How do iconic gestures convey visuo-spatial information? Bringing together empirical, theoretical, and simulation studies

Hannes Rieser[1], Kirsten Bergmann[1,2], and Stefan Kopp[1,2]

[1] Collaborative Research Center 673, "Alignment in Communication", Bielefeld University
[2] Center of Excellence in "Cognitive Interaction Technology" (CITEC), Bielefeld University
{kbergman,skopp}@techfak.uni-bielefeld.de
hannes.rieser@uni-bielefeld.de

Abstract. We investigate the question of how co-speech iconic gestures are used to convey visuo-spatial information in an interdisciplinary way, starting with a corpus-based empirical and theoretical perspective on how a typology of gesture form and a partial ontology of gesture meaning are related. Results provide the basis for a computational modeling approach that allows us to simulate the production of speaker-specific gesture forms to be realized with virtual agents. An evaluation of our simulation results and our methodology shows that the model is able to successfully approximate human gestural behavior use of iconic gestures, and moreover, that gestural behavior can improve how humans rate a virtual agent in terms of eloquence, competence, human-likeness, or likeability.

1 Introduction

The question how co-speech iconic gestures are used to convey visuo-spatial information is still relatively unexplored [1]. In this paper we will mainly focus on two topics, first, how gesture simulation is grounded in an empirical gesture typology and a partial ontology and second, how gesture simulation can be used methodologically, looping back to the empirical data on which both, simulation and theoretical modelling are based. See Fig. 1 for an overview of our interdisciplinary methodology. We meet this challenge with an interdisciplinary methodology combining the empirical study of speech and gesture use, the elaboration of theoretical reconstructions and the formulation of generation models that enable the simulation of such communicative behaviour with virtual agents. Recently, new options for gesture typology have arisen due to systematically collected and annotated data such as the Bielefeld Speech And Gesture Alignment (SAGA) corpus which contains approximately 5000 iconic/deictic gestures used in natural dialogues in a spatial communication task combining direction-giving and sight description (for details see [9]). Corpus-based empirical methods proceed from rated annotations to classification of recurrent structures and ultimately to an investigation of its generalizability supported by statistical investigations [6]. Computational simulation opens up new possibilities enriching this set of methods in many ways. Obviously, gesture simulation has its independent goals in endowing virtual agents with human-like expressiveness. In addition, we use it as a methodological device, more specifically for the post-hoc evaluation of decisions made at various levels of the theory construction process, in other words, as a method of Popperian falsification. As an illustration of every aspect of our methodology, we will discuss a church-window-example from the SAGA corpus shown in Figure 1a (church window datum) and 1b (gesture datum) throughout the paper (restricted to the top of the window).

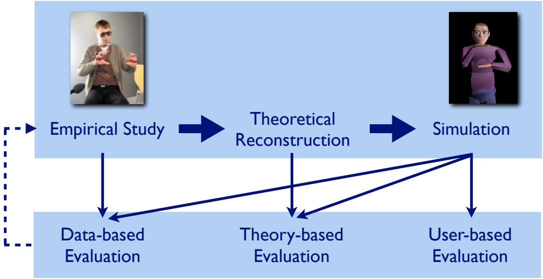


Fig. 1. Our methodology to study iconic gesture combines empirical study, theoretical modeling and computational simulation. The model is evaluated in different ways: (1) by comparing simulated gesture forms with empirically observed gestural behavior, (2) by comparing the gesture's semantics with theoretical semantic reconstructions, and (3) by investigating how the simulation with a virtual agent is judged by human recipients in a user study. At this stage, one might return to the empirical data to start a new pass in order to extract improved models of communicative behavior.

In the following, we will start with an empirical and theoretical perspective on how gesture form and meaning can be described and mapped onto each other under consideration of gestural representation techniques (Sect. 2 and 2). These concepts provide the basis for a computational simulation approach with virtual agents (Sect. 3). Finally we will show how we evaluated the model with respect to empirical data (Sect. 4 and 4) and regarding the degree to which the automatically generated gesturing behavior is able to improve how a virtual agent is rated by human observers (Sect. 4).

