A Computational Model to Connect Gestalt Perception and Natural Language

by

Sheel Sanjay Dhande

Bachelor of Engineering in Computer Engineering,
University of Pune, 2001

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN MEDIA ARTS AND SCIENCES

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2003

Author .....
Program in Media Arts and Sciences
August 8, 2003

Certified by .....
Deb K. Roy
Assistant Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Andrew Lippman
Chairperson
Department Committee on Graduate Students
Program in Media Arts and Sciences

A Computational Model to Connect Gestalt Perception and Natural Language

by

Sheel Sanjay Dhande

Submitted to the Program in Media Arts and Sciences
on August 8, 2003, in partial fulfillment of the
requirements for the degree of
MASTER OF SCIENCE IN MEDIA ARTS AND SCIENCES

Abstract

We present a computational model that connects gestalt visual perception and language. The model grounds the meaning of natural language words and phrases in terms of the perceptual properties of visually salient groups. We focus on the semantics of a class of words that we call conceptual aggregates e.g., *pair*, *group*, *stuff*, which inherently refer to groups of objects. The model provides an explanation for how the semantics of these natural language terms interact with gestalt processes in order to connect referring expressions to visual groups.

Our computational model can be divided into two stages. The first stage performs grouping on visual scenes. It takes a visual scene segmented into block objects as input, and creates a space of possible salient groups arising from the scene. This stage also assigns a saliency score to each group. In the second stage, visual grounding, the space of salient groups, which is the output of the previous stage, is taken as input along with a linguistic scene description. The visual grounding stage comes up with the best match between a linguistic description and a set of objects. Parameters of the model are trained on the basis of observed data from a linguistic description and visual selection task.

The proposed model has been implemented in the form of a program that takes as input a synthetic visual scene and linguistic description, and as output identifies likely groups of objects within the scene that correspond to the description. We present an evaluation of the performance of the model on a visual referent identification task. This model may be applied in natural language understanding and generation systems that utilize visual context such as scene description systems for the visually impaired and functionally illiterate.

Thesis Supervisor: Deb K. Roy
Title: Assistant Professor of Media Arts and Sciences

A Computational Model to Connect Gestalt Perception and Natural Language

by

Sheel Sanjay Dhande

Certified by . . . . . .

Alex P. Pentland
Toshiba Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Certified by . . . . . .

John Maeda
Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Acknowledgments

I would like to thank my advisor Deb Roy for his valuable guidance for this work. I would also like to thank the members of *cogmac*, especially Peter Gorniak, for helpful discussions and advice. I am greatful to my readers, Sandy Pentland, and John Maeda, for their helpful suggestions and advice.

I met a lot of great people at MIT and I would like to collectively thank them all. Finally, I would like to thank my family for their never ending support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Each day, from the moment we wake up, our senses are hit with a mind boggling amount of information in all forms. From the visual richness of the world around us, to the sounds and smells of our environment, our bodies are receiving a constant stream of sensory input. Never the less we seem to make sense of all this information with relative ease. Further, we use all this sensory input to describe what we perceive using natural language. The explanation, we believe, lies in the connection between visual organization, in the form of gestalt grouping, and language.

Visual grouping has been recognized as an essential component of a computational model of the human vision system [4]. Such visually salient groups offer a concise representation for the complexity of the real world. For example, when we see a natural scene and hear descriptions like, *the pair on top* or *the stuff over there*, intuitively we form an idea of what is being referred to. Though, if we analyze the words in the descriptions, there is no information about the properties of the objects being referred to, and in some cases no specification of the number of objects as well. How then do we dismabiguate the correct referent object(s) from all others present in the visual scene? This resolution of ambiguity occurs through usage of the visual properties of the objects, and visual organization of the objects in the scene. These visual cues provide clues on how to abstract the natural scene, composed of numerous pixels, to a concise representation composed of groups of objects.

This concise representation is shared with other cognitive faculties, specifically language. It is the reason why in language, we refer to aggregate terms such as *stuff*, and

pair, that describe visual groups composed of individual objects. Language also plays an important role in guiding visual organization, and priming our search for visually salient groups.

A natural language understanding and generation system that utilizes visual context needs to have a model of the interdependance of language and visual organization. In this thesis we present such a model that connects visual grouping and language. This work, to the best of our knowledge, is one of the first attempts to connect the semantics of specific linguistic terms to the perceptual properties of visually salient groups.

1.1 Connecting gestalt perception and language

Gestalt perception is the ability to organize perceptual input [31]. It enables us to perceive wholes that are greater than the sum of their parts. This sum or combination of parts into wholes is known as gestalt grouping. The ability to form gestalts is an important component of our vision system.

The relationship of language and visual perception is a well established one, and can be stated as, *how we describe what we see*, and *how we see what is described*. Words and phrases referring to implicit groups in spoken language provide evidence that our vision system performs a visual gestalt analysis of scenes. However, to date, there has been relatively little investigation of how gestalt perception aids linguistic description formation, and how linguistic descriptions guide the search for gestalt groups in a visual scene.

In this thesis, we present our work towards building an adaptive and contex-sensitive computational model that connects gestalt perception and language. To do this, we ground the meaning of English language terms to the visual context composed of perceptual properties of visually salient groups in a scene. We specifically focus on the semantics of a class of words we term as conceptual aggregates, such as *group*, *pair* and *stuff*. Further, to show how language affects gestalt perception, we train our computational model on data collected from a linguistic description task. The linguistic description of a visual scene is parsed to identify words and their corresponding visual group referent. We extract visual features from the group referent and use them as exemplars for training our model. In this

manner our model adapts its notion of grouping by learning from human judgements of visual group referents.

For evaluation, we show the performance of our model on a visual referent identification task. Given a scene, and a sentence describing object(s) in the scene, the set of objects that best match the description sentence is returned. For example, given the sentence *the red pair*, the correct pair is identified.

1.2 Towards Perceptually grounded Language understanding

Our vision is to build computational systems that can ground the meaning of language to their perceptual context. The term grounding is defined as, acquiring the semantics of language by connecting a word, purely symbolic, to its perceptual correlates, purely non-symbolic [11]. The perceptual correlates of a word can span different modalities such as visual and aural. Grounded language models can be used to build natural language processing systems that can be transformed into smart applications that understand verbal instructions and in response can perform actions, or give a verbal reply.

In our research, our initial motivation was derived from the idea of building a program that can create linguistic descriptions for electronic documents and maps. Such a program could be used by visually impaired and functionally illiterate users. When we see an electronic document we implicitly tend to cluster parts of the document into groups and perceive part/whole relationships between the salient groups. The application we envision can utilize these salient groups to understand the referents of descriptions, and create its own descriptions.

There are other domains of application as well, for example, in building conversational robots that can communicate through situated, natural language. The ability to organize visual input and use it for understanding and creating language descriptions would allow a robot to act as an assistive aid and give the robot a deeper semantic understanding of conceptual aggregate terms. This research is also applicable for building language enabled

intuitive interfaces for portable devices having small or no displays e.g., mobile phones.

1.3 The Big Picture

The diagram shown in Figure 1-1 shows our entire model. It can be divided into two stages. The first stage, indicated by the *Grouping* block, performs grouping on visual scenes. It takes a visual scene segmented into block objects as input, and creates a space of possible salient groups, labeled *candidate groups*, arising from the scene. We use a weighted sum strategy to integrate the influence of different visual properties such as, *color* and *proximity*. This stage also assigns a saliency score to each group. In the second stage, visual grounding, denoted in the figure by the *Grounding* block, the space of salient groups, which is the output of the previous stage, is taken as input along with a linguistic scene description. The visual grounding stage comes up with the best match between a linguistic description and a set of objects. The parameters of this model are learned from positive and negative examples that are derived from human judgement data collected from an experimental visual referent identification task.

1.4 Organization

In the next chapter, Chapter 2 we discuss relevant previous research related to perceptual organization, visual perception, and systems that integrate language and vision. In Chapter 3 we describe in detail the visual grouping stage, including a description of our grouping algorithm and the saliency measure of a group. In Chapter 4 we describe how visual grounding is implemented. We give details of our feature selection, the data collection task, and the training of our model. The chapter is concluded with a fully worked out example that takes the reader through the entire processing of our model, starting from a scene and a description to the identification of the correct referent. In Chapter 5 we give details of our evaluation task and the results we achieved. In Chapter 6 we conclude, and discuss directions of future research.

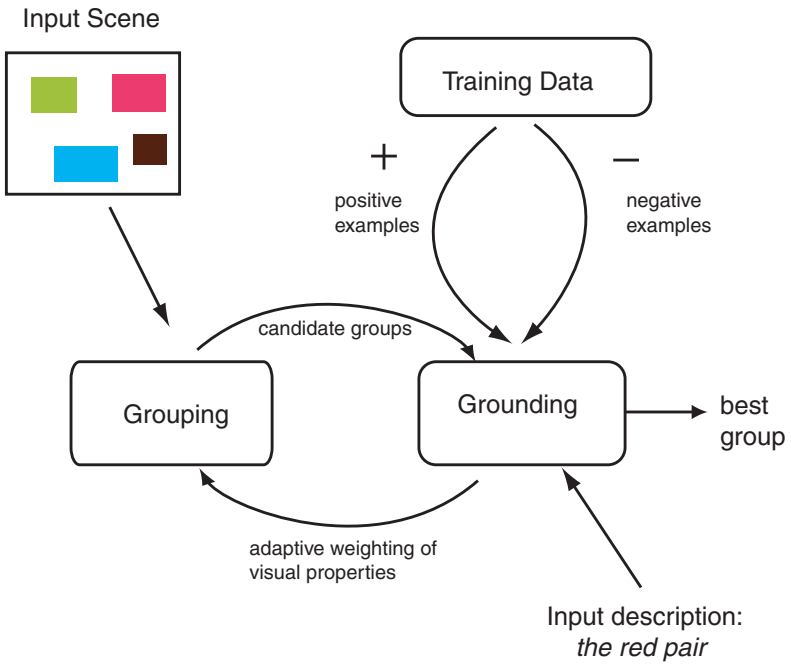


Figure 1-1: Model Architecture

1.5 Contributions

The contributions of this thesis are:

- A computational model for grounding linguistic terms to gestalt perception
- A saliency measure for groups based on a hierarchical clustering framework, using a weighted distance function
- A framework for learning the weights used to combine the influence of each individual perceptual property, from a visual referent identification task.
- A program that takes as input a visual scene and a linguistic description, and identifies the correct group referent.

