions suitable for use with sensor systems, by drawing performable and measurable motions directly from users. It uses reinforcement which rewards originality to induce users to explore the space of motions they can perform. A decomposition of movements into *motion primitives* is constructed, among which a meaningful originality metric can be defined. Because the originality measure is defined in terms of the sensed input, the resulting space contains only movements which can both be performed and sensed. We show how this can be used to evaluate the relative performance of different joint user-sensor systems, providing objective analyses of gesture lexicons with regard to the technical limitations of sensors and humans. In particular, we show how the space of motions varies across the arm for a body-mounted inertial sensor.

ACM Classification Keywords

H.5.2 [Information Interfaces And Presentation]: User Interfaces - Input devices and strategies;

Author Keywords

motion; gesture; reinforcement; originality, inertial; novelty.

INTRODUCTION

> *"At Sea Life Park we shaped creativity in two dolphins [...] by reinforcing anything the animals did that was novel and had not been reinforced before. Soon the subjects caught on and began "inventing" often quite amusing behaviors."*

This quote, from Karen Pryor's book [13] on reinforcement training for animals, illustrates an unusual application of positive feedback: promoting creativity by rewarding novel behaviour. Reinforcement is often used to shape behaviour into specific forms, narrowing behaviours down to a template. But it can be used for the opposite effect; to promote the discovery of new behaviours by rewarding originality. In this paper we explore how this strategy can be used to map the capabilities of a *joint user-sensor system* by extracting a space of useful motions directly from users themselves.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

The purpose of these explorations is to derive a systematic method for studying the capabilities of input devices in context. There is now an enormous and ever-expanding diversity of devices which capture movement for the purposes of controlling a computer. From familiar pointing devices like touch screens and mice to more esoteric vision based systems, pressure-sensing and inertial-sensing hardware, these devices are incorporated into feedback loops where communication between human and computer can take place. With the exception of brain-computer interfaces and a few other bio-sensors (such as galvanic skin response), all of these input devices, indirectly or otherwise, measure the effects muscle activity for the purpose of determining intention.

It is challenging to work out how to most effectively design new input systems, or how to best exploit the capabilities of existing devices. There are a number of well-developed specific interaction paradigms for different classes of input device. Pointing movements, for example, have been extensively studied (e.g. [17], [11]) and there exists a range of techniques for analysing performance and designing interactions where pointing is practical. Static pattern-based "gesture recognizers", which match defined sensor sequences to a set of symbols, have been widely explored for non-pointing devices, especially inertial sensor and vision-based systems. Scoditti et. al. [15], for example, lays out a taxonomy of movements for accelerometer control from a symbolic language perspective. Still other interaction paradigms use general dynamical systems (of which pointing systems are an elementary instance), such as the motion selection techniques of Williamson et. al. [18] and Fekete et. al. [2]. Many input techniques are under-exploited, however; capacitive sensor arrays are often used to detect touches as a set of contact points, but there is much more usable information that could be extracted (e.g. from pose variations).

While these specific paradigms are very powerful and are essential for the construction of working, usable systems, it is interesting to consider how input mechanisms can be quantified in a broader sense, and how factors which influence the use of these mechanisms can be analysed within a coherent framework. This paper looks at the essential characteristics of one aspect of input mechanisms – determining the space of motions which are distinguishable and controllable, and develops a very general technique for exploring, comparing and quantifying these spaces.

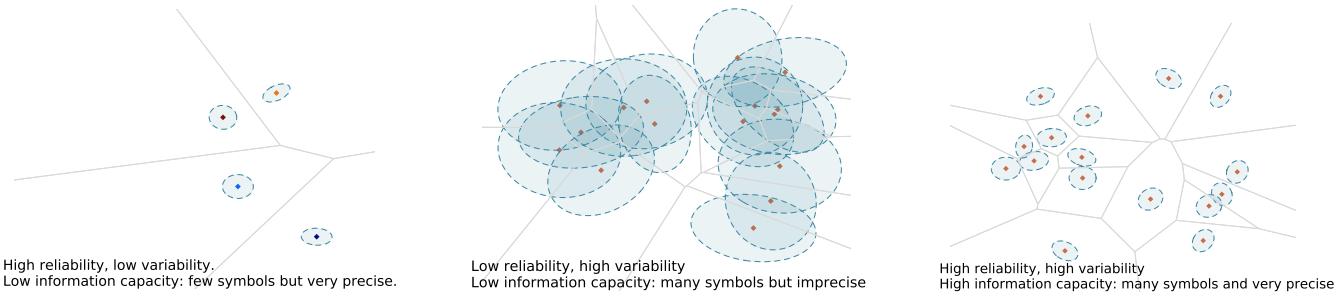

Figure 1: A sketch of variability and reliability. Imaginary distributions of sensor sequences projected onto a 2D plane are shown, where each distribution represents one "symbol" and the shaded area shows the range of measured value for repeated performances. To have high information capacity, there must be both variety and precision, so that there is a partition of the space which separates the symbols well.

Information, semantics and input devices

The quality of a user input system, in terms of its purely utilitarian benefits, can be quantified by its information throughput. The more information that is communicated, the better the interface. This is the basis of many input device studies in HCI, such as those using Fitts' Law [3] to estimate the information capacity of pointing devices. This narrow focus on information transfer rates excludes many important factors in the quality of an interaction, including ergonomics, aesthetics, social acceptability, cost and physical dimensions, but it captures the essential quality which distinguishes effective input systems from ineffective ones.

In our stance, there is a clear separation between the process of communicating intention and the associated semantics. Analysing interaction this way ignores the relation between movements themselves and the meaning a system ascribes to them; there is no consideration of whether flicking a finger is a more suitable operation than shaking a fist for some particular purpose. In building concrete systems, exploiting preconceived notions and metaphors inherent in human activity is essential method to address aesthetic and subjective aspects of interaction. But here we focus on a high-level analysis of the *limits* of an input mechanism without regard to metaphor or idiom. In this way we can separate, control and measure factors such as encumbrance or social context which impact upon the usability of an interactive system.

To be more precise, the sense of "input system" or "input mechanism" as used here, is a system which encompasses the combination of the involved human parts and the physical sensors. A system which uses accelerometers for sensing motion of the hand is quite different from one that is attached to a knee, and different still from a system which uses infrared proximity to sense motion of the hand, and these all form separate input mechanisms.

