Grounding Concepts as Emerging Clusters in Multiple Conceptual Spaces

Roberto Pirrone and Antonio Chella

Dipartimento dell'Innovazione Industriale e Digitale (DIID)
Università degli Studi di Palermo
{roberto.pirrone, antonio.chella}@unipa.it
http://www.unipa.it/dipartimenti/diid

Abstract. A novel framework for symbol grounding in artificial agents is presented, which relies on the key idea that concepts "emerge" implicitly at the perceptual level as *clusters* of points with similar features forming homogeneous regions in multiple *perceptual Conceptual Spaces* (pCS). Such spaces describe percepts such as color, texture, shape, and position that in turn are the properties of the objects populating the agent's environment. Objects are represented in a suitable *object Conceptual Space* where all their features are composed together again using clustering in pCSs. Symbols will be learned from such a tensor space. A detailed description of both the framework and its theoretical foundations are reported and discussed in this work.

Keywords: Symbol Grounding · Conceptual Spaces · Clustering · Tensors

1 Motivation and Theoretical Background

Symbol grounding [9] is a fundamental research topic in both Cognitive Systems and Artificial Consciousness.In recent years, such a topic received great attention in the field of Human Robot Interaction (HRI) and Social Robotics, due to the development of a huge number of robotic architectures aimed at collaborating with humans [10]. Indeed, artificial agents engaged in highly interactive tasks do need a grounded i.e. "internal" representation of their percepts, despite of their embodiment, and the field of application [2,3]. Eventually, we can state that the only way an artificial agent can have a private and subjective experience of the world that is a *quale*, it is through a set of quantitative measurements from its sensors even if there is a heated debate in philosophy and cognitive sciences about the properties and even the existence of qualia. Moving from the previous considerations, we present here a novel framework for symbol grounding based on the theory of Conceptual Spaces [5] where clustering is used as the main perceptual process to devise homogeneous regions w.r.t. different sets of low level visual features like color, texture, shape, and position in the environment. Such regions are mapped as prototypical points in multiple *perceptual Conceptual Spaces* (pCS) describing each property with the same set of features.

Again, clustering in pCSs devises concepts as dense sets of points: cluster centers can be devised as the concept prototype, while the convex hull of each cluster represents the boundary of each concept. Objects in the environment are represented as tensors in a higher level *object Conceptual Space* (oCS) where the Kronecker product is used to represent the relation between an object and its visual features as well as the way in which such features are related with each other when forming the object's percept. Our framework is inherently motivated by the need of building a robot that is able to interact seamlessly with humans when performing a collaborative task. One of the core elements of consciousness is language so it is crucial to provide an artificial agent with the ability of grounding both the lexicon and the meaning related to the objects in the environment.

Many theories address language as one of the main traits of consciousness. In turn, language has to be *grounded* to the phenomenal experience to provide a meaning to words. In the Higher Order Syntactic Thought (HOST) theory of consciousness [14], the conscious thoughts (in linguistic form) about thoughts on the world take place only if "first order" linguistic processing manipulates grounded symbols. This way, one feels that he/she is reflecting upon something in the world. Luc Steels [15] supports the idea that consciousness is strictly related to language: language re-entrance is a semiotic circle where the speaker is also the hearer when he listens to a "inner voice". Steels proposes a symbol grounding procedure [16] based on playing "language games" where two embodied artificial agents (i.e. two robots) generate the symbols for the topics they are talking about. Conceptual spaces are a widely accepted formalism to represent conscious qualia, and grounding them to the perception [4]. Nevertheless, while a CS describes well sensory perceptions like color or shape, claiming that the experience of a bird can be modeled using a "birds CS" defined as a subspace in $\mathbb{R}^n$ where many heterogeneous but interrelated perceptions are simply juxtaposed, is a controversial position. Augello et al. [1] claim that subjective experiences of the world (i.e. qualia) are inherently non linear, so they can not be suitably represented in linear vector spaces as CS are. In this work we support the position that complex perceptions composed by different sensory features have to be expressed "composing" the corresponding CSs, and we model symbol grounding as a learning process taking place in a tensor space generated by the Kronecker product of the feature vector spaces.

