Automatic Nonverbal Behavior Generation from Image Schemas

Socially Interactive Agents Track

Brian Ravenet
LTCI, Télécom ParisTech
Paris, France
brian.ravenet@telecom-paristech.fr

Chloé Clavel
LTCI, Télécom ParisTech
Paris, France
chloe.clavel@telecom-paristech.fr

Catherine Pelachaud
CNRS, ISIR, Sorbonne University
Paris, France
catherine.pelachaud@upmc.fr

ABSTRACT

One of the main challenges when developing Embodied Conversational Agents is to give them the ability to autonomously produce meaningful and coordinated verbal and nonverbal behaviors. The relation between these means of communication is more complex than a direct mapping that has often been applied in previous models. In this paper, we propose an intermediate mapping approach we apply on metaphoric gestures first but that could be extended to other representational gestures. Leveraging from previous work in text analysis, embodied cognition and co-verbal behavior production, we introduce a framework articulating speech and metaphoric gesture invariants around a common mental representation: *Image Schemas*. We establish the components of our framework, detailing the different steps leading to the production of the metaphoric gestures, and we present some preliminary results and demonstrations. We end the paper by laying down the perspectives to integrate, evaluate and improve our model.

KEYWORDS

social behaviors; embodied cognition; text analysis; virtual agents

ACM Reference Format:
Brian Ravenet, Chloé Clavel, and Catherine Pelachaud. 2018. Automatic Nonverbal Behavior Generation from Image Schemas. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 8 pages.

1 INTRODUCTION

When humans talk, they usually accompany their discourse with co-speech gestures that contribute to conveying the desired communicative intentions. In [35], the authors gave an extensive review of work that investigated co-speech gestures, from psychology studies to computer systems. Their results highlighted how closely tied together speech and gesture are (in terms of meaning and timing). In order to design Embodied Conversational Agents that can act as social partners in conversation, they need to be able to produce these co-verbal behaviors as well. We aim at building a system capable of autonomously planning gestures (in terms of timing and shape) using the textual surface discourse of the agent augmented with prosodic information (e.g. pitch accents). Such a system could have multiple applications such as computing animations for automatically generated dialogs, serving as a baseline for 3D character animators or even be used as a pre-visualization tool in games and movies. Following Mc Neill's topology of gestures [25], there exist four main categories of gestures: beat, deictic, iconic and metaphoric gestures. Beat gestures are not tightly coupled with the content of the speech (rather with its rhythm and tempo) but the other categories are; deictic gestures are gestures locating an entity in the physical space (pointing at something, self-touching when talking about self...); iconic gestures replicate physical properties of an object (making a spherical shape with the hands while talking about a ball for instance); and metaphoric gestures give similar physical properties but to abstract entities (describing an improvement with a raising movement for instance). When observing how people gesticulate while talking, one can notice and acknowledge this intricate relationship between the speech and the corresponding gestures. Mc Neill, in his *Growth Point* theory [24], proposed an explanation of this phenomenon as, according to him, speech and gestures would be produced around a common mental imagery and, therefore, are two channels for the same cognitive process.

Our objective is to capture such a mental imagery in order to use it as the link between speech and gestures for our behavior generation system. As a first step at building a formal framework for representational gesture generation, we take inspiration from work that aimed at building metaphoric gesture generation systems. In [21], the authors have used conceptual metaphors identified in the text as an input to a system that produces corresponding metaphorical gestures. Our work is based on this approach but we are interested in going further. We aim at building a more general representation that could be later naturally extended for other representational gestures (such as iconic gestures). This paper is organized as follows: in Section 2 we detail the theoretical foundations of our approach, in Section 3 we review previous work related to our challenges, in Section 4 we describe the different components of our framework and our preliminary results; finally in Section 5 we conclude by discussing the perspectives of our work.

2 BACKGROUND

As mentioned in the introduction, Mc Neill argued in his *Growth Point* theory for a common mental representation used in both the verbal and the nonverbal channels [24]. Kendon also argues for the verbal and nonverbal channels to be two aspects of the one process of utterance [12]. According to the *Growth Point* theory, speech and gestures are planned simultaneously at specific moments of a person's discourse (like pauses) using its common representation. In other words, both the verbal and the nonverbal channel are results of a same cognitive process based on a common representational structure or language. In order to be able to map the concepts from the text of the agent's speech to specific gesture components, we could use such a common representational language. Since we were inspired to generate metaphoric gestures first, we looked at

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

previous work in the domain of linguistic and embodied cognition focusing on metaphorical reasoning and identified an interesting representation called *Image Schemas*.

