Vision-Language Integration in AI: a reality check

Katerina Pastra and Yorick Wilks[1]

Abstract. Multimodal human to human interaction requires integration of the contents/meaning of the modalities involved. Artificial Intelligence (AI) multimodal prototypes attempt to go beyond technical integration of modalities to this kind of meaning integration that allows for coherent, natural, "intelligent" communication with humans. Though bringing many multimedia-related AI research fields together, integration and in particular vision-language integration is an issue that remains still in the background. In this paper, we attempt to make up for this *lacuna* by shedding some light on how, why and to what extent vision-language content integration takes place within AI. We present a taxonomy of vision-language integration prototypes which resulted from an extensive survey of such prototypes across a wide range of AI research areas and which uses a prototype's integration purpose as the guiding criterion for classification. We look at the integration resources and mechanisms used in such prototypes and correlate them with theories of integration that emerge indirectly from computational models of the mind. We argue that state of the art vision-language prototypes fail to address core integration challenges automatically, because of human intervention in stages during the integration procedure that are tightly coupled with inherent characteristics of the integrated media. Last, we present VLEMA, a prototype that attempts to perform vision-language integration with minimal human intervention in these core integration stages.

1 INTRODUCTION

One of the characteristics that define multimedia systems is *integration*, the ability to combine all necessary hardware and software components in the same system forming a whole that allows the use of multimodal input and the subsequent presentation, storage, transmission and technical processing of this multimodal information. State of the art multimedia systems are extremely efficient in this, achieving not only "off-line" integration, but also real-time, over the network one. Apart from this technical integration though, integration of the *meaning* carried by each modality is important in multimodal situations. The computational integration of the content of multiple modalities and of visual and linguistic ones in particular, is an Artificial Intelligence (AI) aspiration that goes back to the very early days of the field[2]. However, the level computational vision-language integration reached since and the means used for accomplishing it are issues that have not been addressed thoroughly and remain still to be answered. In performing a "reality check" —and therefore addressing these issues— one may bring content integration to the foreground and render it the common perspective from which "intellimedia" [16] spanning a wide range of AI research areas from cross-modal information retrieval to robotics are brought together. The gains from doing so are significant; tendencies, practices, lacunae and needs can be identified more easily and directions for dealing with existing problems can emerge directly.

In this paper, we attempt to perform such a "reality check" of vision-language content integration in AI. We first present four categories of vision-language prototypes, which provide an *extensional* definition of vision-language content integration. Then, we present the *intensional* definition of integration as it emerges from the indication of the integration resources and mechanisms used in these systems. Based on these findings, we correlate AI integration practices with theories of integration, which emerge from computational models of the mind indirectly. We argue that state of the art vision-language integration systems fail to perform real integration of these modalities, because they rely on human created data for compensating for the features vision and language inherently lack. We conclude by presenting VLEMA, a prototype that points to directions vision-language integration research could take for addressing integration challenges with minimal human intervention.

2 DEFINING INTEGRATION THROUGH CLASSIFICATION

Vision-language integration has been attempted in AI research fields for a wide variety of tasks and application domains. There is only one survey published in the mid-nineties that attempts to present the state of the art in vision-language integration systems [26]. However, apart from the fact that this survey refers to limited work in the field from the eighties up to 1994, it actually mixes theoretical suggestions and implemented prototypes commenting only sparsely on how integration is attempted. The research reviewed in that survey is classified according to the medium of the input, *i.e.* language, images or both; this classification though does not capture classification dimensions of the systems that go beyond application tasks and domains. Most importantly, it includes research that does not actually perform content integration, such as work on text-based image retrieval and work on prototypes that perform what one could call *quasi-integration*.

The latter refers to cases of fusion of results obtained by sequential modality-dependent processes that one may easily mistake for some form of content integration. A characteristic example comes from the video retrieval, summarization and skimming research. We refer to cases when both natural language and image processing technologies are used, the former for topic detection within video files and the latter for identifying the exact boundaries of the video segments that present each specific topic and the key frames within each segment, cf. for example [19]. In such cases, the results obtained from analysing one modality, *i.e.* speech or text, create the neces-

[1] Department of Computer Science, University of Sheffield, Sheffield, United Kingdom. Email: {katerina,yorick}@dcs.shef.ac.uk

[2] Cf. for example Kirsch who was one of the first to mention that logic could be used as an internal stage of communication for the parallel analysis of text and pictures and indicated problems in developing grammars for the syntactic analysis of drawings similar to the ones for text analysis [14].

