# IMAGE UNDERSTANDING AND SCENE MODELS: A GENERIC FRAMEWORK INTEGRATING DOMAIN KNOWLEDGE AND GESTALT THEORY

*N.Zlatoff* *B.Tellez* *A.Baskurt*

LIRIS (Laboratoire d'InfoRmatique, Image, Système d'information) CNRS FRE 2672
Université Claude Bernard Lyon 1
69622 VILLEURBANNE CEDEX FRANCE

## ABSTRACT

Because of the semantic gap between low-level descriptors and semantic concepts, image understanding systems need to use domain knowledge in order to map regions of a segmented image into semantic objects. However, such systems are often strongly dependent on an application since domain knowledge is not separated from procedures. We therefore argue for a generic grouping mechanism, based on the vision procedures of Gestalt theory, able to evaluate the relevance of its groupings with domain knowledge. In order to be used in a flexible way, such domain knowledge is stored independently in ontologies. We emphasize the need for domain knowledge to be thought as scene knowledge on a multi-level of composition, in order to be efficiently used. We also present the whole framework of our system and an application example within a specific domain: image understanding of thessalian graves' images.

## 1. INTRODUCTION

During the last ten years, constant growth of computers' storage capacity and heavy use of interconnected networks such as the Internet have lead to a huge amount of numeric data, especially images.

Indexing systems, allowing users to find images which are *relevant* to their queries are therefore strongly needed. Nowadays, a lot of content-based systems have come out, describing an image with low-level features such as color, texture or even shape of one main object (See [1] for a complete survey). However, such systems are usually not relevant as one may search an image based on what it depicts (its semantics) and not on its color or textural aspect.

Trying to extend content-based tools in order to derive semantics from low-level features may be impossible without any prior knowledge, since such a link does not exist without any ambiguity. This is what is often called the semantic gap.

Thus, we need to integrate additional knowledge related to the application domain, in order to find what a part of a picture is likely to depict in a given context.

### 1.1 Related work

A lot of knowledge-based image understanding systems is available. For instance SIGMA [2] or Schema [3] perform image understanding tasks on aerial images, based on several *descriptions* of objects which are bound to appear. Other works such as [4] deal with object recognition in macroscopic or microscopic biological images.

Nevertheless, as pointed out by [5] such systems are strongly domain-dependent as they integrate prior knowledge about the scene or several objects in the algorithms of image understanding. In fact, the domain knowledge is not clearly separated from the procedures.

That's why some works have risen about generic grouping algorithms: some of them try to group altogether some regions of a segmented image, regardless of any domain of application, based on the maximization of the groupings' likelihood [6]. Most of time though, Gestalt theory [7] is used, which argues that human vision performs domain-independent groupings (called *Gestalts*) based mainly on five properties: proximity, similarity, closure, continuity and symmetry (See Figure 1). These may reflect some statistical regularities of our visual environment. Thus, [8] and [9] implement a Gestalt-based grouping process of segment lines. In [8], gestalts' relevance is evaluated by the Dempster-Schafer theory of evidence, while [9] controls them using Markov Random Fields. In both cases however, the generic grouping algorithm is totally disconnected from the knowledge about objects it handles.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

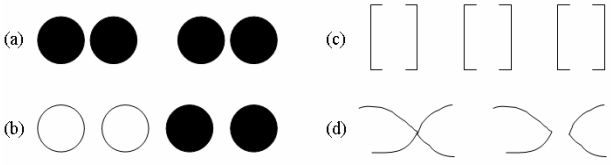


*Figure 1- Gestalt properties: patterns are likely to be grouped according to proximity (a), similarity (b), closure (c) and good continuation (d).*

### 1.2 Our solution

On the contrary, we argue that an image understanding system should be aware of (though not fully dependent on) what it treats. Consequently, we propose a generic process of grouping segmented regions, under the control of domain knowledge. Thus, the system is able to interface with different domain knowledge, depending on the image under treatment, and can extract from it all information needed to assist the grouping task (See Figure 2).

