Wave Like an Egyptian — Accelerometer Based Gesture Recognition for Culture Specific Interactions

Matthias Rehm
University of Augsburg
Eichleitnerstr. 30
D-86159 Augsburg, Germany
+49 (0)821 598 2343
rehm@informatik.uni-augsburg.de

Nikolaus Bee
University of Augsburg
Eichleitnerstr. 30
D-86159 Augsburg, Germany
+49 (0)821 598 2338
bee@informatik.uni-augsburg.de

Elisabeth André
University of Augsburg
Eichleitnerstr. 30
D-86159 Augsburg, Germany
+49 (0)821 598 2341
andre@informatik.uni-augsburg.de

ABSTRACT
The user's behavior and his interpretation of interactions with others is influenced by his cultural background, which provides a number of heuristics or patterns of behavior and interpretation. This cultural influence on interaction has largely been neglected in HCI research due to two challenges: (i) grasping culture as a computational term and (ii) infering the user's cultural background by observable measures. In this paper, we describe how the Wiimote can be utilized to uncover the user's cultural background by analyzing his patterns of gestural expressivity in a model based on cultural dimensions. With this information at hand, the behavior of an interactive system can be adapted to culture-dependent patterns of interaction.

Categories and Subject Descriptors
H.1.2 [Models and Principles]: User/Machine Systems — human factors, human information processing; **I.5.5 [Pattern Recognition]:** Implementation — interactive systems

General Terms
Design, Human Factors.

Keywords
Gesture recognition, Bayesian network modeling, Cultural interactions

1. INTRODUCTION
Our cultural backgrounds largely depend how we interpret interactions with others, which aspects we find relevant, and what kind of behavior is deemed annoying or insulting. Culture is pervasive in our interactions and influences for instance how we negotiate or how close we stand to each other during an interaction. Figure 1 exemplifies typical hand/arm postures of German (crossed arms) and Japanese subjects (joined hands).

If we take the evidence from the literature seriously that users from different cultures interact based on such culture dependent heuristics, then it is necessary to acknowledge these differences for the design of interfaces. In this paper we focus on embodied conversational agents, which serve as anthropomorphic communication devices and thus create even more severe expectations regarding their behavior (verbal as well as nonverbal). Embodied conversational agents as an interface metaphor have a great potential to realize culture specific interaction behavior in several fields of human computer interaction: (i) Information presentation: By adapting their communication style to the culturally dominant persuasion strategy, agents become more efficient in delivering information or selling a point or a product. (ii) Entertainment: Endowing characters in games with their own cultural background has two advantages. It makes the game more entertaining by providing coherent behavior modifications based on the cultural background and it let's characters react in a believable way to (for them) weird behavior of other agents and the user. (iii) Serious games: For educational purposes, experience-based role-plays become possible, e.g. for increasing cultural awareness of users or for augmenting the standard language textbook with behavioral learning.

Figure 1: Typical difference in posture for German (crossed arms) and Japanese (joint hands) [32].

In this paper we address the question if cultural differences in multimodal behavior can be utilized for human computer interaction. We claim that we have to tackle two challenges to this end. On the one hand we have to identify the user's cultural background and on the other hand we need a model on how to use this information in our interactive system, i.e. on how to use culture as a computational notion. Both challenges are

addressed in this paper. We present our approach to automatically uncover the user's cultural background based on his gestural activity. To this end, we make use of the Wiimote controller, which provides acceleration data for the three spatial axes. This informtion on the user's gestural activity is then used for adapting the behavior of virtual characters to reflect behavior patterns found in the user's culture. Some words of caution are in order here. The user's behavior does of course not only depend on the user's cultural background but also on a number of other personal and contextual influences, e.g. on the user's personality, current emotional state, etc. Triandis and Suh [37] for instance review work on cultural influences on personality and culture and give an excellent overview of their interrelations. Thus, in the long run, an integrated model is needed that combines cultural variables and other influence factors. Nazir and colleagues [30] e.g. propose a first model that relates culture and personality in a cognitive architecture.

2. RELATED WORK

Culture has been in the focus of attention relating to design approaches. Marcus and Gould [27] analyze websites from all over the world and show that they are tailored to cultural preferences and differ largely in the features that are deemed necessary for the entry point of a web presence. Gould and colleagues [9] present an additional in-depth comparative analysis of US and Malaysian websites based on the identity and hierarchy dimension of Hofstede [15]. On Malaysian websites, information on an organization and its staff is given in a prominent place, often on the front page of a webpresence. According to Gould and colleagues, this reflects the high power distance of the Malaysian culture. On US websites on the other hand it is difficult to find this information. Instead, websites focus on the task an individual user wants to achieve. Hisham and Edwards [13] take age as an additional variable into account in their case study about Malaysian elderly users.

Choi and colleagues [5] investigate in detail how usability for interfaces on mobile devices depends on the user's cultural background. To this end they utilize two of Hofstede's dimensions (uncertainty and identity) and Hall's [12] notions of context and time perception. By relating these cultural variables to certain interface instantiations they are able to present some links between interface attributes like preference for large amount of information and cultural variables like high uncertainty avoidance.

Whereas the above mentioned studies are concerned with information presentation on websites or other graphical interfaces, others have focused on the relation between the cultural background of a user and interaction styles and interface use. Massey and colleagues [28] examine preferred interaction styles for global virtual teams, which have an enhanced need for efficient communication. Based on Hofstede's dimensions of identity and uncertainty, interaction styles like direct vs. indirect, instrumental vs. affective were examined in relation to the capabilities of different communciation devices like video (conferencing) vs. telephone vs. email. Cultural differences were exemplified with a case study on the use of an asynchronous, text-based online forum, which was in conflict situations more in accordance with the indirect, group oriented style of participants from a collectivistic culture. On the other hand, users with this background experienced difficulties in expressing their opinion only by text as this form of communication deletes most of the contextual clues of face to face communication. The same result is described by Kayan and colleagues [20] for the satisfaction in the use of instant messaging. They show that multi-party audio-video chatting is more popular in collectivist cultures. They relate this effect to the fact that these technologies provide more contextual clues than simple text-based systems. Ford and Gelderblom [7] present a thorough evaluation of cultural effects on interface use, first identifying characteristic cultural dimensions, then defining interfaces in line with opposite ends of these dimensions and then measuring the effect of these interfaces on speed, accuracy and satisfaction levels of users.