sary conditions for processing the other (video frames) effectively and refining the final output. Though both technologies are used, no integration of their output takes place. Either the intersection of the results produced by each process is the final output of the system or the results from the one process are used to constrain the search space for the other, reducing actually the set of possible image/text candidates. Optical Character Recognition (OCR) technology could also be thought of as a kind of integration performing technology that goes from the image of a text to raw text for input to a language processor. Though this technology may make use of knowledge of the linguistic system to recover from errors, it does not perform any kind of content integration. It performs pattern recognition in order to extract the characters from the image, as it would for extracting any texture from an image. These patterns are not transformed in a format suitable for language processing, they *are* in the right format anyway and the task is exactly that, to extract them in a precise and effective way.

In a recent introductory chapter on natural language and multimodal AI systems, references to multimodal integration have been made and some categories of multimodal systems involving language have been presented [2]. However, being an introduction to the field, the chapter is quite general, the boundaries of the categories mentioned are not clearly defined, integration issues and in particular vision-language integration are not in focus at all (most examples involve gestures and language) and multimodal integration is briefly mentioned as a multimodal input fusion function, which is a very vague way of defining integration. So, how could integration be defined and which types of prototypes that involve vision and language do actually perform integration?

In reviewing AI prototypes for which vision-language integration is a *sine qua non* feature for performing a task, we indicated four general categories of such systems. The criterion guiding this classification was the *integration purpose* served by each system and therefore, the four categories identified express four general integration purposes served by the corresponding prototypes. Classifying multimodal prototypes from this perspective contributes in an *in depth* understanding of both:

- the forms vision-language integration has taken in AI, and
- the level state-of-the-art integration systems have reached and the integration means they have used

The former provides an *extensional* definition of integration, while the latter points to an *intensional* definition. Furthermore, the integration purpose criterion provides the common perspective required for classifying systems, which have been developed in diverse AI research areas.

2.1 An extensional definition

Table 1 illustrates the four system categories we have identified in a survey of more than fifty prototypes that span many decades, from SHRDLU [29], which has been considered to be the first vision-language integration system [26], to conversational robot prototypes of the new millennium [23, 25]. These categories provide a general enumeration of the processes the term *integration* applies to, they describe what it is that actually constitutes *integration* and could therefore be thought of as its *extensional definition*. In particular, the first category includes *performance enhancement* integration prototypes such as PICTION [27] [3]; these systems analyse information

[3] We give only a few examples of prototypes that belong to each category because of the space constraints of this paper

Table 1. The four categories of integration prototypes

SYSTEM TYPE	INTEGRATION PROCESS
Performance Enhancement	Medium_x analysis => Medium_y analysis
Media Translation	Source medium analysis => Target medium generation
Multimedia Generation	Abstracted data => Multimedia generation
Situated Multimodal Dialogue	Multimedia Analysis => Medium/multimedia generation

expressed in a specific medium in order to extract information that will give more accurate specifications for the analysis of information in another medium. Integration takes place mainly when the output of the analysis of one medium is used for constraining or even guiding the analysis of another medium. Most of the systems in this category use natural language derived information to enhance their image understanding capabilities.

The second category includes *media translation systems*, *i.e.* systems that allow for the generation of one medium given another. These systems treat the image-language relation as one of "mutual exclusion": they allow for a specific task to be performed using only language or only images and not a combination thereof. Therefore, they go from medium analysis to medium generation, translating the information expressed in the source medium into a target medium. Going from one medium to another with the less possible loss of information requires that the source information be expressed in a common content description format and that mechanisms for dealing with the specific nature of the target medium exist and this is where the integration challenge lies for this type of systems. With systems such as SOCCER [3] that go from images to language, one gets involved with Natural Language Generation, while with ones going from text to images, such as WordsEye [7], one crosses over into the field of Computer Graphics.

Multimedia generation systems in their turn, start from a specific format of the message to be conveyed and realise this message with the best possible media combination. These are the so-called Intelligent Multimedia Presentation Systems (IMMPS) [16]. There are IMMPSs, such as SAGE [21], that starting from a goal to be attained and tabular, numerical data they generate information graphics and associated text; these systems visualise and describe abstract data. On the other hand, systems such as MAGIC [17], start from a knowledge representation of the message to be conveyed and realise it through a media combination. Integration in both cases is more or less supported via both a content representation formalism, that is itself a means of performing media integration, and the media co-ordination phase that aims at a coherent and consistent output. Since decisions and choices for achieving the multimedia presentation and the co-ordination are made throughout the whole process and are dependent upon the output of various modules, the architecture of these systems is such that serves integration purposes too.