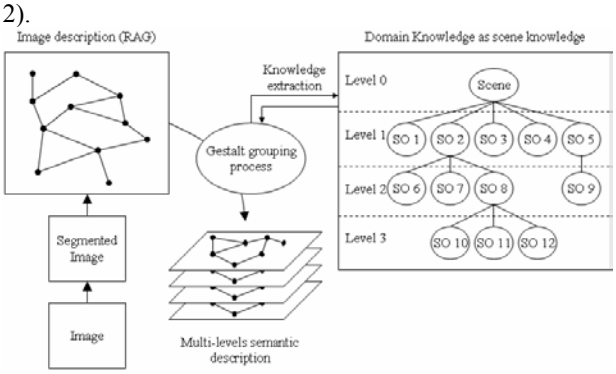


*Figure 2- Generic framework*

Such domain knowledge is modeled by ontology and should easily be mapped into the region-based description (that is a topological graph with additional descriptors). That's why domain knowledge should be modeled as scene knowledge, dealing about how objects appear in images. Besides, considering that semantic objects do not all deal with the same level of details (some are global components of the scene while others are more detailed components), a given object is not *relevant* at all levels of details and should thus not be perceived at all levels. Consequently, we model the notion of scene composition, in order to structure the scene. Hence, semantic objects are organized at different levels. Grouping process will then be performed at each level of the hierarchy, from the most global and then recursively at each sub-level. This process provides a finer context at each level of understanding and also leads to a multi-level description of the image.

## 2. GENERIC GROUPING ALGORITHM

Such an algorithm should allow the iterative grouping of regions from segmented images. The way the regions are grouped should be independent on its own from the domain while being able to extract from the ontology any information needed to assist the task. We propose a grouping mechanism based on Gestalt theory under the control of the domain knowledge.

Gestalt theory was introduced at the beginning of the XX$^{th}$ century by Wertheimer [7]. Gestaltists consider that during vision process, the whole is predominant in an image: we perform some grouping mechanisms mainly based on five properties: proximity, similarity, closure, continuity, symmetry. Nonetheless, such properties deal with quite high-level notions and their implementations are not straightforward.

According to Gestaltists, proximity is one of the most important grouping properties. That's why we choose to work first on a region adjacency graph (RAG) and try to reduce it by iteratively grouping adjacent regions based on several other Gestalt criteria. We implemented at this stage the similarity, the closure and the continuity properties in basic ways. These implementations should be detailed at a second stage.

More precisely, our grouping algorithm (based on [10] and extended in a Gestalt way) computes for each edge between two regions i and j of the RAG, a Gestalt distance $GD_{ij}$ which takes into account several Gestalt properties:

$$GD_{ij} = S_{ij} \text{ x } CL_{ij} \text{ x } CO_{ij}$$

$S_{ij} = \sqrt{\sum_{k=0}^{n} (d_{j,k} - d_{i,k})^2}$ where $d_{j,k}$ is the k$^{th}$ descriptor of region j and n the total number of descriptors. $S_{ij}$ measures the similarity between two regions: when two regions are similar from the descriptors' point of view (color, texture), $S_{ij}$ tends to 0.

$CL_{ij} = \frac{\min(P_i, P_j)}{4P_{ij}}$ where $P_i$ is the perimeter of region i and $P_{ij}$ the common perimeter between region i and j. Since $CL_{ij}$ was first introduced by Schettini in [11] in order to measure the mutual nesting between two regions, we considered it as a closure parameter from the Gestalt point of view.

The continuity property ($CO_{ij}$) is far more difficult to implement, as it involves the combination of several high-level concepts, such as shape and direction. We have chosen at this stage to use a basic criterion for continuity, based on size, which tends to better group a small region in another one:

$$CO_{ij} = \begin{cases} \varepsilon & if \quad (N_i < MinNbPix \quad or \quad N_j < MinNbPix) \\ 1 & \end{cases}$$

with $N_i$ the number of pixels of region i. In order to give $CO_{ij}$ the same weight as $CL_{ij}$, we set ε=0.25.

For each iteration, the regions linked to the edge with the minimum Gestalt distance are grouped. [10] suggests two ways for the algorithm to be stopped: all distances could be better than a fixed value or the total number of groupings less than a fixed parameter. Hence, this process involves one parameter (*MinNbPix*) and two control values: *MaxGD* (maximum Gestalt distance allowed) and *MinGest* (minimum number of expected gestalts). We propose that the system should extract from the domain ontology all the information needed to control the grouping process and to give a value to the several parameters.