For a one-dimensional signal, Shannon's classic formula relates the signal to noise ratio and bandwidth to the throughput in an analog channel subject to additive white Gaussian noise:

$$C = B \log_2 \left(1 + \frac{S}{N}\right),$$

giving a transfer rate (capacity) of C bits per second from a bandwidth (*variability* of the signal) B and a signal-to-noise ratio (*reliability* or repeatability of the signal) S/N. Although this simple relation is only true for one-dimensional channels of a very specific type, the principles of variability and reliability generalise to more complex interfaces. We can thus split the analysis into two subproblems (Figure 1):

- **Variability** The "size" of the space of possible movements that a user can achieve, or more specifically the subset of the sensor measurements a user can reach. The larger the *repertoire* of movements available, the more efficient the communication will be.
- **Reliability** The precision with which actions can be reproduced at will. The more precisely state sequences can be reproduced – *from the perspective of the sensors* – the more quickly information can be conveyed. If movements have a variation which is not under the control of a user, the information capacity of the channel will be diminished as some movements will be indistinguishable.

Two further concepts flow from these two basic factors:

- **Transmissibility** How quickly and easily the skills required to interact with the system can be communicated or discovered. This is affected by feedback from the system and by the ease of learning movements from others, and depends on factors such as similarity to previously encountered interactions, and the affordances of the input device.
- **Recognisability** How effectively the software can map sensor sequences into classes that represent user intention. Even if the user can reliably generate a huge variety of sensor sequences quickly and accurately, there must be systematic way of mapping these onto useful actions.

Each of these aspects must be addressed when designing an input mechanism or when evaluating its performance. Although many of these issues can be addressed for specific interaction styles (e.g. via Fitts' law measurements for pointing devices), it would be useful to have more general techniques which presuppose less about the interaction style to be used. This is particularly the case for sensors for which "traditional" mappings such as targeting are not appropriate.

This paper focuses on solving the first problem: establishing the repertoire of motions which can be used for input. This problem is tackled by constructing a decomposition of movements into "motion primitives" which efficiently encode the current state and its local temporal variation. A

similarity metric between these primitives is defined, and an experimental protocol which rewards users for performing actions which maximise this metric is constructed, and thus rewards users for exploring more of the potential motion space. The result is a map of possible motion primitives for a specific user and input mechanism.

Gesture design

The problems of designing "gesture-based" interactions – which often means any interaction involving non-traditional input devices – have been studied widely. Many papers have laid out strategies for creating gesture systems, sometimes involving gesture creation by designers [20] and sometimes involving user-generated gestures [10], [19]. Although these studies are valuable in constructing viable gesture lexicons, they often proceed without reference to the technical constraints of recognition.

What distinguishes the work presented here is the principled analysis of spaces formed by the combined constraints of users and the sensing hardware, but without regard to specific domains. The motion exploration techniques specifically identify the technical limitations of input mechanisms. Any design process will have to incorporate domain-specific constraints, but the analytic process we present provides a solid foundation from which to work.

EXPLORING THE SPACE OF GESTURES

One of the major issues with any new input technique is establishing what we term a *joint user-sensor space*. In general, the space that is measured by a set of sensors (for example, a vector of pixel brightnesses, in a camera-based motion capture system) is very different from the one in which users are expressing themselves (which might be the position of their fingertip relative to their torso). Motions, as measured by the sensor are sequences in *sensor space* – the space defined by the vector of values that a given sensor produces. Users perform actions in *user space*, which is a more difficult concept to pin down; it clearly involves the muscular system but motions are not usually *imagined* or *controlled* as variations in joint angles or other physiological measurements. It might best be defined as the space in which users *think* they are controlling (i.e. an intentional space), and can vary from task to task. The role of proprioception, optical flow, and other sources of feedback are more relevant in the control of action than the resulting physical state changes.

Many parts of the sensor space correspond to impossible real-world configurations (in a camera system, imagine an image of a user with multiple heads), and many motions that can be performed cannot be sensed (for a camera, the position of the finger when the relevant hand is occluded by the other arm). Only those motions which are both possible and can be sensed are communicative. These set of these motions form the *joint user-sensor space* and ideally a system would use only motions within that set and would use motions from the whole of this set. These criteria will optimise the information transfer, by minimising useless signals, and by maximising the set of distinct signals that can be used.

Motion primitives: elements of gesture

Using a very liberal definition, a gesture can be considered a deliberate movement for the purposes of communicating [8]. The problem to be addressed can then be stated as identifying what set of "gestures" can be performed and sensed reliably; those gestures which lie in the joint user-sensor space. By quantifying this space, we can obtain the superset of all movements which could be used in a gesture-based system. Gestures, in this general sense, are unbounded in length. A gesture could last a few hundred milliseconds or several minutes, and treating "complete" movements is therefore difficult. The space of length-unbounded gestures is obviously very large, and there is not an obviously correct way to make like-for-like comparisons between motions of different lengths.

If we imagine that gestures could be broken down into a set of elementary units, so that any movement could be analysed as the composition of those elementary "motion primitives", the problem of establishing a consistent and meaningful gesture space would be much simpler. The simplest dissection is to divide motions into equal length sections and analyse these separately. For example, all motions could be split into equal 0.1 second blocks and analysis carried out on these sections, and assuming that gestures can be resynthesized by concatenating these sections, the gesture space would become a catalogue of these sections. In essence, we can form a *codebook* of motion vectors. This concatenative approach is merely the simplest of many approaches that could be envisaged, and for some types of motion (e.g. rhythmic motion, as discussed by Ijspeert et. al. [5]) there are more elegant decompositions. The subdivision of generable motions into components on which metrics can be defined is the key to creating processes that can reward originality.

JOINT USER-SENSOR SPACES

There is not one constant joint user-sensor space; there are many spaces and they are very variable in nature. Even if the sensing hardware and interpreting software remain constant, the human element of the interaction varies significantly. Different people have quite different motions that they are capable of and willing to make. Those who are highly skilled in manual tasks may well have a richer range of motions they can draw upon and greater precision in their control of movements. An expert in martial arts might, for example, have much greater control over the velocity profiles of their movements than someone who lacks such training. Those who have motor impairments will have access to a smaller subset of the possible motion space. Individuals will also vary over time, both on short scales (for example as consequence of exhaustion) and on longer time scales (as an effect of growth, weight change or age-related reduced mobility). Capturing a systematic picture of the accessible regions of the gesture space across population groups gives designers a powerful tool with which to customize interfaces to satisfy specific needs. We would like both to identify the set of communicative gestures, and to map the qualities of those within that set. These qualities might include the strenuousness of movement, or the subtlety of the gesture.