Whereas static presentations like e.g., websites can be easily tailored to culture-specific demands during the design process (given that the designer recognizes the challenge), interactive systems pose an additional challenge because they have to react dynamically to situational and contextual factors. To make such systems adaptable to cultural differences in interaction behavior, one needs a set of parameters or rules that allow for influencing the system processes. Most approaches in this area concentrate on learning environments or interactive role-plays with virtual characters. Khaled and colleagues ([22];[23]) focus on cultural differences in persuasion strategies and present an approach of incorporating these insights into a persuasive game for a collectivist society. Maniar and Bennett [25] propose a mobile learning game to overcome cultural shock by making cultural differences aware to the user. Johnson and colleagues [19] describe a language tutoring system that also takes cultural differences in gesture usage into account. The users are confronted with some prototypical settings and apart from speech input, have to select gestures for their avatars. Moreover they have to interpret the gestures by the tutor agents to solve their tasks. Warren and colleagues [39] as well as Rehm and colleagues [31] aim at cross-cultural training scenarios and describe ideas on how these can be realized with virtual characters. Jan and colleagues [17] describe an approach to modify the behavior of characters by cultural variables relying on Hofstede's dimensions. The variables are set manually in their system to simulate the behavior of a group of characters.

To sum up, most of the above mentioned approaches rely on a dimensional theory of culture, which is presented in detail in the next section.

3. ENCULTURATING HUMAN COMPUTER INTERACTION

To integrate culture as a contextual factor into the human computer interaction, two tasks have to be solved. On the one hand, the system's behavior has to be adapted to the user's cultural background. Therefore, culture specific system behavior has to be defined. On the other hand, the user's cultural background must be known to the system either by telling it directly or by inferring this background from the interaction. Before we present our prototype, it is necessary to have a closer look on what we mean by culture and how culture can be exploited for human computer interaction.

3.1 Definitions of Culture

To allow culture to be used in a computational way, it is necessary to build on a concept of culture that features a way to measure the impact of different cultures on behavior or expressivity. The definition of culture is not an easy task and there are many fuzzy definitions of this notion around. Nevertheless there is one theoretical school which claims that culture can be defined as a set of values and norms that members of a certain group adhere to. Kluckhorn and Strodtbeck [24] for instance distinguish between five different value orientations ranging from people and nature over time

sense to social relations. Although this is a first classification of possible values, the impact on behavior is more of an anecdotal character not allowing for an operationalizable model. A similar, value-oriented approach is presented by Schwartz and Sagiv [35]. Values are defined as goals that serve as guiding principles of behavior. These values are based on three universal requirements (biological needs, coordinated social interaction, and group functioning). Cultures now differ in which values, i.e. goals, they relate to these universal needs and how they prioritize different values. It remains to be shown how these different goal structures can be reflected in specific interaction behaviors. Hall ([10];[11];[12]) concentrates in his work mainly on three different dimensions: space, time and context. Accordingly, he defines high- and low-contact cultures refering to spatial behavior, monochronous and polychronous cultures referring to time perception, and low- and high-context cultures referring to aspects of group membership and associated patterns of communication. Hall associates different behavior patterns with the three categories, e.g. high-contact cultures are those in which people display considerable inter-personal closeness and immediacy.

A more recent representative of this line of thinking is Hofstede [15], who defines culture as a dimensional concept. His theory is based on a broad empirical survey that gives detailed insights in differences of value orientations and norms. Hofstede defines five dimensions on which cultures vary. Thus, a given culture is defined as a point in a five-dimensional space.

1. Hierarchy: This dimension describes the extent to which different distribution of power is accepted by the less powerful members. According to Hofstede more coercive and referent power (based on personal charisma and identification with the powerful) is used in high-H societies and more reward, legitimate, and expert power in low-H societies.
2. Identity: Here, the degree to which individuals are integrated into a group is defined. On the individualist side ties between individuals are loose, and everybody is expected to take care for himself. On the collectivist side, people are integrated into strong, cohesive ingroups.
3. Gender: The gender dimension describes the distribution of roles between the genders. In feminine cultures the roles differ less than in masculine cultures, where competition is accepted and status symbols are of importance.
4. Uncertainty: The tolerance for uncertainty and ambiguity is defined in this dimension. It indicates to what extent the members of a culture feel either uncomfortable or comfortable in unstructured situations which are novel, unknown, surprising, or different from usual. Whereas uncertainty avoiding cultures have rules to avoid unknown situations, uncertainty accepting cultures are more tolerant of opinions different from what they are used to and they try to have as few rules as possible.
5. Orientation: This dimension distinguishes long and short term orientation. Values associated with long term orientation are thrift and perseverance whereas values associated with short term orientation are respect for tradition, fulfilling social obligations, and saving one's face.

According to Hofstede, nonverbal behavior is strongly affected by cultural affordances. The identity dimension for instance is tightly related to the expression of emotions and the acceptable emotional displays in a culture. Thus, it is more acceptable in individualistic cultures like the US to publicly display negative emotions like anger or fear than it is in collectivistic cultures like Japan. Based on Hofstede's dimensions, Hofstede, Pedersen, and Hofstede [16] define synthetic cultures as representations of the end points of the dimensions and show how specific behavior patterns differ in a principled way depending on where a culture is located. Table 1 presents a summary for the acoustic and spatial behavior of these synthetic cultures, which serve as a starting point for our parametrized model of cultural variation.

Table 1: Synthetic cultures and corresponding patterns of behavior for low (L) and high (H) values [16].

Dimension	Synthetic culture	Sound	Space
Hiearchy	L: Low power H: High power	Loud Soft	Close Far
Identity	L: Collectivistic H: Individualistic	Soft Loud	Close Far
Gender	L: Femininity H: Masculinity	Soft Loud	Close Close
Uncertainty	L: Tolerance H: Avoidance	Soft Loud	Close Far
Orientation	L: Short-term H: Long-term	Soft Soft	Close Far

Similar cultural differences are found for the use of gestures and gestural activity. Argyle [1] distinguishes between different types or qualities of movements. Movements that accompany speech, conventionalized movements, i.e. emblems or sign language, movements that give information about emotional states and movements that give information about personality traits. Another taxonomy is defined by McNeill [29], who distinguishes between non-speech related gestures, speech related gestures, and conventionalized gestures, which are called emblems. Such emblems have usually been assigned an arbitrary meaning, which makes them a likely factor of intercultural misunderstandings. The American ok-sign for example is interpreted as an insult in Italy ([36]).