Contrary to multimedia generation systems, *situated dialogue integration systems* perform both *medium analysis* and *medium generation*. Most of the systems allow for natural language interface with the user with whom they also share a visual environment. The systems are able to react to the input by providing multimodal or monomodal answers or/and modifying their visual environment. Therefore, the *dialogue* that takes place refers always to a visually perceptible *situation* shared by both the user and the machine. To be included in this category, the system must be able to

track changes in the visual environment or analyse it automatically; question-answering regarding a visually shared, *a priori* known and unchangeable environment is not covered, since it does not involve visual analysis or generation capabilities in any way, just mere interaction with information on the visual environment pre-stored in a database. We view dialogue from the wider perspective of human-machine interaction, which does not necessarily involve verbal reactions/answers on the part of the machine, but any kind of reaction, in any modality, which demonstrates efficient communication with the user. SHRDLU belongs to this type of systems as well as conversational robots do.

2.2 An intensional definition

The visual and linguistic modalities analysed/generated by vision-language integration prototypes are quite different (e.g. 3D graphics, photographs, drawings and text, speech respectively), as the application tasks and domains they cover are too. Table 2 provides details on the example systems mentioned in section 2.1.

Table 2. Details of some integration prototypes

SYSTEM	**CASSIE**
INPUT	Speech (EN), blocksworld (3D)
INTEGR. RESOURCES	KL-PML association lists
INTEGR. MECHANISMS	Unification
OUTPUT	Object identification, limited conversation
SYSTEM	**MAGIC**
INPUT	Choice of patient file
INTEGR. RESOURCES	Schemas
INTEGR. MECHANISMS	Schema instantiation, media conductor module,co-reference
OUTPUT	Speech/text (EN), graphics (animation)
SYSTEM	**PICTION**
INPUT	Photographs (people) and captions (EN)
INTEGR. RESOURCES	Integrated Knowledge Base (lexical info, object schemas)
INTEGR. MECHANISMS	Semantic Networks
OUTPUT	Face identification
SYSTEM	**SOCCER**
INPUT	Image sequences (soccer-video), manually extracted trajectories
INTEGR. RESOURCES	GSD per frame, event models (objects+course diagram)
INTEGR. MECHANISMS	Event model instantiation (labelled directed graphs), event selection, verbalisation history
OUTPUT	Text event description (GER)

These example systems have been classified into different categories; however, they all seem to make use of integration resources that capture similar information, though represented and instantiated differently. The integration resources associate visual and corresponding linguistic information in various forms; CASSIE for example has association lists that link conceptual terms and their corresponding instances and attributes to perceptual-motor related lexemes and their low-level n-ary tuples of visual feature-values and functions (the latter implement action-denoting expressions). Similarly, other systems use integrated knowledge bases consisting of lexical information and corresponding object schemas, associations between object names, properties and the visual state of a scene, event models and corresponding geometric scene descriptions (GSD) etc. The knowledge representation format used for encoding such associations determines (or is determined by) the association instantiation mechanism; schemas, frames and semantic networks are just a few of the mechanisms used. In most cases, integration mechanisms go beyond the instantiation of vision-language associations to inference depiction/verbalisation rules, media allocation, discourse history and cross-modal reference generation modules, layout coordination modules and even incremental or blackboard architectures that allow for constant communication between media-specific modules; all these mechanisms facilitate integration and compliment any association instantiation mechanisms.

The ways through which integration is performed provide the *genus et differentia* required for an *intensional* definition of the term. Computational vision-language integration is the *process* (genus) of establishing associations between visual and linguistic pieces of information (differentia). The means for establishing such associations vary —as mentioned earlier in detail— according to the integration purpose served by a system.

2.3 EMERGING THEORIES OF INTEGRATION

It is true that the idea of using vision-language integration resources with information regarding the name and visual features of objects depicted in images along with links to the image regions they occupy has been proposed at least since the seventies [4]. Recent work on the construction of multimodal thesauri relies on similar, manually constructed associations of image segments, image feature-vectors and the corresponding lexical items and concepts [28]. However, in searching for a theoretical framework underlying the way AI prototypes have approached integration issues, one realises that not only there is no integration theory within AI, but integration has only indirectly been described in cognitive science too[4]. Fodor's computational model of the mind is a characteristic example of such a lack of integration theories. In this model it is argued that the outputs of both language and vision (perceptual systems) are expressed in a format appropriate for the central (cognitive) processes into which they are fed [8]. However, the nature of this format and the modality-specific processes required in both vision and language for encoding their output accordingly is something Fodor not only avoids to explain, but he also argues that it is impossible to explain with current scientific paradigms [9].