Physical constraints

Temporary physical constraints will also affect the space of motions accessible. As an example, consider trying to operate a gesture recognizer while carrying a heavy bag in the gesturing arm. The effect of physical constraints, such

as being seated versus standing, wearing constrictive clothing, or being constrained within a vehicle can also be measured. Some constraints will affect the reliability of movement, rather than variability, by adding noise to the process, and this will not be captured in a motion map. But many constraints limit the range of movement and can be mapped out in the joint user-sensor space.

Social contexts

There are also social constraints on gestures that can be performed in different situations in front of different audiences (see for example Rico and Brewster [14]). Movements which are quite acceptable in private at home may not be so acceptable on the pavement of a busy street. The effect of these various constraints on the input device at hand can be precisely examined by comparing spaces where audiences are varied and identifying which motions are excluded. For example, a mobile phone with an accelerometer-based interface can be evaluated to identify which motions are usable in a private setting and which are then unreasonable in a public setting.

False positives and common movements

The consideration of the joint-user space also leads to examination of regions of the space rendered unusable because they are too frequently performed inadvertently. Motions which occur as part of everyday behaviour might be subtle and unobtrusive, but it can be very hard for a system to distinguish between these movements executed as a deliberate attempt to communicate and those which were not intended to be recognised. Foot tapping is a good example; this is a simple and unobtrusive movement, but a poor choice for communication if used alone. People tap their feet idly and suppressing this behaviour to restrict it only for input purposes is burdensome. We will not treat this in the studies presented, but it is well within the scope of the analytical techniques presented here to map out such reserved spaces by capturing motions during everyday activity and excluding them from the available space.

Intuitiveness

The word "intuitive" is often used to describe interfaces without a clear definition of what it implies. If we take the meaning of intuitive to describe something which is familiar *before* it has been used, we can analyze input systems in terms of which movements are spontaneously generated. By capturing the exploration *process*, not just the set of communicative motion primitives, a picture of the motions which the interaction inspires can be built up. If a new input mechanism is proposed, and we wish to choose some set of motions to use as controls, those which occur "naturally" – those which are afforded by the physical construction of the input devices or the feedback provided – will minimise the burden of learning to control the device. This "seed set" can be identified from the motions which are generated early in the motion exploration process. Those motions which, across a population, are only generated after other movements have been exhausted are presumably less obvious than those which are generated at the beginning. Primitives can be ranked according to their "time of invention" and this ranked repertoire can then be used either to design gestures – by sequencing elements with low rank – or to test the intuitiveness of a proposed gesture by computing the average rank of its component primitives.

REINFORCEMENT FOR EXPLORATION

The basis of the exploration process is reinforcement learning, where the aim is to reward the novelty of a behaviour, rather than its similarity to a desired behaviour. Rather than shaping behaviour, we seek to draw candidate motions from users' minds, encouraging them to vary their movement as much as possible. This is reminiscent of the "superstitious perception" process discovered in Skinner's early work with pigeons. This demonstrated that animals could be conditioned into generating very peculiar behaviours, if a reinforcer such as food, was released in a manner completely unrelated to the behaviour [16]. Recent work by Gosselin et. al. [4] built on this to obtain images of archetypal objects by averaging the result of a selection process where subjects culled a set of purely random images to include only those resembling the imagined object. The Cybernetic Serendipity system [12] is a very early example of a system – in this case an installation – which responds specifically to novelty in input, triggering the movement of mobiles when musical input was sufficiently different from previous patterns the system had "heard".

Our system rewards users for novelty. The hypothesis underlying this approach is that if users are reinforced when they do something novel, they will eventually exhaust the range of distinguishable behaviours. Obviously, the ability to explore the space is dependent on the imagination of the participants, and since the only feedback is related to novelty, any variation must by driven by the user.

MODELLING

Any analysis of the joint user-sensor space must rely upon some model of that space. Such a model is an assumption about how we expect the user-sensor system to behave, and this choice will constrain the results that can be obtained. Although any analysis cannot be assumption free, we can make the exploration as general as possible by making the modelling assumptions as unrestrictive as possible.

The representation of the movement section should compactly capture the nature of the movement; a trivial recording of sensor values is very simple to implement but for most sensors will require a great deal of data to represent what is a much simpler underlying process. Identifying a reasonable set of latent variables – compressing the representation – will reduce the potential for "over-counting" the set of available motions. The more compactly a representation can capture the space of movements without distortion, the more meaningful distance metrics in that space become.

Sequencing of primitives

The motion sequencing approach could, in its most trivial form, be reduced to analysis of a sequence of static poses; in the case of a camera system, this would be analysing video as a "bag of frames". This, however, ignores the fundamentally *dynamic* nature of human movement. Human movement is a continuous physical process driven by muscular motion and a methodology which takes a static approach greatly over-counts the space of possible motions, because

feasible movements lie along smooth, achievable trajectories. A representation which captures temporal aspects is essential to modelling the space of motions well.

Simply modelling the transitions between static poses is an unsatisfactory way to map out the gesture space. We could for example, identify a range of limb poses and ask users to transition between random pairs. Even if the poses were well-selected, this analysis would be dominated by transient phenomena as static poses are left and entered, and would fail to model the co-articulation between dynamic movements. Dynamical models of human movement often approximate limb motion as a spring-like system ([6], [1]) which have clearly continuous motion; a hand in oscillation cannot be well approximated by transitions between static poses in that cycle.

Polynomial approximation

One elementary and compact dynamical model is to represent the current state state along with (smoothed) time derivatives. A limb measurement, for example, which included position, velocity, acceleration and jerk (the 3^{rd} order derivative) can represent a wide variety of motions, and this representation captures the physical essence of motion, if we assume limb motion can be well-modelled by a linear differential equation over short periods of time.

We generally do not, however, want to assume a mapping from sensor space to the physical movement space, as this must necessarily be quite specific, and limits the generality of the exploration technique. Instead, the motion vectors can be represented directly as sensor values, along with their various derivatives. By dividing each one-dimensional signal into segments, and fitting a polynomial to those segments, motions can be represented in a simple form which encodes the time-evolution in a physically-relevant way.

