

Instructing People for Training Gestural Interactive Systems

Simon Fothergill¹, Helena M. Mentis², Pushmeet Kohli³, & Sebastian Nowozin³

¹Computer Laboratory
Cambridge University, UK
jsf29@cam.ac.uk

²Socio-Digital Systems & ³Machine Learning & Perception
Microsoft Research, Cambridge, UK
{hementis, senowozi, pkohli}@microsoft.com

ABSTRACT

Entertainment and gaming systems such as the Wii and Xbox Kinect have brought touchless, body-movement based interfaces to the masses. Systems like these enable the estimation of movements of various body parts from raw inertial motion or depth sensor data. However, the interface developer is still left with the challenging task of creating a system that recognizes these movements as embodying meaning. The machine learning approach for tackling this problem requires the collection of data sets that contain the relevant body movements and their associated semantic labels. These data sets directly impact the accuracy and performance of the gesture recognition system and should ideally contain all natural variations of the movements associated with a gesture. This paper addresses the problem of collecting such gesture datasets. In particular, we investigate the question of what is the most appropriate semiotic modality of instructions for conveying to human subjects the movements the system developer needs them to perform. The results of our qualitative and quantitative analysis indicate that the choice of modality has a significant impact on the performance of the learnt gesture recognition system; particularly in terms of correctness and coverage.

Author Keywords

Natural gesture recognition; machine learning; data collection; instructing movement

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Miscellaneous

General Terms

Experimentation, Performance, Verification, Standardization.

INTRODUCTION

The last few years have seen a tremendous increase in the popularity of body-movement based interfaces, which offer a convenient and engaging experience of touchless interaction. Traditionally, touchless body movement interfaces

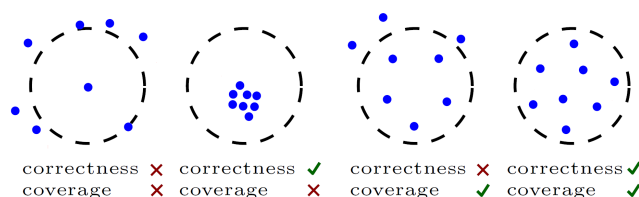


Figure 1. Four collections of gestural performances (blue dots) that illustrate desirable properties of training data sets.

were based on video cameras and were limited to use in applications targeted at a very specialized set of users [6]. The availability of body movement sensing technology in commodity entertainment and gaming systems such as the Nintendo Wii, Sony Playstation Move, and Microsoft Kinect has now made such interfaces available to a much larger audience. Whilst popular for controlling entertainment consoles, applications of such systems also exist in tutoring [6], security [20] and healthcare [30].

These motion sensing systems enable the estimation of movement of various body parts from raw inertial motion or depth sensor data. However, the interface developer is still left with the challenging task of creating a system that recognizes these movements as embodying meaning.

During development of games driven by human movements, developers generally tackle this problem using a trial and error approach. They start by defining a map from body part movement to a set of gestures. This is generally done by specifying a set of rules or conditions on the movements of the body parts under which a particular gesture would be deemed to have happened. An example of such rule would be: 'if both feet simultaneously move upwards, then a *jump* gesture should be detected'. These initial rules are refined by hand by testing their performance on a set of test subjects [24]. This approach does not scale well to more complex gestures and is also not guaranteed to lead to a continual increase in system accuracy or performance.

The machine learning (ML) approach for the gesture recognition problem requires the collection of data sets that contain examples of movements and their associated gesture label. In essence, a machine learning algorithm tries to teach the system what movements can represent a particular gesture. The accuracy of the system is therefore influenced by the anthropomorphic and behavioral kinematic variation in the set of example gestures that are used to train it.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

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

Copyright 2012 ACM 978-1-4503-1015-4/12/05...\$10.00.

For an accurate and responsive form of interaction, not only must the set of performers providing training data be representative of the target population but the dataset of movements used for training the system must reflect what is ideally or most likely to occur during system deployment. In other words, not only must there be examples of only desired gestures (we refer to this property as *correctness* of the dataset), but in order to cope with a wide array of users and their corresponding abilities the dataset must include common, desired variants of the particular movements associated with the gestures (referred to as *coverage* of the dataset).

We explain the desirable properties of a training dataset using Figure 1. The circle in the figure represents the space of a movement that the system developer wants to recognize (e.g. our jump example from before - both feet simultaneously moving upwards). In the left-most picture, the movements performed by eight human subjects lie outside the circle. In other words they do not belong to a common space of movements. In the center-left picture, the movements lie within the circle but do not cover it (e.g. where ‘jump’ movements are collected from only athletic subjects, and there were no movements which would reflect the ‘jump’ gesture made by less mobile users). Thus, they are correct, but with little coverage. In the center-right picture, the movements cover the space of movements, but some of them lie outside the circle and are an inaccurate reflection of the gesture. They have coverage, but some are outside of the space of movement and, thus, would be perceived as incorrect. In the right-most picture, the set of movements collected from the eight human subjects are accurate and also cover a space of movements. They are correct and have coverage.

Developers usually use human subjects to generate the data used to train and test the machine learning system. To convey the body movements a designer associates with each gesture, they give the subjects some instructions. These instructions or signs can be of different semiotic modalities including text, images, video, and combinations of the above [11]. However, there has been no study of what biases are introduced by these different modalities on correctness and coverage. In this paper, we investigate the questions:

R1: Does the semiotic modality of instructions for collecting training data affect the performance of the gesture recognition system?

R2: In what way does semiotic modality of instructions affect correctness and coverage?

We investigate what is the most appropriate semiotic modality of instructions for conveying to human subjects the movements the system developers want to associate with particular gestures in order to achieve (1) Correctness and (2) Coverage (see Figure 1). We analyze the questions from both the performers’ perspective and through the accuracy of the trained gesture recognition system. We present our findings on how different semiotic modalities affect the performers’ understanding, freedom of expression, and, ultimately, the inter and intra modality generalization performance of the

trained gesture recognition system. This study is timely, given the expanding market for gestural interactive applications; conventions will begin to be revealed and established both for gestures and semiotic modalities of instruction.