2.1 Conceptual Metaphors and Image Schemas

The conceptual metaphor theory by Lakoff and Johnson [16] describes how humans use metaphorical reasoning as part of their natural thought process and in the language production. A conceptual metaphor is expressed as **TARGET IS A SOURCE** (for instance *LOVE is a JOURNEY*) and allows mapping properties from the source domain (journey) to the target domain (love). In particular, the authors describe how interactions in the physical environment shape these metaphors. Following the idea that metaphors can be embodied concepts, build from our personal physical experience, Johnson suggested that humans use recurring patterns of reasoning, called *Image Schemas*, to map these conceptual metaphors from an entity to another [11]. For instance, the *Image Schema* CONTAINER gives to an entity typical properties of a container like having a border with elements that are within and elements that are outside. For instance, we can think of a country as a CONTAINER when we think about it in terms of people that are inside this country, and people that are not. This would give an explanation on how humans transfer their reasoning about their physical reality onto abstract concepts, thus giving physical attributes to abstract entities. Could this linguistic structure intervene in the gesture production as well? In [28], the author describes how a gesture (mimicking the shape of a box in the example) can represent the Image Schema OBJECT or CONTAINER, itself being linked to the conceptual metaphor IDEAS are OBJECTS. In another work, Cienki conducted an experiment to study if *Image Schemas* (a subset) could be used to characterize gestures [7]; his conclusions showed positive results. In [6], the authors revealed evidence of the use of conceptual metaphors, spatial ones, in gesture production for mandarin speakers. Another experiment by Lücking and his colleagues tried to find recurrent gestures features in the expression of particular *Image Schemas* [23]. Their results showed that people, for some of the *Image Schemas*, spontaneously used similar gestures features. Finally, in [26], the authors developed a gesture-based interface for an interactive museum system that is based on *Image Schemas* as a basis for their gestural grammar as they pointed out their potential for creating a bridge between natural language and gesticulation.

2.2 Ideational Units and gesture shapes

Calbris argues for the existence of *Ideational Units* in the verbal and nonverbal channels [3]. Ideational Units are units of meaning that give rhythm to the discourse of a person and during which gestures show similar properties. Calbris explains that a gesture can have invariant properties that are critical for the meaning of the gesture. For instance, a gesture representing an ascension would probably have an upward movement or an upward direction but the handshape may not be of particular relevance (for the ascension meaning). Calbris explains that within an Ideational Unit, successive gestures needs to show significant changes to be distinguished. However, she explains that invariant properties of a gesture would be transferred to the variant properties of the next gestures in order to carry the meaning further in the gesticulation of a person. This mechanism allows gestures to be combined together by mixing their invariant properties. Leveraging from this definition, we will establish associations between *Image Schemas* and specific gesture invariants in order to be able to combine and produce them following the model of Calbris, like in [36]. Therefore, gestures produced within the same *Ideational Unit* will be tightly connected and will share certain characteristics such as the shape of the hands or their directions.

2.3 Timing relationship between gesture and speech

In regards to timing and rhythm, are co-speech gestures and speech perfectly aligned then? While one might think so, exposing definitive evidences for this claim has been difficult. According to [35], several work showed that gesture and speech timings seem not to be exactly simultaneous but rather close to each other. Results from [19] or [22] acknowledge the correlation between gesture phases and prosodic markers while accepting slight variations. In the particular case of beat gestures, which are not constrained by meaning, the peak of the stroke seemed to be closer to the pitch emphasis. For representational gestures, it would seem that the gesture anticipates the prosodic markers of the discourse. In [12], Kendon states that the stroke of a gesture precedes or ends at, but does not follow, the phonological peak of the utterance. In her work, Calbris also identified that when constructing thoughts in a discourse, gestures tend to slightly anticipate the speech [3]. Inspired by these results and in order to propose a mechanism to align the produced gestures with the speech, we take into account these findings in our framework.

3 RELATED WORK

Our approach faces three main challenges that are the identification of our common representational language, using *Image Schemas*, in the text content of the agent's speech, the association between this representation and gesture invariants and finally the combination of gesture invariants through *Ideational Units*. In this section, we present work that have investigated these three challenges.

3.1 Text analysis for extracting Image Schema

According to the literature, several researchers have made the hypothesis that speech and gesture can come from the same cognitive process [22, 24]. We saw that the metaphorical process demonstrates this phenomenon through the use of common sources of reasoning called *Image Schemas* [11]. Other previous work also highlighted how language can be structured around *Image Schemas* [33]. In [1], the authors even proposed a formalism to identify *Image Schemas* in the discourse.

