# Image-Language Association: are we looking at the right features?

**Katerina Pastra**

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

**Abstract**

The ever growing popularity and availability of multimedia information has rendered automatic image-language association essential in a number of multimedia integration applications. Bridging the gap between the two media requires an appropriate feature-set for describing their common reference; one that will be both distinctive of the entities referred too and feasible to extract automatically from visual media. In this paper, we suggest an alternative –to current approaches- feature set, which has been used in OntoVis, a domain model for a prototype that describes three-dimensional (3D) indoor scenes. We argue that it is worth employing this feature-set in a larger scale for image-language association and investigating the feasibility of doing so and of detecting such features automatically even beyond 3D visual data, in 2D images.

## 1. Introduction

Internet Protocol Television, image and video-blogs and image and video-search engines are just a few of the latest technology-trends which become more and more popular, rendering digital multimedia content pervasive. Within such a context, the need for intelligent tools for efficient access to multimedia content has boosted research efforts and interest in automatic image-language association. The research issue is not new, of course; it spans a number of decades and a wide range of application areas, from Winograd's SHRDLU system in 1972, which verbalized visual changes in a 2D blocks scene, to medium translation systems (e.g. automatic sports commentators) and to conversational robots of the new millennium (cf. a review in Pastra and Wilks 2004 and Pastra 2005, ch.3).

Association in these multimedia prototypes took mainly the form of correlating visual information and accompanying text/speech or translating one modality into another (i.e. verbalizing visual information or visualizing linguistic information). In most cases, the systems made use of *a priori* known vision-language associations or used simple inference-mechanisms on small-scale association resources, resorting at the same time to either manually abstracted visual information or just working with miniworlds/blocksworlds. Lack of scalability and heavy human intervention for the task was among the most significant criticisms (cf. Pastra and Wilks 2004).

In this paper, we focus on the features used for describing/detecting common visual and linguistic references to entities in real-world scenes. We first look into the limitations of the features used in state of the art image-language association mechanisms, and then present three different types of features used in OntoVis, a feature augmented ontology for logic-based verbalization (description) of 3D indoor scenes in the VLEMA prototype (Pastra 2006). We discuss the possibilities of scaling the use of the suggested feature set and of detecting it automatically in 2D visual data.

## 2. Image-Language Association approaches

In the last few years, image-language association mechanisms as such are being developed for automatic image/keyframe annotation, with the vision of being, at some point, mature enough for being embedded in multimedia prototypes and mainly in indexing and retrieval prototypes. The approaches are either probabilistic (Barnard 2003, Wachsmuth et al. 2003) or logic-based (Dasiopoulou et al. 2004, Pastra 2005, ch. 5.). Learning approaches require properly annotated training corpora (Lin et al. 2003, Everingham et al. 2005) for learning the associations between images/image regions represented in feature-value vectors and corresponding textual labels, while symbolic logic approaches rely on feature-augmented ontologies (Dasiopoulou et al. 2004, Simou et al. 2005). Srikanth et al. (2005) report also on the use of both training corpora and ontologies for achieving automatic image annotation.

In all these cases, the features used for describing image content are low-level ones, such as shape, colour, texture, position (2D coordinates of image region), size (portion of image covered by image region), i.e. features used by image analysis components for automatic object detection. Justification of this choice is obvious: the need for relying on such features for automating object detection within image-language association tasks. However, is it a coincidence that all these approaches are exemplified in mini-worlds (e.g. soccer games, where the identification of the ball and the playground is quite straight-forward through shape and colour descriptors)? How distinctive could such features be for more complex objects, such as e.g. furniture (which comes in many different shapes, colours, textures) scenes (i.e. configurations of objects) and therefore, how scalable could the corresponding image-language association mechanisms be? Actually, what kind of object features/properties could one use, so that:

- they are distinctive of object classes (i.e. they allow differentiation among a large number of object types), and
- their values can be detected and used by an image analysis module for automatic object segmentation?

These questions lead back to an old problem in cognitive linguistics, that of the use of features for conceptual representation (Lakoff 1987; Barsalou and Hale, 1993); in meaning analysis/decomposition, feature-based methods define finite sets of conditions or attributes which determine the reference of a word. As pointed out in

criticisms of feature-based representations, no set of features can fully represent an entity, cf. ch. 7 in (Lakoff, 1987). However, from within many possible abstractions/features a certain feature-set can be more or less successful in fixing the reference of a concept.