Regarding the quality of gestures, Argyle cites Effron's work on comparing qualities of gestures like spatial extent or speed in different groups of immigrants. Similar results are described by Ting-Toomey [36] for differences in gestural frequency and spatial extent between Southern European and Northern European cultures. Thus, the quality of gestures i.e. how a gesture is realized constitutes a cultural pattern of nonverbal behavior. McNeill [29] analyses the dynamics of gestures in more detail and defines three phases of movements. In the pre-stroke phase the hands are brought into the gesture space, the gesture itself is done during the stroke phase; afterwards, the hands are retracted in the post-stroke phase. If gestures are realized with high frequency one after the other, these phases may blend into each other and gestures may be affected by each other. Thus, we can for instance expect to find more of such effects in the Southern than in the Northern European cultures. To sum up, cultures differ in their gesture usage on different levels like the meaning of a gesture or the quality of the movements like speed and spatial extent. Thus, by relating this information from the literature to Hofstede's cultural dimensions we are able to model cultural effects on gestural patterns of behavior. In our model, we concentrate on the

quality of the movement to infer the user's cultural background..

3.2 A Bayesian Network Model of Cultural Influences

Cultural influences manifest themselves on different levels of behavior as we have seen above. Thus, the information about the cultural background of an interlocutor is only indirectly available and has to be derived from observations of other variables. To this end, the user's multimodal communicative behavior like eye gaze, spatial behavior, or gestural expressivity has to be analyzed.

Fortunately, there are already quite sophisticated recognition methods available for different modalities on which the inference of the user's cultural background can rely. Nevertheless, the necessary knowledge for this inference is unsure and unreliable because on the one hand recognition engines are far from perfect, on the other hand there might be a prototypical behavior for a given culture but still a specific user might deviate from this behavior. Thus, the model has to cope with this unreliable information which makes Bayesian networks well suited for the task.

Bayesian networks as described in [18] are a formalism to represent probabilistic causal interactions. For instance, they have already been successfully applied to model emotions for virtual agents ([2];[3]). In the domain of culture they are also very suitable, for the following reasons:

1. Bayesian networks handle uncertainty at every processing step. This is very important for our purpose because the link between culture and nonverbal behavior is a many to many mapping. By using a rule based system, we would get in trouble if one individual is not acting exactly in a way coherent to his cultural background.
2. Because the links in a Bayesian network represent the relations between causes and effects, they are intuitively meaningful. The theoretical effect of the gender dimension of culture on the volume (loudness) of the voice, for example, is represented by a link between these two nodes. The phenomenon that with increasing masculinity the volume of the voice is also rising is easy to realize. The exact probabilities may still be difficult to define, but as we use relatively isolated effects and their relations with the cultural dimensions, we can use tendencies of behavior described in the cultural science, especially in Hofstede's synthetic cultures.
3. Bayesian networks allow for different types of inferences depending on where evidence is introduced in the network. Thus, in the model given in Figure 2, a causal inference can be drawn from evidence regarding the cultural dimensions to nonverbal behavior, which can be used to set culture-specific behavior patterns of virtual characters. On the other hand, diagnostic inferences can be drawn if evidence for the specific nonverbal behavior is at hand, for instance to infer the user's cultural background i.e. his position on the five dimensions, based on his nonverbal behavior.

Our first model is based on Hofstede's ideas of synthetic cultures, which define stereotypes for the five dimensions. In the long run, these stereotypical values will have to be replaced by specific empirical data. To this end, a large comparative corpus study was done to retrieve enough data in prototypical situations for at least two cultures, Germany and Japan [32]. Figure 2 gives an overview of the realized model. The middle layer defines Hofstede's dimensions, the bottom layer consists of nodes for nonverbal behavior that can either be registered from the user or set for a given agent. The top node which is labeled "Culture" is just for demonstration and interpretative purposes. It mainly translates the results from the dimensional representation of cultures into a probability distribution for some exemplary cultures.

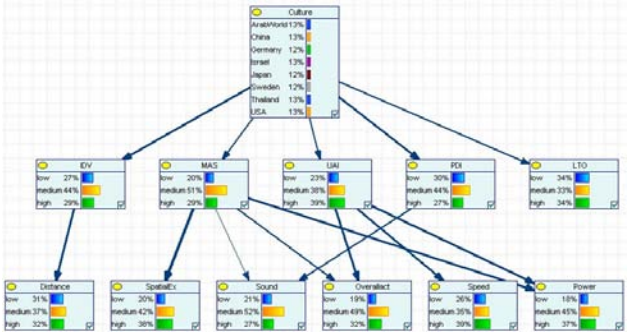

Figure 2: Bayesian Network modeling the interrelation between cultural dimensions and nonverbal behavior.

There are some arguments if this is a valid approach because all of the above mentioned theories describe culture as a social (group) phenomenon and not as aspects of single individuals like it is with personality traits (see e.g. Rojas [33]). Others argue that it can be viewed as a cognitve model in the Vygotskyan sense (see e.g. Vatrapu and Suthers [38]). For our purposes we have to distinguish between the two ways in which we are using the Baysian Network.

1. Infering the user's cultural background: Making use of the cultural dimensions allows abstracting from the specific culture of the user to a distribution on the five dimensions. Thus, deviating behavior of the user, i.e. behavior that is not in accordance to known patterns of behavior for the user's culture, results in a different interpretation of the single user's position on the cultural dimensions. For instance, the user might be from culture A but shows behavior that is more in accordance with culture G. In this case his cultural background is inferred as G and the system reacts relative to this interpretation. Consequently, the behavior of an individual is interpreted by known patterns of behavior found for certain cultural groups. It remains to be shown if the user is then irritated by the system's behavior which is not in accordance with his "real" cultural background.
2. Setting the agent's nonverbal behavior: In this case, the Bayesian network delivers information about dominant patterns of behavior in a culture that is found at the corresponding locations of the cultural dimensions, for instance low on hierarchy, low on identity, high on gender, medium on uncertainty, high on orientation. This results in a probability distribution for each behavior e.g. for volume the probabilites are 70% high, 29% medium, and 1% low. In our first prototype (see Section 5), this information is used directly to set the behavior of a group of agents, who will then speak with high volume.