For sensors with very high inherent dimension (like cameras), this is not a practical approach, and some pre-processing must be applied to extract relevant features (e.g. by tracking objects). Lower-dimensional sensors, such as accelerometers, pressure sensors or styli can be directly mapped without feature extraction. An extension of this technique would be to automatically learn a low-dimensional manifold from a series of sensor measurements, and then use this learned low-dimensional mapping as the input to the motion exploration system. Assuming that the mapping extracted relevant latent variables, this would extend the technique to very high-dimensional systems without having to construct a manual model.

Similarity metrics

The exploration of the space hinges on rewarding users for doing something novel. This requires a careful definition of novelty, as once chosen, the experimental results will depend upon this originality metric. One simple model is the Euclidean distance between motion vectors. This, however, performs poorly when the scales of different elements of the sensor vector are widely distributed.

To calculate the novelty of the movement, a suitable similarity metric is required. The choice of metric is an assumption; we have chosen here to use the mean distance of the nearest k neighbours using the Mahalanobis metric (or generalized interpoint distance). The Mahalanobis distance is effective because it accounts for the differences in scale of the various variables and linear interdependencies between them (see Figure 2). Using the mean distance to the k nearest neighbours mitigates the impact of outliers that might affect the results if distance to the nearest point was simply used.

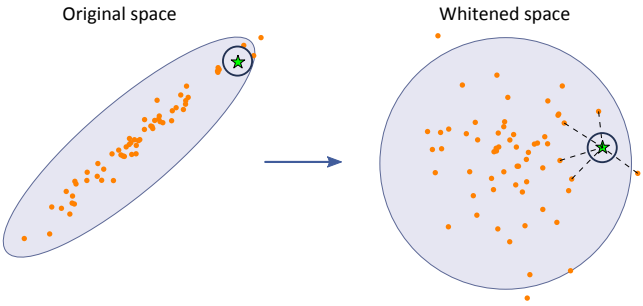

Figure 2: The Mahalanobis metric gives get a scale-independent measure. This is equivalent to decorrelating and rescaling data to a sphere with unit covariance before comparing. The nearest k neighbours under this metric are found, and the mean distance to them becomes the originality value of that point. If this value is large enough then M^* (green star) joins the repertoire.

Other scale-invariant metrics could also be used; the cosine metric $\left(s = \cos(\theta) = \frac{M_i \cdot M_j}{\|M_i\| \|M_j\|}\right)$ might be effective for high-dimensional spaces, but the Mahalanobis distance is well-understood, robust and compensates for linear correlations.

Testing the quality of models

Making assumptions about the nature of movements is inescapable in the construction of models. However, it is possible to evaluate the quality of these assumptions systematically. Given a concatenative model with fixed windows, a data set can be processed with a range of windows and movement models, and the model error can be compared against the compression achieved (e.g. by considering the rate-distortion curve). A model which concisely represents the motions with minimal error provides an optimal space in which to judge originality.

CONCRETE IMPLEMENTATION

To apply these techniques, we built a system for performing motion exploration using an inertial sensor. This sensor can be attached to various body parts to sense motions of those parts, and the system can map out motions available across users and sensor locations. This system uses only audio feedback as a reinforcer, so that participants do not have to divert their visual attention. In our system, we have defined a motion primitive as the measurement of vector of sensor values at a point in time, along with a number of the estimated derivatives of this action – the polynomial approximation described above. This combined vector represents a snapshot of the evolution of the sensor values at a particular time. A 4th order model was used, which captures the current state of a sensor stream and its first three derivatives. A least-squares fit of a polynomial of appropriate order is performed on a sliding window. Derivatives of order greater

than three are likely to be subject to poor fitting, and also have less obvious significance as high order derivatives are not something that humans are accustomed to controlling.

In the measurements presented, a 0.65 second running window was used, as a compromise between capturing the finest details of movements and capturing movements of sufficient magnitude to be useful for communication. This choice is obviously an assumption, but this assumption is numerically justified in the results section.

Implementation

The novelty detector captures data from a SHAKE SK7[1] inertial sensor, which provides 3-axis accelerometer, gyroscope and magnetometer outputs. The accelerometer, gyroscope and magnetometer measure in the range -6g – 6g, -900 – 900 degrees/second and -0.2mT – 0.2mT respectively, all with 12 bit resolution.

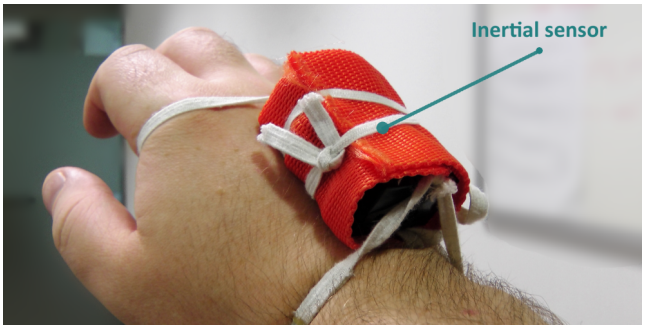

Figure 4: The SK7 inertial sensor, as mounted on the hand. The elastic mounting minimises slippage.

Each sample therefore consists of a 9 element vector $S = [a_x, a_y, a_x, g_\theta, g_\phi, g_\psi, m_x, m_y, m_z]$. We do not attempt to map these values into an alternative space (e.g. to map to device orientation), partly to reduce the assumptions required, but also to capture all motions that the hardware can sense (so including motions involving linear acceleration such as flicks or lunges).

The inertial sensors are sampled synchronously at 32Hz; the sensor streams are filtered on the hardware to bandlimit them before decimating from an original sampling rate of 1024Hz. A 32Hz rate is sufficient to capture the details of most limb movement. At a 0.65s window length one motion primitive is therefore extracted from 21 samples of data (the window time is adjusted slightly to be exactly 21 samples in length). A bank of Savitsky-Golay filters are applied to the signal, each designed to estimate one smoothed derivative efficiently. These coefficients are concatenated into a single vector $M^* = \left[\hat{a_x}, \frac{d\hat{a_x}}{dt}, \frac{d\hat{a_x}}{dt^2}, \ldots\right]$ with $q = 9 \times 4 = 36$ dimensions, which represents the motion in sensor space.

