An Alternative Suggestion for Vision-Language Integration in Intelligent Agents

Katerina Pastra

Institute for Language and Speech Processing,
Artemidos 6 and Epidavrou, Maroussi, 151-25, Greece
kpastra@ilsp.gr

Abstract. State of the art artificial agents rely heavily on human intervention for performing vision-language integration; apart from being cost and effort effective, this intervention deprives artificial agents from the ability to react intelligently and to show intentionality when engaged in situated multimodal communication. In this paper, we suggest an alternative way of building vision-language integration prototypes with limited human intervention. The suggestions have emerged from the development of such a prototype for the verbalisation of visual scenes in a property-surveillance task.

1 Introduction

In Artificial Intelligence, vision-language association has been defined as an *integration* task that artificial agents perform for a number of applications, such as situated multimodal dialogue [1]. Vision-language integration has been theoretically explained as a *double-grounding* case [2], in which vision grounds language to its intended referents in the physical world (cf. the symbol grounding debate [3, 4]), while at the same time language grounds vision to intentional aspects of the mental world (i.e., it renders e.g. object salience and token-type distinctions explicit). Double-grounding suggests that, ideally, a completely autonomous artificial agent with vision-language integration abilities can demonstrate human-like intentionality in certain multimodal situations. In a less ideal situation, the more "independent" from human intervention for performing integration an artificial agent is, the more intentional will appear to be. This autonomy requires that an agent is able to interpret its representations with computational mechanisms that go beyond mere instantiation of known association facts to *inferring* associations, so that the agent is able to generalise within and across domains and to recover from unexpected situations [5].

However, state of the art integration prototypes seem to be far from performing vision-language integration on their own. Through an extensive review of such systems, it has been shown that state of the art prototypes have major *scalability* problems; some of them deal with blocksworlds or miniworlds and therefore fail to generalise not only across but also within application domains [1], while others rely on *manually abstracted visual data*. Furthermore, their

G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 588–591, 2006.

association abilities are *restricted to instantiations* of *a priori* known image-language associations, failing therefore to deal with unexpected data even if common-sensical [1].

Is it feasible to shift all three drawbacks of state of the art integration prototypes, *i.e.*, build a prototype that will work with real world visual data (rather than blobs), that will analyse its visual input automatically and will have inference-making mechanisms that will allow for more scalable vision-language associations? How far can we currently go in achieving computational vision-language integration, minimising human intervention within core integration stages? VLEMA, a vision-language integration prototype developed for a property serveillance task has attempted to address these questions [5]. In what follows, we present suggestions on the direction research could take towards the goal of restricting human intervention in vision-language integration mechanisms as they emerged from the development of this prototype.

2 Suggestion One: Virtualised Reality Images as Visual Input

Object segmentation in 2D images of real world scenes is still quite restricted, with algorithms performing well only when the object(s) to be identified are very specific (e.g face recognition). Therefore, if one is to perform automatic analysis of real-world visual scenes with —more or less complex— configurations of a variety of objects, one is led to a dead-end; this is the most important reason why researchers usually resort to manual abstraction of visual data or to the use of blocksworlds when developing intelligent prototypes that require vision-language integration abilities.

There is, however, a recently emerged research field in which computer vision and computer graphics advances meet: that of *virtualised reality* [6]. Instead of relying on simplistic CAD models for building virtual worlds manually, image reconstruction algorithms construct virtual models of a real scene from multiple static or dynamic images of the scene. Simply put, the process involves the recording of a visual scene/event through (multiple) wide-range cameras, the recovery of the 3D geometry and/or photometric information of the scene from the images and the translation of this scene description into computer graphics models. The latter preserve the scene geometry through depth maps and they may also preserve the scene texture through mapping of the original 2D images of the scene onto the resulting 3D model. Novel views of the virtualised scenes are then easily synthesised on the fly in existing virtual reality hardware.