Thus, for the diagnostic inference it remains to be shown if the user is irritated by the system's adapted behavior because his behavior is interpreted with patterns derived from group interactions. For the causal inference this is no problem because the behavior of a group of agents is regulated by the information derived from the network (see Section 5).

4. ACCELEROMETER BASED GESTURE RECOGNITION

We employ Nintendo's Wii remote controller (Wiimote) to capture the user's gestural behavior. The Wiimote uses accelerometers to sense its movements in 3D space. The controller is able to connect via Bluetooth to a common PC. The acceleration data is gathered for each direction (x: left/right, y: back/forth, z: up/down) with a sampling rate of 100Hz. Figure 3 gives an impression of the Wiimote, how to handle it, and a typical signal for the three accelerometers. To allow for fast and simple use of the Wiimote in a number of different applications, we developed the WiiGLE[1] environment (Wii-based Gesture Learning Environment). It allows defining gesture classes for an application, selecting features for the classification task, training and comparing classifiers, and using it as the classification component of an application. It provides a programming interface to define own features and classifiers. For the use in our prototype system, we integrated classifiers from the Weka data mining toolkit [40].[2] Some approaches already exist to classify gestures from acceleration data. Most of these use HMMs for the classification task. Schlömer and colleagues [34] describe a recent approach also using the Wiimote. We claim that fast and simple classifiers like Nearest Neighbor or Naïve Bayes are also suitable for the task. To this end, we compare the results of a HMM-based approach with the results from the WiiGLE in Section 4.2.3.

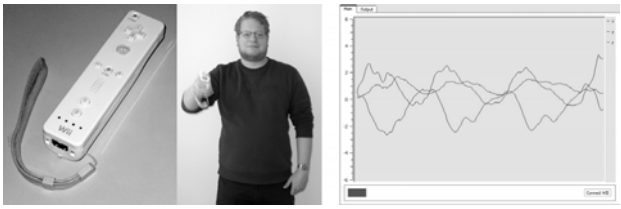

Figure 3: The Wiimote (left), a user handling the Wiimote (middle), and the signal for the three accelerometers (right).

In principal, we can distinguish between two ways of interpreting gestural behavior of the user: (i) how a gesture is done by the user, and (ii) what gesture is done by the user.

How a gesture is done can be described by what Gallaher [8] calls expressivity or expressive style. Gallaher categorizes gestural style by a number of expressivity parameters, e.g. how fast a gesture is done, how much space one uses to perform a gesture, and links expressive style to personality traits. Bevacqua and colleagues [4] describe how these parameters can be exploited to analyse the behavior of a user and use the results to vary the behavior of a virtual character. Some of Gallaher's parameters are also described in the literature on culture-specific gestural behavior. For instance, Southern Europeans are said to do more, bigger, and faster gestures than Northern Europeans (e.g. Ting-Toomey [36]), which are described by the parameters overall activation (number of gestures per time interval), spatial extent, and speed. Thus, we claim that at least some of the expressivity parameters can also be linked to culture and not only to personality. In our first model we integrated the three parameters already mentioned plus the additional parameter power.

Can the recognition of which gesture is performed by the user inform a system about the cultural background of the user? As stated above, emblems have clearly defined forms, convey a communicative meaning, and are culture-dependent. An example from the German emblems is "Waving a hand in front of one's eyes", which communicates the opinion that the addressee is stupid. Thus, certain gestures could be used as an additional source of information to infer the user's cultural background.[3] Consequently, the next two subsections present not only the recognition of gestural expressivity but also of discrete gesture classes.

4.1 Recognizing Gestural Expressivity

Expressivity recognition can either be realized by calculating the expressivity parameters as features on the raw signal which has the advantage of continous results. Or it can be realized by classifying the signal in discrete classes like low, medium, high for each expressivity parameter.

4.1.1 Expressivity Recognition as a problem of feature calculation

To obtain the selected expressivity features from the user's gesture in a format we can use, we first must transform it from raw data to feature values. For a better readability, we define two variables S and L that are calculated on the raw data:

$$S = \sum_{i=0}^{n} a_x^2 + \sum_{i=0}^{n} a_y^2 + \sum_{i=0}^{n} a_z^2 \quad (1)$$

$$L = \sum_{i=0}^{n} |a_x| + \sum_{i=0}^{n} |a_y| + \sum_{i=0}^{n} |a_z| \quad (2)$$

Power (3) is derived straightforward as it is equivalent with energy and can thus be calculated in the usual way.

$$Power = \frac{1}{n} S \quad \text{where n denotes the signal length (3)}$$

To find formulas for the expressivity parameters spatial extent and speed we used an experimental approach. We recorded 20 similar gestures from one person, 10 with big and 10 with low spatial extent, to find a reliable formula. We found that the signal's power (3) divided by the signal's sum of its absolute values is a good representation of spatial extent derived from the acceleration data (see formula 4).

$$SpExt = \frac{S}{L} \quad (4)$$

For finding the formula for speed, we also recorded 20 similar gestures from one person, 10 with fast speed and 10 with low speed. We found that a light variation of the formula for spatial extent (4) by multiplying instead of dividing the signal's power (3) by the sum of its absolute values, gives a good approximation of the gestures speed (see formula 5).

[1] http://mm-werkstatt.informatik.uni-augsburg.de/project_details.php?id=46

[2] http://www.cs.waikato.ac.nz/ml/weka/

[3] For more information on German emblems see the online version of the Berlin dictionary of everyday gestures: http://www.ims.uni-stuttgart.de/projekte/nite/BLAG/ (25th April 2008).

$$Speed = \frac{1}{n^2} SL \quad \text{where n denotes signal length} \quad (5)$$

Table 2: Results of feature calculation

	Power	Speed	SpatialExt.
Low	100%	94.8%	62.4%
High	100%	67.1%	59.5%
Overall	100%	81%	61%

Our gesture set for finding these formulas was very limited. The gestures were all similar and from one single person to avoid any gesture- or user-dependent side effects. In the meantime we recorded a large set of gestures. We asked seven subjects to write three numbers (1, 5 and 8) in the air in front of themselves with the Wiimote and different expressivity. Each gesture was performed 10 times with 6 different expressive styles: high and low power, high and low speed, and high and low spatial extent. In total we recorded 1260 gestures, 420 per class.