## 3. The OntoVis suggestion

OntoVis is a domain model (domain ontology with corresponding knowledge-base) for interior scenes. It has stemmed out of OntoCrime, a domain ontology for indoor and outdoor scenes (Pastra et al., 2003), built through priming with the Common Data Model of the UK Police Information Technology Organisation (PITO), the latter being an attempt to standardize the wording used in all tasks that involve police forces. OntoVis includes the part of OntoCrime which refers to indoor scenes, augmented with properties for a number of entities that one can find in sitting-rooms in particular. The ontology is implemented in the form of a directed acyclic graph (DAG) through the use of the XI Knowledge Representation Language (Gaizauskas and Humphreys, 1996).

The same ProLog-based representation language is used for the OntoVis knowledge-base. The object-property assertions in the latter form a kind of "object-profiles" at the "basic-level" of categorization (Rosch 1978, Lakoff 1987), which cover for each object all following types of features/properties:

- *physical structure*:

the number of parts into which an object is expected to be decomposed in different dimensions, e.g. a sofa is always decomposed into more than one parts along its X dimension (each one corresponding to a seat) as opposed to a chair.

- *visually verifiable 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 and object parts (e.g. whether an object could be *on* the floor or not), the dimension according to which size comparisons would be meaningful etc.

Here is an example of the property profiles of two quite similar objects, both of which belong to the same class, that of "*furniture*":

| |
|---|
| props(sofa(X),[has_xclusters_moreThan(X,1)]). |
| props(sofa(X),[has_yclusters_equalMoreThan(X,2)]). |
| props(sofa(X),[has_ yclusters_equalLessThan(X,4)]). |
| props(sofa(X),[has_ zclusters_equalMoreThan(X,2)]). |
| props(sofa(X),[has_zclusters_equalLessThan(X,3)]). |
| props(sofa(X),[on_floor(X,yes)]). |
| props(sofa(X),[has_surface(X,yes)]). |
| props(sofa(X),[size(X,XCLUSTERS)]). |

Table 1: part of the "sofa" object profile

| |
|---|
| props(chair(X),[has_xclusters (X,1)]). |
| props(chair(X),[has_ yclusters_equalMoreThan(X,2)]). |
| props(chair(X),[has_ yclusters_equalLessThan(X,4)]). |
| props(chair(X),[has_zclusters_equalMoreThan(X,2)]). |
| props(chair(X),[has_zclusters_equalLessThan(X,3)]). |
| props(chair(X),[on_floor(X,yes)]). |
| props(chair(X),[has_surface(X,yes)]). |
| Props(chair(X),[size(X,XCLUSTER_YValue,TableYDIM _UpperConstraint)]). |

Table 2: part of the "chair" object profile

Looking at tables 1 and 2, one realises that the two objects (sofas and chairs) are similar in most of their properties; both of them intersect with the floor[1], they have a surface on which other objects may be placed, and they can structurally be decomposed into 2 or 3 parts in their Z dimension (these being the back, the seat+legs part that touches the floor, and optionally the arms, if there are any). Similarly, they can be decomposed into 2-4 parts along their Y dimension (back, seat, arms, and legs, the last two are optionally present). However they differ in their decomposition along their X dimension: a sofa has always more than one X-parts (more than one seats), while a chair may have only one seat. Size is a variable (changeable) property for sofas, and it is actually determined by the number of seats that the object has, while size for chairs makes normally sense only in terms of the height of the chair (e.g. short chairs for children, tall stool-like chairs etc.).

An "armchair" has the same object profile with a chair, apart from the fact that it will always have three or four Y-clusters (back, seat, arms and optionally legs), and always three Z-clusters (back, seat, arms), *i.e.*, arms are not optional, they must be present. Furthermore, an armchair's relative size does not make sense to be expressed in terms of its height; it is so for chairs, because they are expected to "co-locate" with tables/bars (the height of which may vary considerably), and the "chair's" height is constrained by a table's height. Table 3 presents part of the object profile of "tables":

| |
|---|
| props(table(X),[has_xclusters(X,1)]). |
| props(table(X),[has_yclusters(X,2)]). |
| props(table(X),[has_zclusters(X,1)]). |
| props(table(X),[on_floor(X,yes)]). |
| props(table(X),[has_surface(X,yes)]). |
| props(table(X),[size(X,YDIM,XDIM, Relative_to_Room_YXDIM)]). |

Table 3: part of the "table" object profile

A table has a surface (table-top) which can be identified along its X dimension, it has two yclusters (table-top and legs) and one zcluster (the whole table). Its length and height are relative to the corresponding dimensions of the room it is found in.