2 Empirical and Theoretical Perspective

How can we Describe a Gesture's Form? The analysis of physical gesture form is an indispensable prerequisite of any account of gesture generation. We have developed a typological grid for gestures accompanying noun phrases based on the SAGA corpus to specify and characterize the physical form of co-speech iconic gestures [10].This typology specifies a hierarchy of so-called *annotation*

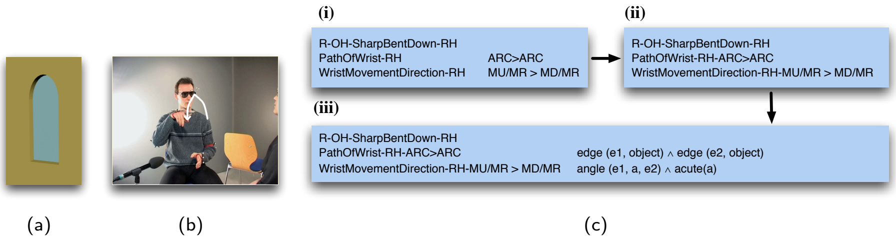

(i)
R-OH-SharpBentDown-RH
PathOfWrist-RH ARC>ARC
WristMovementDirection-RH MU/MR > MD/MR
(ii)
R-OH-SharpBentDown-RH
PathOfWrist-RH-ARC>ARC
WristMovementDirection-RH-MU/MR > MD/MR
(iii)
R-OH-SharpBentDown-RH
PathOfWrist-RH-ARC>ARC edge (e1, object) ∧ edge (e2, object)
WristMovementDirection-RH-MU/MR > MD/MR angle (e1, a, e2) ∧ acute(a)
(a)
(b)
(c)

Fig. 2. (a) Church window datum (b) Empirical gesture datum (c) Mapping of gesture form description onto semantic representation: (i) gesture type and attribute-value pairs relevant for semantics; compare this to Table 1. (ii) gesture type and attribute value pairs of (a) mapped onto composite functions indicating the composing elements ARC>ARC and MU/MR>MD/MR as affixes. These composite functions are used for the next step: (iii) composite functions and their values in terms of logical form. '→' denotes an onto-mapping.

predicates including the four major gestural form features (handshape, hand orientation, position, and movement characteristics) which are widely accepted in gesture research. Our typology, however, goes beyond these features in that recurrent gesture events are classified according to dimensions which have semantic impact. We consider indexings to objects as 0-dimensional, the idea being that in these cases no particular feature is depicted, it is no more than a mere indication of objects. Next come one-dimensional entities, lines, which can be straight or bent. We have composites of lines enclosing an angle. This is exactly what we need for our example: two bent lines meeting in an apex. There are all sorts of two-dimensional shapes, some like geometrical forms, some like fuzzy locations or regions. Two-dimensional entities can also form composites and be embedded in three-dimensional space. A similar story can be told for three-dimensional entities and arbitrary composites for entities of all dimensions. The full church window datum, for instance, combines a three-dimensional corner and a base with bends in the manner described.

The mapping from an annotation predicate to its value is laid down in an attribute value matrix (AVM). The number of predicates used in the respective AVM is determined by the need to capture the most characteristic features of the gestural representation as produced by the motor behaviour of the hand in a time interval fixed by the gesture phases, especially the gesture stroke. Hence, gesture typology looks for recurrent manifestations of motor behaviours and collects them into sets. Assembling into sets is of course done with an eye on semantics, as will become clear soon. Nevertheless, there is no air of circularity in this as the grouping together could be carried out in a completely arbitrary way leaving the well formedness decision to semantic constraints.

