Viewing Vision-Language Integration as a Double-Grounding Case

Katerina Pastra
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, U.K.
katerina@dcs.shef.ac.uk

Abstract

While vision-language integration is important for a wide range of Artificial Intelligence (AI) prototypes and applications, the notion of integration has not been established within a theoretical framework that would allow for more thorough research on the issue. In this paper, we attempt to explore the reasons that dictate this content integration by bringing together Searle's theory of intentionality, the symbol grounding problem, as well as arguments regarding the nature of images and language developed within different AI subfields. In doing so, the Double-Grounding theory emerges which provides an explanatory theoretical definition for vision-language integration. In correlating the need for vision-language integration with inherent characteristics of the integrated media and in associating this need with an agent's intentionality and intelligence, the work presented in this paper aims at providing a theoretically established —and therefore solid— common ground for currently isolated and scattered multimedia integration research in AI subfields.

Introduction

Multimodal human to human communication involves constant integration of visual and linguistic information; while not essential[1], this integration renders communication more effective. In Artificial Intelligence (AI), the ability to perform vision-language integration has been considered an important feature of machines for exemplifying —to an extent— *human-level intelligence* (Waltz 1981). Vision-language integration prototypes span diverse research areas ranging from multimodal retrieval and intelligent multimedia presentation systems to robotics; still, multimodal integration challenges are not systematically dealt with. More importantly, it is on human intervention that vision-language integration prototypes rely for *abstracting* visual information and for *associating* visual and linguistic information (Pastra & Wilks 2004). The field lacks a unifying theoretical framework for computational vision-language integration, one which will actually bring multimedia-related areas closer and will provide the common ground needed for dealing with computational integration challenges systematically.

Bringing together Searle's theory of intentionality and speech acts (Searle 1969; 1983), the *symbol grounding problem* (Harnad 1990) as well as arguments regarding the nature of images and language developed within different AI subfields, we will attempt to address this issue. All these sources —that only partially and indirectly address vision-language integration— are synthesised in formulating what we have called the *double-grounding theory*. The latter argues that vision-language integration is a case of *double-grounding* in multimodal situations, where visual representations *ground* language in physical aspects of the world and linguistic ones, in their turn, *ground* vision in mental aspects of the world. Double-grounding is needed for compensating for inherent features the integrated media lack and it actually endows an agent with the intrinsic intentionality required for exemplifying intelligence in vision-language multimodal situations.

The notion of integration

While used extensively in prototypes which combine vision and language, vision-language content integration was only recently defined within AI: it is the *process of establishing associations between visual and corresponding linguistic representations* (Pastra & Wilks 2004). This *intensional* definition was complimented by an *extensional* definition of the term, an enumeration and description of the processes the term integration applies to, which resulted from the classification of a wide range of vision-language integration prototypes. The criterion for this classification was the *integration purpose* served by the prototypes; four general classes were indicated, each corresponding to a distinct integration procedure (Pastra & Wilks 2004):

1. *performance enhancement*: the analysis of one medium for guiding or enriching the analysis of another
2. *medium translation*: the translation of a source medium into the target medium
3. *multimedia generation*: going from knowledge representation to the generation of multimodal documents, and
4. *situated dialogue*: the analysis of multimodal input and the generation of a reaction/multimodal answer.

[1] Cf. cases of blind people, who can do without vision-language integration and rely on other senses for perceiving reality.

However, while this descriptive definition sheds some light on the issue of integration, it does not actually provide any justification of *why* computational vision-language integration is needed. An indirect justification found within image-language interaction corpus-based studies is that interaction (which is served through integration) serves communicative goals, it establishes coherence in multimodal communication (André & Rist 1993). However, what does coherence in vision-language multimodal situations actually mean, what is it that each of the modalities involved lacks, that justifies the need for integration? In other words, what does language need from vision and what does it offer to vision in its turn? We believe that looking at the inherent characteristics of the integrated modalities might hold a more elaborate answer to this question.

A parallel exploration of Vision and Language

In viewing vision and language as *processes* that make use of modular representation systems and transformation algorithms, Marr's influential theory of vision (Marr 1982) and Jackendoff's structural computational model of natural language (Jackendoff 1983), provide the common ground required for a parallel exploration of the nature of these faculties. Within these theories, images and language are described as systems of iconic and symbolic representations respectively. They meet, therefore, through the notion of representation; they take a form and have a meaning. Both the visual and the linguistic representations stand for a reference object which may be a mental (conceptual) entity or/and a physical one. Leaving analogies and other differences aside, differences in the *nature* of the reference object they can stand for emerge from such a parallel exploration.