This gesture set was used to evaluate the above obtained formulas (see Table 2). Power can be detected without any problems. We couldn't find any overlap within the calculated features of the two classes high and low. Therefore it is no suprise to get a highly significant result from a two-tailed t-test ($t(418) = 25.6$; $p < 8*10^{-88}$). Speed gives as a total recognition result of 81%, whereas the recognition for low speed is much more accurate than for high speed. We optimized the threshold to achieve the best recognition rate. The significant difference for low and high speed is still very high ($t(418) = 13, 8$; $p < 4*10^{-36}$). The recognition results for spatial extent are very poor and cannot really be used at all. Although the difference between high and low spatial extent is still highly significant ($t(418) = -4.0$; $p < 6*10^{-5}$), we cannot find a threshold to differ spatial extent.

4.1.2 Expressivity Recognition as a Classification Problem

As we have seen above, calculating expressivity parameters directly on the acceleration data only works well for power, which can be derived in a straightforward way from the raw signal. The recognition results for the other parameters (with the exception of high speed) are not accepatable. But expressivity recognition can also be defined as a classification problem, making it available to standard recognition methods. Three classifiers are needed for this task, one for each parameter that is trained on the two-class problem of distinguishing between low and high values for the expressivity parameters.

First of all, features are calculated on the raw signal. For the acceleration data, we calculated the length of the signal, the minimum and maximum for each axis, the median and mean for each axis, and the gradient for each axis. The same training set that was described above was used for this method. A ten-fold cross-validation of a Naïve Bayes (NB), Nearest Neighbor (NN) and Multilayer Perceptron (MLP) classifier were done.

The results of this approach are given in Table 3. For the two-class problem, all classifiers deliver acceptable results. The best recognition rate can be seen for the Multilayer Perceptron but because the calculation is faster for the Nearest Neighbor classifier, the latter is preferred for the application.

4.2 Recognizing Discrete Gestures

Accelerometer based gesture recognition is also possible for discrete gestures, i.e to determine which gesture was done by the user. The above proposed recognition engine is well suited to recognize such gestures, which could be employed to convey conversational meaning, special input symbols, or system control parameters. To this end, WiiGLE was tested on three gesture sets: (i) digits from 0 to 9, (ii) a set of German emblems, (iii) VCR control gestures. Digits were chosen because they are a standard problem of handwriting recognition and present a complex (10-class) closed problem space. The set of German emblems exemplifies the usefulness of recognizing such conventionalized gestures as an additional source of information on the user's cultural background. The VCR control gestures at last were chosen for the reason of allowing the comparison of a HMM based approach to the simpler and less costly techniques integrated in WiiGLE.

Table 3: Recognition results for expressivity classification.

	Power		
	L	H	All
NB	99.5%	100%	99.8%
MLP	100%	100%	100%
NN	100%	100%	100%
	Speed		
	L	H	All
NB	93.8%	93.3%	93.6%
MLP	97.1%	97.1%	97.1%
NN	95.2%	94.3%	94.8%
	Spatial Extent		
	L	H	All
NB	91.9%	91.4%	91.7%
MLP	98.6%	99%	98.8%
NN	97.1%	97.1%	97.1%

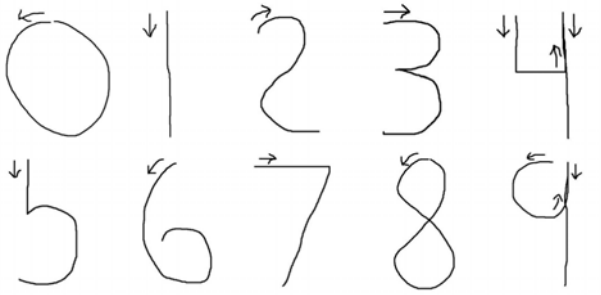

Figure 4: Gesture set one: digits from 0 to 9.

4.2.1 Gesture set one: Digits

This set has the advantage of being conceptually closed, easy to grasp by the user, and having some classes that are very similar in regard to shape and motion like 0 and 6 to make the classification problem difficult enough. Seven users were recorded doing ten gestures for each digit (see Figure 4 for an overview of the gestures in this set). Thus, for each class (digit) 70 examples were collected. The same set of features was employed in this task, i.e. 16 features were calculated (see above). Recognition accuracy for the classifiers was evaluated under two conditions. In the user-independent condition the whole training set was employed. Recognition accuracy was assessed by a ten-fold cross-validation. Results are given in Table 4. Due to the bad result for the Naïve Bayes and Multilayer Perceptron classifiers, the data sets of three random

users were taken to test user-dependent performance on the gesture sets, i.e. for each user 100 samples were available, 10 per digit. As can be seen in Table 4 performance in this condition incresases significantly and is even optimal for the third user.

Table 4: Results for gesture recognition with WiiGLE for gesture set one: digits from 0 to 9.

Classifier	User-independent	User-dependent		
	7 users	1	2	3
NB	58.1%	90.2%	99%	100%
MLP	69.9%	93.1%	99%	100%
NN	100%	100%	100%	100%

4.2.2 Gesture set two: German emblems

Emblems are conventionalized gestures in a given culture and thus provide additional information on the user's cultural background. From the "Berlin Dictionary of German Everyday Gestures" we chose eight gestures[4] (Table Table 5). One user prepared ten training samples for each class. Again, the Naïve Bayes, the Nearest Neighbor and the MLP classifier were trained on this set and tested with a test set consisting of 5 instances per class. Recogniton results are comparable to the previous problem (Table 6). All classifiers had problems with emblem A13, which was misclassified as A23.

Table 5: Gesture set three: seven German emblems.

ID	Description	Movement
A01	Reproaching someone for stupidity	Waving a hand in front of one's eyes
A02	Requesting someone to hurry up	Indicating to one's wrist
A04	Refusing an offer	Moving hands horizontally back and forth
A05	Asking for something to drink	Drinking from a container
A13	Requesting calm	Repeatedly lowering downward facing palms
A21	Expressing existential crisis	Cutting one's throat
A23	Expressing distrust	Rotating one's hand back and forth

Table 6: Results for gesture recognition with WiiGLE for gesture set two: German emblems.