The task of identifying *Image Schemas* in the discourse of the agent requires some form of *Natural Language Processing*. Lately, it has become quite common to rely on machine learning techniques to perform such tasks. Sequential learning has been proven to be effective to extract semantic information from text [10]. However, these techniques usually require corpora of annotated data to learn a model that can be difficult to collect. The only automated method for *Image Schemas* detection we are aware of is the work of Gromann and Hedblom [9]. In this work, the authors use a clustering method on the Europarl corpus to obtain clusters of *verb-preposition* couples. Then, using a semantic role labeling tool built in [32] with the PropBank corpus, they labeled their clusters with a semantic role. Finally, using two annotators, they assigned an *Image Schema* to each cluster based on the semantic role previously identified. While their approach seems interesting, they limited themselves to a subset of spatial *Image Schemas*.

3.2 Association between speech and gestures

Previous work attempted to build computational models that give to an agent the ability to couple gesture generation with speech generation in an automatic way.

In [17], the authors proposed a system that detects keywords through a surface analysis of what the agent would say. It then maps associated nonverbal features to them, mainly head movement and eyebrow movements. For instance, the system was capable of detecting the words *yes* and *no* and would insert a head nod or a head shake at their location in the behavior of the agent. The same authors used a machine learning approach in [18] to automatically produce head movements according to the dialog acts and the affective state of the agent.

In [2], the authors attempted to learn from an annotated corpus of spatial descriptions a Bayesian model used to predict the shape of a speaker's iconic gestures used to describe the shapes of objects situated in a virtual environment (like a church). However, they propose a decomposition of gestures into two features only: representational technique (like placing or drawing for instance) and handshape. Additionally, they also limited themselves to the specific context of giving directions.

In [5], the authors learned two models; a first one using *Conditional Random Fields* for associating the audio, using prosodic information, with gesture elements and a second one using *Gaussian Process Latent Variable Models* for producing the motion based on these elements. However, their models do not take into account the semantic behind the speech but focus solely on producing gesturing activity without meaning.

These works either proposed a system limited to a specific context or tried to produce gestures synchronized on timing but not shaped to represent the meaning that is being conveyed. In our work, we aim at proposing a system that takes into account both the timing and the meaning while being compatible with any context.

The work that was the most inspiring for us was the work of Lhommet and Marsella [21]. In this work, the authors proposed a logic-based model that maps the communicative intentions of an agent to primary metaphors in order to build a mental state of *Image Schemas*. This mental state is then used to produce corresponding gestures using a second layer of reasoning. While their approach is really interesting, they consider only a limited subset of *Image Schemas* and they associated specific prototypical gestures to each action on the mental state. In our work, we want to propose a finer mechanism by associating more *Image Schemas* to gesture invariants that would allow richer combinations in creating gestures.

3.3 Combination of gestures into Ideational Units

We saw that in most of the existing work about producing automatically representational gestures, the extracted meaning is mapped to a prototypical gesture library. In our work we are interested in the combination of gesture properties and their transfer among each other in order to be able to combine Image Schemas together and to have a more flexible system. Such an approach was used in [4] where the authors established *Image Descriptive Features* IDF (conceptually close to *Image Schemas* but used to describe geometrical and spatial features of concrete entities) and how they relate to gesture features. Their context was a direction-giving task. They analyzed a corpus of interaction between person giving directions and exposed evidences of correspondence between the gesture features and the spatial features of the object being described. While their system allows combining multiple IDFs to form one gesture, they do not consider the transfer of properties throughout the utterance of the agent.

In [36], the authors take inspiration from Calbris' work on *Ideational Units* to propose a model that constrains the production of metaphoric gestures following the *Ideational Units* of the agent's utterances. For instance, the agent should not go back to a rest position between gestures that are within the same *Ideational Unit*. Their system does not address the automatic extraction of semantic elements within the agent utterances.

We take inspiration from this previous work to propose a more complete system that relies on capturing an intermediate language between speech and gestures and on decomposing gestures into finer features to combine them within *Ideational Units*. Our approach is 3 fold. First, we develop a new method for extracting automatically *Image Schemas* from the text. Secondly, we propose a dictionary of gesture invariants, associated to *Image Schemas*, that can be used to compose dynamically representational gestures for any context. Lastly, we integrate the whole process within an Ideational Unit compatible behavior realizer where the invariant mechanism allows the combination of gestures and the transfer of the *Image Schematic* meaning throughout the discourse of the agent.