Chapter 2

Background

Our aim is to create a system that connects language and gestalt perception. We want to apply it to the task of identifying referents in a visual scene, specified by a linguistic description. This research can be connected to two major areas of previous work. The first area is visual perception and perceptual organization and the second area is building systems that integrate linguistic and visual information.

2.1 Perceptual Organization

Perception involves simultaneous sensing and comprehension of a large number of stimuli. Yet, we do not see each stimulus as an individual input. Collections of stimuli are organized into groups that serve as cognitive handles for interpreting an agent's environment. As an example consider the visual scene shown in Figure 2-1. Majority of people would parse the scene as being composed of 3 sets of 2 dots, in other words 3 pairs rather than 6 individual dots. Asked to describe the scene, most observers are likely to say *three pairs of dots*. This example illustrates two facets of perceptual organization, (a) the grouping of stimuli e.g., visual stimuli, and (b) the usage of these groups by other cognitive abilities e.g., language.

Wertheimer in his seminal paper [30] on perceptual organization coined the term *Gestalt* for readily perceptible salient groups of stimuli, *wholes*, to differentiate from the individual stimulus, *parts*. He also postulated the following set of laws for forming groups from individual stimuli:

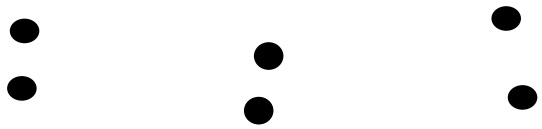


Figure 2-1: Three pairs of dots

- Proximity - two individually perceived stimuli that are close to each other are grouped together.
- Similarity - two stimuli that are similar along a perceptual dimension, tend to be grouped together.
- Common Fate - stimuli sharing a common direction of motion tend to be grouped together.
- Continuity - two stimuli that fit the path of an imaginary or discontinuous straight line or smooth curve tend to be grouped together.
- Closure - two stimuli are grouped together so as to interpret forms as complete. This is due to the tendency to complete contours and ignore gaps in figures.
- Figure ground separation - a set of stimuli appear as the figure (positive space), with a definite shape and border, while the rest appear as background (negative space).
- Goodness of form (*Pragnanz*) - a set of stimuli *a* will be perceived as a better group compared to another set of stimuli *b* if *a* is a more regular, ordered, stable and balanced group than *b*.

Though not stated as a separate rule, using a combination of these principles simultaneously to visually parse a scene, is another important and as yet unresolved [31] element

of perceptual organization. Wertheimer further gave examples from visual and auditory perception to show the applicability of these principles across different modalities.

Previous work related to perceptual organization has focused on computational modeling of the laws of perceptual organization. This involves creating a mathematical model that given all the current stimuli as input, will return groups of stimuli as output. The groups should be the same as formed by human observers. In the case of visual stimuli this amounts to forming groups from elements perceived in a visual scene. The formation of these groups could use one or all of the rules listed above. Each law of perceptual organization, in itself, has given rise to whole bodies of work of which we give a few examples here.

Zobrist & Thompson [33] presented a perceptual distance function for grouping that uses a weighted sum of individual property distances. We use a similar distance function, but have a different weight selection method that learns the probability of usage of specific weight combinations by training on human group selection data. This is discussed in detail in Section 3.3 and Section 4.2. Quantifying similarity, especially across different domains has been discussed by Tversky [27], and Shepherd [21], and has given rise to the set theoretic and geometrical functions for similarity judgment. In our work we use geometrical functions. A comprehensive discussion on similiarity functions can also be found in Santini & Jain [18].

For detecting curvilinear continuity, the concept of local saliency networks was introduced by Sha'shua & Ullman [20]. Curved shapes can be detected by defining a saliency operator over a chain of segments. Optimization over all possible chains using a dynamic programming method leads to the detection of salient curves. For the principle of figure-ground segmentation and object detection, work has been done by Lowe [14] to use visual grouping for object recognition, and by Shi & Malik [22] for scene segmentation using normalized-cuts. One of the first methods for quantifying goodness of a group, in information theoretic terms, was presented by Attneave [9]. This work quantified goodness of a group through a measure of the simplicity of the group (the smaller the description of a group in some encoding scheme, the simpler it is). Amir & Lindenbaum [1] introduced a domain independent method for quantifying grouping. It abstracts the set of all elements

into a fully connected graph, and uses a maximum likelihood framework to find the *best* partition of the graph into groups. Using perceptual organization principles in applications has been an area that has received relatively lesser attention. Saund [19] presents a system for deriving the visual semantics of graphics such as sketches using perceptual organization. Carson et. al. [5] present another example in which gestalt grouping analysis is used as part of an image query system.

The references listed above tackle identifying salient groups as a purely visual problem, and try to model characteristics of low level human vision. In this thesis we wish to present work that uses not only visual, but linguistic information as well, to identify salient groups. We however, abstract the grouping problem to a higher level, where object segmentation of a scene is given, and the problem is to combine objects into groups that are referred to in the scene description.

2.2 Language and Perception

Research in the field of language and perception attempts to identify how what we perceive is affected by linguistic context, and how word meaning is related to perceptual input. The close link between language and perception has been studied [cf. 29] and modeled [cf. 15] in the past Special attention has been devoted to research on computationally modeling the relation between language and *visual* perception. It involves work related to building systems that use visual information to create language descriptions, language descriptions to create visual representations or systems that simultaneously deal with visual and linguistic input [23].

In our work we are primarily interested in investigating how the semantics of language can be derived through visually grounding linguistic terms. The problem of building grounded natural language systems using visual context, has been addressed in previous work such as the work done by [12], [10], [17] and [16]. The VITRA system [12] is a system for automatic generation of natural language descriptions for recognized trajectories of objects in a real world image sequence such as traffic scenes, and soccer matches. In VITRA verbal descriptions are connected to visual and geometric information extracted

from the real-world visual scenes.

Gorniak & Roy [10] describes a system, Bishop, that understands natural language descriptions of visual scenes through visual grounding of word meanings and compositional parsing of input descriptions. It also provides a list of different strategies used by human subjects in a scene description task. In this list of strategies grouping is stated as an important but not fully resolved (in the paper) part. We have attempted to extend this work by specifically tackling the problem of understanding scene descriptions in which grouping is used as a descriptive strategy.

Roy [17] presents DESCRIBER, a spoken language generation system that is trained on synthetic visual scenes paired with natural language descriptions. In this work, word semantics are acquired by grounding a word to the visual features of the object being described. There is no *a priori* classification of words into word classes and their corresponding visual features, rather the relevant features for a word class, and the word classes themselves are acquired. The system further goes on to generate natural and unambiguous descriptions of objects in novel scenes. We have used a similar word learning framework as [17]. Work done by Regier [16] provides an example of learning grounded representation of linguistic spatial terms.

The main point of difference in our work is the handling of conceptual aggregate terms. Most language understanding systems handle linguistic input that refers to a single object, e.g., *the blue block*. We attempt to extend this work by trying to handle sentences like, *the group of blue blocks*, or *the blue stuff* using a visually grounded model.

How language connects with perception has been studied extensively in the fields of psychology, and linguistics as well. The connection of language and spatial cognition and how language structures space was discussed by Talmy [24]. In it he stated that language schematizes space, selecting certain aspects of a referent scene, while disregarding others. Tversky [28] discusses the reverse relation of how space structures language. She presents an analysis of the language used in a route description task with an emphasis on discovering spatial features that are included or omitted in a description. These analyses support the notion of a bi-directional link between language and vision. Landau and Jackendoff [13] explore how language encodes objects and spatial relationships and present a theory of

spatial cognition. They further show commonalities between parsing in the visual system and language.

2.3 Initial Approaches

In this section we discuss, in brief, our initial approaches, and the insights we gained that led us to define and solve our final problem. As mentioned previously in Section 1.2, the initial motivation for tackling the problem of grouping was from the perspective of building an application for document navigation and description. As we envisage the document description program, it will visually parse a given scene and then generate a natural language description of the scene, or read out information from a referent specified through a language description.

Documents can be handled by abstracting them to scenes composed of blocks encompassing each letter and image. In this abstract form the problem resolves to finding salient groups in the document that correspond to how a reader might segment the document. The ideal segmentation would result in a set of groups at different levels of detail. At the lowest level letters will cluster to form words, and at the highest level paragraphs will cluster to form articles, or sentences will cluster to form lists. The descriptor program[1] should perform the following four steps:

1. Segmentation
2. Grouping
3. Scene Hierarchy
4. Natural Language Generation

Of these four steps we worked on the first two, segmentation and grouping in web pages.

[1] A document discussing such a system can be found at http://web.media.mit.edu/∼sheel/publications/dyd-paper-final.pdf

2.3.1 Segmentation of web pages

To analyze a document from the visual perspective it must first be segmented into atomic units that form the lowest level of the document hierarchy. This lowest level could be the pixel level, or even the letter level, depending on context. Thus, segmentation involves clustering at multiple levels of detail. The web pages were captured as images and converted to black and white. Run length smoothing followed by connected component analysis [6] was used to create bounding boxes for entities in a document e.g., letters and images.

2.3.2 Grouping

We implemented four different grouping/clustering algorithms. The first algorithm by Thorisson [25] was used only on hand segmented images. The other three were used on images that went through the segmentation process detailed above. The three algorithms were (a) K-means, (b) Gaussian Mixture Modeling, and (c) Hierarchical clustering. Each took the block segmented document image as input and returned salient groups.

2.3.3 Discussion on initial approaches

Some of the issues that arose in our initial experimentation were:

- The need for a robust definition of similarity. Within the domain of input that we considered, black and white block segmented images, simple proximity between the centroid of two blocks was sufficient. But, for colored images with a greater variety in shape and size, this would not suffice. Hence, there was a need to define a distance function that combined the distances along various perceptual properties e.g., proximity, color, size, and shape.

- Multiple levels of grouping. The level of grouping is dependent on context, especially linguistic context. The level decides the granularity of the elements to be grouped e.g., pixels, or objects. With this condition in mind, hierarchical clustering stood out as the best method.

- The language used for visual description. In describing web page images the language used is composed of higher level terms such as *articles* and *lists*. These terms are domain specific higher level examples of words whose semantics imply a conceptual aggregate. This adds a further layer of complexity to the problem. To avoid that, in our final framing of the problem, the input visual scene was simplified to a randomly generated synthetic scene made up of blocks. Note, as this was coupled with increasing the number of perceptual properties used in grouping there was a trade off in simplifying one aspect of the problem while complexifying another.

These issues helped define our final problem - building a model that looks at visual scenes composed of block objects having random location, size, and color, and trying to identify the correct referent based on a linguistic description. In the next chapter we present details of the gestalt grouping stage of our model.