In particular, the reference of a visual representation may be a *physical object or scene* whose mere existence triggers the vision generation process (visual perception). The goal to infer (or even go beyond) this reference object (literal meaning) drives the recognition process (image understanding). When the vision process is triggered by a *mental representation* (concept), then one crosses over into visualisation. In this case, images actually transform the status of their reference object from mental into physical, so that one cannot tell for sure —by looking at the image alone— whether the object depicted is a physically existing entity or whether it stands as a visual example of a general class of entities. Figure 1 illustrates this view of vision.

On the other hand, the input to the language generation process is a mental representation (conceptualisation) whose levels of abstraction or genericness can be perfectly conveyed by the linguistic representations. While language retains the non-physical status of its input, it has no access to the physical world. It can only indirectly refer to specific instances of physical objects (through e.g. proper names, deictic references etc.), relying on the use of perceptual processes for a direct access to these objects. Figure 2 illustrates this view of language.

Inherent media characteristics

Language has been characterised as *general* (Bernsen 1995): it goes beyond things that can be depicted to abstract ideas and reasoning. Furthermore, it has subtle mechanisms for controlling attendance to details or focus (Minsky 1986; Bernsen 1995). On the contrary, images are *specific*, *i.e.*, they are exhaustive representations of what they stand for, (providing, for example, direct measurements of the entities depicted). In addition, they have no inherent means for efficiently indicating the focus (Bernsen 1995) and salience of what they depict. In fact, images which result from visual perception representations have no mechanisms of their own for indicating such aspects[2]; this characteristic of vision has led researchers to the inclusion of indexical-deictic mechanisms in computer vision theories (Ballard *et al.* 1997). This way, the generation of visual representations of all the positions and properties of all the objects depicted in a visual scene is avoided and instead, indication of the focus and salience intended is achieved[3]. In its turn, images which result from visualisation have some limited mechanisms for indicating focus and salience (e.g. through highlighting, use of pointing arrows etc.) (André & Rist 1993; Bernsen 1995); still, a very accurate and efficient metalanguage for expressing such information is missing[4].

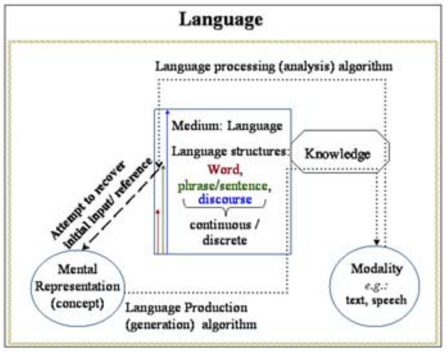
Language
Language processing (analysis) algorithm
Medium: Language
Language structures:
Word,
phrase/sentence,
discourse
continuous / discrete
Knowledge
Attempt to recover initial input/ reference
Mental Representation (concept)
Language Production (generation) algorithm
Modality
e.g.:
text, speech

Figure 1: The vision process: its media and modalities

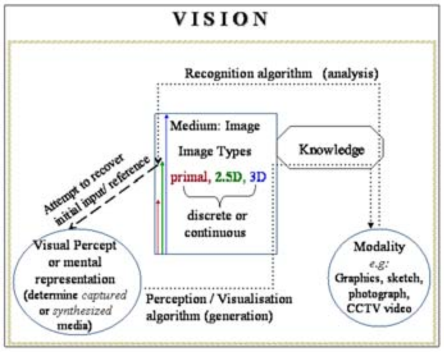
VISION
Recognition algorithm (analysis)
Medium: Image
Image Types
primal, 2.5D, 3D
discrete or continuous
Knowledge
Attempt to recover initial input/ reference
Visual Percept or mental representation (determine captured or synthesized media)
Perception / Visualisation algorithm (generation)
Modality
e.g:
Graphics, sketch, photograph, CCTV video

Figure 2: Language as a process: its media and modalities

[2] It is fixation and its neural analogue (attention) which indicate focus in visual perception (Ballard *et al.* 1997).

[3] Cf. also a cognitive theory of vision in (Pylyshyn 2001).

[4] In mono-modal situations, the context of images gives clues for inferring such information; still, no *direct*, explicit, indication

Furthermore, both visual perception and visualisation representations are very specific in nature, they always depict individual tokens, even when the intention of the agent who generates these representations may be to express a representative of a class of objects, a representative of a specific level of abstraction (e.g. a dinner table standing as an example of "domestic furniture"). It has been argued that token:type distinctions cannot actually be indicated by visual representations (Jackendoff 1987); though this view has been recently challenged (Barsalou *et al.* 2003), placing something depicted to the appropriate level in a rich conceptual system has not been proven feasible with visual mechanisms alone. In both cases of focus/salience and token:type distinctions, visual representations require ways of precisely indicating the intentions of the agent who generates the visual representations. A fine-grained meta-linguistic function —inherently lacking in visual representations— is required to indicate intentional aspects of what is depicted.