Classifier	NB	MLP	NN
Result	88.6%	91.4%	94.3%

Thus, classifying emblematic gestures poses no principled problem for our accelerometer-based approach. It remains an issue of discussion if the Wiimote is the suitable device for capturing the necessary data as it has to be grasped to perform the gestures. Currently, we are experimenting with a different device that is less obstrusive and can be attached to the forearm. Combining the information derived from the classification of the emblematic gestures with other types of information about the cultural background of the user like the expressivity of the gesture, his proxemics behavior, etc. can be employed to disambiguate problematic gestures like the above mentioned American ok-sign, which is interpreted as an insult in Italy.

4.2.3 Gesture set three: VCR control

Gesture set three was used to compare the results of the Wii-based approach with an approach described in the literature that also relies on accelerometer data but classifies gestures with HMMs ([21];[26]). The raw acceleration data is quantified and then used for training the HMM models, i.e. no higher level feature calculation is done on the gestures. In principle HMMs could be used for continous gesture recogniton but the test set for the VCR control does not take this advantage into account. Thus, our approach of calculating features on the signal and classify the whole gesture is applicable in this domain (see Figure 5 for the eight gestures in this set).

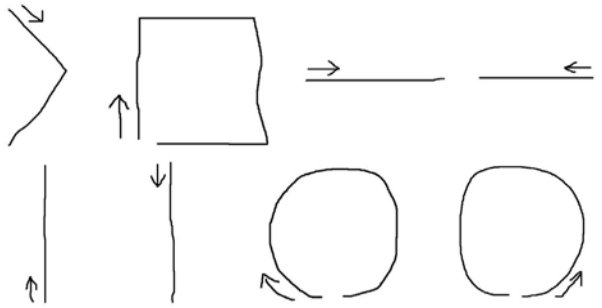

Figure 5: Gesture set three: VCR control. Top row from left to right: gestures for play, stop, next, previous. Bottom row from left to right: gestures for increase, decrease, fast forward, fast rewind.

Mäntijärvi and colleagues [26] test different training procedures to increase the recognition rate of their classifier. The best result they achieve is 97.2% recognition rate. This is taken as the benchmark to compare the WiiGLE toolbox against. Gestures were recorded under the same conditions. One user did 30 gestures per class, which were recorded in two sessions. In each session, 15 gestures per class were performed. Recognition rates were caculated by a 14-fold cross-validation. Results are given in Table 7 and show clearly that the faster, computationally less complex classifiers like Naïve Bayes or Nearest Neighbor are sufficient to solve the recognition task for a given user. All classifiers had a problem with the same gesture. They classified one example of gesture eight (precision: 1, recall: 0.967, F-score: 0.983) as gesture seven (precision: 0.968, recall: 1, F-score: 0.984). The results are satisfying and comparable to the results given for the user-dependent condition of gesture set one. It would be interesting to see test results for the HMM model for the user-independent condition.

Table 7: Results for gesture recognition with WiiGLE for gesture set three: VCR control.

	WiiGLE			Mäntijärvi et al.
Classifier	NB	MLP	NN	HMM
Result	99.6%	99.6%	99.6%	97.2%

[4] Video samples of the emblems can be found on the following website: http://www.ims.uni-stuttgart.de/projekte/nite/BLAG/

5. ADAPTING TO THE USER'S CULTURAL BACKGROUND

The Wiimote serves as the input device for our test application, which we call a cultural mirror. It exemplifies how the user's cultural background can be automatically inferred from his gestural behavior and utilized to adapt the system's reactions. Figure 6 gives an overview of our system architecture. The grey parts have not been integrated so far. In the long run, the user will be equipped with additional input devices allowing for analysing his gaze behavior or his emotional state using the audio signal. The signal from the Wiimote is send to the *Recognition Engine*, which classifies the input. This information is then forwarded to a context sensing toolkit (ACOSAS) [6] and to the *Culture Recognition* component that incorporates the Bayesian network described above. The information from the network will then be available for the interaction planning but at the moment is just passed on to the behavior modification module *Cultural Modification*, which consists of a second copy of the network and allows for setting the cultural dimensions of the agents and then selects correponding behaviors that are displayed to the user.

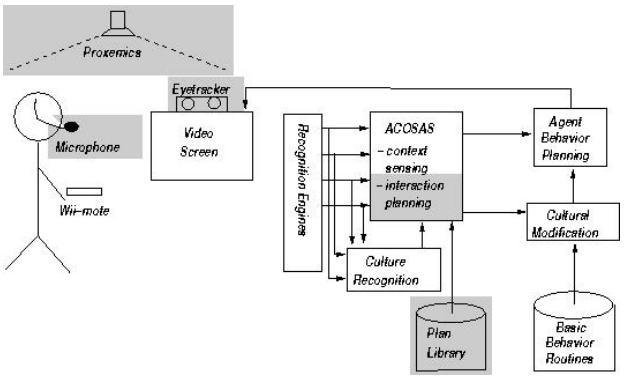

Figure 6: Proposed system architecture. Shaded areas not integrated yet.

The current processing of the probability distribution for the cultural dimensions consists in selecting the value with the highest dimension. For instance, if the result for the hierarchy dimension is 45% high, 33% low, and 22% medium, the value high is selected. For each cultural dimension node in the modification module, evidence is set for the selected value, in this case for high on the hierarchy dimension. Far more sophisticated interpretations of the network's results are possible, which could for instance take the distribution for each node into account to modify the agents' behavior individually and thus reflecting this distribution.

Although the user can only provide input to the system by his gestural expressivity, the use of the Bayesian network allows modifying other agent behaviors as well. Currently, apart from the gestural expressivity the agents' spatial behavior (proxemics) and the volume of speech are influenced. Figure 7 gives an example of different proxemics behavior of a group of agents as a reaction to the user's gestural expressivity.

6. CONCLUSION

In this paper we discussed the challenge of how cultural influences can be parametrized to adapt the behavior of an interactive system to a given user. To this end we suggest to analyze the user's behavior as a contextual clue that allows automatically infering his cultural background. This information is then used to modify the behavior of a group of virtual characters to reflect patterns of behavior known for the infered culture. We claim that in general the challenge of enculturating human computer interaction always has to take these two processing steps into account.