The AVM for the church window example is shown in Table 1: the handshape of the Router's right hand ('RH') is ASL-G, intuitively the handshape used for pointing, delineating or drawing. The palm orientation is downwards ('PDN';

facing the floor), the orientation of the back of the hand is away from the body ('BAB'). The position of the wrist is in the very centre of the Routers gesture space describing two arcs ('CenterCenter'). The wrist movement goes up, does the bend and comes down again ('MU'>MR>MD'). The gesture is large and the wrist position is between the centre of the torso and the elbow ('D-CE'). Furthermore, both hands are involved, the left hand being in a stable position ('LHH'). What we already see here is that we have a set of form features (the attributes) taken from the motor characteristics of the hand movement, the torso position and the relation of hand to torso. 0-values in the AVMs show that, in observational terms, the values of the respective annotation predicates do not reach a critical limit. Hence, they are neglected.

Table 1. Annotation of the empirical gesture datum (Fig. 1) and the simulated gesture datum (Fig. 3).

Annotation predicate	**Empirical gesture datum**	**Simulated gesture datum**
Handshape	ASL-G	ASL-G
– Path	0	0
– Direction	0	0
– Repetition	0	0
Palm Orientation	PDN	PDN
– Path	0	0
– Direction	0	0
– Repetition	0	0
BoH Orientation	BAB	BAB
– Path	0	0
– Direction	0	0
– Repetition	0	0
Wrist Position	CenterCenter	CenterCenter
– Distance	D-CE	D-CE
– Path	ARC>ARC	ARC>ARC
– Direction	MU>MR>MD	MR/MU>MR/MD
– Repetition	0	0
– Extent	Large	Large
Agency	Router	MAX
Handedness	RH	RH
– TwoHandedConfiguration	RFTH>BHA	0
– MovementRelativeToOtherHand	LHH	0

How can we Capture a Gesture's Meaning? Intuitively the meaning of a gesture can be captured in the following way: We assume that meaning is a property of signs. To acquire the status of signs, objects must be conventionalised to some extent, conventionalisation admitting a considerable amount of variation, similar to the pronunciation of words. Hence we have to investigate whether particular hand postures are conventionalised to some extent, and, if provided with some meaning, can align with verbal meaning in a compositional way. To shed some light on this matter is the task of gesture typology extracting types of form features like wrist movement. Clusters as well as types of whole gestures are defined in turn using types of form features. How do we get from these classes, lines, locations and so on to meanings? Instead of applying a feature classification approach as in [2, 7], our idea is that elements of these classes such as bends or lines can be given a fairly non-specific meaning which allows them to combine with verbal meaning. This non-specific meaning is called a *Partial Ontology*. It is partial because it does not fully specify meanings like a lexical definition, remaining hence underspecified, and it yields an ontology because it circumscribes sets of fairly abstract objects.

We explain this reconsidering Table 1 and asking which attribute value pairs might be relevant for determining the semantics of the gesture. Clearly, all 0-values of attributes are non-relevant. The others could all receive different values, consistency presumed, in order to yield the same semantics. Which of them are semantically relevant? Here we rely on the fact that iconic gestures can be subclassified according to different means of representation that are employed.

Several classifications of such representation techniques have been proposed [8]. By and large, they can be unified to the following categories for the description of objects: (1) *(abstract) indexing*: pointing to a position within the gesture space[3]; (2) *placing*: an object is placed or set down within gesture space; (3) *shaping*: an object's shape is contoured or sculpted in the air; (4) *drawing*: the hands trace the outline of an object's shape; (5) *posturing*: the hands form a static configuration to stand as a model for the object itself.

To investigate how the relation of gesture form and meaning is constrained by these techniques of representation we analyzed the SAGA data for characteristics of the very technique broken down in terms of common technique-specific patterns as well as residual degrees of freedom. This analysis revealed that each technique is characterized by particular *technique-specific patterns* as well as *iconic* aspects. Regarding our example—a drawing gesture—this means that some features, namely handedness (typically one-handed), handshape (typically 'ASL-G'), and palm orientation (typically downwards) have technique-characteristic values. The gesture's iconicity is realized only by the type ('ARC>ARC') and trajectory ('MU>MR>MD') of the wrist movement as the gesture's semantically relevant feature values.