Therefore, while both vision and language can express mental entities (concepts), vision has in this case no way to retain the mental/abstract character of what it depicts, which is due to its specificity and lack of effective focus/salience indication means. It lacks *control* over the focus and level of abstraction of what it expresses. In its turn, language is never triggered by the existence of a physical object as vision is; it can only indirectly refer to something that exists in the physical world, due to the always conceptual, mental character of its input. There seems to be an *intentionality* issue here: some aspects of intentionality cannot be expressed by visual modalities effectively and require the use of highly expressive indicators; on the other hand, intentional content referring to specific physical entities relies on perceptual (visual) modalities when expressed through language. We turn to Searle's intentionality theory for exploring this further.

Searle on visual and linguistic representations

A close look into Searle's theory of speech acts (Searle 1969) reveals an intimate relation between language and perception —and therefore visual perception— in many cases. In particular, in the case of *assertives*, an *"independently existing reality"* is supposed to be matched[5], which actually presupposes that this reality is somehow accessed/known by the agent who makes use of the assertive. Access to this reality enables the agent to check whether the conditions of satisfaction of what (s)he expresses through the assertive are met or not. The example case *par excellence* is the one of the indexicals (e.g. demonstratives), which require —according to Searle— visual perception of what is referred to[6]. While the other types of speech acts (excluding expressives) *impose* changes to the world, rather than describing its state of affairs, they are —by nature— causally unconnected to the physical world too[7]. Given the fact that an agent needs to know whether the conditions of satisfaction of what (s)he expresses are met or not[8], a direct access to the world they act upon is required in their case too. Contrary to linguistic representations which have no intrinsic causal relation with the world, visual representations are caused by the state of affairs they represent and can, indeed, give a "direct access" to this physical world that invokes them[9].

So, Searle's theory indicates not only a "lack of direct access to the physical world" that linguistic representations have, but also an *inherent need* on the part of linguistic representations, for such an access. Visual perception representations are shown to be able to assist language in compensating for this. However, Searle does not refer to visualisation at all. Do visual representations that stand for a conceptual reference object (and have therefore not been invoked by a physical entity/scene) provide such an access to language too? Although Searle does not address this issue, some extensions to his theory actually do. Drawing an analogy to speech acts, the notion of *pictorial acts* (Kjorup 1978; Maybury 1993) has been introduced for indicating the double-level of intentionality expressed through visualisation representations. So, Searle's theory refers to both linguistic *speech acts* (Searle 1969; 1976) and visual perception representations and some of its extensions have also addressed the visualisation issue. This is one of the reasons we chose this specific philosophical theory of intentionality. The second reason is that Searle has criticised AI symbol systems for being non-intentional and has linked, therefore, his intentionality theory to AI research, AI also being the perspective from which we explore integration in this paper.

The symbol grounding problem

In Searle's well-known *Chinese-room argument*[10], it is being argued that AI symbol systems lack the *intentionality* required for being artificial minds (Searle 1980). Searle argued that it is *intrinsic intentionality* —as opposed to human-derived one— that is required on the part of the system for the latter to demonstrate human-level intelligence.

The main point was that AI systems consist of programs with formal symbols (encoding), whose meaning is obvious only to humans and not the systems themselves; the artificial mind does not impose intentionality on them, it is rather the human developer who does so. Though computer programs consist of artificial language symbols that encode e.g. linguistic or visual representations (natural language words, visual objects etc.), the meaning of this encoding and the one of the representations themselves resides in humans' mind. The system can manipulate the encoding syntactically, but it

of this information through visual means can be provided.

[5] Cf. pp. 173 in (Searle 1983).

[6] Cf. pp. 218-230 in (Searle 1983).

[7] Cf. pp. 173 in (Searle 1983).

[8] This is required —according to Searle— for attributing real intentionality to an agent, cf. pp. 177 in (Searle 1983).

[9] Cf. pp. 46 in (Searle 1983).

[10] The argument that a human getting input in Chinese and following strict instructions on the manipulation of such symbolic representations can generate native-quality Chinese output with no actual knowledge and understanding of Chinese; Searle argues that similarly, artificial systems able to analyse natural language and generate linguistic output exemplify no real understanding, because they lack intentionality (Searle 1980).

has no idea what the encoding stands for (semantic interpretation) or what the representations encoded actually mean (what a specific word refers to, or what a visual object looks like).

Therefore, mere instantiation of a program, even if it resulted in perfectly structured and coherent output, would not reveal any intentionality on the part of the system (Searle 1980). In other words, Searle accused AI systems, of failing to express intentionality in their simulation of mental states (such as understanding), due to their inability to impose —on their own— meaning on the symbols they used. However, what is required from a mind in order to be ascribed intrinsic intentionality and, therefore, human-level intelligence?