RELATED WORK

The problems of detecting and recognizing gestures from human body movements captured using videos or 3D skeletal data have been extensively studied in the computer vision and machine learning communities [2, 31, 32, 23]. They are primarily focused on the developments of mathematical models that can generalize the semantics-kinematics mapping learned from a set of training examples to unseen data. These studies have generally ignored the problem of how to collect the set of movements associated with a gesture.

There is little work in the Human-Computer Interaction (HCI) and Computer Vision (CV) literature on the problem of how to specify which movements need to be performed by subjects generating data for training a gesture recognition system. A number of datasets of body movements corresponding to different gestures have been collected by the Computer Vision community; however, compared to the *Cambridge Gestural Performance Database 2012* (CGPD12) collected for this study, they do not provide details on how the performers were instructed, as shown in Table 1.

The primary meanings of instructions are to convey the kinematics (the features of motion of the body). We are interested in both the semiotic semantics and immediate pragmatics of the different modalities. The effect they should have on those who interpret them is to instruct these people in the performance of kinematics signified while allowing for them to perform the movement as they feel is most natural.

In the field of semiotics, Peirce [22] outlines three phenomenological categories of signs emphasizing the way it denotes the object of reference: icon, index, and symbol. An icon denotes its object by resembling or imitating its object (e.g. diagrams). An index denotes its object through a connection (e.g. smoke is an index of fire). Finally, a symbol denotes its object solely through interpretation which are usually formed through social convention (e.g. alphanumeric symbols).

There are examples of using iconic and symbolic instructions to convey kinematics pictures, movies, or text that aim for correctness, but not coverage. For instance, methods such as Labanotation provide a detailed approach of describing dance movements, but the complexity of the notation puts it beyond the reach of novices [9]. Teaching dance moves has been investigated [5], but only from the point of view of correctness at deployment when they use correctness. In addition, they only investigate differences in iconic video instructions for sequences of iconic gestures.

In addition to correctness, coverage requires a mechanism that supports the gathering of a wide assortment of performances for each gesture. This has been far less addressed in research as it has less importance in other areas of instruc-

Name	Year	Ref.	Domain (M=metaphoric, I=iconic)	Classes	Actors	Instances	Instructions	Body parts	Annotation precision
CGPD12*†‡◊ (this dataset)	2012		gaming, music, dance (M+I)	12	30	6000	text/video/image (single & composite)	full body	frame
CGD2011*◊	2011	[1]	assorted (M+I)	15	NR	30,000	NR	upper body	performance
HMDB*	2011	[12]	natural (M+I)	51	(6849)	6849	actors or none	full/upper body	performance
Keck military*	2009	[14]	signalling (M)	14	3	294	Handbook pictograms	hands	frame-ranges
UCF YouTube*	2009	[15]	sports (I)	11	(1168)	1168	none	full body	performance
Hollywood2*	2009	[16]	cinema (M+I)	12	NR	884	actors	full body	performance
Hollywood*	2008	[13]	cinema (M+I)	8	NR	430	actors	full body	performance
UCF Sports*	2008	[26]	sports (I)	9	(200)	200	none	full body	performance
Weizmann Actions*	2007	[8]	natural (M+I)	10	9	90	NR	full body	performance
KUG*†‡	2006	[10]	assorted (M+I)	54	20	NR	NR	full-body	performance
KTH Actions*	2004	[28]	natural (M+I)	6	25	2391	NR	full body	performance

Table 1. Popular benchmark datasets for gesture recognition (NR = Not reported, * = video, † = 2D stereo images, ‡ = 3D data, ◊ = depth map data)

tion. But in machine learning, gathering a variety of samples for training data significantly improves the performance of a recognition system during testing and deployment. For instance, the importance of similarity between training and test data has already been recognized for speech recognition. It is understood that a speech recognition system trained using speech samples collected from people in a very nervous or overly excited state would not have good performance on speech samples obtained from people in a relaxed state. Furrui *et al.* [7] show a larger training dataset is required for recognition of spontaneous speech compared to speech generated by reading a manuscript as the sound of spontaneous phonemes take up a smaller spectral space compared to phonemes generated by ‘reading’. Other challenges that arise in data collection for voice and speech recognition are described in [21] where the authors tried to collect continuous and spontaneous speech samples occurring in day-to-day life. They asked human subjects to summarize a written passage in their own words, rather than reading a passage or using isolated words. They further characterise training data in terms of whether there is a human audience or whether the speaker gets feedback. These characteristics are pertinent to our task of detecting human body movements but beyond this paper’s scope although collecting gestures tend to be more contrived as it is harder to record them as they occur in day-to-day life.

STUDY METHODOLOGY

Gestures

In addition to investigating the effect of instruction modality, we are also looking at whether the type of gesture makes a difference in the affect of modality. We introduced two gesture types based on McNeil [17] categorize of gesticulation. The first was Iconic gestures - those that imbue a correspondence between the gesture and the reference. The second was Metaphoric gestures - those that represent an abstract content. For the former, we borrowed six gestures from a first person shooter game (Table 2) and for the latter, we borrowed six gestures for a music player (Table 3).