2 The Proposed Framework

The steps of the proposed symbol grounding procedure are reported and detailed in the following.

Perception : extraction of multiple low level features from the visual input, and clustering of emergent homogeneous regions.

Mapping in *p*CS : cluster centers for each region are mapped as points in multiple CSs where the dimensions are the same as in the features.

Building the object tensor in *o*CS : the vectors representing the object's properties in every pCS are composed in a single tensor through the Kronecker product; multi-part objects are the mean of the tensors representing each part.

Learning symbols : a learning machine is used in this respect to bind tensors to their symbolic representation in a structured knowledge base where the relations between objects and their perceptual properties are made explicit.

Perception While grounding symbols to perception, an artificial agent may behave either in an *instructed* or in an *exploratory* way; in the first case a human points at a ROI, while referring to a symbolic description, and the agent performs symbol grounding explicitly as part of its interactive task. On the other hand, the agent may focus its attention to something new, thus trying to provide a meaning for such a percept. Often in this case, the agent already knows the symbols for a part of the perception (as an example *"a red thing"*). In both cases, the agent does not perform explicit pattern recognition processes, and we can think of its perception in a Gestalt perspective where pre-attentive grouping of low level features occurs, while visual attention intervenes just to constrain the search region in the visual array. We modeled such processes by extracting different low level features like color, shape, texture, and position from the visual input, and clustering them using density based approaches [13] because in general such features are vectors in low dimensional spaces where the notion of distance is well defined. Several clusters emerge in each feature space, and each pixel can be labeled w.r.t. the cluster it belongs to; the image will be then segmented in regions whose pixels exhibit the same set of labels. Gaussian or fuzzy smoothing can be used to avoid little holes and removing outliers.

***p*CS** Perceptual CSs are defined as a set of CSs where each perceptual property is defined by a series of dimensions that are the same features we extracted from the visual input; with this choice we want to address all the considerations made by Gardenfors about the best choice of the "quality dimensions" in a CS to describe sensory input. Features that are strictly linked to psychology of perception will be used, so color may be described using a perceptive color space such as La^*b^*, the principal curvatures (k_1, k_2) can beused to describe shape locally, while a suitable texture description could be obtained using Malik's textons [11]. Each region is mapped onto the set of pCSs as the center of the corresponding cluster where its pixels fall in; in this way clustering maintains its effectiveness to determine similar points in a pCS. Such points can now be regarded as both examples and counterexamples of some property value. Gardenfors partitions a CS using the Voronoi tessellation to create convex regions representing concepts, starting from some prototypes. New incoming examples and counterexamples modify the boundaries of such a tessellation; the Region Connection Calculus (RCC) endowed with a suitable definition of the "crisp relation" is used for reasoning about CSs [7]. In our framework, concepts are simply clusters in a pCS: reasoning using clusters is a more flexible approach than building a new Voronoi tessellation each time a new example is added in the CS. A concept has its prototype in the cluster's center, while it is bounded by the cluster's convex hull. Just a single new example falling away from the boundaries of the already known clusters is sufficient to generate a new one, while similar examples will fall close to each other, and the corresponding convex hull will change accordingly. We do not need particular reasoning primitives apart from the notion of a sample falling inside or outside a cluster, while the closeness of new examples to other clusters allows for describing their meaning in terms of previous knowledge i.e. *"orange is like a bright yellow-red"*.