Chapter 3

A model of Visual grouping

Visual grouping refers to performing perceptual grouping on a visual scene. It enables us to see *gestalts*, i.e. create wholes from parts. In the grouping stage a visual scene is taken as input and the output is a search space populated by groups that are perceptually observed on viewing the scene. In this chapter we frame the grouping problem as an unsupervised classification (clustering) problem, discuss clustering algorithms, give reasons for our choice of using hierarchical clustering, and finally present the details of implementation.

3.1 Clustering algorithms

A visual scene can be viewed at different levels of granularity. At the most detailed level, it is an array of pixels, and at the coarsest level it is composed of objects that compose the foreground, and the background. Grouping occurs at all levels. This is the reason why neighboring pixels combine to form objects, and why objects combine to form groups. The forming of groups in the visual domain can be mapped to the forming of conceptual aggregates in the language domain. For example, a group of birds can be referred to as a *flock*. Many other words in language such as, *stuff* used in the phrase - *that stuff*, *pair* used in the phrase - *a pair of*, reveal semantics that classify such words as conceptual aggregates. Thus, the first step towards connecting language and gestalt perception is to identify the groups in a visual scene that correspond to the conceptual aggregates in language.

In our work we abstract the problem of grouping to object level grouping.We have

selected an experimental task in which a visual scene is composed of rectangular objects or *blocks* . The blocks are allowed to overlap. The problem of grouping can be stated as a clustering problem. Here, the aim is to classify each object as belonging to a class c_i from amongst a possible n classes. The value of n is unknown.

For modeling gestalt grouping we compared two specific classes of clustering algorithms Partitioning, and Agglomerative clustering [7]. Partitioning algorithms tend to form disjoint clusters at only one level of detail. In contrast, Agglomerative algorithms e.g., hierarchical clustering attempt to classify datasets where there is a possibility of sub-clusters combining together to form larger clusters. Agglomerative algorithms thus provide descriptions of the data at different levels of detail.

The ability to perform grouping on a scene at different levels of detail is similar to how our vision system does a multi-level parsing of a visual scene, as evident from the example of *a group of birds* being referred to as a *flock*. Due to this similarity hierarchical clustering was chosen. Another reason for our choice was the ability to define a goodness measure based on the hierarchical clustering algorithm that captures the gestalt property of stability of a group. We discuss this measure in detail in section 3.6.

Before using hierarchical clustering a distance function between objects needs to be defined. The distance function should be such that, if $distance(object_1, object_2) < threshold$, then $object_1$ and $object_2$ are grouped together. This criterion is related to similarity, because in our context similarity is inversely proportional to distance. We formally define our distance function in the next section.

3.2 Perceptual Similarity

To group two objects together there must first be a quantification of how similar they are. We define the problem of quantifying similarity as one of calculating the distance between two objects in a feature space. Similarity can then be extracted as it is inversely proportional to distance. Conversely distance values can be treated as a measure of dissimilarity.

Each object o is defined as a point in a feature space, the dimensions of which are equal to the total number of perceptual properties. For example, the location of an object o_i can

be represented by the object's x-coordinate (f_1) and y-coordinate (f_2) values, with respect to a fixed reference frame, as a vector,

$$\begin{bmatrix} f_1^i \\ f_2^i \end{bmatrix} \tag{3.1}$$

More generally, the perceptual property of an object o_i can be represented by a feature vector $\mathbf{x}$ of size l as,

$$\mathbf{x} = \begin{bmatrix} f_1^i \\ . \\ . \\ f_l^i \end{bmatrix} \tag{3.2}$$

We assume our feature space to be euclidean, hence the distance between two objects is given by,

$$d(o_i, o_j) = \left[\sum_{n=1}^{l} (f_n^i - f_n^j)^2 \right]^{1/2} \tag{3.3}$$

We have used features whose perceptual distances can be approximated well by a euclidean distance in the chosen feature space, and can be termed as metrics because they satisfy the minimality, symmetricity and triangle inequality conditions [27]. The euclidean assumption is invalid for some measurements of perceptual similarity [18], but in this work we do not lay any claim to the universality of our similarity functions beyond the specific domain.

Distance between two groups, denoted by g_n and g_m is defined as the minimum distance between any two objects from the two groups,

$$g_n - \{o_1, o_2, ..., o_v\} \tag{3.4}$$

$$g_m - \{o_1, o_2, ..., o_w\} \tag{3.5}$$

$$d(g_n, g_m) = min\{d(o_i, o_j)\} \qquad \forall i,\ j\ such\ that\ o_i \in g_n\ and\ o_j \in g_m \tag{3.6}$$

Properties	Features	
area	P	number of pixels covered by an object
color	L	lightness
	a	positive values indicate amounts of red and negative values indicate amounts of green
	b	positive values indicate amounts of yellow and negative values indicate amounts of blue
proximity	x	(centroid_x)
	y	(centroid_y)
shape	h	(height)
	b	(width)

Table 3.1: Perceptual Properties and Features

3.2.1 Perceptual properties and Perceptual features

We wish to make a distinction here between perceptual properties and perceptual features. In language we refer to perceptual properties e.g., shape, that are physically grounded to perceptual features e.g., height and width. A perceptual property is calculated from perceptual features. Our discussion until now has dealt with defining the distance between two objects along only one perceptual property e.g., area, color etc. However, an object can be defined by more than one perceptual property. In our model we have currently included four perceptual properties - *area, color, proximity, shape*. Table 3.1 lists the properties and the features associated with each. The individual perceptual property distances are defined in the following sections.

Proximity

Given two objects o_i and o_j, the contours of both objects are calculated. Let C_i and C_j be the set of all points belonging to the contours of o_i and o_j respectively. The distance is

given by,

$$d_p(o_i,o_j) = \forall_{i,j}, min\{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}\} \quad such\ that\ (x_i,y_i) \in C_i\ and\ (x_j,y_j) \in C_j \tag{3.7}$$

Area

Given two objects o_i and o_j, all pixels corresponding to an object are counted. Assuming o_i has a P_i pixel size, and o_j has a P_j pixel size,

$$d_a(o_i,o_j) = |P_i - P_j| \tag{3.8}$$

Color

For representing color we use the 1976 CIEL*a*b* feature space. This is a system adopted by the CIE[1] as a model for showing uniform color spacing in its values. It is a device independent, opponent color system, in which the euclidean distance between color stimuli is proportional to their difference as perceived by the human visual system [32]. The three axes represent lightness (L*), amounts of red along positive values and amounts of green along negative values (a*), amounts of yellow along positive values and amounts of blue along negative values (b*). The perceptual distance between any two colors can be found by calculating the euclidean distance between any two points in L*a*b* space. Given two objects o_i and o_j having color values, L_i, a_i, b_i and L_j, a_j, b_j respectively the perceptual color distance can be stated as,

$$d_c(o_i,o_j) = \sqrt{(L_i-L_j)^2+(a_i-a_j)^2+(b_i-b_j)^2} \tag{3.9}$$

The formula for conversion from r,g,b space to CIEL*a*b* space is given in Appendix C.

[1] Commission Internationale de L'Eclairage (International Commission on Illumination)

Shape

The shape of an object is quantified by its height to width, i.e. aspect ratio. This measure can be calculated for arbitrary shapes by considering the bounding box of a given object. For object o_i having height h_i, width b_i, and object o_j having height h_j and width b_j,

$$d_s(o_i, o_j) = |h_i/b_i - h_j/b_j| \tag{3.10}$$

3.2.2 The combined distance function

The final distance function needs to combine the individual property distances. Our approach to combining perceptual properties is using a weighted sum combination of the individual distances [33] for each perceptual property. The weight assigned to a perceptual property is denoted by w and the the weights for all properties is denoted as a vector $\mathbf{w}$. This approach assumes independence between perceptual properties. That is to say, area distance between two objects is independent of their color distance. In our model we measure area by calculating the number of pixels in the region covered by an object. This generalized method maintains the independence between area and shape, even though for rectangular objects area could be calculated using the height and width features that are used for shape.

The final distance between two objects o_i and o_j, having P perceptual properties is denoted by $d_{all}(o_i, o_j)$ can thus be defined as,

$$d_{all}(o_i, o_j) = [w_1 \dots w_N] \begin{bmatrix} d_1(o_i, o_j) \\ . \\ . \\ d_P(o_i, o_j) \end{bmatrix} \tag{3.11}$$

where,

$$\sum_i w_i = 1 \tag{3.12}$$

In our case,

$$P = 4 \tag{3.13}$$

$$d_1 = d_p, \qquad d_2 = d_a, \qquad d_3 = d_c, \qquad d_4 = d_s \tag{3.14}$$

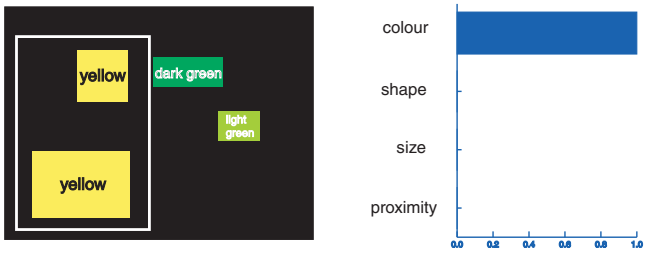
3.3 Weight Selection

Selecting different weight values, for distance calculation, leads to different sets of groupings. As an example, consider the visual scene in Figure 3-1.

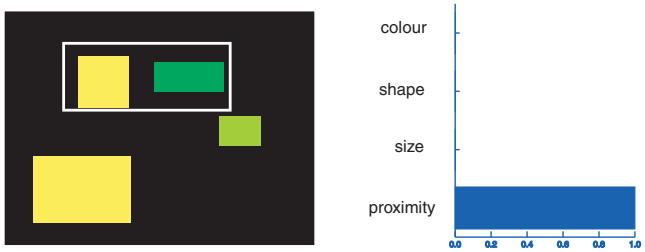
The objects can be grouped together using area to give one set of groups or they can be grouped together using color to give another set. Further still, a combination of color and proximity effects could give yet another set of groups. A point to note here is that none of the resultant groupings are intuitively wrong. On a subjective basis one may be chosen over the other giving rise to a probability for each weight vector. A weight vector having a higher probability will give a more intuitive *pop-out* grouping in comparison to a weight vector with lower probability. We formulated a method to derive the probability values associated with each weight vector using data collected from a group selection task. For each group selected by a subject, in a particular visual scene, we calculated the weight vectors that would generate the same group from our grouping algorithm. For example, if a subject chose the selected group (shown by a white bounding box) in 3-1(a), then all weight vectors that gave the same grouping as output from the grouping algorithm would be recorded. The probability distribution over weight space is created using these recorded weight vectors. Thus, selecting which of the groups is best is similar to asking the question, which of the weight vectors used to form each group is the best.