Instructions Tested

We chose to provide participants with three familiar, easy-to-prepare instruction modalities and their combinations that





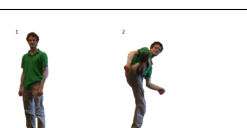

Gesture Outcome	Descriptive Text	Static Images
Crouch or hide	<i>Squat down or crouch</i>	
Shoot with a pistol	<i>Stretching your arms out in front of you and holding your hands together to form a pistol, make a recoil movement</i>	
Throw an object such as a grenade	<i>Using your right arm, make an overarm throwing movement</i>	
Change weapon	<i>Reach over your left shoulder with your right hand and then bring both hands in front of your body as if you are holding something</i>	
Kick to attack an enemy	<i>Karate kick forwards with your right leg</i>	
Put on night vision goggles to change the game mode	<i>Bring your hands up to your eyes as if they were goggles</i>	

Table 2. Descriptive Text and Static Image Instructions for Iconic Gestures.

did not require the participant to have any sophisticated knowledge. The three were (1) descriptive text breaking down the performance kinematics, (2) an ordered series of static images of a person performing the gesture with arrows annotating as appropriate, and (3) video (dynamic images) of a person performing the gesture. We wanted mediums to be *transparent* so they fulfill their primary function of conveying the kinematics. Text is analogous to someone explaining







Gesture Outcome	Descriptive Text	Static Images
Start music / raise volume	<i>Raise outstretched arms</i>	
Navigate to next menu	<i>Slide right hand, palm down in front of you, from left to right</i>	
Wind up the music	<i>Make circular movements with both arms, in front of your body, clockwise with right hand and counter-clockwise with left hand</i>	
Take a bow to end the session	<i>Bend forward at the waist, pause and come back up again</i>	
Protest the music	Pause and rest your hands on your head	
Lay down the tempo of a song	<i>Beat the air with both your hands</i>	

Table 3. Descriptive Text and Static Image Instructions for Metaphoric Gestures.



Figure 2. Example of Dynamic Video Capture for Wind up the Music

a gesture verbally without a bias of a speaker with an accent. Static images correspond with any static drawings and video is also analogous to live demonstration; although these pictorial modalities may again be biased by the performer they depict. The two combinations of modalities were the simultaneous juxtaposition of descriptive text with each pictorial modality.

Textual descriptions with varying degrees of verbosity were possible over all gestures and each description was determined by the authors. The videos were all of the first author performing the gestures as defined by each application's designer and started and stopped with the beginning and end of the gesture. They were filmed in front of a white background ensuring all body movements were within frame (Figure 2 shows an extreme example).

For the static images, individual video frames were extracted at points the designer considered necessary to fully define the gesture. For clarity, the background was removed and

arrows sometimes added indicating direction of movement at the designer's discretion. The frames were presented horizontally and chronologically.

Data Collection Materials

Two types of structured questionnaires (an after-instruction questionnaire and a final ranking questionnaire) and an open-ended final interview were used to gather participant subjective data. The former was administered after every gesture was performed. It consisted of 11 psychometric questions on how well the participant felt they understood the instructions (7 questions) and if the participant felt they were able to perform the gesture freely (4 questions) - all questions were rated on a 4-point Likert scale from Strongly Disagree to Strongly Agree. The latter was administered at the end of the study and consisted of 11 questions on the same concerns as the after-instruction questionnaire; however, the participants were asked to rank the instruction methods (1-5) for each question.

Questions regarding participants' understanding of instructions were posed in terms of the clarity of the movements, importance of body parts, amount of information about desired movements, effort required for their interpretation, their ambiguity, the correctness of performances they allowed, and the amount of practice they required.

Questions regarding participants’ feeling of freedom were posed in terms of feeling of confident whilst performing, feeling inhibited, being in control, and feeling odd.

The performances and final interview was also recorded using a normal video camera for later review and transcription. A markerless motion capture system was used (Microsoft Kinect) to record the 3D position of skeletal joints at 30Hz to within approximately 10cm accuracy [29].

Participants

Thirty participants were recruited from a multicultural, industry research lab and a university computer science department in the UK. Although some of the participants were familiar with the domain of machine learning and computer vision, none of the participants were privy to the workings of the machine learning algorithm of the study we were conducting. The demographics of the participants were 60% male, 93% right-handed, 5’0”-6’6” tall with an average of 5’8”, and 22-65 with an average of 31 years of age.

Each participant performed each gesture based on at least one semiotic modality. Since two of our conditions were combinations of other modalities, we had participants do some gestures in two conditions. For example, a participant would shoot a pistol instructed by descriptive text first and then, after completing an after-instruction questionnaire, be instructed to shoot a pistol with descriptive text plus static images. We ensured that participants did not receive a multimodal instruction followed by a unimodal instruction for the same gesture in order to handle any significant learning effect. In addition, the two-step condition provided us with an opportunity to investigate the ordering of sequentially presented pairs of instruction that increased in modalities used (i.e. we could investigate if text followed by text plus images was a preferential order for eliciting appropriate gestures).

Procedure

The experiment was conducted in a large private space with one experimenter. Participants were told that we were investigating instructions for performing gestures for the sole purpose of training a gestural interaction system.

They then were asked to stand and face a 30” LCD TV with a Kinect sensor in front of it. When they indicated they were ready, the first gesture’s instructions appeared on the screen in a PowerPoint slideshow. At the top of each slide, the application category (e.g. music player) and gesture outcome were displayed (see Table 2 and 3) and below that the instruction was placed using the appropriate modality. The participants had as much time as they desired to read or watch the instructions. Questions were not addressed by the experimenter and instead the participants were told to ‘do what they wanted’.

When they indicated they were ready to begin, they were instructed to perform the gesture ten times and to ensure that there was a pause between each repetition of the gesture. When all ten repetitions of the gesture were completed, the participant returned to sit at the table and completed an after-

instruction questionnaire. They repeated this process for 20 gesture instructions.

At the end of the study, they were asked to complete the ranking questionnaires and then the participants were interviewed by the experimenter with open-ended questions regarding their experience and their opinions of the different semiotic modalities.

DATA ANALYSIS

Questionnaire and Interview Data Analysis

The 11 post-action questions were grouped in two clusters, one for understanding and one for freedom. The reliability analysis shows that the items in each cluster were highly intercorrelated: the Cronbach’s alpha values were .92 for Understanding and .86 for Freedom. For high reliability, Nunnally and Bernstein [19] suggest to use a cut-off of .7, thus, we could compute aggregated scores for each factor.