On the other hand, looking at Minsky's theory of the *society of mind*, one can only find implicit references to integration in his descriptions of how linguistic and perceptual knowledge/information are associated. Minsky seems to see integration as an activation spreading procedure that makes use of *learned* associations between linguistic data and sensory attribute-values [18]. This points to the fact that, similarly, computational integration requires machine-learning acquisition of such associations rather than manually constructed ones. Though there are attempts to automate learning of such associations in AI [22], no research prototype has been developed yet that incorporates such learning mechanisms for use in multimodal scenarios [30].

More specific suggestions regarding vision-language integration have been made by Jackendoff in his own computational model of the mind [11]. Within this model, vision-language integration is described as a set of principles of correspondence between the *conceptual structure level* of language and Marr's *3D model level* of vision. His view of integration implies the use of a kind of integrated knowledge base with both conceptual and visual information along with al-

[4] Most research on cognitive architectures gives a very general picture on how different faculties interact; in this section, we mention only some of the most influencial models of the mind.

gorithms for making inferences and extracting meaning. In suggesting 3D models as the appropriate visual representations of objects for forming associations with corresponding linguistic expressions, Jackendoff points to an important issue in vision-language integration, *i.e.* the representation level at which vision-language associations are more efficiently made. It has been argued that the different nature of the primitive units of the media involved creates a *correspondence* problem between visual and linguistic representations that hinders their integration [26]; if volumetric, object-centred and hierarchical information (as the one provided by 3D models), which is difficult to extract with state of the art image processing techniques, is needed for integration, then one can justify why integration prototypes rely on manually abstracted visual information. Both Minsky's and Jackendoff's emerging theories of vision-language integration point to important integration aspects towards which our own criticisms of current AI vision-language integration prototypes have been directed.

3 THE AI QUEST FOR VISION-LANGUAGE INTEGRATION

Looking at vision-language integration in AI from both an implementation and a theoretical perspective has led to the identification of three important characteristics shared by all state-of-the-art prototypes and which indicate mistakes in current computational vision-language integration practices:

- It is only simulated or manually abstracted visual input that is used in integration prototypes

Whenever the input to the system is visual, it is either simulated (e.g. synthesised graphics instead of video images) or *a priori*, manually abstracted (e.g. geometric scene descriptions); if it is processed automatically, in the best case, it is minimally analysed (e.g. extraction of trajectories). It is no coincidence that most of the visual information involved in the prototypes deals with events or single objects rather than complex static scenes; the latter would require very accurate and detailed visual analysis, while tracking changes in an already known visual environment is less demanding. The case is not the same for natural language input, which is usually analysed automatically to different extents. It has been argued that visual modalities have inherent difficulties in providing indications of their focus and the degree of abstraction of what they depict [5, 12], which hinders their computational analysis. It seems, that by feeding the systems with already abstracted information, a core integration challenge that is related to the nature of the visual modalities is not being encountered automatically, but is, instead, skipped.

A similar criticism was expressed more than a decade ago for AI research in general. Brooks noted that system developers tend to do the abstraction for their systems themselves leaving just some search to the systems, thinking that at some point developments in other AI sub-fields for automatic input acquisition and analysis will be incorporated to their own systems [6]. This leads to our second observation that has also been repeatedly mentioned as a critique for AI research in general and which is related to the assumption, on the part of AI system developers, that things will somehow, sometime scale up [15, 1]:

- All vision-language integration prototypes are restricted to blocksworlds or miniworlds

The applications of the prototypes reviewed involve either blobs in various spatial configurations or extremely restricted real world objects/events such as car models, espresso machines, electric current diagrams, collisions etc. When the application is a real one e.g. soccer games, simplified simulations of the visual input or abstracted data is used instead of the actual input. By simplifying the application, ideal situations are assumed, many factors are ignored or unrealistically simplified increasing the risks of making wrong or unrealistic assumptions. Scaling has been considered important for judging the significance of AI research implemented in a prototype [24]; however, hardly is it served when research is confined to blocksworlds or miniworlds.