3.3.1 Psychological basis

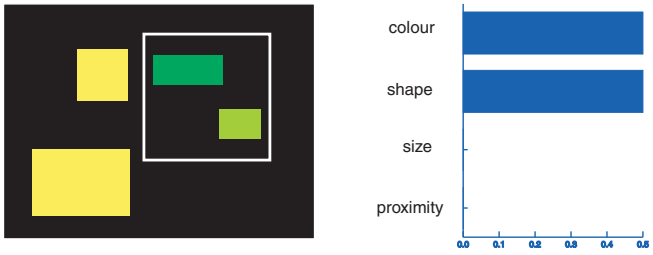
Triesman [26] conducted studies on referent search in visual scenes composed of objects that showed that search for a referent object in certain cases is parallel, hence very fast, while in other cases it is serial, and comparatively slower. Based on this evidence she put forward a visual model that created separate feature maps for salient features and



(a)



(b)



(c)

Figure 3-1: Effect of different weight vectors on grouping

showed that the parallel searches correspond to searches in only one feature map, while serial searches correspond to searches through multiple feature maps.

Our notion of weights for each perceptual property is related to Triesman's feature maps. For example, if a scene referent is identified only on the basis of color, in Triesman's model this would correspond to a search in the color feature map, and in our model this would correspond to a group formed with color having a weight of 1, while all other properties have a weight of zero. Thus, the most distinctive, *pop-out* group will correspond to weight vectors that have a non-zero value only for one of the properties color, or proximity, or area, or shape. This corresponds to the case where referent identification is done using a parallel search. The instance of a serial search is handled when the *pop-out* group corresponds to a more complex weight vector similar to a combination of feature maps. We have expanded on the definition of a pop-out by trying to identify not just a single object, but rather a set of objects, i.e. a group pop-out.

3.4 Notation

Before presenting further details, we wish to familiarize the reader with the following notation which will be used here on. Some notation used previously has been listed again as well, to provide a ready reference for all symbols used in the discussion to follow. Figure 3-2 contains visual examples of some of the notation presented here.

- o - an object in the visual scene
- $\mathbf{O}$ - $\{ o_1, o_2, \ldots , o_N \}$ set of all objects in a given scene
- g - $\{ o_1, o_2, \ldots , o_n \}$ a group made up of one or more objects
- $\mathbf{w}$ - $\{ w_1, w_2, \ldots , w_l \}$ a weight vector composed of weights for each perceptual property. All weight values lie between 0 and 1, and the sum of all weights w_i is 1
- $g_\theta^{\mathbf{w}}$ - $\{ g_1, g_2, \ldots , g_k \}$ a grouping made up of a set of groups created when weight vector is $\mathbf{w}$, and threshold value is θ

- $G^{\mathbf{w}}$ - $\{ g^{\mathbf{w}}_{\theta_1}, g^{\mathbf{w}}_{\theta_2}, \ldots, g^{\mathbf{w}}_{\theta_m} \}$ a grouping hypothesis composed of one or more groupings. Each weight vector **w** corresponds to a unique grouping hypothesis G

- π - $\{G^{\mathbf{w_1}}, G^{\mathbf{w_2}}, \ldots, G^{\mathbf{w_k}} \}$ the entire group space composed of all the grouping hypotheses generated from a specific visual scene

3.5 Hierarchical Clustering

We have used a hierarchical clustering algorithm to generate groups in a visual scene. The choice of this algorithm was motivated by the need to generate groups at different levels of detail, and further to know the relationship between groups at each level. This means generating groups having few objects, when analyzing the scene in detail, or generating groups with a large set of objects, when analyzing the scene at a coarser level.

The grouping module takes **O** as input and returns π as output. The hierarchical grouping algorithm for a given weight vector **w** is given in Figure 3-3.

The grouping model returns all possible groupings (composed of groups) over a range of weight vectors. We sample the value of each weight in the range $[0, 1]$, at intervals of 0.2 to come up with 56 different weight vectors. 56 weight vectors result in 56 grouping hypotheses ($G^{\mathbf{w}}$), together forming the group space π. This is the input to the Visual grounding module.

The reason for carrying this ambiguity to the next stage is to be able to utilize linguistic information. This is an instance of our strategy to resolve ambiguities by delaying decisions until information from all modalities has been collected and analyzed.

3.6 Estimating *pragnanz* of a group

Starting from the seminal papers on Gestalt theory, one of the hardest qualities to define has been the goodness of a group. In traditional literature this is referred to by the term *pragnanz*. Goodness of grouping gives a measure for ranking the set of all possible groups in descending order of quality.

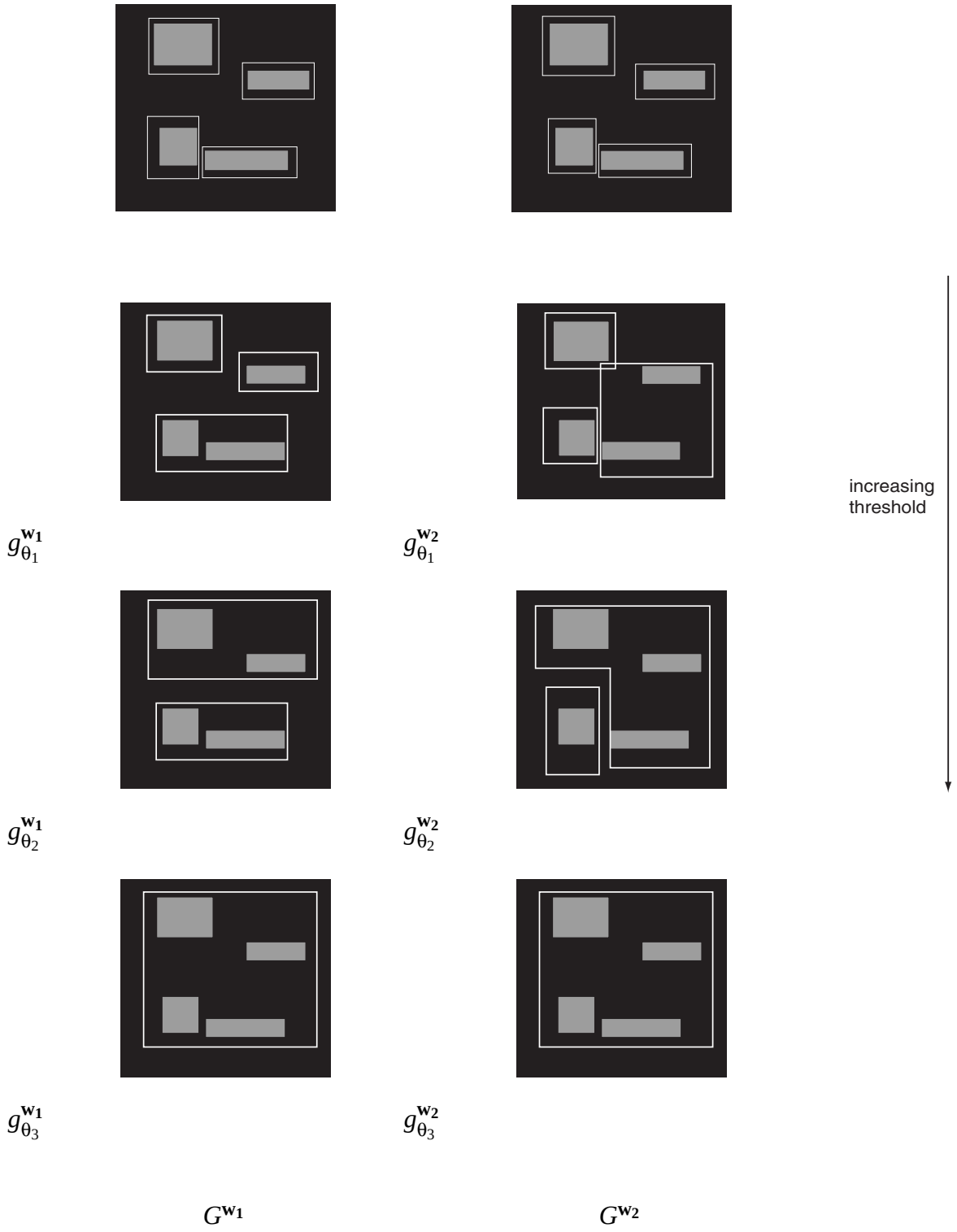


Figure 3-2: Stages of grouping for (a) weight vector w = [1 0] (proximity = 1, shape = 0), and (b) weight vector w = [0 1] (proximity = 0, shape = 1)

variables
$n \leftarrow$ number of groups
$\delta \leftarrow$ threshold increment
$\theta \leftarrow$ threshold
dist() $\leftarrow$ distance function

algorithm
n = N; $\delta = 0.1$; $g_i = \{ o_i \}$, i = 1, ... , N;
for $\theta = 0$ to 1
  $k \leftarrow 0$
  for i = 1 to n
    for j = 1 to n
      if(distance(g_i, g_j) $< \theta$)
        $g'_k \leftarrow$ merge g_i, g_j
        $k \leftarrow k + 1$
      else
        $g'_k \leftarrow g_i$
        $g'_{k+1} \leftarrow g_j$
        $k \leftarrow k + 2$
      end
    end
  end
  $n \leftarrow k$
  $\theta \leftarrow \theta + \delta$
  $g^{\mathbf{w}}_{\theta} \leftarrow \{ g'_1, \ldots , g'_n \}$
end
$G^{\mathbf{w}} = \{ g^{\mathbf{w}}_{\theta 1}, g^{\mathbf{w}}_{\theta 2}, \ldots , g^{\mathbf{w}}_{\theta k} \}$
return $G^{\mathbf{w}}$

Figure 3-3: Hierarchical Clustering

In our grouping algorithm a greedy approach is adopted, where the distance threshold value θ is iteratively increased to initially allow formation of small groups until finally the threshold is large enough to allow all objects to fall into one group. The amount by which the threshold value has to change before any two groups merge to form a larger group is called the goodness/stability of the grouping. The formal definition is as follows.