ML-based Recognition of Gestures and Analysis

The long history of automatically recognising gestures from visual or kinematic measurements is reviewed in [23] [32] [31] [2]. We address only the recognition of relatively simple human gestures of a few seconds, not activities of minutes or hours. For these short gestures, Schindler and van Gool [27] have shown that short windows of measurements are sufficient to obtain state-of-the-art recognition performance. (For a more extensive of the method used, refer to [18]).

Mathematical Notation

We assume a small vocabulary of gestures \mathcal{A} is given. Each gesture $a \in \mathcal{A}$ has associated to it an *action point* that is characteristic for the gesture. As an example, for a punch we can define the action point as the first point in time at which the arm is straight out in front. We further denote by $x_t \in \mathbb{R}^q$ an observation vector at discrete time t , and by $x_{s:t}$ the sequence $(x_s, x_{s+1}, \dots, x_t)$ of observations.

Performance Measure: F-score@ Δ

The performance of the system is measured in terms of *precision* and *recall*. To achieve a high precision, the training data should only contain movements that users of the deployed system will associate with the gesture (earlier referred to as *correctness* of the dataset in Figure 1). To achieve a high recall, the training data should contain all movements that the designer wants to associate with a gesture (earlier referred to as *coverage*).

We assess the quality of our predictions using ground truth annotations. To this end, we define a performance measure that captures the characteristics of the system in an online setting. These are, its *precision*—how often is the gesture actually present when the system claims it is, its *recall*—how many true gestures are recognized by the system, and its *latency*—how large is the delay between the true action point and the systems prediction.

For a specified amount of tolerated latency (Δ ms) we measure the precision and recall for each gesture $a \in \mathcal{A}$, as

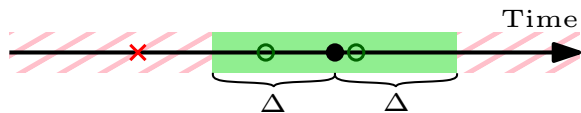


Figure 3. Latency-aware measure of predictive performance for a single gesture: a fixed time window of size 2Δ is centered around the ground truth (marked \bullet) and used to partition the three predicted firing events into correct (marked \circ) and incorrect predictions (marked \times); precision = 0.5, recall = 1.

shown in Figure 3. A balanced F-score between 0 and 1 [25] combines precision and recall. In the experiments we will examine the performance measure for a fixed $\Delta = 333ms$.

Gesture Recognition using Random Forests

We now show how random forest classifiers [3] can be straightforwardly adapted to the problem of recognizing gestures. Our approach is similar to that of Yao et al. [33] but allows online gesture recognition.

At test-time, for a time t , we derive a feature vector $\phi_t = \phi(x_{t:(t-\ell+1)}) \in \mathbb{R}^d$ from the last ℓ observations x_t to $x_{t-\ell+1}$. We use 35 skeletal joint angles, 35 joint angle velocities, and 60 xyz-velocities of joints for a 130-dimensional feature vector at each frame. We use $\ell = 35$ frames, obtaining $d = 4550$. The feature vector ϕ_t is evaluated by a set of M decision trees, where simple tests $f_\omega : \mathbb{R}^d \rightarrow \{\text{left}, \text{right}\}$ are performed recursively at each node until a leaf node is reached. The parameters $\omega \in \Omega$ of each test are determined separately during the training phase, to be described below. Each tree $m = 1, \dots, M$ produces one class decision $y_t^{(m)}$ and the *posterior class distribution*

$$p(y_t = a | x_{(t-\ell+1):t}) := \frac{1}{M} \sum_{m=1}^M I(y_t^{(m)} = a) \quad (1)$$

over gesture classes \mathcal{A} and a background class “None” determines whether a gesture is recognized. If for a gesture class $a \in \mathcal{A}$ we have $p(y_t = a | x_{(t-\ell+1):t}) \geq \delta$, we fire the gesture as being detected at the current time t . We use a fixed value of $\delta = 0.16$ for all experiments. This value has been determined from previous runs.

For training we use the full observations and action point annotations for a set of N sequences, where the n ’th sequence is an ordered list $(x_t^{(n)}, y_t^{(n)})_{t=1, \dots, T_n}$. Our goal is to learn a set of M decision trees that classify the action points in these sequences correctly by means of (1). We use simple “decision stump” tests [3] with $\omega = (i, h)$, $1 \leq i \leq d$, $h \in \mathbb{R}$,

$$f_{(i,h)}(\phi_t) = \begin{cases} \text{left} & \text{if } [\phi_t]_i \leq h, \\ \text{right} & \text{otherwise.} \end{cases}$$

We use the standard *information gain* criterion and training procedure [3]. Hence, we greedily select a split function $f_{(i,h)}$ for each node in each decision tree from a set of randomly generated proposal split functions. The tree is grown until the node is pure, that is, all training samples assigned to that node have the same label.

Instr. Modality	All	Metaphoric	Iconic
Text	0.479 ± 0.104	0.432 ± 0.089	0.708 ± 0.103
Images	0.549 ± 0.102	0.462 ± 0.132	0.742 ± 0.088
Video	0.627 ± 0.053	0.612 ± 0.056	0.683 ± 0.123
Images+Text	0.563 ± 0.045	0.506 ± 0.095	0.750 ± 0.079
Video+Text	0.679 ± 0.035	0.651 ± 0.099	0.765 ± 0.070

Table 4. F-Score at $\Delta = 333ms$ for correctness by training and testing on the same modality. We show the average and standard deviations over ten leave-persons-out runs.