This is fine, but where to encode the semantics? Gesture meaning and word meaning must be integrated in the end and we will need the resulting representations for derivations, soundness proofs, inferences and entailments (see [11] for work in this direction). This deliberation leads to the strategy of associating some type of logical form to the relevant form features. Methodologically speaking, this is an annotation problem and should ideally be solved for the whole corpus. In more detail (see Fig. 1): We have a type "Router's one handed sharp bend down with right hand" (R-OH-SharpBendDown-RH) which tags the AVM. The relevant attributes are PathofWrist-RH with value 'ARC>ARC', i.e. two bends, and WristMovementDirection-RH with value 'MU/MR' (continuously moving right while moving up) turning consecutively into ('>') 'MD/MR' (continuously moving down while moving right). The information extracted from the rated annotation is shown in Fig. 1a. This information, representing an in-

[3] By considering abstract indexing gestures we extend the scope from iconic gestures towards representational, i.e. iconic and deictic, gestures.

termediate state, is mapped onto a complex function made up of the attribute value pair as exhibited in Fig.1b. Finally, Fig. 1c shows the stipulated underspecified semantic representation, strictly speaking in logical syntax terms, to which a model-theoretic interpretation must be given. The wrist movement provides two edges of an object, the up-and-down-movement an angle existing between the two edges. Underspecification exists with respect to the orthogonal axis and the typological dimension. So it could be used for a two-dimensional or a three-dimensional object arbitrarily oriented, upright, slanted, inverted etc. in an embedding three-dimensional space.

3 The Generation Perspective

Based on the empirical and theoretical issues discussed above, we will now address the question how iconic gestures convey visuo-spatial information from a generation perspective. In particular, we will show how a computational content representation implements the partial ontology, and how the simulation of gesture use relies on the representation technique-based based mapping of meaning onto gesture form.

A Computational representation of content. As a prerequisite to generate gesture forms, the nature of the underlying meaning representation is of major importance. In other words, an implementation of the partial ontology of abstract gesture description is required as a semantic representation from which overt gesture forms are to be generated. Here we employ a representation called Imagistic Description Trees (IDT) [12]. Each node in an IDT contains an imagistic description which holds an object schema representing the shape of an object or object part. Object schemas contain up to three axes representing spatial extents in terms of a numerical measure and an assignment value like ‘max’ or ‘sub’, classifying this axis’ extent relative to the other axes as an approximation of shape. Accordingly, the IDT model is able to approximate exactly those 0- to 3-dimensional shapes that are covered by the partial ontology. The boundary of an object is defined by a profile vector that states symmetry, size, and edge properties for each object axis or pair of axes. The size property reflects change of an extent as one moves along another axis; the edge property indicates whether an object’s boundary consists of straight segments that form sharp corners, or of curvy, smooth edges. The links in the tree structure represent the spatial relations that hold between the parts and wholes and are quantitatively defined by transformation matrices. It is thus possible to represent decomposition and spatial coherence. In addition, the IDT model provides the possibility of leaving information underspecified which is an important characterizing feature of the partial ontology. The model is, thus, able to represent both concrete and abstract objects. Fig. 3a illustrates how the church window from our example can be operationalized with the IDT model.

A Computational Model of Gesture Production. To generate gesture forms from the IDT representation we have proposed *GNetIc*, a gesture net

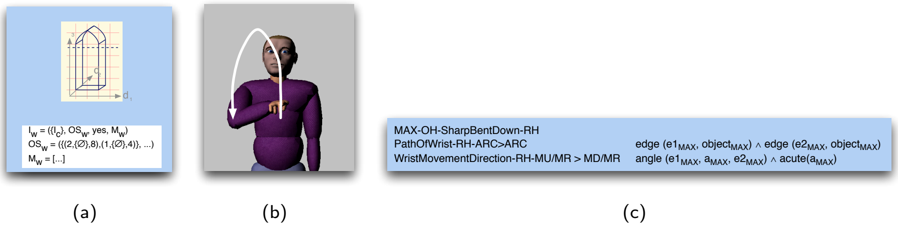
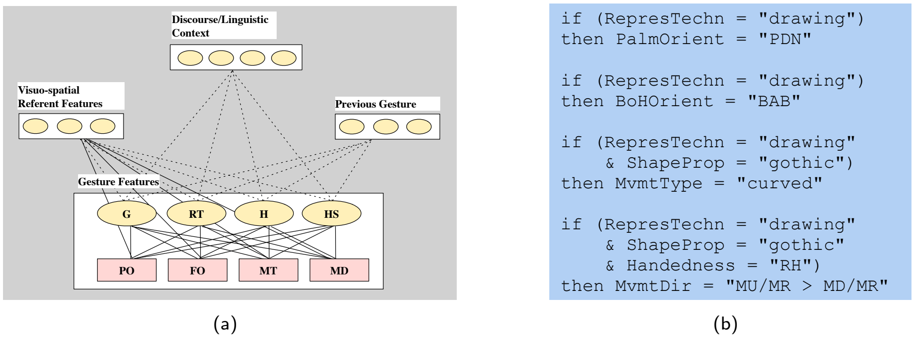