- It is only manually constructed vision-language integration resources that are used

Apart from other knowledge resources, vision-language integration prototypes use *a priori* known associations between conceptual and visual entities. It has been argued that linguistic modalities have no way to directly connect to physical entities, they can only indirectly refer to them [10]; this missing link is conventional, language-dependent and therefore learned. It is exactly this link that is provided by developers to integration prototypes in the form of integrated resources; once again, a limitation related to the nature of one of the modalities, *i.e.* the linguistic modalities, which needs to be dealt by an integration mechanism is tackled by the developers instead of the system.

It seems, therefore, that AI integration prototypes do not address integration problems related to the nature of the modalities integrated at all; instead, they avoid these issues by relying exclusively on human intervention in the form of pre-interpreted input and pre-stored associations. Subsequently, core integration challenges are not dealt with automatically to any extent; computational integration is confined to cross-modal interpreters and co-ordination mechanisms. Though developing such mechanisms is far from trivial, one cannot talk about real integration prototypes, when major integration challenges are left completely up to the humans to deal with. The question then becomes, whether it is currently possible at all to address these challenges constraining human intervention to other integration stages.

4 ENCOUNTERING INTEGRATION CHALLENGES WITH VLEMA

The Vision-Language intEgration MechAnism (VLEMA) is a prototype that attempts to generate natural language descriptions of static visual scenes depicting building interiors. The input to the system is an automatically-generated visual reconstruction of a real static scene encoded in the Virtual Reality Markup Language (VRML). The system parses the file and creates an attribute-value matrix that gathers visual information per object depicted, such as object parts, position, shape and colour/texture. A rule-based ProLog interpreter translates directly extracted visual information from this matrix into natural language, so that all visual information is expressed linguistically rather than numerically (in geometric parameters). On top of that, the interpreter is also able to infer information from the matrix, such as the relative size and spatial configuration of the objects depicted and express it in natural language. The linguistically expressed visual information populates the slots of a natural language template for describing rooms. At this point, the description is analytic, *i.e.* it does not *name* the objects depicted, but it describes them in terms of e.g. shape and colour, spatial or other relations to objects depicted etc. Categorization (type indication) of the objects through linguistic naming (e.g. "the chair" rather than "the brown object consisting of two orthogonal flat surfaces supported by four legs") is a task that we

currently attempt to perform with the use of a domain-specific ontology. The visual features and relations of each object depicted in the scene and the known features and allowable relations for each concept included in the ontology are compared; a probabilistic algorithm is used for choosing the best-match for each object. The object name selected substitutes the part of the template that refers to characteristics of the object which are implied by its class, *i.e.* the category membership determining features.

The motive underlying the —currently ongoing— development of this prototype was to address the vision-language integration problems mentioned in section 3 in a way that would shift human intervention from core integration stages and would actually minimise it. In particular, VLEMA makes use of visual input that has not been manually abstracted. It takes advantage of *virtualised reality* [13], a recently emerged field that bridges the gap between computer vision and computer graphics by using visual sensors to construct virtual models of real visual scenes automatically, preserving the visual details of the latter. The files used as input to VLEMA are part of a corpus of virtualised real building interiors, constructed and encoded in texture-mapped VRML format automatically by a robot-surveyor. Thus, the description of real, complex world scenes is also attempted, rather than blocksworlds or miniworlds. Of course, the fact that the VRML files have been created automatically makes the construction of an interpreter much harder than when one deals with manually built VRML source code; one of the difficulties lies on the way shape primitives are encoded. In manually constructed virtual worlds the source code includes, usually, linguistic references to the shape of an object (e.g. "cube"); in the automatically created ones, the approximate shape can only be inferred from matrices of three-dimensional coordinates of the points in space that form the contour of the object.

As far as the manually constructed integration resources problem is concerned, a vision-language association learning mechanism would be important for avoiding any human intervention for object naming. However, developing such a mechanism requires not only large multimodal resources (training corpora), but also appropriately annotated ones for effective and efficient development of learning algorithms for real visual scenes [20]. VLEMA has opted for a semi-automatic solution, which relies on a manually constructed, visual-feature augmented ontology and on probabilistic matching. Relying on feature-bundles for performing categorization has been criticised extensively in computational and cognitive linguistics. In VLEMA's case though, the facts that categorization is constrained to physical entities and that there is a vision-driven choice of features used to represent categories in the ontology change things totally; admittedly, the immenent evaluation of the prototype holds the answer to how suitable an approach this is.

5 CONCLUSION

This "reality-check" of vision-language content integration in AI has pointed to specific problems for computational integration and tendencies in the field that work against the solution of these problems. Addressing vision-language integration thoroughly and systematically has been the main objective of this paper, along with the hope to provide concrete suggestions on the direction AI research could currently take to address some of the related computational problems.