Assume that for a fixed weight vector $\mathbf{w}$, at a particular stage in the clustering process the distance threshold is θ_1 with an associated grouping $g^{\mathbf{w}}_{\theta_1}$. Further assume, on incrementally increasing the threshold to θ_2 the corresponding grouping is $g^{\mathbf{w}}_{\theta_2}$. If,

$$g^{\mathbf{w}}_{\theta_1} = g^{\mathbf{w}}_{\theta_2} \tag{3.15}$$

$$g^{\mathbf{w}}_{\theta_1-\delta} \neq g^{\mathbf{w}}_{\theta_1} \tag{3.16}$$

$$g^{\mathbf{w}}_{\theta_2+\delta} \neq g^{\mathbf{w}}_{\theta_2} \tag{3.17}$$

$$\tag{3.18}$$

then,

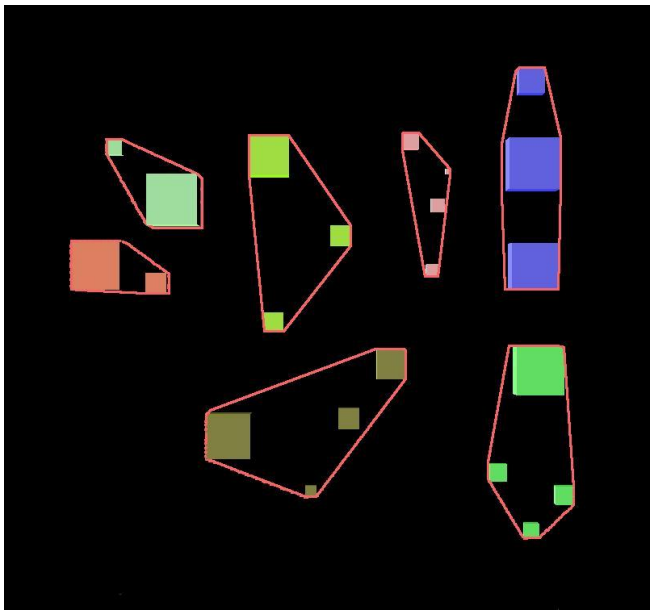
$$goodness(g) = (\theta 2 - \theta 1)/sizeof(g^{\mathbf{w}}_{\theta_1}) \qquad where \qquad g \in g^{\mathbf{w}}_{\theta_1} \tag{3.19}$$

the sizeof() function returns the total groups g in a grouping $g^{\mathbf{w}}_{\theta}$.

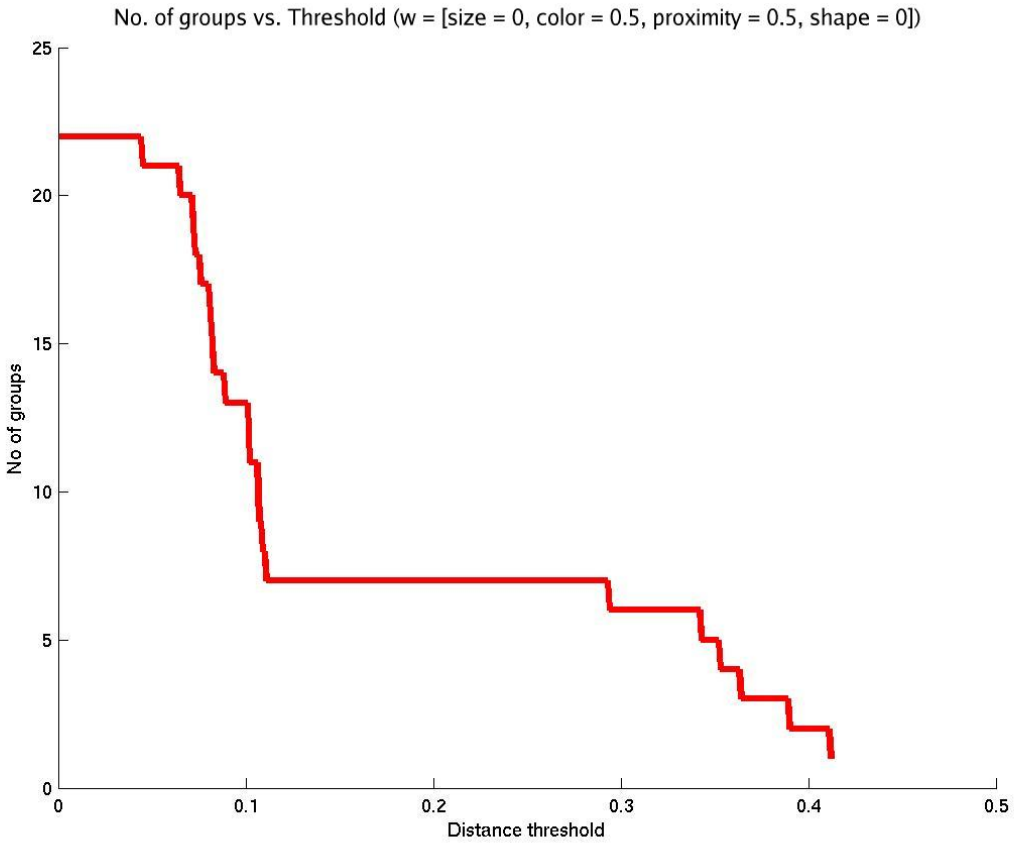
This value can be best described as measuring the gestalt property of stability, goodness, or *pragnanz* of a group. The figure 3-4 plots the formation of clusters versus the the distance threshold. The plateaus represent the points in grouping where changing the distance threshold did not result in the merging of any groups. This is the graphical representation of the goodness of a group. The adjoining image shows the grouping that corresponds to the longest, hence most stable plateau for number of clusters/groups equal to 7.

3.7 Summary

In this chapter we discussed the details of the grouping stage. The input to the grouping stage is a visual scene, composed of a set of block objects. The output, which is passed to



(a)



(b)

Figure 3-4: (a) visual scene, (b) corresponding stability curve

the visual grounding stage, is the set of groups formed using all the weight vectors (π), and a goodness of group/stability value corresponding to each group.

Chapter 4

Visual Grounding

We use the term grounding to mean deriving the semantics of linguistic terms by connecting a word, purely symbolic, to its perceptual correlates, purely non-symbolic [11]. Thus, the word *red* has its meaning embedded in color perception, as opposed to a dictionary, where red would be defined cyclically by other words (symbolic tokens).

The ability to form *gestalts* is a quality of human perception and in this chapter we present a methodology for grounding words to gestalt perception.

4.1 Word Learning

The problem we wish to solve is, given a novel scene and a description phrase, to accurately predict the most likely set of objects, i.e. a group, to which the phrase refers. An example of this is given in 4-1. Our solution involves training word conditional classification models on training pairs made up of a description and a group selection from a visual scene. As detailed in the following sections we first learn the semantics of individual words and then solve a joint optimization problem over all words in the input phrase to calculate a phrase conditional confidence value for each candidate group in the scene.

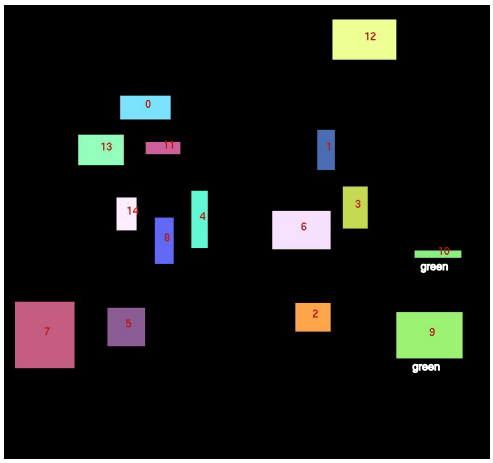


Figure 4-1: *S*: *the green pair on the right*

Word class	Words	Visual Features
grouping	pair, group, stuff	**w** (weight vector), *N* (number of objects in a group)
spatial	top, bottom, left, right	x, y (location)
color	red, blue, brown	L, a, b
area	large, small, tiny	h (height), b (width)

Table 4.1: Word classes, words, and visual features

4.2 Feature Selection

In our model we deal with a limited vocabulary set that is divided *a priori* into four word classes. Through data collected from subjects in a visual selection task, we wish to ground each term in our lexicon to its appropriate set of visual features. Table 4.1 lists the word classes, their corresponding visual features, and example words belonging to each class. The full list can be also be seen in Appendix A.

Our prime focus is on connecting the meaning of words to features of a group. For this purpose, in the following section we define the group features that are connected to a word (belonging to the grouping word class) and present our method for extracting those features from collected data and the current visual scene.

Grouping terms

Visual grouping, in our model, occurs by clustering objects using a weight vector with each element of the vector specifying a weight over a perceptual property distance. Different groups can be generated using different weight vectors. We believe this weighting ability is part of gestalt perception. Thus, the true perceptual correlate to a grouping term is composed of (a) the size of the groups e.g., a *pair* implies a group of size 2, and (b) the combination of goodness of a group as defined in Section 3.6, and the weight vector. This is defined in section 4.5.1.

The size of a group, and the goodness of a group are calculated in the visual grouping module. Here we discuss how to calculate the probability of usage of a weight vector. We choose groups selected by subjects in our collected Dataset and correlate the group selection with the weight combination that gives rise to the same group in our model. Figure 4-2 shows the probability of occurrence of a given weight combination in connection with a grouping selection. This surface has a maximum at proximity=1, color=0, area=0 (only those cases where the weight for shape is zero were counted for the purpose of visualization).

This provides evidence for the general intuition that proximity dominates the groupings that we form. We interpret each point on this surface as the likelihood that a particular weight combination was used to perform a grouping.