It is this type of visual data that VLEMA uses as input and suggests that should be used in similar prototypes, because it allows the system to work with real world, complex visual scenes in a format (*i.e.*, 3D) that is more promising for automatic image analysis due to the wealth of visual information it provides (e.g. there is no partial information on a specific object due to occlusion). While these images seem photo-realistic when rendered on screen by VRML browsers, the source code consists actually of very low-level visual information: indices

of the coordinates of thousands of triangular faces that shape the scene in 3D. The faces are listed in no particular order,and their coordinates are expressed in relation to a common coordinate system (that of the whole scene) While not trivial, object segmentation in such data is feasible: in VLEMA, a face clustering algorithm decides which faces are boundary-forming ones in each dimension and produces three different clusterings of the faces into candidate-objects, according to the dimension that was taken to be the principle one each time. It then compares the different clustering results assuming that the first and last cluster in each clustering/dimension stand always for the *extrema* objects of the scene (i.e. the walls of an indoor scene). It also assumes that a cluster which is identified as such in more than one clusterings/dimensions stands for a single, non decomposable object. The remaining face clusters in each dimension may be clusters standing for the parts of an individual object (analysed differently in each clustering/dimension) or parts of more than one object; this distinction is left to an object naming module [5].

3 Suggestion Two: Naming Through a Feature-Augmented Ontology

VLEMA's suggestion for naming is the development and use of a domain ontology and knowledge-base with feature-based profiles of the entities of the domain. In particular, this resource should record information for each object regarding its:

- *structure*: the number of parts into which the object is expected to be decomposed in different dimensions
- *functionality*: visual characteristics an object may have which are related to its function e.g. whether an object has a surface on which things can be placed/fixed, and
- *interrelations*: these refer mainly to (allowable) spatial configurations of objects (e.g. whether an object could be *on* the floor or not), the dimension according to which size comparisons would be meaningful etc.

There are many arguments in favour of such feature selection for object naming in the literature [7, 8]. Still, one could argue, that relying on such features for drawing object name inferences would be risky when analysing e.g. a scene of a room with many different multi-part objects. Machine learning methods for associating words in an utterance with objects in view represent visual data through feature-value vectors, with features being usually shape, colour or/and texture [9, 10]; these approaches work for naming only very simple physical objects (e.g. a ball) or blobs, the image of which has been carefully stripped from any visual background during learning. However, approximate shape, colour or texture information alone cannot lead to any inferences in complex real world scenes. On the contrary, the suggested types of features allow a prototype's inference mechanisms: a) to go beyond differences in the appearance of similar objects (e.g.

different styles of sofas) and name these objects in the same way and b) to generalise over viewpoint differences; for example, VLEMA can identify a sofa as such even when seen from the side (rather than *en face*).

These are generalisations that machine learning-based object naming algorithms cannot do easily (or at all). In particular, identifying objects which differ in appearance as ones of the same type is something that cannot be achieved even with a very large amount of training data, if similar examples are not present in the training data. Similarly, learning approaches cannot deal with viewpoint differences in the appearance of an object; there is an almost infinite number of different images of the same object which may result from differences in the viewpoint from which the object is seen in a complex scene.

4 Conclusion

In this paper, we suggested that the development of vision-language integration agents could benefit from the combined use of *virtualised reality images* and a specific kind of a *feature-augmented ontology*. The automatic analysis of the former not only satisfies the requirement of working with real world scenes (rather than blocksworlds or one-object images), but it also allows for a wealth of information to be extracted. Combined with a feature-augmented ontology, this information can be used from an object naming module to perform structural, functional and relational (spatial or other) checks for determining the final number and identity of objects depicted in a scene.

References

1. Pastra, K., Wilks, Y.: Vision-language integration in AI: a reality check. In: Proceedings of the 16th European Conference in Artificial Intelligence. (2004) 937–941
2. Pastra, K.: Viewing vision-language integration as a double-grounding case. In: Proceedings of the AAAI Fall Symposium on "Achieving Human-Level Intelligence through Integrated Systems and Research". (2004) 62–69
3. Searle, J.: Minds, brains, and programs. Behavioral and Brain Sciences **3** (1980) 417–457
4. Harnad, S.: The symbol grounding problem. Physica D **42** (1990) 335–346
5. Pastra, K.: Vision-Language Integration: a Double-Grounding Case. PhD thesis, University of Sheffield (2005)
6. Kanade, T., Rander, P., Narayanan, R.: Virtualised reality: constructing virtual worlds from real scenes. IEEE Multimedia **4** (1997) 34–46
7. Minsky, M.: The Society of Mind. Simon and Schuster Inc. (1986)
8. Landau, B., Jackendoff, R.: "What" and "Where" in spatial language and cognition. Behavioural and Brain Sciences **16** (1993) 217–265
9. Kaplan, F.: Talking AIBO: First experimentation of verbal interactions with an autonomous four-legged robot. In: Proceedings of the TWENTE Workshop on Language Technology. (2000) 57–63
10. Roy, D.: Learning visually grounded words and syntax for a scene description task. Computer speech and language **16** (2002) 353–385