3
d2
d1
Iw = ({Ic}, OSw, yes, Mw)
OSw = ({(2,{∅},8),(1,{∅},4)}, ...)
Mw = [...]
MAX-OH-SharpBentDown-RH
PathOfWrist-RH-ARC>ARC
WristMovementDirection-RH-MU/MR > MD/MR
edge (e1MAX, objectMAX) ∧ edge (e2MAX, objectMAX)
angle (e1MAX, aMAX, e2MAX) ∧ acute(aMAX)
(a)
(b)
(c)

Fig. 3. (a) IDT representation of the church window (b) Generated gesture datum realized by the virtual agent MAX (c) Partial ontology of MAX' gesture.

specialized for iconic gestures [3]. These networks implement the representation technique-based form-meaning relationship as described in Sect. 2, and even go beyond it in that they account for empirical findings which indicate that a gesture's form is also influenced by specific contextual constraints like linguistic or discourse contextual factors (e.g., information structure, communicative goals, or previous gesture use of the same speaker) as well as obvious inter-individual differences. The latter become evident in gesture frequency, but also in preferences for particular representation techniques or the low-level choices of gesture form features such as handshape or handedness. We employ a formalism called Bayesian decision networks (BDNs)—also termed *Influence Diagrams* that supplement standard Bayesian networks by decision nodes. This formalism provides a representation of a finite sequential decision problem, combining probabilistic and rule-based decision-making. We are, therefore, able to specify rules for the mapping of meaning onto gesture forms and at the same time we can account for individual patterns in gesture use.



```
if (RepresTechn = "drawing")
then PalmOrient = "PDN"

if (RepresTechn = "drawing")
then BoHOrient = "BAB"

if (RepresTechn = "drawing"
    & ShapeProp = "gothic")
then MvmtType = "curved"

if (RepresTechn = "drawing"
    & ShapeProp = "gothic"
    & Handedness = "RH")
then MvmtDir = "MU/MR > MD/MR"
```

Fig. 4. (a) Schema of a GNetIc network and (b) a set of rules realized in the decision nodes of these networks determining the values for palm and BoH orientation, movement type, and movement trajectory of a drawing gesture.

GNetIc provides a feature-based account of gesture generation, i.e., gestures are represented in terms of characterizing features as their representation technique and form features which correspond to those ones covered by the gesture typology (cf. Table 1). These make up the *outcome* variables in the model which divide into chance variables quantified by conditional probability distributions in dependence on other variables, ('gesture occurrence', 'representation technique', 'handedness', 'handshape'), and decision variables that are determined in a rule-based way from the states of other variables ('palm orientation', 'BoH orientation', 'movement type', 'movement direction'). Factors which potentially contribute to these choices are considered as input variables. So far, three different factors have been incorporated into this model: linguistic/discourse context (communicative goals, information structure, thematization, noun phrase type), features characterizing the previously performed gesture, and features of the referent (shape properties, symmetry, number of subparts, main axis, position). The latter are extracted from the IDT representation.