When computing the novelty of a motion, we represent the *repertoire* (or *codebook*) of distinct motions as a set of vectors $M_1 \ldots M_N$. These are the set of motions which the "novelty detector" has determined are distinct enough to maintain. It is initially empty, and is populated as the motion capture progresses. As described, the mean Mahalanobis distance to the k nearest neighbours is used; thus to compute the distance between a new vector M^* and an existing sample in the repertoire M_j, we compute

$$d(M^*, M_j) = \sqrt{(M^* - M_j)^T \Sigma^{-1} (M^* - M_j)}, \forall M_j \in M,$$

where Σ is the covariance matrix of the repertoire $M_1 \ldots M_N$. We then find the k smallest distances $d(M_j, M^*)$ and the log mean of this, $d_{\mu(k)}(M^*) = \log \frac{1}{k} \sum_{j=1}^{k} d(M_j, M^*)$, is the similarity to the existing samples in the repertoire. If $d_{\mu(k)}(M^*) > \Delta$, where Δ is a threshold setting the "novelty cutoff", then the sample M^* is added to the repertoire. By adjusting Δ and k the sparsity of the space, and thus the time taken to exhaust it, can be controlled. Δ is proportional to a confidence interval on the interpoint distances.

In practice, calculating the matrix Σ, which represents the covariance of the repertoire, is quite expensive, and is done only when a number of new samples have been added to the repertoire. The current implementation computes this covariance matrix for every fifty new vectors added to the repertoire. The matrix is initialised to a covariance matrix estimated from a pilot trial; it is quickly replaced as the new repertoire accumulates. As the covariance matrix is necessarily updated during the motion capture, the similarity metric changes slightly throughout the process. These changes are gradual and result in a smooth variation in the originality feedback. The adaptive nature of the metric maintains reinforcement at a roughly constant rate over individuals.

Audio feedback: originality reward

The "reward" for novel movements is an increase in the apparent liveliness of the audio feedback presented. When a new element is inserted into the repertoire, the distance $d_{\mu(k)}(M^*)$ is accumulated into a leaky integrator, which decays exponentially at a constant rate. This leaky integrator is fed to the audio output to produce the positive feedback for novel movements. We use a flexible granular synthesiser to produce the feedback. The integrator output adjusts the density of grains and the spectral brightness of the source waveforms such that more novel movements make brighter, more intense sounds, which gradually decay away to duller, sparser sounds as originality decreases. Eventually the sound fades away to a low hum when no new vectors are being added to the repertoire. The synthesiser is configured to sample snippets of ≈ 40ms from a short excerpt of jazz. This results in an aesthetically pleasant texture which can be modulated continuously with a degree of subtlety.

EXPERIMENT

As an example of motion space analysis, an experiment was conducted with the inertial sensor based exploration system. The exploration was carried out in two conditions: condition A, with the sensor attached to the dominant hand; and condition B, with the sensor attached to the arm just above the elbow. The hypothesis is that the range of reasonable motions should be smaller for the elbow than for the hand, because the available degrees of freedom are substantially lower. The experiment is designed to test whether the processes presented here could quantify this difference reliably,

[1] http://code.google.com/p/shake-drivers/

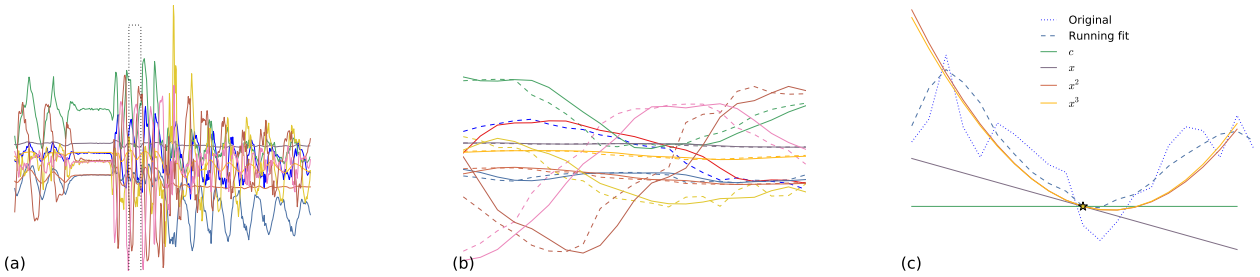

Figure 3: The process of extracting motion primitives. A short window of the signal is extracted (a), lowpass filtered (b) (original dashed, filtered in solid). An estimate of the value and derivatives at the central point is made (c) (for the lower blue curve in (b)). These coefficients become the representation of that movement. Only one axis is shown in (c) for clarity; the motion vector is the concatenation of these coefficients for all axes.

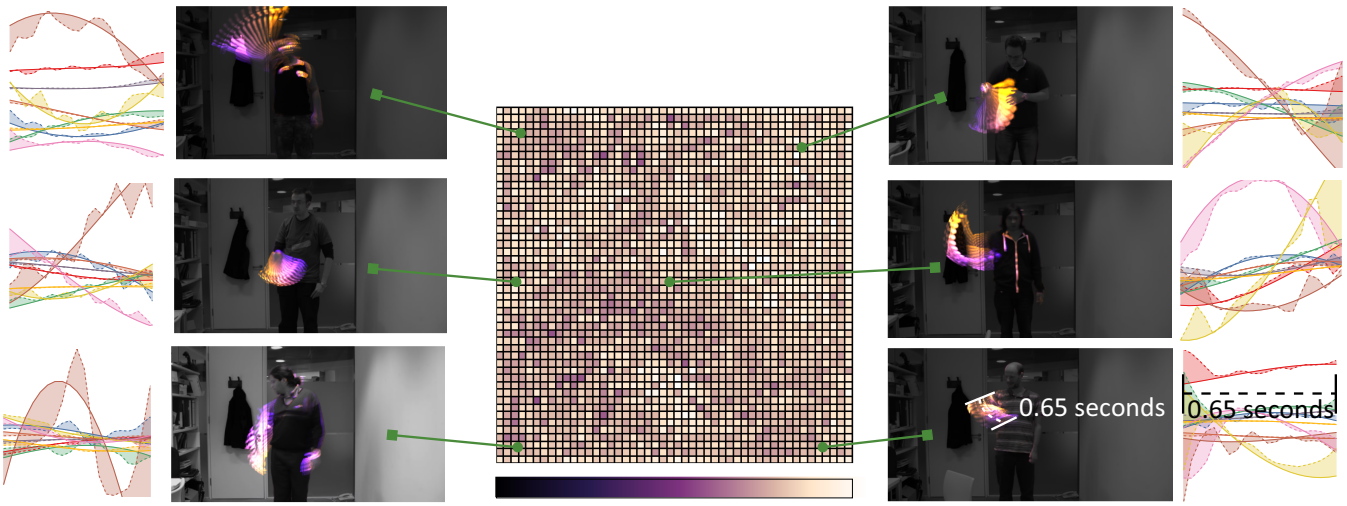

Figure 5: Self organizing map, showing the log density of points on the projected motion space for all participants across all conditions. The breakouts show the video sequence for the nearest vector to that point, and the motion curves for that vector. The fitted curve (from the motion vector coefficients M_i) is shown as a solid line, the true raw data as a dotted line, and the area between is shaded.