***o*CS** Perceiving an object as the composition of its properties can not be modeled using a simple vector representation of all such properties joint together; a first objection is the "curse of dimensionality" in such a vector space, as the notion of *distance* looses significance as the size of such a space increases. As a consequence the main theoretical foundation of CS doesn't hold that is we can no more devise concepts as convex sets in the CS because we can't measure distances properly. Moreover, perceptual features influence each other: a color is perceived as darker or brighter as its surface orientation tilts towards to or away from the light source due to its curvature, and the same holds for the relation between texture and color. We want to address all the previous issues using tensors for representing objects in the environment. If we assume for the sake of simplicity that a generic object is described through a color $c \in C$, a shape $s \in S$, and a texture $t \in T$, the object itself will be represented as:

$$\mathcal{O} = c \otimes s \otimes t$$

where C, S, and T are the pCS for colour, shape, and texture respectively, while $\otimes$ represents the Kronecker product in its usual definition. Given that $\mathbb{M}_{h,k}$ refers to the space of the matrices of order (h, k), and $v \in \mathbb{M}_{m,1},\ w \in \mathbb{M}_{n,1}$, their Kronecker product $v \otimes w \in \mathbb{M}_{m,n}$ is a matrix whose rows are in the form $v_i \cdot w^T,\ i \in 1, \ldots, n$. The previous definition can be extended along multiple products. As it is well known, tensor spaces defined in this way have a vector space structure so an inner product along with an induced norm can be defined, and it is possible to think about an *object Conceptual Space* where the convexity requirement already holds. The $\otimes$ product expresses a way in which properties "modulate" each other. From a computational point of view, even if we will possibly compute distances in the oCS to judge similarity between a couple of objects, we do not need to perform any explicit clustering procedure in such a space because it is encoded by the learning procedure that actually binds symbols to objects. The Kronecker product is a way to express the mutual influence of the perceptual properties in forming the subjective experience of the object. Tensors account also for objects defined by a subset of properties i.e. the symbol *"ball"* will correspond to tensors where the shape dimension is in some sense prevalent, because all such tensors will have their s vectors falling near the same prototype in the shape pCS corresponding to the symbol *"round"*. Eventually, also single properties may have their tensor representation in the oCS thus allowing for the same process being used to learn both property and object symbols. In real cases, an object falling into the ROI investigated by the agent will be segmented in multiple regions i.e. a cup will result in two tensors accounting for both the convex and the concave side of the cup itself, while they will have the same color and texture. In such cases, the tensor resulting from the mean between the parts will be taken into consideration.

Learning symbols in a Structured KB As the artificial agent can be either instructed to learn symbols or it can discovery new objects in the environment, both supervised and unsupervised learning should take place to bind symbols with their tensor representation. There are two main learning schemes in our view that best suit to implement such step in the grounding procedure: Support Vector Machines (SVM) using RBF kernel, and Convolutional Neural Networks (CNN) [8]. SVM are very good classifiers also with several classes; new classifiers can be instantiated when new clusters emerge in the pCSs so that current classifier ensemble starts rejecting examples as outliers in the oCS. Moreover, RBF kernels proved to be very good for learning categories described as a prototype vector along with a bounded region in the feature space [16]. On the other hand, CNN are learning machines devoted to tensors; they are trained only in a supervised way, but the output layer can be arranged to accommodate for learning a limited number of unknown classes. An OWL ontology will be used to store the symbolic knowledge of the agent; here the relation between the objects and their perceptual features will be represented explicitly. It is worth noting that the agent learns frames whose structure is of the form $Object : \langle hascolour, hasshape, hastexture \rangle$. Such structures have been widely investigated in Computational Linguistics to enable verbalization trough the use of Construction Grammars (CxG). Construction poles are a well suited structure to host the bind between symbols as the meaning of a "surface form" made by a numerical embedding representing perception. Some of the authors already proposed an OWL axiomatization process [12] for producing constructions in the Steels' Fluid Construction Grammar (FCG) [17]. The similarity (i.e. closeness) between the embeddings can be used suitably to guide the unify-and-merge procedure, which selects constructions in the FCG production step.