4.2.1 The composite feature, from object properties to group properties

Averaging of group features is not a good method as the average along certain feature sets does not accurately represent the perceptual average of that set. For example, the average color of a group composed of a red, green, and blue object would be a fourth distinct color, which is perceptually wrong. What we really wish to do, is pool the word conditional likelihood of all objects that form a group. The method we employ calculates individual word conditional probabilities for each object and takes a logarithmic sum over those values. The logarithmic sum is equivalent to estimating the joint likelihood over all

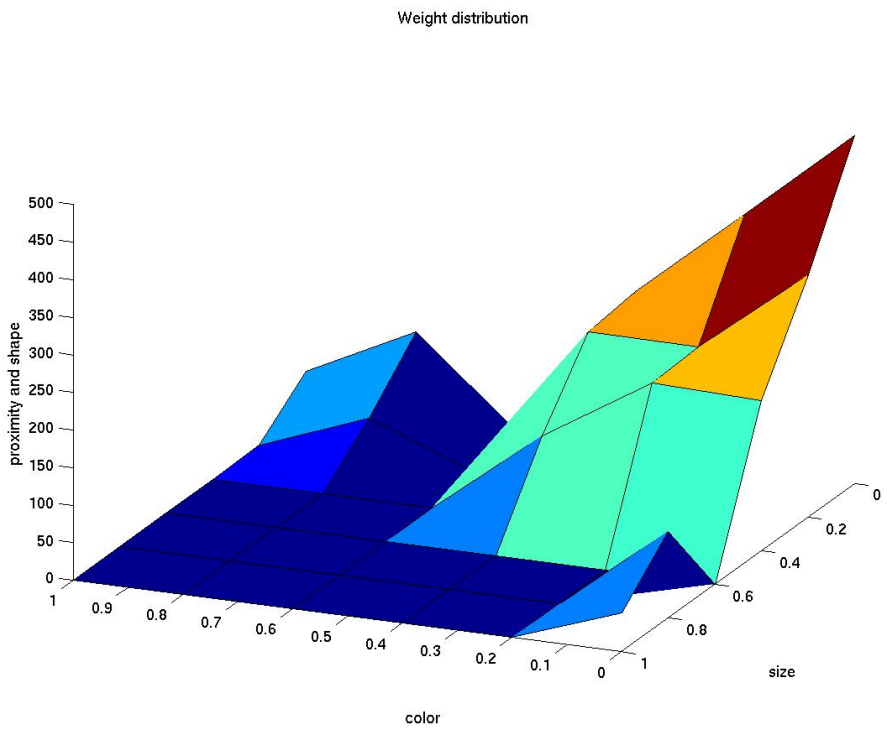


Figure 4-2: Distribution over weight space

objects in a group, where the word-conditional probability of each object is independent of all other objects. Thus, given a word t, and a group g composed of objects o_1, o_2, ... , o_n,

$$\log[p(g|t)] = \log[p(o_1|t)p(o_2|t)...p(o_n|t)] \tag{4.1}$$

4.3 The data collection task

Data was collected in the form of pairs of linguistic descriptions and visual scenes. Subjects were shown two windows on a computer screen, as shown in 4-3.

One window displayed a visual scene composed of 15 rectangular objects, and the other window contained buttons to select description phrases. The properties of the objects in the visual scene e.g., area, position, color were randomly chosen and overlapping was allowed. The subjects were asked to select using a mouse, a phrase from the second window and all sets of objects that matched the phrase description from the first window. A new scene

button could be clicked to proceed to the next scene. Each selected object was highlighted by a white boundary (the background color of all scenes was black). Deselection of an object was allowed. The subjects were told to make a selection if they felt an appropriate set of objects existed in the scene, otherwise to make no selection at all. A *no selection* case was recorded when a phrase was selected but no objects were selected, or a phrase was not selected at all.

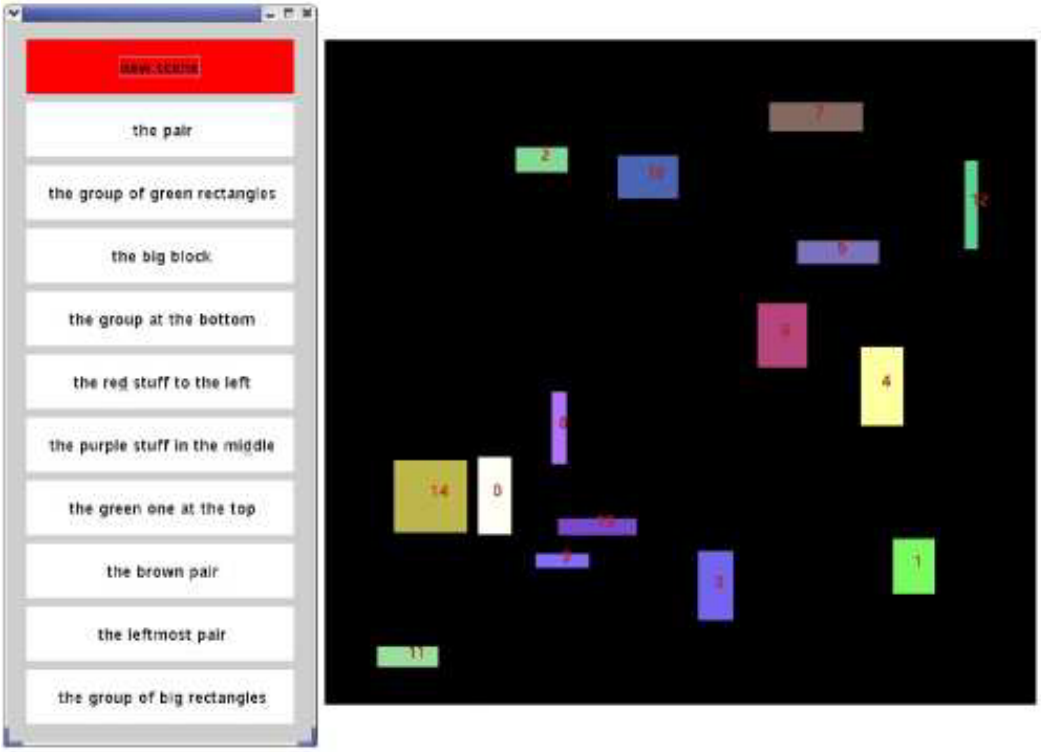


Figure 4-3: The data collection task

The phrases chosen for the task were constructed from words from our lexicon listed in Appendix A. Data was collected from 10 subjects, with each subject being shown a set of 10 phrases. The 10 phrases for each subject were chosen from a set of 48 unique phrases. Some phrases were shown to more than one subject. All the phrases can be seen in Appendix B. Along with the 10 phrases each subject was shown 20 different scenes, of which the first and last five were discarded, leaving 10 scenes, resulting in 100 phrase-scene pairs per subject. Aggregated over 10 subjects, this amounted to a total of 1000

unique phrase-scene pairs. As we use information from the *no selection* case to train our model, we have counted those cases as well. In the initial data collection it was noticed that there were insufficient exemplars for color terms, hence a 11th subject was used to collect data, but only for color description phrases e.g., *the brown one*. The selection of the number of objects in the scene (*15*) was done with an aim to elicit complex grouping choices. Appendix B contains a listing of all the phrases used according to subjects along with a histogram of word-class occurrence.

4.4 Grounding a word to its visual features

The collected data is composed of phrases and group selections. Each word in a description phrase is paired with the visual features of the objects in the corresponding group selection. Which visual features to associate with a word is based on the word class to which the word belongs. Formally, a phrase S_k is composed of words t_i, and is paired with a group selection g_k from scene I. Each word t_i is associated with a visual feature vector $\mathbf{x}$ that is dependent on the word-class of t_i. We wish to estimate a distribution over all examples of a word and the corresponding selection. This in essence is a distribution estimated over all the *good* examples of a word. For example, an object selected as *red* would be a good example of *red*, while one that is not selected as *red* would be a bad example of *red*. Thus we refer to the values estimated over all good examples as the *positive* model parameters for a given word. We use the maximum likelihood estimates of the Gaussian parameters to estimate a word-conditional distribution over the feature space corresponding to the semantics of the word t_i:

$$\mu_{t_i} = \frac{\sum_{k, w_i \in S_k} \sum_{j, o_j \in g_k} \mathbf{x}_j}{\sum_{k, t_i \in S_k} \sum_{j, o_j \in g_k} 1} \tag{4.2}$$

$$\Sigma_{t_i} = E[(\mathbf{x} - \mu_{t_i})(\mathbf{x} - \mu_{t_i})^T] \tag{4.3}$$

$$\theta_{t_i} = \begin{bmatrix} \mu_{t_i} \\ \Sigma_{t_i} \end{bmatrix} \tag{4.4}$$

For each word in our lexicon we also estimate a word-conditional distribution over all *bad* exemplars associated with a word. For example, in a scene where a subject chose a particular object as red, all other objects are classified as bad examples of a red object. The parameters of this distribution are referred to as the background model. Objects with features having a high probability value in the background distribution would qualify as *good* examples of an object that could be described as *not* red.

$$\mu_{\overline{t_i}} = \frac{\sum_{k,t_i \in S_k} \sum_{j,o_j \notin g_k} \mathbf{x}_j}{\sum_{k,t_i \in S_k} \sum_{j,o_j \notin g_k} 1} \tag{4.5}$$

$$\Sigma_{\overline{t_i}} = E[(\mathbf{x} - \mu_{\overline{t_i}})(\mathbf{x} - \mu_{\overline{t_i}})^T]$$

$$\theta_{\overline{t_i}} = \begin{bmatrix} \mu_{\overline{t_i}} \\ \Sigma_{\overline{t_i}} \end{bmatrix} \tag{4.6}$$

4.5 Selecting the best group, using good exemplars and bad exemplars

The decision on how well a word describes a group of objects is based on calculating two confidence values:

1. Confidence that a group of objects is a good exemplar for a word t_i,
2. Confidence that a group of objects is a bad exemplar for a word t_i.

Consider the case shown in 4-4. Even though there is no truly red object in the scene,

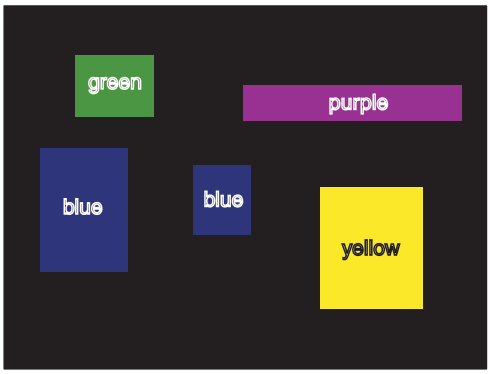


Figure 4-4: Detecting the absence of a red object

we will be forced to make an incorrect decision by choosing the object that has the highest rank, when sorted by confidence score. This raises the question of how to make a decision on the validity of the final answer given by our model. In the case shown in the figure, our model should be able to judge that an object is not red. Rather than building in a hard threshold value, we employ the information gained from the background model for a given word. Thus judging the validity of a group given a word can be framed as a two class classification problem, where the classes are good exemplars given t_i, and bad exemplars given t_i.

Once we have established the validity of a group, we rank the groups using the confidence value that it is a good exemplar for a given word. Our final answer for a group that corresponds to a given word will be the valid group that has the highest confidence value for being a good exemplar for the given word.