4 IMAGE SCHEMA BASED GESTURE GENERATOR

4.1 Architecture

We saw that some researchers acknowledge the theoretical model of gesture and speech being produced by the same cognitive process [12, 24]. Therefore, *Image Schemas* are pattern of reasoning applied at an early stage and should be used as inputs for a verbal and coverbal behavior realizer. However, we place ourselves in the context where the speech of the agent would be already produced. We make this assumption in order to be able to combine our framework with existing speech production systems and to propose a mechanism approximating the link between speech and gestures. Therefore, instead of starting from the *Image Schemas* to generate both the speech and gestures, we start from the text, aiming at identifying the *Image Schemas* that could have led to the production of this speech to generate the corresponding gestures (see Figure 1). We justify

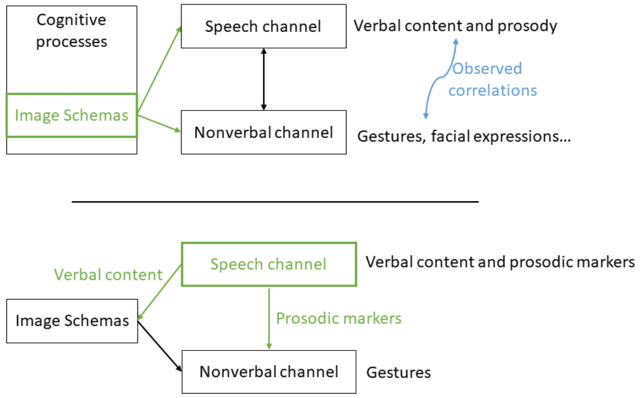

Figure 1: Top: theoretical model according to [11, 12, 24], Image Schemas are used within the cognitive processes as inputs for both channels. Bottom: our framework architecture, the Image Schemas are retrieved from the text and combined with prosodic markers to generate gestures.

this approach through the results of the literature that exposed the correlation that can be observed between speech and gestures.

Our architecture is composed of three levels: an *Image Schema* extractor, a gesture modeler and a behavior realizer supporting *Ideational Units.*

4.2 Image Schema extractor

The *Image Schemas* extraction component has the task of identifying from the surface text of the agent's speech *Image Schemas* and to align them properly with the spoken utterance (for future gesture alignment). There is not a definitive list of *Image Schemas* and different researchers have proposed complementary or alternative ones. We propose our own list adapted from the original list of Johnson [11] and of Clausner and Croft [8]: UP, DOWN, FRONT, BACK, LEFT, RIGHT, NEAR, FAR, INTERVAL, BIG, SMALL, GROWING, REDUCING, CONTAINER, IN, OUT, SURFACE, FULL, EMPTY, ENABLEMENT, ATTRACTION, SPLIT, WHOLE, LINK, OBJECT. This list allows us to manipulate spatial, temporal and compositional concepts (container vs object and whole vs split for instance).

Algorithm 1 ImageSchema extraction using WordNet.

```
for all word do
    ImageSchema = none;
    SynonymSet = Lesk(word);
    while TopNotReached() & ImageSchema == none do
        ImageSchema = getImageSchema(SynonymSet);
        MoveUpFollowingHypernym();
    end while
end for
```

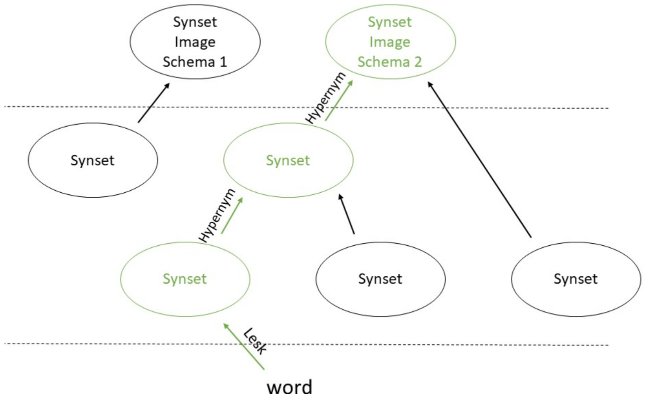

Figure 2: The algorithm follows the path of hypernyms up until it reaches a Synonym Set associated to an Image Schema