3 Conclusions and Future Works

We are currently developing our framework on an Aldebaran Pepper robotic platform using the Python programming language and ROS. Here we report some final considerations about our proposal. Clustering in CSs is an effective technique for manipulating concepts: apart from binding symbols, the geometric relations between clusters in the pCS allow for learning also imprecise expressions like *"a sort of"* or *"similar to"*. Also spatial language can be accounted for, when using a pCS for expressing position. In a HRI scenario, there is no spontaneous lexicon formation. Symbols are already in the mind of the instructor, while new symbols can acquire their meaning through similarity with the properties of other symbols. Our learning through interaction scheme is compliant to the notion of "Meeting of Minds" proposed by Gardenfors [6]: our framework allows the agent to reach a *fixpoint* with the instructor in a scenario where the attention is pointed at something new; the instructor will name the new object, but in general its features will fall into known properties, and the agent will form a tensor representation of the object that is partially similar to something already known. In turn, the meaning will be grounded to known objects with some degree of uncertainty i.e. *"an egg is a sort of white/brown smooth ball"*.

Tensors are a suitable representation for objects in a CS that maintains its algebraic properties, and expresses the influence between quality dimensions i.e. colour, shape, and texture considered as a whole, while avoiding the construction of a high dimensional feature space by means of the mere Cartesian product. Future work will be aimed at deepening the theoretical aspects related to tensor representation of objects, properties, and relations emerging from the *p*CSs.

References

1. Augello, A., Gaglio, S., Oliveri, G., Pilato, G., et al.: Acting on conceptual spaces in cognitive agents. In: AIC@ AI* IA. pp. 25–32 (2013)
2. Bentivoglio, C., Bonura, D., Cannella, V., Carletti, S., Pipitone, A., Pirrone, R., Rossi, P., Russo, G.: Intelligent agents supporting user interactions within self regulated learning processes. Journal of E-Learning and Knowledge Society **6**(2), 27–36 (2010)
3. Chella, A., Dindo, H., Matraxia, F., Pirrone, R.: Real-time visual grasp synthesis using genetic algorithms and neural networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **4733 LNAI**, 567–578 (2007)
4. Chella, A., Coradeschi, S., Frixione, M., Saffiotti, A.: Perceptual anchoring via conceptual spaces. In: proceedings of the AAAI-04 workshop on anchoring symbols to sensor data. pp. 40–45 (2004)
5. Gärdenfors, P.: Conceptual spaces: The geometry of thought. MIT press (2004)
6. Gärdenfors, P.: The geometry of meaning: Semantics based on conceptual spaces. MIT Press (2014)
7. Gärdenfors, P., Williams, M.A.: Reasoning about categories in conceptual spaces. In: IJCAI. pp. 385–392 (2001)
8. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016)
9. Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1-3), 335–346 (1990)
10. Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R.: Artificial cognition for social human–robot interaction: An implementation. Artificial Intelligence **247**, 45–69 (2017)
11. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International journal of computer vision **43**(1), 29–44 (2001)
12. Pipitone, A., Pirrone, R.: Cognitive linguistics as the underlying framework for semantic annotation. In: 2012 IEEE Sixth International Conference on Semantic Computing. pp. 52–59. IEEE (2012)
13. Pirrone, R., Cannella, V., Monteleone, S., Giordano, G.: Linear density-based clustering with a discrete density model. ArXiv e-print arXiv:1807.08158 (Jul 2018), https://arxiv.org/abs/1807.08158
14. Rolls, E.T.: Cerebral cortex: principles of operation. Oxford University Press (2016)
15. Steels, L.: Language re-entrance and the inner voice. Journal of Consciousness Studies **10**(4-5), 173–185 (2003)
16. Steels, L.: The symbol grounding problem has been solved. so whats next. Symbols and embodiment: Debates on meaning and cognition pp. 223–244 (2008)
17. Steels, L., De Beule, J.: Unify and merge in fluid construction grammar. In: Symbol grounding and beyond, pp. 197–223. Springer (2006)