The probabilistic part of the network is learned from the SAGA corpus data by applying machine learning techniques. The definition of appropriate rules in the decision nodes is based on our theoretical considerations of the meaning-form relation via gestural representation techniques and our corpus-based analysis of these techniques. That is, depending on the very representation technique, gesture form features are defined to be subject to referent characteristics as well as other gesture form features. See Fig. 3a for the generation network schema and Fig. 3b for a set of rules to determine the values for palm and BoH orientation, movement type, and movement trajectory of a drawing gesture. With respect to representation technique-specificity, the rules account for the fact that drawing gestures are typically performed with a downwards palm orientation and fingers oriented away from the speaker's body. In addition, regarding movement type, the referent-characteristic shape property 'gothic' is considered in terms of a curved movement with a circle-shaped trajectory.

Generation Example. To illustrate gesture generation on the basis of GNetIc models, the generation of an example gesture for the church window to be realized with the virtual agent Max is described in the following (see Fig. 3b). Generation starts upon the arrival of a message which specifies the communicative intent to describe the window with respect to its characteristic properties: 'lmDescrProperty (churchwindow-1)'. Based on this communicative intention, the imagistic description of the involved object gets activated and the agent adopts a spatial perspective towards it from which the object is to be described. The representation is analyzed for referent features required by the GNetIc model: position, main axis, symmetry, number of subparts, and shape properties. Regarding the latter, a unification of the imagistic churchwindow-1 representation and a set of underspecified shape property representations (e.g. for 'longish', 'round' etc.) reveals 'gothic' as the most salient property to be depicted. All evidence available (referent features, discourse context, previous gesture and linguistic context) is then propagated through the GNetIc network (learned from the corpus data of one particular speaker before) resulting in a posterior distribution of probabilities for the values in each chance node. This way, it is first decided to generate a gesture in the current discourse situation at all, the representation technique is decided to be ‘drawing’, to be realized with the right hand and the pointing handshape ASL-G. Next, the model’s decision nodes are employed to decide on the palm and back of hand (BoH) orientation as well as movement type and direction: as typical in drawing gestures, the palm is oriented downwards and the BoH away from the speaker’s body. These gesture features are combined with a curved movement consisting of two segments (to the right and upwards and to the right and downwards) to depict the shape of the window. All values are used to fill the slots of a gesture feature matrix which is transformed into an XML representation to be realized with the virtual agent MAX (see Fig. 3b).

4 Different Styles of Evaluation

The final step is an evaluation of the generation results. This is done in two ways. First, looping back to empirical data and theoretical reconstructions, we take the simulated gesture as a datum. Its annotation is provided with a partial ontology and compared with the originally annotated and interpreted real-world datum. That is, we compare, first, the annotations of both gestures regarding gesture form, and second, the partial ontology of both gestures with regard to semantics. And second, we evaluate the simulation by accessing to what extent the derived model enables a prediction of empirically observed gestural behavior, as well as the degree to which automatically generated gestures, realized with a virtual agent, are beneficial for human-agent interaction.

Data-based Evaluation of Gesture Forms. Concerning the comparison of gesture forms we computed (for a sub-corpus of 473 noun phrases and 288 gestures) how often the model’s assessment was in agreement with the actual gesturing behavior in the SAGA corpus for five networks learned form the data of individual speakers and one ‘average’ network which was learned from the combined data of those five speakers. In a leave-one-out cross-validation it turned out that for each generation choice the prediction accuracy values clearly outperform the chance level baseline. In total, networks learned from the data of individual speakers achieved an accuracy of 71.3% while the accuracy for the combined network was 69.1% (learning with contraint-based PC algorithm). Mean accuracy for rule-based choices made in all networks’ decision nodes is 57.8% (SD=15.5). Altogether, given the large potential variability for each of the variables, results are quite satisfying. E.g., the mean deviation of the predicted finger orientation (direction of the vector running along the back of hand) is 37.4 degrees, with the worst case, opposite rating corresponding to a deviation of 180 degrees.