4.2.1 Image Schemas detection. Ideally, *Image Schemas* should be associated with a short sequence of text, in order to encapsulate the different words that might be carrying them. Sequence labeling algorithm is a common task in Natural Language Processing as it is used to deal with various problems like Named Entities Recognition for instance [30]. However, this kind of approach requires an annotated corpus of data which we found to be difficult to obtain for *Image Schemas.* To the best of our knowledge, no database of text annotated with *Image Schemas* is available. We started to conceive an annotation schema in order to build our own corpus but our tests showed that the annotators need to be already familiar with *Image Schemas* or need to take too much time to be comfortable with the concepts. Therefore, we decided to use an expert approach using WordNet dictionary [27]. In WordNet, words are organized in synonym sets, each set being connected to other sets by semantic relations. Following the hypernymic relations of a synonym set, one can obtain a synonym set with a more general meaning (for instance a hypernym of *car* is *vehicle*). This organization is similar to a class inheritance system. Our algorithm works as follow (see Algorithm 1): for each word in the text, we use the Lesk method to disambiguate the meaning of the word and find the most likely synonym set for it using WordNet [20]; Then, we follow the hypernym path up in the hierarchy until we find a synonym set corresponding to our *Image Schemas* as depicted in Figure 2 (if none is found, no *Image Schema* is returned). Using the literature on conceptual metaphors and by observing political videos (which are known to be rich in metaphoric gestures) of former U.S.A president Barack Obama, we empirically established a repertoire of synonym sets corresponding to our *Image Schemas.* To establish these correspondences, we follow the hypernym path up on typical instances of the *Image Schema* (like *affinity* for the *Image Schema* ATTRACTION for instance) and we stop when the next hypernym does not carry the meaning anymore. For instance, for the word *affinity*, its next hypernym is *attraction.* Since *attraction* still carries the meaning of ATTRACTION, we proceed to the next hypernym. In the case of *attraction*, the next hypernym is *quality* therefore we stop. Several synonym sets are associated to each *Image Schema* to cover possible variations in meaning.

4.2.2 Syntactic and prosodic selection. Instead of keeping all *Image Schemas* detected for every word, we operate a filtering selection in order to replicate observations from the literature and to avoid exaggerating the gesticulations of the agent. We use OpenNLP chunker [29] to obtain group of words (like noun groups and verb groups) and we tag one *Image Schema* per group as the main *Image Schema* of this group. We use the Stanford POS Tagger [34] to retrieve the syntactic role of each word and we prioritize the *Image Schemas* obtained from modifiers such as adverbs and adjectives [3] as main ones unless a stressed accent is put on a particular word. As we saw earlier in Section 2, gestures can also slightly anticipate the speech [35]. In order to properly align them, we use the prosodic information to ensure that gesture strokes end at or before (up to 200ms) pitch accents [13]. The result is a list of Image Schemas, each one specifying exactly when it starts and ends in the spoken text using time markers.

4.3 Gesture Modeler

Now that we obtain a list of aligned *Image Schemas* for a sequence of spoken text, the gesture modeler will build the corresponding gestures.

The first step is to retrieve the gesture invariants to build the final gestures. According to the literature, the typical features of a gesture are: the handshape, the wrist position, the palm orientation and the arm movement [3]. In [4], the authors proposed to represent gestures using these first three features augmented with a movement information on each of them. Our objective is for each *Image Schema* to find which features are needed to express its meaning and how it is expressed. For this task, we propose a dictionary that maps each *Image Schema* to its corresponding invariants. This dictionary is depicted in Table 1. This dictionary was conceived after a review of work on gesture meaning [3, 13] and contains the minimal features required to express a specific *Image Schema*.

From the invariants, we can compose the final gestures by combining them together into one gesture per group. Conflicting situations might happen and this is why we use the main tag to keep the invariant features of the main *Image Schema* of a group.

4.4 Behavior realizer using Ideational Units

The final layer of our framework has the role of combining the composed gesture obtained through the previous components to produce the final animation of the virtual agent. Note that our system requires the text to be annotated with the delimitations between the *Ideational Units* as we do not deal with their extraction in this work. In our system, we follow the BML specification [14] where gestures are defined by two keyframes, *gesture-stroke-start* and *gesture-stroke-end*. Our animation system interpolates between the keyframes to compute the animation of the virtual agent.

We use a rule-based system that follows the *Ideational Unit* model proposed by Calbris [3] and that follows the work of [36] and adapts it to our animation system. The system operates the following main functions: 1) co-articulating gestures within an *Ideational Unit* by computing either a hold or an intermediate relax poses between them (instead of returning to a rest pose), 2) transferring properties of the main gesture onto the variants properties of the other gestures of the same *Ideational Unit*, 3) ensuring that a meaning expressed through an invariant is carried on the same hand throughout an *Ideational Unit* and 4) finally dynamically rising the speed and amplitude of repeated gestures. More precisely, to compute the relax pose of a gesture, our algorithm lowers the wrist position in 3D space; it also modifies the handshape by using relax position of the fingers rather than straight or closed positions. A gesture phase is held within an *Ideational Unit* when the time between the end of the gesture stroke and the beginning of the next gesture stroke is below a given threshold. To transfer properties of one gesture (here the main gesture) to other ones, we instantiate for each other gesture features that are not indicated as invariant with the corresponding value of the gesture features. To mark the repetition of a gesture, we extend the position of the wrist in 3D space for each gesture stroke position to increase the amplitude of the gesture. We do not modify the timing of the gesture phases but since the position of the arms have been extended, the interpolation speed is increased as a consequence.