REFERENCES

[1] J. Allen, 'AI growing up: challenges and opportunities', *AI Magazine*, **19**(4), 13–23, (1998).

[2] E. André, 'Natural Language in Multimedia/Multimodal Systems', in *Handbook of Natural Language Processing*, ed., R. Mitkov, chapter 36, 650–669, Oxford University Press, (2003).

[3] E. André, G. Herzog, and T. Rist, 'On the simultaneous interpretation of real world image sequences and their natural language description: the system SOCCER', in *Proceedings of the European Conference on Artificial Intelligence*, pp. 449–454, (1988).

[4] R. Bajcsy and A. Joshi, 'The problem of naming shapes: Vision-language interface', in *Proceedings of the Theoretical Issues in Natural Language Processing*, pp. 157–161, (1978).

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

[6] R. Brooks, 'Intelligence without Reason', in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 569–595, (1991).

[7] B. Coyne and R. Sproat, 'WordsEye: An automatic text to scene conversion system', in *Proceedings of the International Conference on Computer Graphics and Interactive Technologies*, pp. 487–496, (2001).

[8] J. Fodor, *The modularity of mind*, MIT Press, 1983.

[9] J. Fodor, *The mind doesn't work that way*, MIT Press, 2000.

[10] S. Harnad, 'The Symbol grounding problem', *Physica D*, **42**, 335–346, (1990).

[11] R. Jackendoff, *Consciousness and the computational mind*, MIT Press, 1987.

[12] R. Jackendoff, 'On beyond Zebra: the relation of linguistic and visual information', *Cognition*, **20**, 89–114, (1987).

[13] T. Kanade, P. Rander, and R. Narayanan, 'Virtualised Reality: constructing virtual worlds from real scenes', *IEEE Multimedia*, **4**(1), 34–46, (1997).

[14] R. Kirsch, 'Computer Interpretation of English Text and Picture Patterns', *IEEE Transactions on Electronic Computers*, **13**(4), 363–376, (1964).

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

[16] *Intelligent User Interfaces*, eds., M. Maybury and W. Wahlster, Morgan Kaufmann Publishers, 1998.

[17] K. McKeown, D. Jordan, B. Allen, S. Pan, and J. Shaw, 'Language generation for multimedia healthcare briefings', in *Proceedings of the Applied Natural Language Processing Conference*, pp. 277–282, (1997).

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

[19] A. Olligschlaeger and A. Hauptmann, 'Multimodal information systems and Geographic Information Systems', in *Proceedings of the Environmental Search Research Institute User Conference*, (1999).

[20] K. Pastra and Y. Wilks, 'Image-Language Multimodal Corpora: needs, lacunae and an AI synergy for annotation', in *Proceedings of the 4th Language Resources and Evaluation Conference*, (2004).

[21] S. Roth and J. Mattis, 'Automating the presentation of information', in *Proceedings of the IEEE Conference on Artificial Intelligence Applications*, pp. 90–97, (1991).

[22] D. Roy, 'Learning visually grounded words and syntax for a scene description task', *Computer speech and language*, **16**, 353–385, (2002).

[23] D. Roy, K. Hsiao, and N. Mavridis, 'Conversational Robots: Building blocks for grounding word meanings', in *Proceedings of the Human Language Technologies Workshop on Learning word meaning from non-linguistic data*, (2003).

[24] R. Schank, 'What is AI anyway?', in *The foundations of Artificial Intelligence*, eds., D. Partridge and Y. Wilks, 3–13, Cambridge University Press, (1990).

[25] S. Shapiro and H. Ismail, 'Anchoring in a grounded layered architecture with integrated reasoning', *Robotics and Autonomous Systems*, **43**, 97–108, (2003).

[26] R. Srihari, 'Computational models for integrating linguistic and visual information: A survey', *Artificial Intelligence Review*, **8**(5/6), 349–369, (1994).

[27] R. Srihari, 'Use of captions and other collateral text in understanding photographs', *Artificial Intelligence Review*, **8**(5/6), 409–430, (1994b).

[28] R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal, 'Automating the linking of content and concept', in *Proceedings of the ACM Conference on Multimedia*, pp. 445–447, (2000).

[29] T. Winograd, *Understanding Natural Language*, Academic Press, 1972.

[30] T. Ziemke, 'Rethinking Grounding', in *Understanding Representation in Cognitive Sciences*, eds., A. Riegler, M. Peschl, and A. von Stein, 177–190, Plenum Academic/Kluwer Publishers, (1997).