The validity of a group is based on satisfying the condition,

$$p(g_i|\theta_{t_i}) > p(g_i|\theta_{\overline{t_i}}) \tag{4.7}$$

where,

$$p(g_i|\theta_{t_i}) = \mathbf{N}(\mu_{t_i}, \Sigma_{t_i}) \tag{4.8}$$

$$p(g_i|\theta_{\overline{t_i}}) = \mathbf{N}(\mu_{\overline{t_i}}, \Sigma_{\overline{t_i}}) \tag{4.9}$$

Selecting the final answer can be formally stated as,

$$g = \underset{i}{\text{argmax}} \quad p(g_i|\theta_{t_i}) \qquad \forall i, g_i \in I \text{ and } g_i \text{ is valid} \tag{4.10}$$

4.5.1 The case of grouping terms

Recall from section 3.3 that each group formed can be a member of multiple groupings, each of which is associated with a goodness measure, and a weight vector. The corresponding weight vectors are used to create a distribution over weight space that gives the probability of a given weight vector being used to form a selected group, labeled a good group, the parameters of which are labeled θ_{t_i}, where t_i is a word belonging to the *grouping* word class. We also estimate a distribution over weight space of the probability of a given weight vector *not* being used to form a selected group, the parameters of which constitute the background model and are labeled $\theta_{\bar{t_i}}$.

We now combine the information from the positive and the negative examples to bias our final decision. Let $\mathbf{w}$ be the weight vector that corresponds to a grouping $g_\theta^{\mathbf{w}}$ of which group g is a part. The group g must satisfy the following conditions:

$$p(g|\theta_{t_i}) > p(g_k|\theta_{t_i}) \quad \forall k, g_k \in \pi, g \neq g_k \tag{4.11}$$

$$p(g|\theta_{t_i}) > p(g|\theta_{\bar{t_i}}) \tag{4.12}$$

As defined in the Chapter 3, π is the entire search space of all possible groups returned from the visual grouping module. We can calculate $p(g|\theta_{t_i})$ and $p(g|\theta_{\bar{t_i}})$ as follows,

$$p(g|\theta_{t_i}) = \sum_{i=1}^{56} p(g|\theta_{t_i}, \mathbf{w_i}) p(\mathbf{w_i}|\theta_{t_i}) \tag{4.13}$$

$$p(g|\theta_{\bar{t_i}}) = \sum_{i=1}^{56} p(g|\theta_{\bar{t_i}}, \mathbf{w_i}) p(\mathbf{w_i}|\theta_{\bar{t_i}}) \tag{4.14}$$

$p(g|\theta_{t_i}, \mathbf{w})$ = goodness of the grouping given word t_i

$p(\mathbf{w}|\theta_{t_i})$ = probability $\mathbf{w}$ was used to make a group selection found in the training data

$p(\mathbf{w}|\theta_{\bar{t_i}})$ = probability $\mathbf{w}$ was used to make a group selection that was not found in the training data

4.6 Accounting for word order in scene descriptions

Word order plays a role in determining the importance of a given word in a phrase. For example, *on the right of the middle one* versus, *in the middle of the right one* in the majority of cases will have two distinctly different referents. The problem can be resolved if each word can be assigned a rank that is a function of its position in the phrase. Gorniak in [10] alludes to how spatial and descriptive terms that occur closer to the noun terms seem to be more important for referent identification. Thus in the phrase *in the middle of the right one* the word *right* will have a higher rank and a filtering process will be performed shortlisting referents that more strongly qualify for the spatial phrase *on the right* and then among those referents look for the one that best qualifies for the spatial phrase *in the middle*.

Given an input phrase S composed of words t_i each having a weight α_i, phrase conditional confidence is defined as,

$$confidence(g|S) = \prod_i \alpha_i \log[p(g|t_i)] \tag{4.15}$$

The weights are a function of the distance of a word in a phrase from the noun term, and the total length of the phrase (only counting words in our lexicon). Thus, the weights α are calculated as,

$$\alpha_i = exp(-d) \tag{4.16}$$

where *d* is the distance in words counted from the noun term.

4.7 Backtracking

Most of the discussion above has dealt with resolving reference to what can be termed good groups. Ones that pop out for obvious reasons such as proximity, color etc. But, there are other cases in which grouping judgments do not correspond to the formation of good groups. For example, consider Figure 4-5 with the accompanying phrase description *the red pair*.

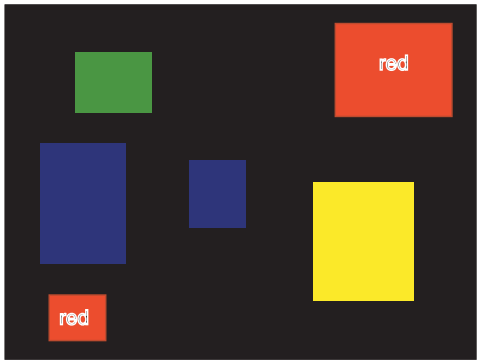


Figure 4-5: *The red pair*, a case handled by backtracking

Even though the two red objects are not a good example of a group (along the dimensions of area, proximity and shape), yet it is clear that they are the set of objects being referred to. This points towards the need to have a mechanism that in the absence of a good answer can backtrack, loosen its search restrictions, and re-populate the set of candidate groups. Our backtracking algorithm can be stated as follows,

search for best group;
if (a group is found as the best fit)
  return the group;
else
  if (restriction1 not removed)

remove restriction1 specified in equation 4.7 for all words excluding words belonging to the grouping word class;
search for best group;

else

if (restriction2 not removed)
remove restriction2 specified in equation 4.12 for all words belonging to the grouping word class;
search for best group;

else

return no valid referent in scene;

end

end

end

4.8 A dry run

To further explain the details of our model, in this section, we work through an example scenario. Consider the visual scene *I* shown in Figure 4-6 and the accompanying description phrase *S*. The processing steps are as follows:

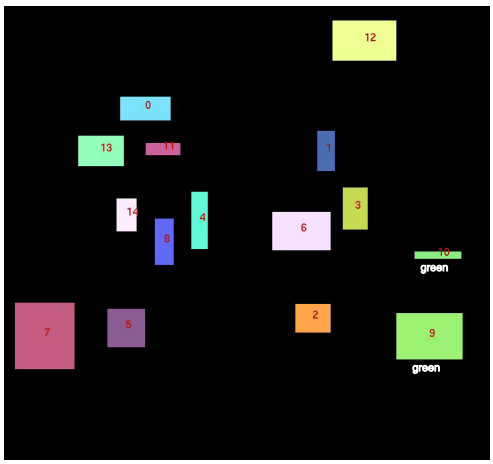


Figure 4-6: *S*: *the top left pair*

1. Visual grouping

   Input: Visual scene I

   Output: Group search space π

   Hierarchical clustering forms the full set of groupings. Two free parameters are changed to give rise to different groupings, the grouping threshold and the distance weight vector $\mathbf{w}$. The full set of groupings, that form the grouping search space π, is returned.

2. Word conditional confidence

   Let the features corresponding to a perceptual property of an object belonging to group g be represented by a feature vector $\mathbf{x_i}$ where o_i denotes the object. Taking the word *top* as an example, the word conditional probability for words belonging to the spatial, color or area word class can be stated as,

$$p(\mathbf{x_i}|\theta_{top}) = \mathbf{N}(\mu_{top}, \Sigma_{top}) \tag{4.17}$$

$$p(g|\theta_{top}) = \prod_i p(\mathbf{x_i}|\theta_{top}) \qquad o_i \in g \tag{4.18}$$

   Taking *pair* as an example, the word conditional probability for words belonging to the grouping word class can be stated as,

$$p(g|\theta_{pair}) = \sum_{i=1}^{56} p(g|\theta_{pair}, \mathbf{w_i}) p(\mathbf{w_i}|\theta_{pair}) \tag{4.19}$$

3. Semantic weighting based on word order

   The word order of a phrase gives an order for semantic resolution. Each word in the phrase is weighted according to its position in the phrase as follows (only words belonging to our specific vocabulary are counted),

$$S : the\ top\ left\ pair$$

$$word\_distance(top, pair) = 2$$

$$\alpha_{top} = exp(-word\_distance(top, pair))$$

4. Phrase conditional confidence values

   After calculating word conditional confidences, the evidence from each word conditional distribution is pooled together to create a composite phrase conditional confidence value,

$$confidence(g_i|S) = \log[p(f_i|top)] \cdot \alpha_{top} + \log[p(f_i|left)] \cdot \alpha_{left} + \log[p(f_i|pair)] \cdot \alpha_{pair} \tag{4.20}$$

5. Final selection

   For a group to be selected it must satisfy the following conditions

   (a) $p(g_i|\theta_{t_j}) > p(g_i|\theta_{\overline{t_j}})$, $\forall\ j$, where $t_j \in S$

   (b) $p(g_i|S) > p(g_j|S)$, $\forall\ j$, where $i \neq j$

6. Backtracking

   If the final selection procedure does not allow any groups to filter through, the search criteria are loosened and can be summarized as,

   - loosen restriction (a) from the Final Selection procedure with respect to spatial, color and area terms. Repeat Final Selection procedure
   - loosen restriction (b) from Final Selection procedure with respect to grouping terms. Repeat Final Selection procedure

Chapter 5

Evaluation

We evaluated the performance of our model on a visual identification task. The task measures the percentage accuracy of our model in identifying the correct referent group for a scene description. A decision is judged correct when it matches the decision of a human judge for the same scene and linguistic description pair. The method of data collection was previously detailed in Section 4.3.

5.1 Task details

The description phrases were composed of terms from four word classes, (1) grouping e.g., *pair*, (2) spatial e.g., *top*, (3) color e.g., *red*, and (4) size e.g., *large*. The entire vocabulary is listed in Appendix A and the the breakdown of phrases used for testing is listed in Appendix B.

We performed leave one out testing on our dataset. We employ this techniques so as to best utilize our limited amount of data and to present results that account for the individual subjectivity in the decisions made by each person from whom data was collected. Data collected from a particular subject was marked as testing and training was done using the data collected from the remaining 9 subjects. In this manner 10 separate evaluation results were calculated that are presented in the Section 5.2.

5.1.1 Evaluation function

We evaluate our model using two different criterion. In the first criterion, labeled *C1*, a decision is considered correct, when the decision made by a human subject is the same as the best decision returned by our model. When a human subject has made say *n* different decisions for a given description, then we compare each decision with the first *n* best groups returned by our model. We employ this method because, the order in which a group is selected by a human subject cannot truly be said to be an indicator of whether the group fits the given description better relative to other groups. It could be chosen first simply because the subject started parsing the visual scene from a particular spot in the scene.

Using the second criterion, labeled *C2*, a decision is considered correct, when the decision made by a human subject is in the set of all final possible groups returned by our model.