As seen in the above examples, there is a whole network of interrelations between objects in OntoVis, the detection and identification of each of which contributes

[1] The floor, as well as other room-parts/walls, is defined, in its turn, as the one-dimensional object (surface) with the lowest Y-values in an indoor-scene.

to the detection and identification of the other. The profiles include also assertions regarding the object parts objects are being formed of (and which are not included in the above tables due to space restrictions). For example, a sofa consists of a back, more than one seats and optionally legs and arms; these object parts are themselves defined in a similar way, using the property types suggested above.

There are many arguments in favour of the suggested feature selection for object naming in the literature. In particular, the need for defining objects through their physical structure and their functionality/purpose has been argued by many researchers, such as Minsky (1986). Structural properties were described by Minsky as ones which do not change “capriciously", while functional ones capture intentional aspects of the objects and both are important when defining visual objects. On the other hand, Landau and Jackendoff have explicitly argued that spatial representations imply properties of the objects involved (1993); for example, an “on" relation between two objects requires that the reference object is one with a surface or line boundary on which the figure object is located.

While not panacea, the suggested feature set could assist scaling image-language associations beyond mini-worlds, and actually allow for:

- going beyond differences in the appearance of similar objects (e.g. different styles of sofas) naming these objects in the same way, and
- generalizing over viewpoint differences e.g. identifying a sofa as such even when seen from the side (rather than *en face*)

These are generalizations that current image-language association algorithms cannot do easily (or at all). 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 (cf. e.g. the visual ontology by Zinger et al. 2005), if a similar example is not present in the training data. Similarly, current approaches cannot deal with viewpoint differences in the appearance of an object and there is an almost infinite number of different images of the same object which may result from differences in the viewpoint (viewing angle and distance) from which the object is seen in a complex scene.

## 4. Using OntoVis

While the effectiveness of each feature type individually has been argued in the literature, their use in conjunction and their incorporation in a domain model has not been attempted before. Actually, in the case of OntoVis, the feature set has been determined by the visual data itself, and the need to perform automatic object naming within an application scenario that goes from vision to language. OntoVis was created for the development of VLEMA, a system that attempts to test the extend to which one may currently “emancipate” a vision-language integration prototype, in order for the prototype to work with *real visual scenes*, to *analyze its visual data automatically*, and have *inference mechanisms for scalable vision and language association* abilities.

The VLEMA prototype works with automatically reconstructed in 3D images of sitting rooms. It includes a module that performs object segmentation in 3D space by extracting physical structure-related information (clusters of faces forming part of an object in each dimension) to detect objects and/or object parts. An object-naming module refines this detection results by either naming a candidate object or/and suggesting the clustering of candidate object parts into one object which it also names. The module relies on OntoVis for drawing inferences for object naming; the inference mechanisms take advantage of the rich visual information that can be extracted in the 3D space (i.e., 3D coordinates of the candidate objects, relative information on their spatial interrelations, size etc., as well as lack of occlusion, registration, viewpoint problems etc.) to check whether the property assertions in the OntoVis object profiles actually hold (cf. ch. 5 in Pastra 2005a and Pastra 2006). This means that the specific feature set suggested in the previous section stemmed out of a prototype that worked on 3D visual data, and it actually includes features that can be more easily identified in 3D space.

In Computer Vision, research on the automatic 3D reconstruction of real indoor and outdoor visual scenes, as well as on the automatic transformation of 2D images into 3D worlds points to optimistic prospects of taking advantage of the rich information one could extract in 3D space, in real-world application scenarios rather than merely in manually built virtual worlds. While OntoVis was used in such a real-world setting and it was applied on visual data that had been reconstructed in 3D automatically, these reconstruction mechanisms and the ones that transform 2D into 3D are still immature. The question then becomes, whether the OntoVis suggestion could be applied to 2D images, on which the vast majority of state of the art vision-language association mechanisms run.

While this is an issue that should be thoroughly explored with computer vision experts, there is some first evidence that automatic techniques for detecting such (or a reduced version of) visual information in 2D images exist. For example, there are methods for detecting spatial relations between objects in 2D images (cf. e.g. the work by Regier and Carlson, 2001), and there is also research on identifying object structure/parts in 2D images and associating them with textual labels (cf. Wachsmuth et al. 2003).