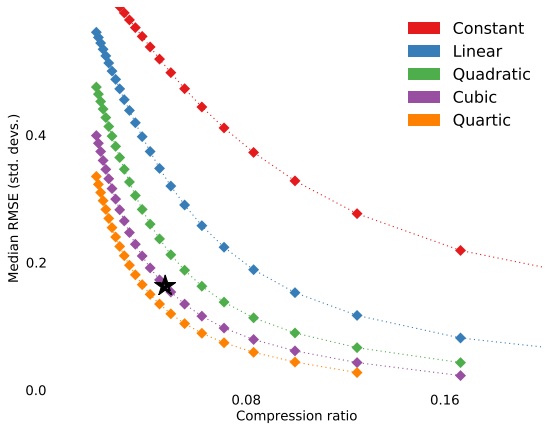

Figure 6: Median root mean squared error of Savitsky-Golay fit versus "compression" ($1/\text{window length}$), computed for the entire dataset. Choices near the origin represent a good trade-off between error and compactness. The cubic, 0.65 second model used is shown as a black star, near the optimum.

and to provide a pool of captured data for exploratory analysis.

The experiment was conducted in a laboratory. There were $N = 20$ participants, 16 male, 4 female, recruited from a pool of computer science students, with an age range of 18–40. All participants gave informed consent. The SK7 sensor pack was affixed with an elastic strap (see Figure 4). This prevented the sensor from being moved relative to the body and so eliminates symmetric motions where the same movement is performed with the sensor axes in a different alignment and also prevents users from performing "motions" such as dropping the device or throwing it into the air. Participants were requested to keep their legs and torso relatively steady and concentrate on moving their dominant arm. We did not aim to completely suppress natural motion of the rest of the body, but to focus intentional movement on the relevant area. The participants motions were also recorded with a video camera. This is not used as input, but the captured video stream is synchronized with the SK7 sensor stream. This provides a visual reference for the movements performed.

All participants completed both conditions in counterbalanced order. Each condition was conducted as a set of three subtrials, lasting three minutes, with a break between each. Each subtrial continued from the point the previous one left

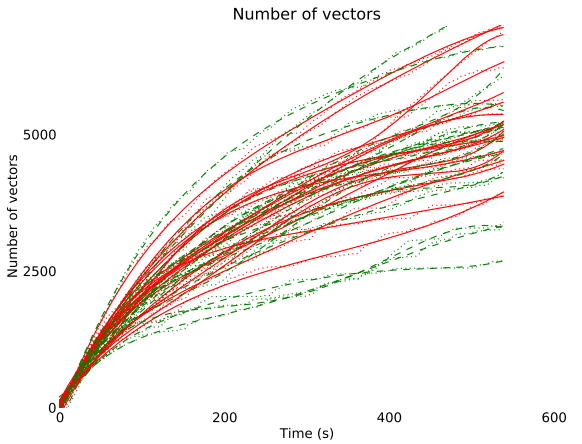

Figure 7: The total number of primitives versus time, for all participants. Measurements for condition A (hand) are shown as solid lines, measurements for condition B (elbow) are shown as dashed. The graph shows a polynomial fit to the true data; the original raw data is shown as a dotted line for comparison.

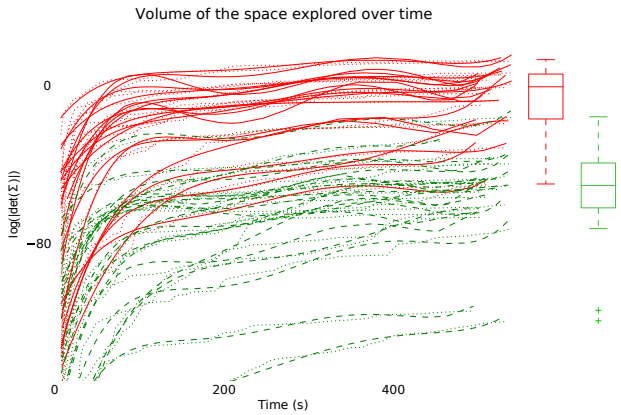

Figure 8: The "size" of space explored. This graph shows the log determinant of the covariance matrix of vectors seen so far for each participant $\log(|\det(\Sigma)|)$. $|\det(\Sigma)|$ is proportional to the hypervolume of the covariance ellipse; larger values indicate more of the space being explored. A value of 0 ($|\det(\Sigma)|$=1) means that the volume of the covariance ellipse was equal to that of the covariance of the entire data set. The Box plot to the right shows the final distribution of values.

off, for a total of nine minutes of exploration in each condition. The conditions were broken into subtrials to combat fatigue from long performances. The implementation uses $k = 5$ nearest neighbours and a value of $\Delta = 1.50$ for the novelty cutoff.

ANALYSIS

In total, across both conditions, there were 203671 36 dimensional motion vectors captured from a total of exactly 6 hours of exploration. Figure 6 shows error against model size for a range of polynomial orders and window lengths; a 0.65 second cubic model is close to the optimum for low-order polynomial models.

The number of unique vectors generated by each participant in each condition is shown in Figure 7. The vectors generally asymptotically approach a value which varies by participant, with a mean value ≈ 5000. Because the process

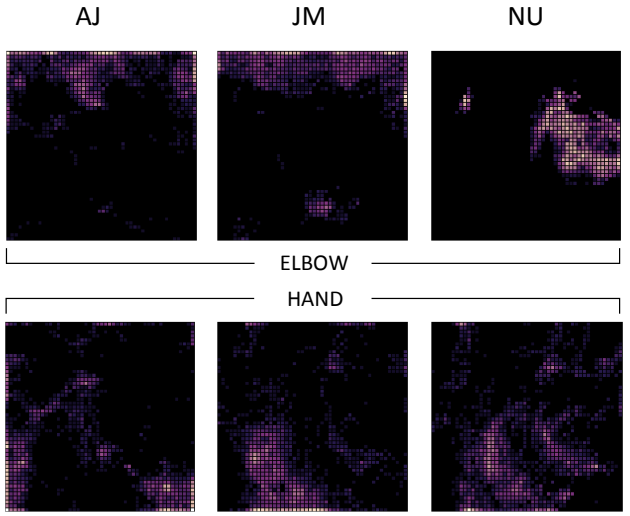

Figure 9: Individual (log) densities for a selection of participants. Upper row shows the maps for the elbow condition, lower shows the map for the hand condition.