The prevailing solution suggested for this problem was the bottom-up grounding of symbolic representations to perceptual ones that provide a causal link to the world; knowing this link is what Searle has required for attributing intentionality to agents. In answering to Searle's criticisms of AI systems, the need for *grounding* linguistic representations to perceptual ones has been advocated within the so-called *symbol grounding problem* (Harnad 1990). The latter is interested in computational techniques for grounding the meaning of symbolic representations (such as the linguistic ones) to non-symbolic ones. Hybrid systems making use of both symbolic and connectionist techniques for learning the link between linguistic symbols and their physical (non-symbolic) visual references have been suggested (Harnad 1990; Cangelosi, Greco, & Harnad 2000), as well as pure connectionist techniques (Jackson & Sharkey 1996). Recent cutting-edge research on machine learning algorithms for teaching an agent how to name visual objects is an empirical follow-up of this symbol grounding related literature (Kaplan 2000; Roy & Pentland 2000; Roy 2002; Bredeche *et al.* 2003; Vogt 2003).

Symbol grounding has been addressed in AI from positions of wider or narrower scope (cf. for example work by Ziemke 1997 and Vogt 2002). Parts of the symbol grounding problem have even emerged as AI research sub-fields on their own right; *anchoring*, for example, has been recently described as a new challenge in robotics (Coradeschi & Saffiotti 2003)[11]. In all its manifestations though, the symbol grounding debate has focused on lower or higher representation levels in AI systems.

While Searle's arguments were targeted originally against AI symbol system programs as realised through formal symbols/encoding/artificial language symbols, the illustration of his claims through the Chinese-room argument shifted the debate to natural language symbolic representations in general[12]. Subsequently, natural language understanding as an intentional state exemplified through natural language generation has monopolised the debate as the grounding case *par excellence* in Searle's arguments. All symbol grounding computational solutions suggested followed the same direction and referred, therefore, to ways of developing causal relations between *natural language* and corresponding *visual percepts* through bottom-up grounding.

It seems that symbol grounding —as treated in the literature so far— is characterised by *a predominant bottom-up direction*, *i.e.*, it focuses on the grounding of natural language symbolic representations in visual perception ones, a grounding that provides linguistic representations with the direct access to the physical world they inherently lack. What about visualisation representations though? Do they ground linguistic representations too? More importantly, could it be that the nature of the grounding process is bi-directional, rather than one-directional?

The Double-Grounding Theory

In this section, we attempt to address the above mentioned questions, which leads us to the formulation of the *double-grounding theory*.

Visual grounding of language in the physical world

Figure 3 illustrates symbol grounding as traditionally advocated in AI research. The figure presents the physical world and the mental world; no Cartesian dichotomy between the two is implied though. We follow Searle's perception of the world as consisting of physically existing entities (the outside reality) and mental entities (an agent's inside reality) (Searle 1983). Visual and linguistic representations stand in between, since though they are mental, they can also be physically realised (cf. notion of modalities). Visual perception representations have a direct access to the physical world, since their generation is triggered by physical entities. In their turn, linguistic representations are triggered by mental entities (and therefore have a direct access to the mental world). When associated to visual perception representations, they get grounded in the physical world, *i.e.*, they acquire a direct access to physical referents. Can the same access be provided to language by visualisation representations though?

As mentioned earlier, some researchers have drawn a parallel to speech acts for describing intentionality in visualisation (Kjorup 1978; Maybury 1993). Similarly to speech acts, *pictorial acts* are not invoked causally by physical entities. This is an important similarity they share with linguistic representations. The similarity points —logically— to the fact that visualisation representations lack and need access to the physical world, exactly as linguistic representations do. However, a basic difference in the nature of these two affects this need greatly: visualisation representations are not always (and merely) symbolic, as language is; they are iconic in nature and therefore, resemble —more or less— the physical world. Some of them are highly realistic/iconic, others are more symbolic[13].

[11] The *anchoring problem* involves artificial agents establishing and maintaining correspondences between the ID/names of individual physical objects and their corresponding visual percepts.

[12] This is justified by the fact that AI and natural language symbolic representations have been found to be alike; cf. for example an extensive debate on this in (Nirenburg & Wilks 2001).

[13] Even highly symbolic images e.g. a pie chart presenting the results of an election-related exit poll, has iconicity; it has been chosen from among other information graphics types for its iconic properties.