## 5. Future Plans

Currently, OntoVis includes “object profiles” for twenty basic-level objects (with their corresponding parts); our plans for the immediate future are to extend this resource to concrete-objects of indoor and outdoor scenes and test their discriminative power in a corpus of manually-constructed virtual reality scenes. Mechanisms for detecting the specified object features in these scenes automatically for object naming purposes will also be applied, as an extension to the work done in the VLEMA prototype.

Given the advantages that could be gained, we believe that it is also worth investigating the possibility of using the suggested feature set in state of the art image-language association mechanisms for 2D images; it is towards this

direction that we tend to head our research efforts towards.

## 6. Conclusions

In this paper we presented a feature-set for the representation of real world objects and scenes, within tasks that attempt to bridge the gap between low-level visual information and high-level (conceptual) linguistic descriptions of entities. The suggestion has been implemented in OntoVis, a domain model for building-interior scenes; the suggested features have been detected automatically in 3D visual data and have been used for the verbalization of this data. We argue that the feature set could be an alternative or complimentary one to feature sets used in state of art image-language association mechanisms and would like to invoke cooperation and collaboration towards this direction of research.

## 7. References


Barnard K., Duygulu P., Forsyth D., de Freitas N., Blei D., Jordan M. (2003), "*Matching words and pictures*", in Machine Learning Research, 3:1107-1135.

Dasiopoulou S., Papastathis V., Mezaris V., Kompatsiaris I., Strintzis M. (2004), "An ontology framework for knowledge-assisted semantic video analysis and annotation", in Proceedings of the International workshop on Knowledge markup and semantic annotation, International Semantic Web Conference.

Everingham M., Van Gool L., Williams C. and A. Zisserman (2005), "*PASCAL Visual Object Classes Challenge Results*", Technical Report, PASCAL Network of Excellence, http://www.pascal-network.org/challenges/VOC/voc/results_050405.pdf

Gaizauskas, R. and K. Humphreys, (1996), "*XI: A Simple Prolog-based Language for Cross-Classiffication and Inheritance*", In *Proceedings of the 7th International Conference in Artifficial Intelligence: Methodology, Systems, Applications*, pages 86-95.

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

Lakoff, G., (1987). "*Women, Fire, and Dangerous Things"*. The University of Chicago Press.

Landau, B. and R. Jackendoff (1993) "*What" and "Where" in spatial language and cognition", Behavioural and Brain Sciences*, 16:217-265.

Lin C., Tseng B. and J. Smith (2003), "*Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets*", in online Proceedings of TRECVID 2003.

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

Pastra K. (2006), "*An alternative suggestion for vision-language integration in intelligent agents"*, in Proceedings of the International Hellenic Artificial Intelligence Conference, Athens, Greece.

Pastra K. (2005), "*Vision-Language Integration: a Double-Grounding Case"*, PhD thesis, Department of Computer Science, University of Sheffield.

Pastra K. and Y. Wilks (2004), "*Vision-Language Integration in AI: a reality check*", in Proceedings of the 16th European Conference on Artificial Intelligence (ECAI), pp. 937-941, Valencia, Spain.

Pastra, K., H. Saggion, and Y. Wilks, (2003), "Intelligent indexing of crime-scene photographs", IEEE Intelligent Systems, 18(1):55-61.

Regier, T. and L. Carlson, (2001), "*Grounding spatial language in perception: An empirical and computational investigation*". *Journal of ExperimentalPsychology*, 130(2):273-298.

Simou N., Tzouvaras V., Avrithis Y., Stamou G., Kollias S. (2005), "*A visual descriptor ontology for multimedia reasoning*", in Proceedings of the Workshop on Image analysis for Multimedia Interactive Services (WIAMIS).

Srikanth M., Varner J., Bowden M., Moldovan D. (2005), "*Exploiting ontologies for automatic image annotation*", in Proceedings of SIGIR.

Wachsmuth S., Stevenson S., Dickinson S. (2003), "*Towards a framework for learning structured shape models from text-annotated images*", in Proceedings of the HLT-NAACL workshop on Learning word meaning from non-linguistic data.

Zinger S., Millet C., Mathieu B., Grefenstette G., Hede P., Moellic P. (2005), "*Extracting an ontology of portrayable objects from WorNet*", in Proceedings of the MUSCLE/Image CLEF workshop on Image and Video Retrieval Evaluation.