4.5 Integrated system and examples

We implemented our whole framework within the agent platform Greta [31], the Image Schema extractor, the gesture modeler and the *Ideational Unit* compatible behavior realizer. The framework takes as input a BML document with the textual speech of the agent marked with prosodic and *Ideational Unit* information and produces the complete animation with the audio using Text-To-Speech component.

As a first assessment of the quality of our approach, we selected a video showing a politician displaying metaphoric gestures (the original video can be found at https://youtu.be/0ggic7bDNSE); we transcribed the textual speech and the prosodic information from the videos and let our system produce the corresponding gestures. Our goal was to observe if the automatically generated gestures were close to the ones from the source videos or, if they differed, we wanted to assess if, nevertheless, they carry similar meanings as the original videos. An example of simulation is shown in Figure 3 and the video can be watched at https://youtu.be/47QLONZS5zw. The output results are quite similar to the input video. For each metaphoric gesture of the video, our animation replicated a gesture with similar shape and timing. For instance, at the beginning, the politician says "we have to get back to harvesting the wisdom of crowds" while moving his arms in a circle like he is gathering the wisdom. Our algorithm captured the *Image Schema* ATTRACTION in the word *harvesting* and therefore, produced a gesture where the agent is pulling something towards him, in this case *wisdom*. Another interesting example from our video happened when the politician said "good ideas rise to the surface": in the video, the politician do a gesture mimicking something going up, to accompany the verb "rise". In our output, the *Image Schema* SURFACE, extracted from the word "surface", was identified as the main *Image Schema* of the group rather than the UP one (that was extracted for the "rise" word). This choice resulted in the agent doing a gesture with an horizontal wipe (see Figure 4). Despite being different in shape, the gesture produced by the agent is still coherent with the content of the speech. It highlights another element of speech. In the original video, the temporal relationship between speech and gestures varies, with gestures being perfectly in sync and others being a little bit more ahead of the speech, as described in the literature. Our system aligns gestures with speech. However it does not produce that much variability in the temporal relationship between

Table 1: Association between Image Schemas and gesture invariants

Image Schemas	handshape	wrist position	palm orientation	arm movement
UP		up		
DOWN		down		
FRONT		front		
BACK		close	back	
LEFT		left		
RIGHT		right		
NEAR		close center		
FAR		away	frontward / downward	
INTERVAL	flat		inward	
BIG	open	away	inward	
SMALL	mid-closed	close center	inward	
GROWING				from SMALL to BIG
REDUCING				from BIG to SMALL
CONTAINER	bowl-shape		inward	
IN	picking-shape		downward	
OUT	open spread		outward	
SURFACE	flat		downward	horizontal wipe
FULL	closed fist			
EMPTY	open spread			
ENABLEMENT	open			frontward
ATTRACTION	closed fist			backward
SPLIT	flat		inward	abrupt downward
WHOLE	open		inward	
LINK	hold			translation
OBJECT	conduit shape			

speech and gesture. This results in gesture having closer temporal relationship with speech in our output than in the original video. Understanding what causes this temporal variability in human communication in order to model it is another challenge that could be addressed in future work. We can notice that the output of our system did not systematically reproduce the exact gestures seen in the source video as it may select other *Image Schema* to highlight with a gesture; but, nevertheless, it was able to generate animated sequences that are coherent in terms of speech-gesture synchronization. This difference in selecting which gestures to produce comes from the selection of the 'important' *Image Schema*. At the moment we consider one utterance and its prosody profile, we do not take into consideration other contextual factors such as what has already being said, if they are contrastive elements... Another explanation of this difference could be that our system has a limited set of gesture invariants built from the literature and, despite being able to produce coherent gestures, it cannot capture the variations or style of a speaker. An interesting alternative could be to build a stochastic model of invariants learned from a corpus of gesture data for a given speaker. This would allow introducing more variability and reproducing a "speaker style" parameter as input.