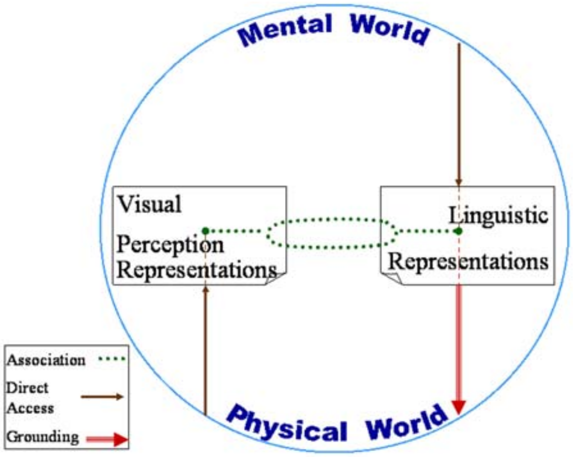

Figure 3: Visual perception representations ground language in the physical world

In the first case, physical resemblance to the world is a safe criterion for associating the images to visual perception representations of the physical world, while in the case of highly symbolic images this association is conventional (as in the case of language). Visual perception representations can ground both symbolic and realistic visualisation representations. The difference in these two groundings lying in the types of associations needed. Symbolic images (e.g. a rectangle standing for a vehicle in a route diagram) require a learned/imposed association, while realistic images (e.g. a portrait) map directly to corresponding visual perceptions.

While grounded visualisation representations can ground —in their turn— language too, their nature is such that even when ungrounded themselves, they attempt to ground linguistic representations in the physical world or —to be more precise— to the **intended** physical world. This is so, because even when realistic visualisations express something mental, they give a believable physical appearance to it, so that the viewer cannot tell —by looking at the image alone— whether the image resulted from visual perception or visualisation. In this case, visualisation *transforms* a mental entity into a physically looking one; if used to ground linguistic/symbolic representations, it provides access to (an AS IF) THE physical world which is intended (by the agent who generates the visualisation) to be thought of as THE physical world, in a specific communication context. For example, photo-realistic three-dimensional graphics depicting what the interior of a house will look like, when its renovation is complete, provide the corresponding linguistic description of the house with a direct access to (an AS IF) THE physical world; the latter, is not the visually perceptible physical world, since it depicts a future state of affairs, but within the specific communication context, it is supposed to be THE physical world of interest, in which the linguistic discourse needs to be grounded. It is in this sense, that realistic visualisations can provide linguistic representations with a direct access to the physical world, no matter whether themselves grounded or not. Figure 4 illustrates the two cases of grounding linguistic representations through grounded or/and ungrounded realistic visualisations.

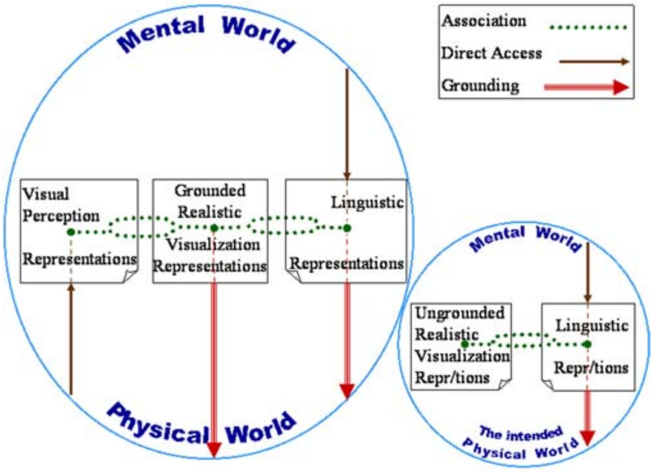

Figure 4: Realistic visualisations ground language in the physical world

Based on all the above, we argue that visual representations in general —*i.e.*, visual perception, grounded symbolic visualisations and grounded or ungrounded realistic visualisations— provide linguistic representations with a direct access to the physical world, *i.e.*, they ground language in the physical world. This takes place through association and it is an important ability for artificial systems, if the systems are to be ascribed intrinsic intentionality.

Linguistic grounding of visual representations in the mental world

The very same characteristic of visual representations which provides a direct access to the physical world, *i.e.*, being always specific, has further implications though, which are related to their ability to *indicate* intentional aspects of the content they express. While not referring to the ability of visual representations to indicate various distinctions, Searle does stress the importance of expressing such distinctions. According to Searle's theory, representations have —by definition— an *intentional aspect* that denotes the relevance of what is expressed to a situation; its role is significant in that it points to the appropriate background capacities and knowledge needed for determining the conditions of satisfaction of the intentional mental states represented[14]. The Peircian semiotic theory, refers to the existence of such an aspect too, using the term *ground*; the "ground" of a sign being the idea in respect of which a representation takes place (Peirce 1960). This aspect could denote anything that is situation-specific, such as the time and place of the communication event, the domain/perspective from which something is approached, the register, focus/salience of what is referred to etc.

[14] Cf. pp.13 and chapter 5 in (Searle 1983).