	Text	Images	Video	Images+Text	Video+Text
T	$.621 \pm .041$	$.537 \pm .020$	$.602 \pm .034$	$.611 \pm .048$	$.644 \pm .070$
I	$.504 \pm .048$	$.561 \pm .034$	$.700 \pm .034$	$.646 \pm .093$	$.611 \pm .060$
V	$.524 \pm .032$	$.551 \pm .041$	$.673 \pm .020$	$.588 \pm .070$	$.707 \pm .069$
IT	$.651 \pm .020$	$.582 \pm .023$	$.647 \pm .042$	$.629 \pm .051$	$.680 \pm .074$
VT	$.583 \pm .051$	$.571 \pm .038$	$.702 \pm .021$	$.607 \pm .065$	$.709 \pm .047$

Table 5. F-Score at $\Delta = 333ms$ for coverage, using all combinations of training and testing modalities for all gestures (metaphoric and iconic). Rows correspond to a single training modalities, columns to the test modality.

Recognition Model Performance Assessment

We are interested in measuring the inter-person generalization performance of our gesture recognition system. To this end, we follow a “leave-persons-out” protocol: for each instruction modality we remove a set of people from the full data set (of 30 people) to obtain the minimum test set that contains performances of all gestures. The remaining larger set constitute the training set. After training on this set the generalization performance is assessed on the people in the test set. This is repeated ten times for fixed disjoint sets of test persons. The average test performance over the ten runs is a good estimator of the generalization performance of the system trained on this instruction modality. We perform two separate experiments as follows.

Test 1. We assess the intra-modality generalization performance: training and testing using the same instruction modality. Hence we take only those sequences for training and testing that originate from performances with the respective instructions. As results we obtain five F-scores, one for each modality, and each being an average over all 10 repetitions and 12 gestures (see Table 4). We also report separately the F-scores achieved on the first-person-shooter gestures (six gestures) and the music player gestures (six gestures).

Test 2. We assess the inter-modality generalization performance: training on one modality, for example text, but testing on a different modality, for example videos. We evaluate all possible training-testing combinations where training and testing modalities differ. As a results we report there sets of 5-by-5 average F-scores, one for each gesture set: all (Table 5), music (Table 6), and first-person-shooter (Table 7). The five rows correspond to the training modalities, and the five columns correspond to the testing modalities.

RESULTS

Correctness of Gestures

In the following analyses, the F-scores from 10 runs were compared with one-way, between subjects ANOVAs between the five conditions: Text, Images, Video, Images+Text, and Video+Text.

	Text	Images	Video	Images+Text	Video+Text
T	.621 ± .059	.387 ± .024	.488 ± .031	.573 ± .079	.619 ± .090
I	.461 ± .074	.441 ± .069	.684 ± .058	.566 ± .120	.521 ± .122
V	.461 ± .029	.382 ± .037	.552 ± .025	.494 ± .076	.696 ± .094
IT	.673 ± .041	.423 ± .049	.558 ± .032	.574 ± .096	.607 ± .097
VT	.622 ± .038	.404 ± .042	.583 ± .040	.571 ± .089	.698 ± .092

Table 6. F-Score at $\Delta = 333ms$ for coverage using all combinations of training and testing modalities for the metaphoric gestures. Rows correspond to a single training modalities, columns to the test modality.

	Text	Images	Video	Images+Text	Video+Text
T	.756 ± .053	.838 ± .041	.843 ± .033	.807 ± .075	.807 ± .083
I	.663 ± .035	.810 ± .020	.861 ± .037	.812 ± .090	.805 ± .076
V	.711 ± .015	.821 ± .061	.825 ± .050	.808 ± .090	.842 ± .074
IT	.682 ± .025	.899 ± .030	.858 ± .021	.808 ± .069	.833 ± .064
VT	.685 ± .023	.857 ± .037	.883 ± .034	.778 ± .091	.783 ± .060

Table 7. F-Score at $\Delta = 333ms$ for coverage, using all combinations of training and testing modalities for the iconic gestures. Rows correspond to a single training modalities, columns to the test modality.

For all gestures, there was a significant difference between the five condition means, $F(4, 45)=10.768$, $p \leq .01$ (see Figure 4 and Table 4). Tukey’s HSD post-hoc analyses revealed that Video alone was more effective than Text alone, and Video+Text was more effective than Text alone, Images alone, or Images+Text (all at $p \leq .01$). In addition, for the metaphoric gestures, there was a significant difference between the five condition means, $F(4, 45)=9.643$, $p \leq .01$. A post-hoc analysis revealed that Video alone was more effective than Text or Images and Video+Text was more effective than Text, Images, or Images+Text (all at $p \leq .01$). However, for the iconic gestures, there was no significant difference between the five condition means. Thus, although the instructions’ semiotic modality made a difference in F-scores for the metaphoric gestures, it made no difference in the F-scores for the iconic gestures.

A series of 2x2 ANOVAs between Static vs. Dynamic (Images vs. Videos) and Text Added vs. No Text Added further shows that Dynamic Images (Videos) were better than Static Images (Images) despite whether any text was added for All Gestures ($p \leq .01$) and metaphoric gestures ($p \leq .01$). However, there was no difference for the iconic gestures.

Thus, Video alone and Video+Text were better than the other semiotic modalities in terms of achieving correctness in performing the metaphoric gestures. However, Video alone was statistically just as effective as Video+Text. On the other hand, there is no discernible difference between the instructions’ semiotic modality for the iconic gestures.

In the analysis of the questionnaire data on Understanding the instructions, a series of one-way within subjects ANOVAs were performed on All gestures, metaphoric gestures, and iconic gestures. All Gestures showed that the means are significantly different between the five conditions ($F(4,116) = 18.866$, $p \leq .01$) (Figure 5). A series of paired t-tests with a Bonferroni correction confirmed this in that it showed that Video alone was better understood than Images or Text; Images+Text was better understood than Images or Text; and Video+Text was better understood than Images or Text (all at $p \leq .01$). Furthermore, a one-way within subjects ANOVA

on Understanding of only metaphoric gestures showed that the means are significantly different between the five conditions ($F(4,56)=4.466$, $p \leq .01$). A series of paired t-tests with a Bonferroni correction showed that Video+Text was better understood than Text and Images+Text, Video was better than Images and Text, and Images+Text was better than just Text (all at $p \leq .01$). In addition, a one-way within subjects ANOVA on Understanding of iconic gestures showed that the means are significantly different between the five conditions ($F(4,56) = 3.439$, $p \leq .01$). A series of paired t-tests with a Bonferroni correction showed that Video+Text was better understood than Images and Text, Images+Text was better understood than Text, and Videos were better understood than Text (all at $p \leq .01$).