5 CONCLUSION AND PERSPECTIVES

In this paper, we presented a framework investigating the potential of an intermediate semantic representation, inspired by embodied

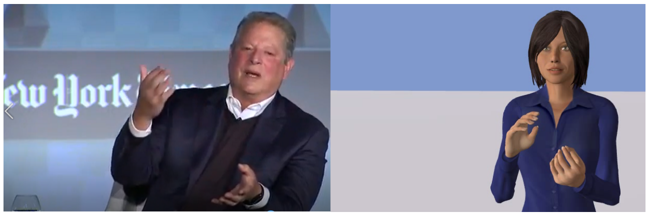

Figure 3: Left: the politician is expressing two ideas, one on each hand. Right: the agent is replicating this metaphor in displaying similar hand gestures

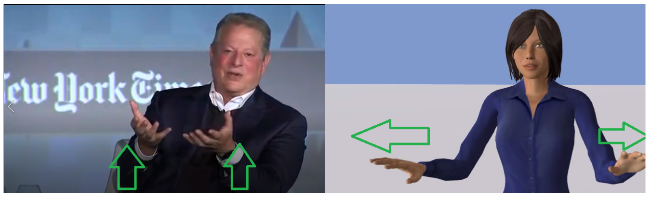

Figure 4: The case of the sentence "good ideas rise to the surface". Left: the politician illustrates his speech with a rising gesture. Right: the agent choses to illustrate the surface concept and thus displays an horizontal wipe gesture

cognition and nonverbal behavior theories, for synchronizing spoken meaning with gestural meaning. Through the investigation of different NLP tools and techniques, we were able to propose a method to automatically extract *Image Schemas* from spoken texts. An additional set of rules inspired by lexical and prosodic models ensure the proper selection and alignment of the *Image Schemas* with the text. Secondly, a dictionary for translating *Image Schemas* into gesture invariants was proposed as the core of a gesture modeler capable of combining these invariants into complete gestures. Finally, an *Ideational Unit* compatible behavior realizer was conceived in order to handle the combination and transfer of properties of the generated gestures through the discourse of the agent. The whole system was implemented within our agent platform and preliminary examples were generated from political videos in order to assess the quality of the framework. Throughout the examples we generated, we observed that the system produces coherent gestures, in line with the content of the agent's speech. Nonetheless, our current work faces several limitations.

Even though we are proposing a more complete list of *Image Schemas* than previous works do, our list is far from complete. Since we focused on metaphoric gestures, we proposed a representational gesture system that works mainly for abstract entities. Many iconic gestures can be produced as well since some concrete entities can also carry *Image Schema* (a box is a CONTAINER for instance). To encompass larger set of iconic gestures, a first addition could be to compute the underlying geometrical shapes iconic gestures should depict (see work by [2, 4]).

While our WordNet based system to extract Image Schema produces relatively correct outputs, this part was built following empirical methods. In order to improve it, more synonym sets could be integrated to cover all the variations in meaning. However this process can be costly and difficult to set up. An alternative method based on sequential learning algorithm could be considered in order to take advantages of the flexibility of such models. This would require to come up with a proper annotating schema (such as the BIO coding system used for Conditional Random Fields for instance [15]) and to apply it to a corpus of videos. The improved model should be properly evaluated in experimental conditions then.

Even though we wanted to move away from the library of prototypical gesture approach, the gesture invariant dictionary only considers one possibility for each invariant of an *Image Schema*. One way to extend it could be to use a machine learning approach to capture a stochastic model for each considered invariant. Again, to our knowledge, there is no available corpus for this task.

Finally, we only address representational gestures, with a particular emphasis on metaphoric ones. One of the main challenges that will rise in the future will be the combination with other gestures such as beats for instance. This will require us to improve our *Ideational Unit* compatible behavior realizer system to ensure that the whole framework is capable of handling the simultaneous, different and complementary communicative intentions of the agent, like expressing emotions for instance.

Through this work, we highlight the potential of *Image Schemas* to capture part of the semantic shared between the verbal and the nonverbal channel in order to automatically compute representational gestures. In order to pursue our effort at creating a fully autonomous nonverbal behavior generator, two main tasks should be completed in the future: an evaluation of our current system, that may shine the light on additional limits of our approach, and the construction of a corpus of speech and gestures annotated with the corresponding *Image Schemas*, aimed at exploring machine learning approaches to improve our framework.

ACKNOWLEDGMENTS

The authors would like to thank Stacy Marsella for the preliminary ideas and discussions that lead to this work.

REFERENCES

[1] Benjamin Bergen and Nancy Chang. 2005. Embodied construction grammar in simulation-based language understanding. *Construction grammars: Cognitive grounding and theoretical extensions* 3 (2005), 147–190.