## 3. ONTOLOGY OF DOMAIN KNOWLEDGE

Modeling a domain knowledge leads to create a domain ontology. Even if there is no unique way to create ontology, we shall conceive it regarding on its future use.

### 3.1 Modeling knowledge as scene knowledge

Here, the ontology should help to map regions of a segmented image into semantic concepts. Given that the only description one can automatically extract from images is a topological graph (augmented with low-level descriptors on each region), domain ontology should fit into such a description. That's why we argue that domain knowledge should be modeled as scene knowledge. Hence, the ontology is composed by concepts (semantic objects of the domain) and spatial relationships between these concepts (inclusion and adjacency with different sub-cases: top, bottom, right, left, right-bottom and so on). In addition, both semantic objects and spatial relationships are described taken into account their own properties.

As explained before, considering the domain knowledge from the scene point of view leads the notion of scene composition in order to know what object should be perceived given an understanding level.

### 3.2 Using domain knowledge to control grouping step

First of all, as the grouping process embodies vision knowledge, trying to group all regions into correct semantic objects seems to be unrealistic. We should rather aim at grouping several regions into perceptual groupings under the constraint of the domain knowledge. Hence, some of the semantic objects should still be decomposed into several regions. Consequently, we suggest two kinds of control: one during the grouping process and another one after, able to derive a more semantic description.

Concerning the grouping process, and given that it should occur at each level of composition of the scene, the system can infer from the domain knowledge some constraints the gestalts may check: minimum or maximum semantic objects, relative size, spatial relationships and shape. Hence, such constraints can be used to control the grouping. See Table 1 for examples.

| Parameter | Value |
|---|---|
| *MinGest* | (number of objects necessarily present)*2<br>OR: (number of objects bound to appear)*2 |
| *MinNbPix* | (Minimum size of object bound to appear)/10 |

*Table 1- Examples of parameters' settings*

The procedure to set *MaxGD* is more complex. Taking into account that *NbGroupMin* already prevents the grouping process from making too many iterations, we suggest to set *MaxGD* with:

$$MaxGD = \text{mean}(S_{ij}) + \alpha \, \text{std}(S_{ij})$$

Where std is the standard deviation and α reflects the granularity of expected grouping, set heuristically for each understanding level of a given domain.

Concerning the post-grouping control, we suggest extracting from the domain knowledge all geometric constraints that semantic objects should check, in order to infer a semantic interpretation of the gestalts.

Hence, we need some generic tools, able to automatically extract needed information from the ontology for each control. Such tools may correspond to first order logic, and could therefore be implemented using expert systems.

## 4. EXAMPLE OF APPLICATION AND CURRENT RESULTS

Our work is still in progress. However, we will present here some current results considering a specific domain: archaeology. More precisely, the *Maison de l'Orient et de la Méditerranée* (MOM) owns about 3000 images of thessalian graves issued from the digitization of photographs. A thessalian grave is a carved and painted stone, used in ancient times to show a burial place.

We have chosen to use texture as low-level descriptors since it can deal both with color and grayscale images. Using Laws texture description [12] followed by a clustering algorithm (K-Means), we obtain segmented images composed of about 600 regions.

Domain ontology has been modeled involving the MOM experts and leads to a frame-based model. These formalism groups the knowledge about a single concept into block called frame, with some properties (slots) optionally constrained by facets.

The ontology is composed of about 300 class frames, stored in the *Protégé-2000* ontology editor from the Stanford University [13]. Generic processes able to extract information from ontology have been implemented using the expert shell Jess, via a Protégé-2000 plug-in, allowing an automatic mapping between frames from the ontology and facts in Jess formalism. A set of 40 rules implements the generic extraction of constraints, and allows the settings of several parameters of the grouping step.

Experiments have been made on 20 images, involving a grouping step at the first level of detail. Segmented images are composed by about 600 regions, while post-grouping images own between 7 and 20 gestalts, relevant from a semantic point of view. Our grouping mechanism is thus able to significantly help the understanding process.

Table 2 shows several results (α=3.5; *MinGest*=7; *MinNbPix*=200) for 4 images in first level, displaying both segmented images and post-grouping images. In (a) and (b) grouping step has been stopped by *MaxGD* parameter, thus leading to more gestalts as in (c) and (d), which were controlled by *MinGest* parameter. At first level of understanding, three semantic objects were expected: the top component, the body component and *Geison*: thick component between top and body ones. Such component has been missed in (c) and (d) because of the low difference between its texture descriptors and its neighbor's ones. Note that, by extracting geometric constraints on semantic objects from ontology (shape), it may be possible to find all semantic objects from post-grouping images of (a) and (b). Moreover, artifacts present in body component in (d) or (a) could also easily be removed.