A series of 2x2 repeated measures ANOVAs on Understanding further showed that the means are different for Text vs. No Text and Static vs. Dynamic for All Gestures (both at $p \leq .01$) and metaphoric gestures (both at $p \leq .05$), but there were no significant differences to report for the iconic gestures (both at $p \leq .05$). Thus, like the F-score analysis, we see that the modality of Video (plus Text) yielded a better understanding of what was to be performed.

This analysis is corroborated by a review of the interviews. Participants generally related that the videos were the clearest and one knew exactly what to do. In addition, many of the participants appreciated that the addition of text specified exactly what was the important aspect of the gesture for the system recognition.

“I would say the video [is the clearest] because the text really wasn’t necessary because then we got all the information throughout the video.”

“The text provided a specific of what the sensor was going to pick up, so if I saw a video it was not always clear what was important in the video but once I had a text to go with it, it seems much clearer.”

However, participants also explained that the videos were not as necessary for the iconic gestures since they felt they could understand what was being requested of them from previous experiences.

“The kicking was alright, the throwing was fine, they are gestures you do in everyday life, in sports for example.”

Coverage of Gestures

In the following F-score analyses, the algorithm was trained on one modality (e.g. Text) and then tested on a set of data from each of the five modalities ten times each. Thus, we present the analyses from the F-scores averaged across the five testing modalities.

For All gestures, a one-way, between subjects ANOVA showed a significant difference between the five condition means, $F(4, 45)=7.327$, $p \leq .01$ (see Figure 6 and Table 5). Tukey’s HSD post-hoc analyses revealed that training on Videos+Text was more effective than Text alone ($p \leq .01$), Images ($p = .01$),

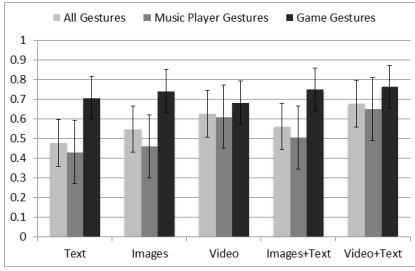


Figure 4. Correctness F-Scores between all modalities and their combinations for all, metaphoric and iconic gestures

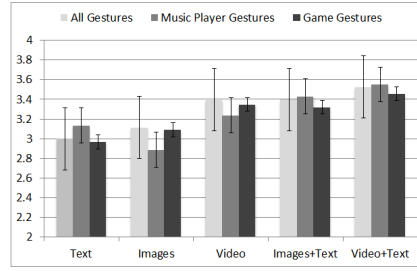


Figure 5. Means for Understanding Questions between the three modalities and their combinations for all, metaphoric and iconic gestures

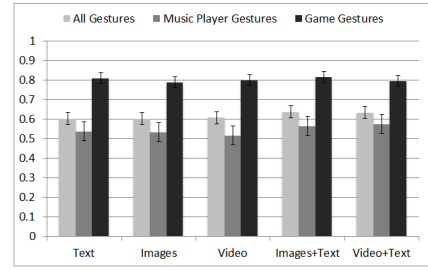


Figure 6. Coverage F-Scores between all modalities and their combinations for all, metaphoric and iconic gestures

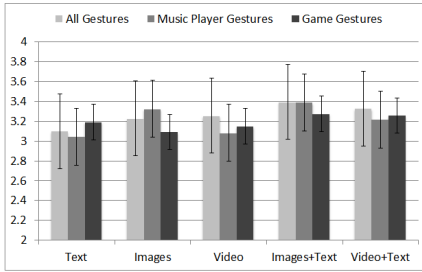


Figure 7. Means for Freedom Questions between the three modalities and their combinations for all, metaphoric and iconic gestures

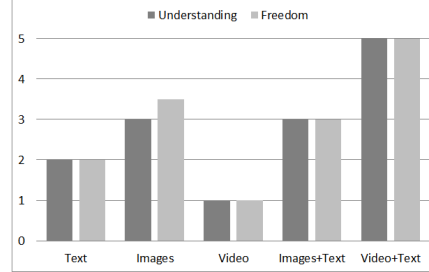


Figure 8. Average ranking of modalities for Understanding and Freedom, retrospectively over all gestures

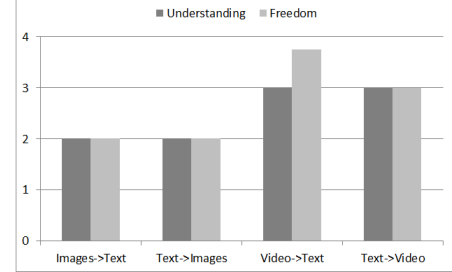


Figure 9. Average ranking of order of modalities for multimodal instructions for Understanding and Freedom, retrospectively over all gestures.

or Videos alone ($p=.04$) and Images+Text was more effective than Text ($p\leq.01$), Images alone ($p\leq.01$), or Videos ($p=.02$). In addition, for the metaphoric gestures, a one-way, between subjects ANOVA showed a significant difference between the five condition means, $F(4, 45)=6.604$, $p\leq.01$ (Table 6). Tukey's HSD post-hoc analyses revealed that Video+Text was more effective than Video ($p\leq.01$), Images ($p=.03$), or Text ($p=.05$) and Images+Text was more effective than Videos alone ($p\leq.01$). However, for the iconic gestures, there was no significant difference between the five condition means (Table 7).