Theory-based Evaluation of Gesture Semantics. Even gestures whose form features are partly classified as mismatches, may very well communicate adequate semantic features. Therefore, we employ another comparison consisting of the model-theoretic interpretation of the annotations. We explain this with regard to our example. Comparing Figs. 1c (iii) and 3c we see that the semantics MAX gesturally represents is, if considered in terms of intended models, equivalent to the one the Router represented. Using conjunction in the standard way and double brackets '[[,]]' for semantic values we get:

$$[[edge(e1, object) \wedge edge(e2, object) \wedge angle(e1, a, e2) \wedge acute(a)]]^{M,g} \Leftrightarrow \tag{1}$$

$$\begin{gathered}[[edge(e1_{MAX}, object_{MAX}) \wedge edge(e2_{MAX}, object_{MAX}) \wedge \\ angle(e1_{MAX}, a_{MAX}, e2_{MAX}) \wedge acute(a_{MAX})]]^{M,g}\end{gathered} \tag{2}$$

Quantifying over Models M and assignments g we get that any model satisfying (1) will also satisfy (2) and vice versa. In other words, the simulation yields the same semantics as the one deduced from the corpus and is hence adequate. Note, in order to assess the result you have to keep in mind that MAX' gesture was generated using a different methodology, namely Bayesian decision networks (see Sect. 3). In our ongoing work we apply this method to a larger data sample of simulated and empirically observed gestures.

User-based Evaluation. Finally, going beyond the purely communicative functions of gestures, another goal is to explore the user acceptance of the GNetIc-generated gestures, as well as to investigate how the virtual agent itself is judged by human users [4]. Five different conditions were designed differing solely with respect to which GNetIc network was used in the architecture: two individual conditions (*ind-1* and *ind-2*) with GNetIc networks learned from the data of individual speakers, a *combined* condition with a network generated from the data of five different speakers, and two control conditions (*no gestures* and *random* choices at the chance nodes in the network). Note that in all conditions, gestures were produced from identical input and accompanied identical verbal output.

In a between-subject design, a total of 110 participants (22 in each condition), aged from 16 to 60 years (M = 23.85, SD = 6.62), took part in the study (44 female/66 male). Participants received a description of a church by the virtual human MAX, produced fully autonomously with a speech and gesture production architecture containing GNetIc. Immediately after receiving the descriptions, participants filled out a questionnaire to rate quantity and quality of MAX' gestures, quality of the overall presentation and their person perception of the virtual agent in terms of items like 'polite', 'authentic', or 'cooperative'.

Results can be summarized in four major points (for details see [4]). First, MAX' gesturing behavior was rated positively regarding gesture quantity and quality, and no difference across gesture conditions was found concerning these issues. That is, building generative models of co-verbal gesture use can yield good results with actual users. The fact that gesture quality was rated more or less equal across conditions rules out the possibility that other effects of the experimental conditions were due to varying quality of gesture use and realization in the virtual agent. Second, both individual GNetIc conditions outperformed the other conditions in that gestures were perceived as more helpful, overall comprehension of the presentation was rated higher, and the agent's mental image was judged as being more vivid. Similarly, the two individual GNetIc conditions outperformed the control conditions regarding agent perception in terms of likeability, competence, and human-likeness. Third, the *combined* GNetIc condition, notably, was rated worse than the individual GNetIc conditions throughout. This finding underlines the important role of inter-individual differences in communicative behavior and implies that the common approach to inform behavior models from empirical data by averaging over a population of subjects is not necessarily the best choice. Finally, the *no gesture* condition was rated more positively than the *random* condition, in particular for the subjective measures of overall comprehension, the gesture's role for comprehension, and vividness of the agent's mental image. That is, with regard to these aspects it seems even better to make no gestures than to randomly generate gestural behavior even though it is still considerably iconic.