Language is extremely efficient in making all such aspects clear (cf. theme/topic theories in linguistics, controlled sub-languages for expressing domain of interest, phenomena of code-switching, idiolects etc.). However, can visual representations indicate and express such intentional aspects clearly and efficiently?

Earlier, we argued that visual representations *lack inherent ways of indicating focus/salience and type:token distinctions*. Visual perception representations have no direct access to the mental world, since their generation is triggered only by physically existing entities; in representing the mental state of visual experience though, they need to express intentional aspects of their referents too, a task for which they have no visual means. In their turn, visualisations have a direct access to the mental world, since they are triggered by mental entities, they have some mechanisms of their own for indicating focus and salience (e.g. highlighting, zooming, pointing arrows etc.), but they have no efficient mechanisms for such a task and no mechanisms at all for indicating type:token distinctions and different levels of abstraction in general. Since visual representations strip the intentional content they express from any mental aspects it might contain, an agent analysing the representations cannot grasp the intended meaning of such representations fully (relying on visual means only). So, how could the agent who generates visual representations indicate any related intentional aspects?

Going back to the symbol grounding problem, one could argue that establishing vision-language associations could play another role too; by associating visual tokens with language units one is actually able to label/annotate the visual tokens making use of language's expressive power in indicating subtle distinctions. Therefore, associating a visual object with a specific word/phrase can indicate the focus of what is depicted (by referring exactly to what is of interest, leaving aside background visual elements) and its token or type status and in particular its exact conceptual status (level of abstraction - e.g. the use of a male proper name vs. the use of the word "man" vs. the use of the word "individual" etc.). It seems therefore, that in vision-language associations it is not only vision that plays a significant role in grounding language in the physical world; language, in its turn, provides the means for vision to indicate intentional aspects of what it depicts, providing visual representations with a *control* over their referents, and *grounding* —this way— visual representations in the mental world (since intentional content and aspect are mental entities according to Searle). In particular, in such associations, language assists in distinguishing between cases when:

- Different visual representations (because of differences in resolution, point of view, occlusions etc.) are used to refer to the same object (Vogt 2003). By associating the same word to the different visual representations an agent learns that they all refer to the same object.
- Different visual representations are used to refer to different objects that belong though to the same class (family resemblance cases) (Vogt 2003). Language denotes the class of the representations, determining the conceptual level at which they meet.
- Same visual representations for the same object are assigned different words/labels. In this case language may denote differences in register, domain (terminology) and level of genericness (type-token distinctions).
- Last, language assists in referring/pointing at exactly that which is of interest/in focus in a visual representation, leaving aside things that are necessarily depicted but are not of interest (Vogt 2003).

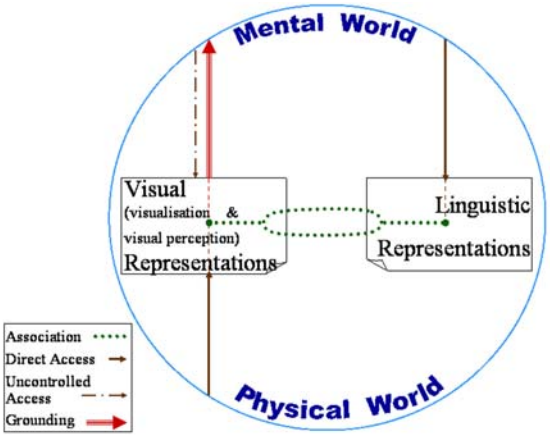

Figure 5: Language representations ground vision in the mental world

Therefore, we argue that *language grounds visual representations in the mental world*, providing them with a *controlled access* to mental referents and in particular, providing a way for indicating levels of abstraction, focus and relevance in what is depicted. Figure 5 illustrates this. Visual representations have access to the mental world (visualization ones are triggered by mental entities, while perceptual ones express the visual experience mental state); however, they have no control over this access. They need a controlled grounding, but they have no (or they have feeble) inherent means of achieving it. *One way* of achieving this grounding is through association of visual representations with linguistic ones.

Seen from a wider scope of grounding representations in general (regardless of whether they are symbolic or iconic), the grounding direction is not one- but rather bi-directional. Natural language symbolic representations can ground visual ones too, by functioning as *intentional aspect indicators*.

The synthesis

The parallel exploration of vision and language and Searle's theory of intentionality have both pointed to important, inherent characteristics of visual and linguistic modalities:

- Language lacks direct access to the physical world, whereas
- Vision lacks controlled access to the mental world (it lacks effective ways of indicating mental aspects of the world)

Although the former has been indicated and addressed computationally for solving the symbol grounding problem, the latter has been largely ignored. Extending the notion of symbol grounding to representation grounding in general and indicating the bi-directionality of grounding as a process, we argue that:

- Visual representations ground (provide direct access to) linguistic representations in the physical world, while
- Linguistic representations, in their turn, ground visual representations in the mental world, providing a way for indicating degrees of abstraction, focus and relevance in what is depicted

Figure 6 illustrates this *double grounding*. *Association* is the *sine qua non* constituent of grounding in either direction. In fact, it is what bridges the gap between the two quite different types of representations, it is bi-directional itself, and therefore renders grounding bi-directional too.