If we look at each testing modality separately we see that some training modalities are better than others for covering the variation. For instance, for All gestures, testing on a data set instructed through Text or Images, we found that Images+Text yielded a significantly better result (both at $p\leq.01$). However, testing on a data set instructed on Videos, we found that Videos+Text yielded a significantly better results ($p\leq.01$). If we specifically look at Text as an instruction modality for testing with the assumption that it is most representative of executed gestures in actual system use (i.e. the most variation), then we see that, for the metaphoric gestures, testing on a data set instructed through Text we found that Images+Text yielded a significantly better results ($p\leq.01$) and, for the iconic gestures, we found that when testing on a data set instructed through Text the trend signifies that the best results came from training on Text (n.s.). Thus, we see that in terms of capturing natural variation less information was optimal.

In the analysis of the questionnaire data, a one-way within subjects ANOVA on Freedom Questions showed that the means are significantly different between the five conditions ($F(4,116) = 5.390$, $p\leq.01$) (Figure 7). A series of paired t-

tests with a Bonferroni correction showed that Images+Text are significantly less inhibiting than Text ($p\leq.01$), and Video+Text provides a greater sense of freedom than Text (both at $p\leq.01$). Although not a strong difference between instruction methods, the general trend is that as more information was provided, the sense of inhibition lowered. But the instructions' semiotic modality with the most information, Video+Text, yielded a slightly less sense of freedom than Images+Text. Freedom Questions for the metaphoric gestures showed that the means are significantly different between the five conditions ($F(4,56) = 3.469$, $p=.013$) and a series of paired t-tests with a Bonferroni correction showed that Images+Text are significantly less inhibiting than Text ($p\leq.01$). Again, not a very strong difference, but Images+Text provide a sense of more freedom than the other semiotic modalities. Finally, Freedom Questions for iconic gestures showed that the means are the same between the five conditions. However, if one looks at the graph, again, we can see the trend is for Images+Text to be providing a greater sense of freedom. A 2x2 repeated measures ANOVA on Freedom for All, metaphoric, and iconic gestures showed that the means are the same for Text vs. No Text Added, Static vs. Dynamic, and the interaction between dynamicity and text-added.

From the interviews, we start to understand what was occurring: Participants asked about Freedom were referring less to an ability to improvise than to not feeling apprehensive or embarrassed from potentially misunderstanding or disagreeing with what was being asked of them. In this latter case feeling more free would retract undesirable *coverage* from awkward gestural performances and may even instill confidence that encourages more controlled improvisation.

“Some of them I felt I had prior knowledge of what you were conveying so then you can add a little bit of embellishment.”

When the discussion turned to doing the gesture as they saw fit to give better *coverage*, then they generally agreed that less information allowed this more effectively for metaphoric gestures. This is primarily as a reaction to the videos, which prescribed exactly what to do whereas other mechanisms were seen as being more open for interpretation.

“I think when you have the visual description or the visual images, you really try to do what is on the picture, really. Video. So I wasn’t thinking of any creativity at all.”

“For the hand slide [instructed by the Video+Text] I really tried to match the speed you did it with. I tried to match the violent movement you had. Whereas for the ... in the text [only], the gesture I had for the slide was sort of clear but I could still do it my way and I felt confident.”

“I think just the words on themselves are a lot more ambiguous in a way that allowed you to do how you feel.”

Correctness and Coverage of Gestures

Using Video+Text modalities promotes *correctness* but Images+Text does seem to be a good modality for *coverage*. The ranking questionnaire allows us further investigation of the best recommendation by forcing performers to give a preference over modalities whilst recalling their performances of all gestures. Taking the median of all questions relating to understanding or all questions relating to Freedom and then taking the median of the rankings of modalities over all participants gives a final, non-gesture specific ranking of modalities (see Figure 8). For both Understanding and to a less extent for Freedom, Video+Text is preferred. One reason for this could be the appeasing nature of the participants to accept being told what to do and their lack of a desire or ability to improvise.

Additionally, the participants suggested providing a two tier instruction, no doubt as an outcome of the way we ran the study. Participants experienced, for example, Text alone followed by Video+Text or Images alone and then with added Text for the same gesture. They felt their performances changed with the addition of a modality but were divided over which should come first during the interviews. Some preferred initial room for interpretation, others preferred constraints to be relaxed.

“When you have all the good texts that started, you do, you understand what to do and then when you add the image or the video, you try to do exactly what you are saying rather than what you understand from the description.”

“You get the general picture from the text and then you can fill in the things you are not certain about from the video.”

“Just because there is more interpretation in it for the image, it makes you feel like I have to be like the image is or like the video is so it breaks you. It breaks you in your freedom and text is more free ...”

However, in the overall ranking analysis, Video followed by Text was preferred on average by the participants with respect to both Understanding and Freedom, as shown in Figure 9. Unfortunately, we can not speak to whether this ranking differs for metaphoric or iconic gestures.

DISCUSSION

This study aimed to shed light on how the importance of using a particular semiotic modality to instruct participants can play a significant part in the development of gestural interaction applications. Namely, to understand the robustness of different ways of eliciting movements based on *correctness* and *coverage* requirements of a corpora of machine learning training data that must be collected. Such understanding allows developers to choose a training data collection methodology that suits their needs, without having to empirically justify their chosen kinematics. We investigated three different semiotic modalities - Descriptive Text, Static Images, and Dynamic Video - as well as combinations of Images with Text and Video with Text.

We have shown that different semiotic modalities of an instruction alter the amount of variation in a set of training performances. That this factor has not been addressed in any prior literature now questions the precision of all their results to date and future data sets should include this characterisation. For a robust system, it is necessary to balance accurate recognition with a need to generalise recognition over an unknown population whilst matching the degree of flexibility the application designer’s gestural definitions’ allow.