5 Conclusion

In this paper we provided an interdisciplinary view on the question how co-speech iconic gestures convey visuo-spatial information combining empirical study, theoretical modeling and computational simulation (see Fig. 1). Empirical data is used for establishing a gesture typology which rests on gesture form features like handshape, palm-direction or wrist-movement extracted from systematic corpus annotations. Clusters of features then provide entities of different dimensions such as lines, regions, partial objects and composites of these which are provided by a partial ontology. Founding the simulation on empirical study and theoretical reconstructions is then accomplished with a computational content representation that implements the partial ontology, and with a simulation model of gesture use that realizes the mapping of meaning onto gesture form. The computational generation approach with GNetIc is, however, not only driven by features of the referent object, but also takes into account the current discourse context and the use of different gestural representation techniques. Finally, in terms of an evaluation two mappings are established between the gesture in the original datum and the generated gesture. Its annotation is provided with a partial ontology and compared with the originally annotated and interpreted real-world datum. The model was shown to be able to successfully approximate human gesture use of iconic gestures, and gestural behavior can increase the perceived quality of object descriptions as well as the perception of the virtual agent itself in terms of likeability, competence and human-likeness as judged by human recipients.

We are aware that our results also reveal deficiencies, which mark starting points for further refinements. For instance, we restricted our work to gestures used in object descriptions for simplified VR objects, so far. The description of more realistic entities or other forms of gesture use, like verb-phrase aligned gestures, e.g., pantomime gestures or typical direction-giving gestures as in 'turn right', pose further challenges. Another focus of our future work is an extension towards gesture use in *dialogues*. This includes the consideration of dialogue phenomena like gestural mimicry, but also the use of gestures to regulate the organization of the interaction, e.g., in terms of gestural acknowledgements or turn allocation gestures [5]. We are confident that the interdisciplinary methodology we have demonstrated in this paper, with several points of interaction between the involved disciplines, has the potential to also deal with these issues.

Acknowledgements. This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 “Alignment in Communication” and the Center of Excellence in “Cognitive Interaction Technology” (CITEC). The authors want to thank Florian Hahn for joint work on gesture typology.

References

1. Bavelas, J., Gerwing, J., Sutton, C., Prevost, D.: Gesturing on the telephone: Independent effects of dialogue and visibility. Journal of Memory and Language 58, 495–520 (2008)
2. Beattie, G., Shovelton, H.: An experimental investigation of the role of different types of iconic gesture in communication: A semantic feature approach. Gesture 1, 129–149 (2001)
3. Bergmann, K., Kopp, S.: GNetIc—Using Bayesian decision networks for iconic gesture generation. In: Proceedings of IVA 2009, pp. 76–89. Springer, Berlin/Heidelberg (2009)
4. Bergmann, K., Kopp, S., Eyssel, F.: Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans. In: Proceedings of IVA 2010. pp. 104–117. Springer, Berlin/Heidelberg (2010)
5. Bergmann, K., Rieser, H., Kopp, S.: Regulating dialogue with gestures—towards an empirically grounded simulation with virtual agents. In: Proceedings of SigDial 2011. ACL, Portland, Oregon (2011)
6. Hahn, F., Rieser, H.: Explaining speech gesture alignment in mm dialogue using gesture typology. In: Lupowski, P., Purver, M. (eds.) Proceedings of SemDial. pp. 99–111. Polish Society for Cognitive Science (2010)
7. Holler, J., Beattie, G.: How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? Semiotica 146/1, 81–116 (2003)
8. Kendon, A.: Gesture—Visible Action as Utterance. Cambridge Univ. Press (2004)
9. Lücking, A., Bergmann, K., Hahn, F., Kopp, S., Rieser, H.: The Bielefeld speech and gesture alignment corpus (SaGA). In: Proceedings of the LREC 2010 Workshop on Multimodal Corpora (2010)
10. Rieser, H.: On factoring out a gesture typology from the bielefeld on factoring out a gesture typology from the bielefeld speech-and-gesture-alignment corpus (saga). In: Kopp, S., Wachsmuth, I. (eds.) Gesture in Embodied Communication and Human-Computer Interaction. Springer, Berlin/Heidelberg (2010)
11. Rieser, H.: How to disagree on a church-window’s shape using gesture. In: Hölker, K., Marello, C. (eds.) Dimensionen der Analyse von Texten und Diskursen. pp. 231–247. LIT Verlag, Münster (2011)
12. Sowa, T., Wachsmuth, I.: A model for the representation and processing of shape in coverbal iconic gestures. In: Proceedings of KogWis 2005. pp. 183–188 (2005)