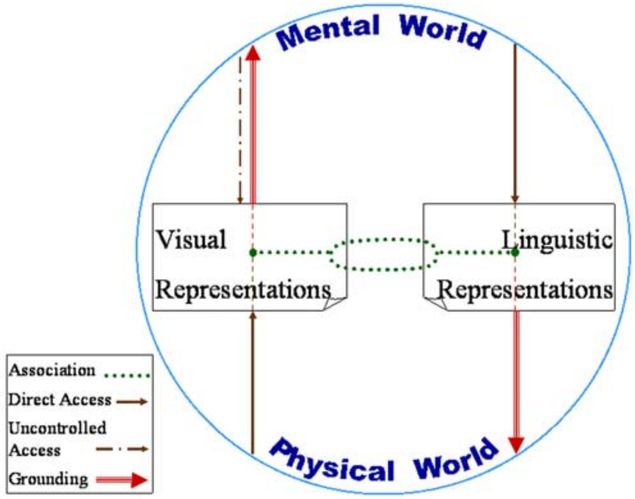

Figure 6: Vision-Language Integration: a double grounding case

Visual and linguistic representations *need* to be directly grounded in both physical and mental aspects of the world. Visual representations *inherently need* a *controlled* access to the mental world, in order to indicate intentional aspects of their physical or/and mental referents efficiently. In their turn, linguistic representations *inherently need* a *direct* access to the physical world, in order to maintain coherence in communication, when performing the intentionality act they are employed to. Generally speaking, the agents who analyse or/and generate visual and linguistic representations, need to be able to ground these representations on their own, if they are to be ascribed intrinsic intentionality.

If associating visual and linguistic representations is the essence of double-grounding, then, re-visiting the descriptive AI definition of vision-language integration (Pastra & Wilks 2004) that:

> “vision-language integration is the process of establishing associations between visual and linguistic pieces of information”

leads us to the conclusion that:

> “Vision-language integration is *a case of double-grounding* of visual and corresponding linguistic representations.”

The conclusion provides a theoretically established explanatory justification of why integration of visual and linguistic information takes place in multimodal situations: so that one medium compensates for the features that the other lacks, or, to put it generally, for serving coherence in communication. Vision-language integration abilities allow an agent to express and understand intentionality in specific multimodal situations[15].

At this point, we need to note that we are far from claiming that *all* language can be grounded in/integrated with vision; language in many cases refers to abstract, non-perceptible entities and states of affairs. We do not claim that language *always* needs grounding. Neither that it is *only* vision that can ground language in the physical world[16]. Similarly, we do not claim that *all* images can be grounded in/integrated with language (there are cases when one is short of words for describing something, cf. e.g. surrealistic paintings)[17]. Neither do we claim that images *always* need to be grounded (communication context may clearly indicate the intended intentional aspect that an image can’t indicate on its own). We do not claim that it is *only* language that can ground images; gestures, for example may do this to an extend, though language is undoubtedly extremely efficient in denoting things that images can’t denote with visual means.

We look into situations in which the content of linguistic representations can, indeed, be associated with the content of visual ones. We claim that in such situations, vision and language are both used, because their integration (double-grounding) allows the representations of one to compensate for features the representations of the other inherently lack; in these situations, lack of integration causes coherence gaps in communication and is a sign of lack of intrinsic intentionality on the part of the agent involved. Which are these situations though? When is integration needed?

In identifying four types of vision-language integration processes, work reported in (Pastra & Wilks 2004) has also indicated four corresponding integration purposes served by the AI prototypes reviewed. These purposes could be thought of as very general situations in which vision-language integration is a *sine qua non* process (as indicated within AI, *i.e.*, for computational purposes):

[15] We need to note that our work has an AI perspective; justifying vision-language integration in humans would require a further correlation of our theory with cognitive science related literature which goes beyond the scope of this paper. Some preliminary correlations can be found in (Pastra & Wilks 2004).

[16] Other perceptual processes could substitute vision’s role. Cf. the case of blind people who rely on other senses, such as audition and haptics.

[17] Note, also, that we refer mainly to cases of images of objects/scenes; however, there are images of e.g. text or gestures too. We do not address these special cases of images in this paper because of space constraints.

- when the analysis of a medium needs disambiguation through the analysis of another medium,
- when the content of one medium needs to be expressed in another,
- when a multimodal answer needs to be generated, or
- when situated multimodal dialogue takes place.