Our analysis revealed considerations for developers on how and when instructions’ semiotic modality makes a difference. Intra-modality F-scores show Video (with or without Text) is best for *correctness* and Understanding is promoted most by Video + Text. Inter-modality F-scores show Image + Text is best for *coverage* and it also gives the strongest feeling of Freedom. We also learned a bit regarding sequences of instructions. Overall, Video followed by Text was preferred in ranking, but this was only for All Gestures. In addition, multimodal instructions are favored over unimodal, which is supported by [5]. Although we still think the two-step process yielded interesting results, we also realize that the learning method of enaction may also have had an effect on the improvements under both combinations [4]. In other word, the process of enacting the gesture the first time provided a learning advantage to enacting the gesture the second time - despite our lack of giving feedback as to the correctness of the first gestural enactment.

In addition to these general findings we learned more regarding the difference in effect of modality on different types of gestures. Few significant differences existed for the Iconic gestures for which people possessed a priori associations of kinematics. This familiarity from real-life experience could

have increased clarity and confidence so it dominated any differences between modalities. *Correctness* and *coverage* of performances of metaphoric gestures whose semantics-kinematics mapping had to be 'taught' benefited from being able to read the description as well as see an example. However, our results reflect our societies current conventions based on a moderate number of participants with wide demographics. They may not be applicable in 20 years.

We also saw evidence of a performer's tolerance of ambiguity as playing a roll. Those with initially high inhibitions could be more apprehensive and effort should be made to lower them for good *coverage*; this is encouraged by [5] who discourages using mirrors. It may be sensible to let the performer choose the order of a sequential, multimodal instruction. This would be an area of further investigation, though.

ACKNOWLEDGEMENTS

We would like to thank the staff, students and interns of Microsoft Research, Cambridge and the University of Cambridge for participating in our user study. We would also like to thank Olivia Nicell for help with data gathering and tabulation.

REFERENCES

1. Chalearn gesture dataset (cgd2011), chalearn, california, 2011.
2. Aggarwal, J., and Ryoo, M. Human activity analysis: A review. *ACM Computing Surveys* (2011). To appear.
3. Breiman, L. Random forests. *Machine Learning* 45, 1 (2001).
4. Bruner, J. *Toward a theory of instruction*. Belknap Press of Harvard University Press, 1966.
5. Charbonneau, E., Miller, A., and LaViola, J. Teach me to dance: Exploring player experience and performance in full body dance games.
6. Fothergill, S., Harle, R., and Holden, S. Modelling the model athlete : Automatic coaching of rowing technique. In *Structural, Syntactic, and Statistical Pattern Recognition*, vol. 5342 of *LNCIS* (2008), 372–381.
7. Furui, S., Nakamura, M., Ichiba, T., and Iwano, K. Why is the recognition of spontaneous speech so hard? In *Text, Speech and Dialogue*, V. Matouek, P. Mautner, and T. Pavelka, Eds., vol. 3658 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2005, 747–747.
8. Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence* 29, 12 (December 2007), 2247–2253.
9. Guest, A. H. *Labanotation, or, Kinetography Laban: The System of Analyzing and Recording Movements*. Dance Books, 1996.
10. Hwang, B.-W., K. S., and Lee, S.-W. A full-body gesture database for automatic gesture recognition. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, FGR '06*, IEEE Computer Society) (2006), 243–248.
11. Kress, and van Leeuwen. *Reading Images: Grammar of Visual Design*. Routledge, 1996.
12. Kuehne, H., J. H. G. E. P. T., and Serre, T. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2011).
13. Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. Learning realistic human actions from movies. In *CVPR*, IEEE Computer Society (2008).
14. Lin, Z., Jiang, Z., and Davis, L. S. Recognizing actions by shape-motion prototype trees. In *ICCV*, IEEE (2009), 444–451.

15. Liu, J. G., Luo, J. B., and Shah, M. Recognizing realistic actions from videos 'in the wild'. In *CVPR* (2009), 1996–2003.
16. Marszalek, M., Laptev, I., and Schmid, C. Actions in context. In *CVPR*, IEEE (2009), 2929–2936.
17. McNeil, D. *Hand and Mind, What Gestures Reveal about Thought*. The University of Chicago Press, 1992.
18. Nowozin, S., and Shotton, J. Action points: A representation for low-latency online human action recognition.
19. Nunnally, J. C., and Bernstein, I. H. *Psychometric Theory*. McGraw-Hill, 1994.
20. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., and et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR* (2011).
21. Padmanabhan, M., Ramaswamy, G., Ramabhadran, B., Gopalakrishnan, P. S., and Dunn, C. Issues involved in voicemail data collection. In *DARPA Hub 4 Workshop* (1998).
22. Peirce, C. On a new list of categories. *Proceedings of the American Academy of Arts and Sciences* (1867).
23. Poppe, R. A survey on vision-based human action recognition. *Image and Vision Computing* 28, 6 (2010), 976–990.
24. Quinn, D. Personal communication with David Quinn (RARE, UK), August 2011.
25. Rijsbergen, C. J. V. *Information Retrieval*. Butterworths, 1979.
26. Rodriguez, M. D., Ahmed, J., and Shah, M. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, IEEE Computer Society (2008).
27. Schindler, K., and Gool, L. J. V. Action snippets: How many frames does human action recognition require? In *CVPR*, IEEE Computer Society (2008).
28. Schüldt, C., Laptev, I., and Caputo, B. Recognizing human actions: A local SVM approach. In *ICPR* (2004), 32–36.
29. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. Real-time human pose recognition in parts from a single depth image. In *CVPR* (2011).
30. Stone, E., and Skubic, M. Evaluation of an inexpensive depth camera for passive in-home fall risk assessment. In *Pervasive Health Conference* (2011).
31. Turaga, P. K., Chellappa, R., Subrahmanian, V. S., and Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn* 18, 11 (2008), 1473–1488.
32. Weinland, D., Ronfard, R., and Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. Tech. rep., INRIA, February 2010.
33. Yao, A., Gall, J., Fanelli, G., and van Gool, L. Does human action recognition benefit from pose estimation? In *BMVC* (2011).