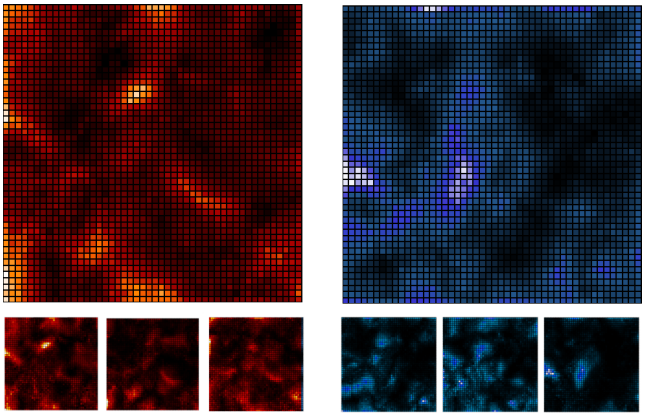

Figure 10: (left, red) Average total acceleration across the space of motions. The three smaller images below show the individual maps for the x, y, z axes respectively. (right, blue) Average total angular velocity of motions. The lower images show the maps for each of the three axes.

is adaptive (the covariance matrix is continuously recomputed), there is not a huge variation between the conditions. This is to be expected, as the feedback is designed to fall off at a roughly constant rate regardless of the absolute motion space explored, in order to maintain user motivation. However, Figure 8 shows a plot of the estimated "volume" of space explored in each condition; it is clear from this that there was a much smaller portion of the space explored in the elbow condition. This clearly supports the hypothesis that the range of available motions is reduced in the case of the elbow as compared to the hand.

Motion maps

The motion vectors forming the repertoire are difficult to visualise directly, primarily because of their high dimension. The space can be more easily visualised by performing dimensional reduction to a 2D layout. In our analysis, we have used a standard self-organizing map ([7], [9]) on the whitened feature vectors to reduce the 36-dimensional vectors to a 2D map. The maps used here are discrete 48×48 arrays with a rectangular neighbourhood function. Each point

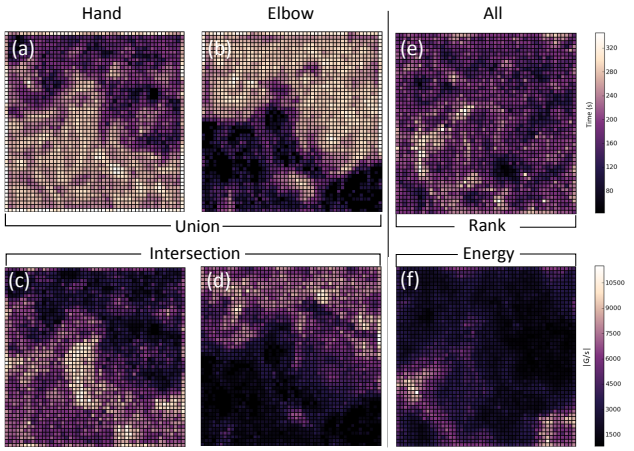

Figure 11: (a,b,c,d) Union and intersection of all motion maps for the two conditions. Each map is the sum/product of the normalised densities for each participant. (e) Map coloured by average time of invention; brighter areas were discovered later in the process. (f) Map coloured by average energy in $|G/s|$. Brighter colours involve more violent motion.

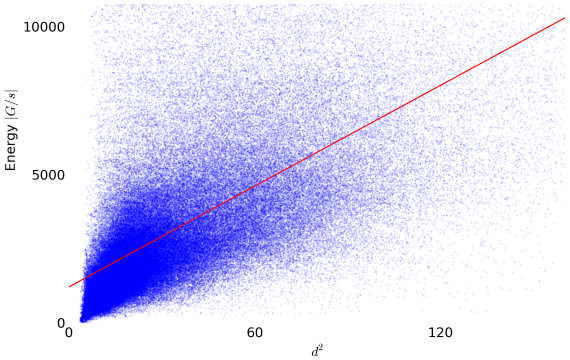

Figure 12: Movement energy (root-mean-square derivative of acceleration), in G per second, versus Mahalanobis distance (linear regression as a red line). There is a strong inverse correlation between how violent a movement is and how likely users are to perform it.

on this map represents a vector in the original high dimensional space, and the points are organized so that nearby vectors on the 2D map are nearby in the high-dimensional space. For consistency between plots, the dimensional reduction was computed from all data recorded (across all participants, and all conditions), and each separate condition/participant map was obtained as a subset of this master projection. All of the visualisations are thus directly comparable.

We constructed an interactive tool which uses the synchronized video recordings to show snapshots of motions. This tool finds the point on the map nearest the cursor, looks up the index in the video and plays the relevant section, or generates a composite image for that motion. The composite images show the moving parts of the frame in colour, fading from purple to yellow, against a monochrome background.

Figure 5(a) shows a map of the space explored by all users. Maps for individual participant/condition pairs can be produced; some examples are shown in Figure 9. The maps can be coloured by objective measurements; Figure 10 shows maps coloured by total acceleration and by total angular velocity. Motions which involve twisting around the forward axis can, for example, easily be picked out from these maps. Figure 11(e) colours movements by time-of-invention (movements that were generally found later are coloured brighter). Comparing with Figure 10 it can be observed, for example, that movements involving twisting tend to be discovered later. Figure 11(f) shows a map the energy of movements estimated from the root mean square derivative of acceleration, where the derivatives are estimated from the *raw* data corresponding to the vectors at that position. Figure 12 plots this energy measure against Mahalanobis distance (i.e. how close the motion was to the overall distribution of the repertoire). It can be seen that energetic movements are distinctly rarer than calmer ones. Densities can also be formed from the (continuous) intersection, union or difference of a pair of maps; for example Figure 11(a) shows all motions that users performed in the hand condition, while Figure 11(c) and (d) shows movements performed by all users in the hand and elbow condition respectively.