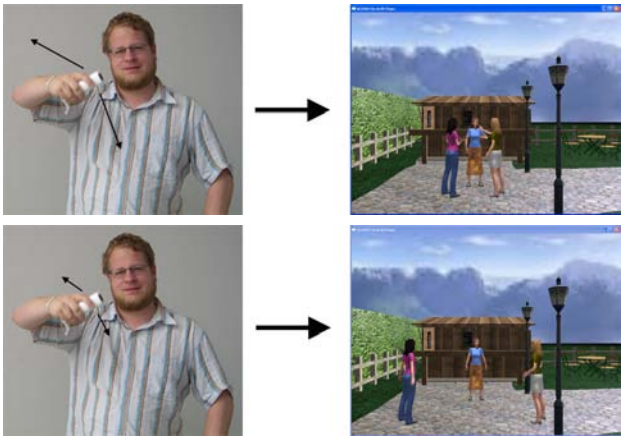

Figure 7: Modifying the agents' behavior as a reaction to high spatial extent and high speed (above) vs. low spatial extent and low speed (below).

1. Recognizing the user's cultural background
2. Modifying the system's behavior

This was exemplified with the use of a novel interface device featuring acceleration sensors and by virtual characters interacting in a virtual environment. But the same holds true for other forms of human computer interaction like the traditional website as was shown in Section 2. It would be interesting to see a dynamic adaptation of the website design based on the cultural preferences of the user. An adhoc approach could infer the user's culture simply by his IP-address. Then, a network similar to the one described here could be used to adapt the design. The output nodes then would model design guidelines for culture preferences.

Our current approach is deficient in two ways. (i) The model can only be as good as the data that is used to specifiy the probabilities. The first approrach mainly relies on Hofstede's ideas of synthetic cultures. Currently, the results from a large-scale corpus study [32] in two different cultures are integrated into the model. The corpus was recorded for three prototypical situations present in every culture and provides a rich source of empirical data for updating the model. (ii) So far we have concentrated on the technical aspect of analyzing the user input and applying it to the behavior selection of the agents. The evaluation of the input techniques was presented in this paper. The next step of course is to show if users are satisfied with the reactions of the system and if they can interpret the behavior of the agents coherently.

The envisioned application for our work is an experience-based training of cultural communication skills following suggestions by Hofstede [14] who presents three steps for such a training endeavour.

1. Awareness: The first step of gaining intercultural competence is being aware and accepting that there are differences in behavior. To realize this step in a learning system with virtual characters, the trainee is confronted with a group of characters displaying the behavior routines of the target culture. With the knowledge of the trainee's cultural background, the agents could also contrast the behavior of the target culture with the behavior of the trainee's culture.

Comparing the behavior patterns the trainee recognizes that there are differences but might not be able to pin them down.

2. Knowledge: In the second step, the trainee's knowledge of what exactly is different in the behavior is increased, which can be interpreted as getting an intellectual grasp on where and how one's own behavior differs. For instance the trainee might have felt a little bit uncomfortable in step one due to a different pattern of gaze behavior. In step two, he will gain the knowledge on how his patterns differ from the patterns of the target culture and what the consequences are. In the learning system, the user is confronted with reactions to his behavior by his interlocutors. For instance, the agents could move away if the user comes too close. Moreover, the agents could replay specific behavior routines of the user and contrast them to the behavior routines of the target culture, pointing out where exactly the user's behavior deviates from the target culture.
3. Skills: Hofstede argues that the first two steps are sufficient to avoid most of the obvious blunders in cross-cultural communication. If the trainee has the ambition to blend into the target culture and adapt his own behavior, a third step is necessary: the training of specific nonverbal communication skills. If e.g. avoiding eye contact in negotiations is interpreted as a sign of disinterest in the target culture, it might be a good idea to train sustained eye contact for such scenarios. Again, virtual characters can play a vital role in this learning endeavour.

Such an application can be interpeted as an augmentation of the standard language textbook to allow for a deeper understanding of the communication processes than could be achieved by just learning the grammar and the words.

7. ACKNOWLEDGMENTS

The work described in this paper is partially supported by the German Research Foundation (DFG) under research grant RE 2619/2-1 (CUBE-G) and by the European Community (EC) in the eCIRCUS project IST-4-027656-STP and the CALLAS project IST-34800. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

The Bayesian network model was created using the GeNIe modeling environment and integrated with SMILE developed by the Decision Systems Laboratory of the University of Pittsburgh (http://dsl.sis.pitt.edu).

The gesture recognition toolbox for the Wiimote (WiiGLE – Wii-based Gesture Learning Environment) is available from: http://mm-werkstatt.informatik.uni-augsburg.de/project_details.php?id=46

8. REFERENCES

[1] M. Argyle. Bodily Communication. Methuen & Co. Ltd., London, 1975.

[2] E. Ball. A Bayesian Heart: Computer Recognition and Simulation of Emotion. In R. Trappl, P. Petta, and S. Payr, editors, Emotions in Humans and Artifacts, pages 303–332. MIT Press, 2002.

[3] N. Bee, H. Prendinger, A. Nakasone, E. André, and M. Ishizuka. AutoSelect: What You Want Is What You Get: Real-Time Processing of Visual Attention and Affect. In E. André, L. D. W. Minker, H. Neumann, and M. Weber, editors, Perception and Interactive Technologies (PIT 2006), pages 40–52, Springer: Berlin, Heidelberg, 2006.

[4] E. Bevacqua, A. Raouzaiou, C. Peters, G. Caridakis, K. Karpouzis, C. Pelachaud, and M. Mancini. Multimodal sensing, interpretation and copying of movements by a virtual agent. In E. André, L. D. W. Minker, H. Neumann, and M. Weber, editors, Perception and Interactive Technologies (PIT 2006), pages 164–174, 2006.

[5] B. Choi, I. Lee, Y. Jeon, and J. Kim. A qualitative cross-national study of cultural influences on mobile data service design. In CHI 2005, pages 661–670, 2005.

[6] D. Erdmann, K. Dorfmüller-Ulhaas, and E. André. Integrating VR-Authoring and Context Sensing: Towards the Creation of Context-Aware Stories. In: Technologies for Interactive Digital Storytelling and Entertainment (TIDSE), 2006.

[7] G. Ford and H. Gelderblom. The Effects of Culture on Performance Achieved through the use of Human Computer Interaction. In Proceedings of SAICSIT, pages 218–230, 2003.