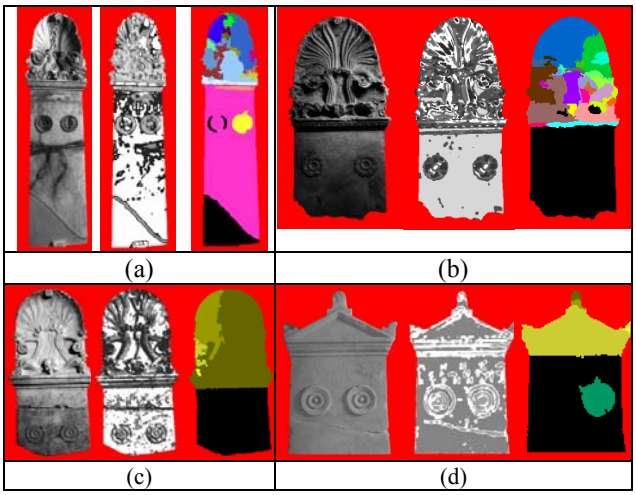


*Table 2- Results on first level of understanding*

## 5. CONCLUSION AND PROSPECTIVES

We have proposed a generic grouping framework, integrating Gestalt properties of vision under the control of domain knowledge, in a flexible way and show promising results on a specific domain.

We are currently working on more precise implementations of Gestalt properties, especially the continuity one involving a shape descriptor.

Moreover, we will also test our framework on other domains, such as aerial images.

## 6. REFERENCES

[1] A. Smeulders, M. Worring, S. Santini et al, "Content-Based Image Retrieval at the End of the Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22(12), pp. 1349-1380, 2000.

[2] T. Matsuyama, V. Hang, *SIGMA: A Framework for Image Understanding Integration of Bottom-up and Top-down Analyses*, Plenum, New-York, 1990.

[3] B. Draper, A. Colins, J. Brolio et al, "The Schema System", *International Journal of Computer Vision*, Vol. 2(3), pp. 209-250, 1989.

[4] C. Hudelot, M. Thonnat, "An Architecture for Knowledge-Based Image Interpretation", *Workshop on Computer Vision System Control Architecture*, Austria, 2003.

[5] D. Crevier, R. Lepage, "Knowledge-Based Image Understanding Systems: a Survey", *Computer Vision and Image Understanding*, Vol. 67(2), pp. 161-185, 1997.

[6] A. Amir, M. Lindenbaum, "A Generic Grouping Algorithm and its Quantitative Analysis", *IEEE Transactions on Pattern Analysis and Machine Acquisition*, Vol. 20(2), pp. 168-180, 1998.

[7] M. Wertheimer, "Principles of Perceptual Organization", *Readings in Perception*, pp. 115-135, 1958.

[8] P. Vasseur, C. Pégard, M. Mouaddib et al, "Perceptual Organization Approach by Dempster-Schafer Theory", *Pattern Recognition*, Vol. 32, pp. 1449-1462, 1999.

[9] A. Maßmann, S. Posch, G. Sagerer et al, "Using Markov Random Fields for Perceptual Grouping", Proc. Of *International Conference on Image Processing*, Vol. 2, pp. 207-210, 1997.

[10] K. Idrissi, G. Lavoué, J. Ricard, "Object of Interest based Visual Navigation, Retrieval and Semantic Content Identification System", *Computer Vision and Image Understanding*, 2004, in press.

[11] R. Schettini, "A Segmentation Algorithm for Color Images", *Pattern Recognition Letters*, Vol. 14, pp. 499-506, 1993.

[12] K. Laws, *Textured Image Segmentation*, PhD Thesis, University of Southern California, 1980.

[13] N. Noy, R. Fergerson, M. Musen, "The Knowledge Model of Protégé-2000: combining Interoperability and Flexibility", *Knowledge Engineering and Knowledge Management: 12th International Conference EKAW 2000*, *Lecture Notes in Artificial Intelligence*, Springer-Verlag, pp. 17-32, 2000.