[2] Kirsten Bergmann and Stefan Kopp. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 361–368.

[3] Geneviève Calbris. 2011. *Elements of meaning in gesture*. Vol. 5. John Benjamins Publishing.

[4] Justine Cassell, Stefan Kopp, Paul Tepper, Kim Ferriman, and Kristina Striegnitz. 2007. Trading spaces: How humans and humanoids use speech and gesture to give directions. *Conversational informatics* (2007), 133–160.

[5] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 781–788.

[6] Kawai Chui. 2011. Conceptual metaphors in gesture. *Cognitive Linguistics* 22, 3 (2011), 437–458.

[7] Alan Cienki. 2005. Image schemas and gesture. *From perception to meaning: Image schemas in cognitive linguistics* 29 (2005), 421–442.

[8] Timothy C Clausner and William Croft. 1999. Domains and image schemas. *Cognitive linguistics* 10 (1999), 1–32.

[9] Dagmar Gromann and Maria M Hedblom. 2017. Kinesthetic Mind Reader: A Method to Identify Image Schemas in Natural Language. In *Advances in Cognitive Systems*, Vol. 5. Cognitive Systems Foundation, Paper–9.

[10] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whatâĂŹs next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[11] Mark Johnson. 2013. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press.

[12] Adam Kendon. 1980. Gesture and speech: two aspects of the process of utterance. *Nonverbal Communication and Language* (1980), 207–227.

[13] A. Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

[14] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International Workshop on Intelligent Virtual Agents*. Springer, 205–217.

[15] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[16] George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The journal of Philosophy* 77, 8 (1980), 453–486.

[17] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*. Springer Berlin Heidelberg, 243–255.

[18] Jina Lee and Stacy Marsella. 2009. Learning a model of speaker head nods using gesture corpora. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 289–296.

[19] Thomas Leonard and Fred Cummins. 2011. The temporal relation between beat gestures and speech. *Language and Cognitive Processes* 26, 10 (2011), 1457–1471.

[20] Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*. ACM, 24–26.

[21] Margot Lhommet and Stacy Marsella. 2016. From embodied metaphors to metaphoric gestures. In *Proceedings of Annual Meeting of the Cognitive Science Society*.

[22] Daniel P Loehr. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3, 1 (2012), 71–89.

[23] Andy Lücking, Alexander Mehler, Désirée Walther, Marcel Mauri, and Dennis Kurfürst. 2016. Finding Recurrent Features of Image Schema Gestures: the FIGURE corpus. In *Proceedingsofthe10thInternational Conference on Language Resources and Evaluation. LREC.*

[24] David McNeill. 1985. So you think gestures are nonverbal? *Psychological review* 92, 3 (1985), 350.

[25] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press. http://books.google.fr/books?id=3ZZAfNumLvwC

[26] Alexander Mehler, Andy Lücking, and Giuseppe Abrami. 2015. WikiNect: image schemata as a basis of gestural writing for kinetic museum wikis. *Universal Access in the Information Society* 14, 3 (01 Aug 2015), 333–349. https://doi.org/10.1007/s10209-014-0386-8

[27] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[28] Irene Mittelberg. 2008. Peircean semiotics meets conceptual metaphor: Iconic modes in gestural representations of grammar. *Metaphor and gesture* 3 (2008), 115–154.

[29] Thomas Morton, Joern Kottmann, Jason Baldridge, and Gann Bierner. 2005. OpenNLP: A Java-based NLP Toolkit. (2005). http://opennlp.sourceforge.net.

[30] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

[31] Florian Pecune, Angelo Cafaro, Mathieu Chollet, Pierre Philippe, and Catherine Pelachaud. 2014. Suggestions for extending saiba with the vib platform. In *Workshop Architectures and Standards for IVAs, International Conference of Intelligent Virtual Agents.* 16–20.

[32] Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* 34, 2 (2008), 257–287.

[33] Daniel C Richardson, Michael J Spivey, Shimon Edelman, and AD Naples. 2001. Language is spatial: Experimental evidence for image schemas of concrete and abstract verbs. In *Proceedings of the twenty-third annual meeting of the cognitive science society.* Erlbaum, 873–878.

[34] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.* Association for Computational Linguistics, 173–180.

[35] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014), 209–232.

[36] Yuyu Xu, Catherine Pelachaud, and Stacy Marsella. 2014. Compound gesture generation: a model based on ideational units. In *International Conference on Intelligent Virtual Agents.* Springer, 477–491.