[8] P. E. Gallaher. Individual Differences in Nonverbal Behavior: Dimensions of Style. Journal of Personality and Social Psychology, 63(1):133–145, 1992.

[9] E. W. Gould, N. Zakaria, and S. A. M. Yusof. Applying culture to website design: a comparison of Malaysian and US websites. In Proceedings of IEEE professional communication society international professional communication conference and Proceedings of the 18th annual ACM international conference on Computer documentation: technology & teamwork, pages 161–171, 2000.

[10] E. T. Hall. The Silent Language. Doubleday, 1959.

[11] E. T. Hall. The Hidden Dimension. Doubleday, 1966.

[12] E. T. Hall. Beyond Culture. Doubleday, 1976.

[13] S. Hisham and A. D. N. Edwards. Incorporating Culture in User-interface: A Case Study of Older Adults in Malaysia. In HT'07, pages 145–146, 2007.

[14] G. Hofstede. Cultures and Organisations — Intercultural Cooperation and its Importance for Survival, Software of the Mind. Profile Books, 1991.

[15] G. Hofstede. Cultures Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations. Sage Publications, Thousand Oaks, London, 2001.

[16] G. J. Hofstede, P. B. Pedersen, and G. Hofstede. Exploring Culture: Exercises, Stories, and Synthetic Cultures. Intercultural Press, Yarmouth, 2002.

[17] D. Jan, D. Herrera, B. Martinovski, D. Novick, and D. Traum. A Computational Model of Culture-Specific Conversational Behavior. In C. Pelachaud et al., editors, Intelligent Virtual Agents (IVA'07), pages 45–56, Springer: Berlin, Heidelberg, 2007.

[18] F. V. Jensen. Bayesian Networks and Decicion Graphs. Springer, 2001.

[19] W. Johnson, S. Choi, S. Marsella, N. Mote, S. Narayanan, and H. Vilhj´almsson. Tactical Language Training System: Supporting the Rapid Acquisition of Foreign Language

and Cultural Skills. In Proc. of InSTIL/ICALL — NLP and Speech Technologies in Advanced Language Learning Systems, 2004.

[20] S. Kayan, S. R. Fussell, and L. D. Setlock. Cultural Differences in the Use of Instant Messaging in Asia and North America. In CSCW'06, pages 525–528, 2006.

[21] J. Kela, P. Korpipaää, J. Mäntyjärvi, S. Kallio, G. Savino, L. Jozzo, and S. D. Marca. Accelerometer-based gesture control for a design environment. Pers. Ubiquitous Computing, 10:285–299, 2006.

[22] R. Khaled, P. Barr, R. Fischer, J. Noble, and R. Biddle. Factoring Culture into the Design of a Persuasive Game. In OZCHI'06, pages 213–220, 2006.

[23] R. Khaled, R. Biddle, J. Noble, P. Barr, and R. Fischer. Persuasive interaction for collectivist cultures. In W. Piekarski, editor, The Seventh Australasian User Interface Conference (AUIC 2006), pages 73–80, 2006.

[24] F. Kluckhohn and F. Strodtbeck. Variations in value orientations. Row, Peterson, New York, 1961.

[25] N. Maniar and E. Bennett. Designing a mobile game to reduce culture shock. In Proceedings of ACE'07, pages 252–253, 2007.

[26] J. Mäntyjärvi, J. Kela, P. Korpipää, and S. Kallio. Enabling fast and effortless customisation in accelerometer based gesture interaction. In Proceedings of MUM'04, pages 25–31, 2004.

[27] A. Marcus and E. W. Gould. Crosscurrents: Cultural Dimensions and Global Web-User Interface Design. ACM Interactions, 7(4):32–46, 2000.

[28] A. P. Massey, Y.-T. C. Hung, M. Montoya-Weiss, and V. Ramesh. When Culture and Style Aren't About Clothes: Perceptions of Task-Technology Fit in Global Virtual Teams. In GROUP'01, pages 207–213, 2001.

[29] D. McNeill. Hand and Mind — What Gestures Reveal about Thought. The University of Chicago Press, Chicago, London, 1992.

[30] A. Nazir, M. Y. Lim, M. Kriegel, S. Enz, and C. Zoll. Culture-personality based affective model. In IUI-Workshop Enculturating Interfaces, Gran Canaria, 2008.

[31] M. Rehm, E. André, Y. Nakano, T. Nishida, N. Bee, B. Endrass, H. H. Huang, and Michael Wissner. The CUBE-G approach – Coaching culture-specific nonverbal behavior by virtual agents. In Proceedings of ISAGA 2007, in press.

[32] M. Rehm, Y. Nakano, H.-H. Huang, A. A. Lipi, Y. Yamaoka, and F. Grüneberg. Creating a Standardized Corpus of Multimodal Interactions for Enculturating Conversational Interfaces. In Proceedings of the IUI-Workshop on Enculturating Interfaces (ECI), 2008.

[33] J. Rojas. The cultural construction of ubiquitous computing. In Culture and Collaborative Technologies Workshop (CHI 2007), 2007.

[34] T. Schlömer, B. Poppinga, N. Henze, and S. Boll. Gesture Recognition with a Wii Controller. Proceedings of TEI, pages 11-14, 2008.

[35] S. H. Schwartz and L. Sagiv. Identifying culture-specifics in the content and structure of values. Journal of Cross-Cultural Psychology, 26(1):92–116, 1995.

[36] S. Ting-Toomey. Communicating Across Cultures. The Guilford Press, New York, 1999.

[37] H. C. Triandis and E. M. Suh. Cultural influences on personality. Annual Review of Psychology, 53:133–160, 2002.

[38] R. Vatrapu and D. Suthers. Culture and Computers: A Review of the Concept of Culture and Implications for Intercultural Collaborative Online Learning. In T. Ishida, S. R. Fussell, and P. T. J. M. Vossen, editors, International Workshop of Intercultural Collaboration (IWIC), pages 260–275. Springer, Berlin, Heidelberg, 2007.

[39] R. Warren, D. E. Diller, A. Leung, W. Ferguson, and J. L. Sutton. Simulating scenarios for research on culture and cognition using a commercial Armstrong, and J. A. Joines, editors, Proceedings of the 2005 Winter Simulation Conference, 2005.

[40] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 2005.