One notable qualitative result is that the motions on video bear little relation to the sensor measurements. Arm movements with very different scales can result in very similar inertial sensor sequences and vice versa; the mapping between inertial space and human-centered space is complex and non-obvious. This partially explains the difficulty in constructing inertial sensor based interfaces beyond simple metaphors such as shaking and tilting.

These analyses demonstrate how movement spaces can be efficiently categorised and compared. Motions can be analysed with cross-referenced objective metrics to identify features of interest or relevant trends. This study focused on variations across body locations, but other variables of interest, such as sensor types, demographics, social context, or physical encumbrances are equally open to exploration.

Estimating information transfer rates

Our process estimates the repertoire of possible motions. It identifies the *variability* of the channel, under the restrictions of the motion primitive model. Measuring the information capacity of an input system would require measures of the *variability* and the *reliability*. Measuring reliability (the precision with which motions can be controlled or repeated) is beyond the scope of this paper, but an analysis which could identify reliability of motions across a motion space could be combined with the techniques presented here to provide bounds on information transfer rates.

Application

In applying these techniques, there are two essential elements: defining novelty and providing reward. We have shown that simple modulation of responsiveness is a sufficient reward to drive users to explore different behaviours. The major challenge is deriving suitable novelty metrics, and the key issue there is constructing a decomposition of sensed data into comparable units which also capture the essence of the communication. With many simple input mechanisms the fixed time polynomial representation given here would be suitable (e.g. with a pressure sensor) but other modalities will require more careful definition of primitives which compress sensed data into meaningfully comparable units.

CONCLUSIONS

The exploration of potential input mechanisms is often necessarily driven by ingenuity in imagining metaphors. When a new avenue for input presents itself, interactive systems designers have to conceive of ways of building an interaction around the mechanism; conversely, input devices are often designed to fit an established metaphor. While this is still essential, the techniques given here extract maps of motion primitives with only very limited assumptions and create a platform for comparing and analysing input mechanisms. Reinforcement of novelty is a systematic and straightforward way of identifying joint user-sensor spaces.

This approach does not provide a complete strategy for designing for new input devices. Meaning must still be ascribed to gestures, which requires the imagination of designers in creating idioms; recognition techniques must still be created to categorize movements accurately; and the role of feedback in the interaction process is unaddressed. However, the techniques laid out here present designers with a palette of suitable elements which can be woven into an interaction; provide measures for comparing the capabilities of input mechanisms under different constraints; and give implementers a suite of tools with which to rate and dissect proposed gesture sets. A full information theoretic analysis will require the development of techniques for systematically analysing the precision of movements across a whole space. Probabilistic methods are increasingly being used to analyse interaction, but such methods often require a space over which measurements can be normalised. This has been previously generally inaccessible, but the process given here provides an efficient way of extracting a suitable global space of movements.

We have shown here how the exploration process can be applied to an inertial sensor but these techniques are easy to apply to other sensors. The techniques provide designers with a rich set of exploratory and analytical tools for building movement-based systems. The decomposition of continuous gestures into simple primitives makes it possible to systematically compare complete gesture spaces. The outputs of these analyses are well-founded objective measures which necessarily take into account both the restrictions of input devices and the limitations of the humans using them. These can be used to compare contexts, populations, devices and metaphors within a coherent framework.

ACKNOWLEDGEMENTS

This work was supported by a fellowship from the Scottish Informatics and Computing Science Alliance (SICSA) and by the PASCAL2 Network of Excellence. All (non-video) data from this project is freely available online at `www.dcs.gla.ac.uk/~jhw`.

REFERENCES

1. R. M. Alexander. *Principles of animal locomotion*. Princeton University Press, 2003.
2. J.-D. Fekete, N. Elmqvist, and Y. Guiard. Motion-pointing: target selection using elliptical motions. In *CHI '09*, pages 289–298. ACM, 2009.
3. P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *J. Exp. Psych.*, 47(6):381–391, 1954.
4. F. Gosselin and P. G. Schyns. Superstitious Perceptions Reveal Properties of Internal Representations. *Psychological Science*, 14(5):505–509, 2003.
5. A. J. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. In *NIPS'03*, pages 1523–1530. MIT Press, 2003.
6. R. J. Jagacinski and J. M. Flach. *Control theory for humans : quantitative approaches to modeling performance*. L. Erlbaum Associates, 2003.
7. J. Kangas, T. Kohonen, and J. Laaksonen. Variants of self-organizing maps. *IEEE Tran. Neural Networks*, 1(1):93 –99, 1990.
8. A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.
9. T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
10. C. Kray, D. Nesbitt, J. Dawson, and M. Rohs. User-defined gestures for connecting mobile phones, public displays, and tabletops. In *MobileHCI '10*, pages 239–248, 2010.
11. I. S. MacKenzie. Fitts' law as a research and design tool in human-computer interaction. *Hum.-Comput. Interact.*, 7:91–139, 1992.
12. G. Pask. A comment, a case history and a plan. In J. Reichardt, editor, *Cybernetics, Art and Ideas*, pages 76–99. New York Graphic Society, 1971.
13. K. Pryor. *Don't Shoot the Dog: The New Art of Teaching and Training*. Ringpress Books, 2002.
14. J. Rico and S. Brewster. Usable gestures for mobile interfaces: evaluating social acceptability. In *CHI '10*, pages 887–896, 2010.
15. A. Scoditti, R. Blanch, and J. Coutaz. A novel taxonomy for gestural interaction techniques based on accelerometers. In *IUI '11*, pages 63–72. ACM, 2011.
16. B. F. Skinner. 'superstition' in the pigeon. *Journal of Experimental Psychology*, 38(2):168–172, 1948.
17. R. W. Soukoreff and I. S. MacKenzie. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *Int. J. Hum.-Comput. Stud.*, 61:751–789, 2004.
18. J. Williamson and R. Murray-Smith. Pointing without a pointer. In *CHI '04*, pages 1407–1410. ACM, 2004.
19. J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *CHI '09*, pages 1083–1092, 2009.
20. K. Yatani, K. Tamura, K. Hiroki, M. Sugimoto, and H. Hiromichi. Toss-it: Intuitive information transfer techniques for mobile devices using toss and swing actions. *IEICE - Trans. Inf. Syst.*, E89-D:150–157, 2006.