Conclusion

In this paper, we attempted to present an explanatory, theoretical framework for vision-language integration in AI[18]. In pointing out that this type of integration is dictated by the inherent characteristics of visual and linguistic representations, double-grounding aims at correlating the notion of media *inter-dependence* with the task of integration. Furthermore, in associating issues of intrinsic intentionality (and intelligence) with the vision-language integration abilities of an agent, the theory emphasises the AI *objective* served through computational vision-language integration. Through both these aspects, double-grounding provides a theoretically grounded view of collaboration between AI subfields as a *sine qua non* requirement for vision-language integration.

Acknowledgments

The author would like to thank Yorick Wilks for his valuable comments on previous drafts of this paper. Part of this work has been funded by the GR/M89676/01 EPSRC grant.

References

André, E., and Rist, T. 1993. The design of illustrated documents as a planning task. In Maybury, M., ed., *Intelligent Multimedia Interfaces*. AAAI Press/MIT Press. chapter 4, 94–116.

Ballard, D.; Hayhoe, M.; Pook, P.; and Rao, R. 1997. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences* 20(4):723–767.

Barsalou, L.; Simmons, W.; Barbey, A.; and Wilson, C. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences* 7(2):84–91.

Bernsen, N. 1995. Why are analogue graphics and natural language both needed in HCI? In Paterno, F., ed., *Interactive Systems: Design, specification and verification. Focus on Computer Graphics*. Springer Verlag. 235–251.

Bredeche, N.; Chevaleyre, Y.; Zucker, J.; Drogoul, A.; and Sabah, G. 2003. A meta-learning approach to ground symbols from visual percepts. *Robotics and Autonomous Systems* 43:149–162.

Cangelosi, A.; Greco, A.; and Harnad, S. 2000. From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science* 12:143–162.

Coradeschi, S., and Saffiotti, A. 2003. An introduction to the anchoring problem. *Robotics and Autonomous Systems* 43:85–96.

Harnad, S. 1990. The Symbol grounding problem. *Physica D* 42:335–346.

Jackendoff, R. 1983. *Semantics and Cognition*. MIT Press.

Jackendoff, R. 1987. On beyond Zebra: the relation of linguistic and visual information. *Cognition* 20:89–114.

Jackson, S., and Sharkey, N. 1996. Grounding computational engines. *Artificial Intelligence Review* 10:65–82.

Kaplan, F. 2000. Talking AIBO: First Experimentation of Verbal Interactions with an Autonomous Four-legged Robot. In *Proceedings of the TWENTE Workshop on Language Technology*, 57–63.

Kjorup, S. 1978. Pictorial Speech Acts. *Erkenntnis* 12:55–71.

Marr, D. 1982. *Vision*. San Francisco: W. H. Freeman.

Maybury, M., ed. 1993. *Intelligent Multimedia Interfaces*. AAAI Press/MIT Press.

Minsky, M. 1986. *The Society of Mind*. Simon and Schuster Inc.

Nirenburg, S., and Wilks, Y. 2001. What's in a symbol: ontology, representation and language. *Journal of Experimental and Theoretical Artificial Intelligence* 13(1):9–23.

Pastra, K., and Wilks, Y. 2004. Vision-Language Integration in AI: a reality check. In *Proceedings of the 16th European Conference on Artificial Intelligence*.

Peirce, C. 1960. *Collected Papers of Charles Sanders Peirce*, volume 1 and 2. Belknap Press of Harvard University Press.

Pylyshyn, Z. 2001. Visual indexes, preconceptual objects, and situated vision. *Cognition* 80:127–158.

Roy, D., and Pentland, A. 2000. Learning Words from Sights and Sounds: A computational model. *Cognitive Science* 26(1):113–146.

Roy, D. 2002. Learning visually grounded words and syntax for a scene description task. *Computer speech and language* 16:353–385.

Searle, J. 1969. *Speech acts: an essay in the philosophy of language*. Cambridge University Press.

Searle, J. 1976. A classification of illocutionary acts. *Language in Society* 5(1):1–23.

Searle, J. 1980. Minds, Brains, and programs. *Behavioral and Brain Sciences* 3(3):417–457.

Searle, J. 1983. *Intentionality: an essay in the philosophy of mind*. Cambridge University Press.

Vogt, P. 2003. Anchoring of semiotic symbols. *Robotics and Autonomous Systems* 43(2):109–120.

Waltz, D. 1981. Generating and understanding scene descriptions. In Joshi, A.; Webber, B.; and Sag, I., eds., *Elements of discourse understanding*. Cambridge University Press. 266–282.

[18] The theory is part of a larger investigation of vision-language integration in AI, which also includes empirical work on testing the limits of current AI technology for performing vision-language integration, minimising human intervention in core integration stages.