5.2 Results

To provide a comparative analysis we evaluated three methods for grouping. In method 1, labeled *MR*, for any given input phrase a set of objects is randomly chosen as the answer. A program was created to generate random answers (only one answer per description, hence it is compared using criterion $C1$ only), and the values shown are an average over 10 trials using *MR*. The theoretical probability for choosing a group in a scene is $1/2^n$, where n is the number of objects, as there can be 2^n possible subsets. The actual probability however is significantly increased by the availability of linguistic information. Method 2, labeled *MP*, is a variant of our model, in which distance is calculated using only one perceptual property, *proximity*. We implement this, by fixing the weight vector of our distance function to have a weight value of 1 for proximity and a weight value of 0 for all other perceptual properties. Method 3, labeled *MG*, uses our model with all weight vectors taken into consideration.

All results are shown for two sets, (a) $S1$, in which all answers given by a subject, including *no selection* answers are counted, and (b) $S2$, in which only those answers in which an actual selection was made are counted. The average percentage of selections made by subjects was 59.5% of the total phrase-scene pairs shown, and the standard deviation was

	*S*1			*S*2	
Subject	Random (*MR*)	Model (*MG*)	Proximity (*MP*)	Model (*MG*)	Proximity (*MP*)
1	2.2	46.6	57.3	26.1	30.3
2	3.5	28.9	36	9.1	10.4
3	4.5	32.7	45.5	16.7	16.7
4	5.4	24.1	37.5	6.8	21.6
5	3.6	22.3	20.5	11.4	6.8
6	1.3	15.8	9.9	15.8	9.9
7	6.1	36.6	42.6	34.4	26.3
8	4.1	26.2	34	14.7	19.6
9	5.3	22	27.27	8.3	7.1
10	4	10.9	18.8	9.6	14.4
Average	4	27.5	32.9	15.3	16.3

Table 5.1: Evaluation results per subject using criteria *C*1

	*S*1		*S*2	
Subject	Model (*MG*)	Proximity (*MP*)	Model (*MG*)	Proximity (*MP*)
1	49.5	59.2	39.1	39.1
2	36.8	36	20.7	10.4
3	57.4	54.5	41.7	25
4	42.9	42.9	35.1	29.7
5	33	21.4	25	7.9
6	24.8	15.8	24.8	15.8
7	37.6	42.6	39.4	28.1
8	46.6	42.7	49.2	34.4
9	25	27.3	13.1	7.1
10	31.7	26.7	34.9	24.1
Average	38.5	36.9	32.3	22.2

Table 5.2: Evaluation results per subject using criteria *C*2

26. This shows that the selection process in our task has a lot of subjective variability as regards the presence or absence of a described set of objects.

Table 5.1 shows the percentage accuracy results for the referent identification task, using evaluation criterion *C*1. Proximity alone does better in this evaluation. This indicates that the effect of proximity on group selection is disproportionately larger compared to other perceptual properties. Table 5.2 shows the percentage accuracy results using evaluation criterion *C*2. In this evaluation *MG* does better than *MP*. The better performance reflects the accuracy with which using multiple perceptual properties helps to shortlist the correct referent group. In Figure 5-1 we present average values for the results using eval-

uation criteria $C1$ with error bars (one standard deviation above and below). Figure 5-2 presents a similar graphic for evaluation criteria $C2$.

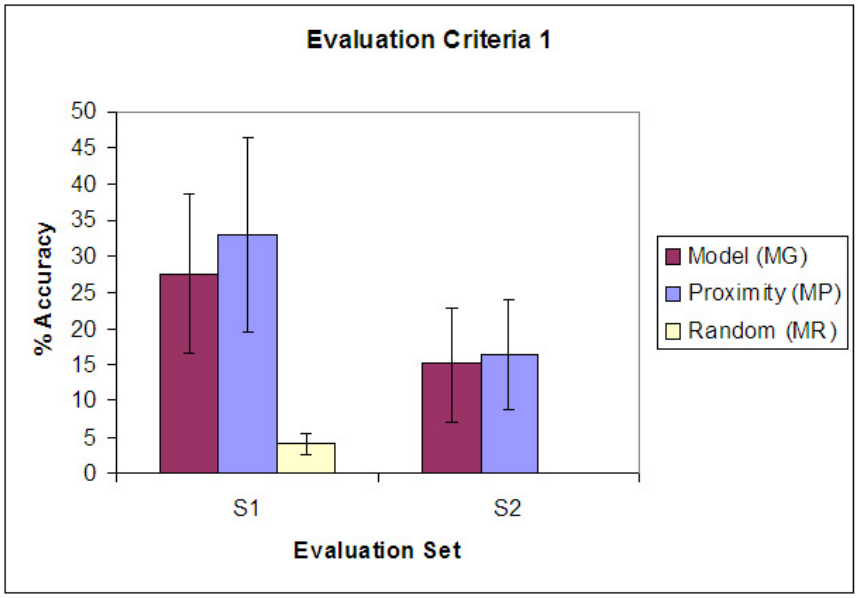


Figure 5-1: Average values of results calculated using evaluation criteria $C1$

5.3 Discussion

Our dataset was composed of randomly created synthetic scenes, with all visual properties selected from a continous feature space. Hence, regular patterns or groups of objects with absolute similarity along a visual property occurred very few times. This imposed a greater degree of hardness on the task, that is reflected in the low accuracy values in the results.

The results show that as expected our model performs better than random selection *MR*. The similar performance of *MG* and *MP* indicates the influence of proximity in the integration of visual properties. The possible conclusions to be drawn from this result are that, (a) proximity alone is enough for a model of grouping, and (b) that proximity plays a disproportionately larger role in influencing a grouping decision but does not suffice by itself. The first explanation is invalid because simple test cases can be constructed in which proximity will be unable to pick out the right group. Such an example is given in Figure 5-3. For the linguistic description *the pair*, the most intuitive choice is selecting the two

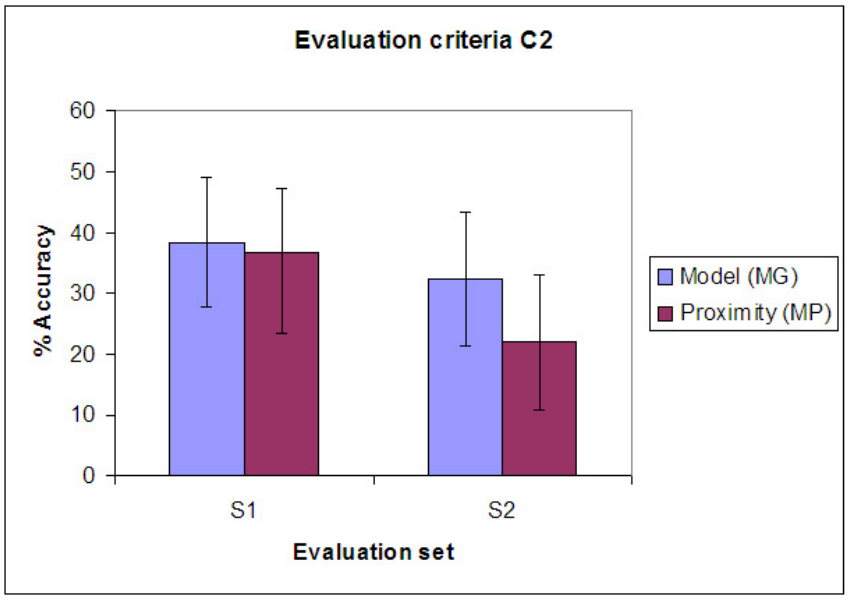


Figure 5-2: Average values of results calculated using evaluation criteria *C*2

white blocks. But any proximity measure will group the red and white block together first, thus preventing the formation of the correct pair. As we used randomly generated scenes the probability of creating two objects with the same shape, size and color was very low. In the absence of strong similarity along these perceptual properties, proximity is used as the most stable visual property.

Figure 5-3: Visual grouping scenario in which proximity alone fails

The second conclusion, that proximity plays a disproportionately larger role is in accordance with our results. Our initial hypothesis, that preference for different weights, derived from the collected data, will indicate a preference for using proximity is evident in the distribution derived over all weights shown in Figure 4-2. Our results indicate though

that the bias towards proximity, in our model, is not in a proportion sufficient enough to show significant gains over using proximity alone. This could have resulted due to lack of sufficient training data. Another reason for this could have been the inability, within the framework of our task, to assume a direct causality between a particular weight combination and a grouping. Hence, all weight combinations that produce the grouping must be taken into account, thus spreading out the derived weight distribution. In future, we wish to explore methods for learning the weight distribution with a mutual inhibition mechanism, so as delineate the effect of a perceptual property relative to all others. Ecological statistics collected for gestalt grouping phenomenon [8], [3] also support the explanation, that proximity is more important than other perceptual properties. As a future direction, we wish to devise an experimental task to collect data from which we can elicit more accurate statistics for how the influence of individual perceptual properties is integrated, and use that data for building our distribution over weight space.

The difference in results when using evaluation set $S1$ and $S2$ indicates that our model performs well in selecting the absence of a correct referent. This provides a measure of performance for the method of using a positive and a background model for word conditional classifiers to judge the validity of a referent for a scene description.

As a first step towards modeling the interdependence of language and gestalt grouping, these results show promise, and provide a critical analysis of the various challenges posed by the problem.

Chapter 6

Conclusion

6.1 A summary

We have devised and implemented a computational model that grounds linguistic terms to visual gestalt perception, and uses data collected from a linguistic description task to train a program to identify salient groups in a visual scene.

We presented our model in two parts, (1) visual grouping, and (2) visual grounding. In visual grouping, taking a visual scene as input, we implemented a hierarchical clustering algorithm to form a hypothesis set of groupings (π). As our distance function, we used a weighted combination of the distances between two objects, along individual perceptual properties e.g., color. We created a distribution over weight combinations from data collected in an experimental visual group selection task. This distribution represents the confidence that a given weight combination is used to form a salient grouping. We defined a stability measure for quantifying the goodness of a grouping as, the change in distance threshold value from the formation of a group to its merging with another group during the visual grouping of a scene. We introduced an overall saliency score of a group that combines the goodness of a group with the confidence of the weight combination that was used in forming the group. This gives us a hypothesis set of all groups with a saliency score for each group.

The second part of our model, visual grounding, dealt with the problem of resolving the semantics of a scene description. The entire lexicon of words is divided *a priori* into four

word classes. Each word class has a vector of features associated with it. We learn word conditional probability models over scene-linguistic description pairs. For each word, we create a positive model trained on the features of objects that are selected as exemplars for the word, and a background model over the features of all objects that are not selected as exemplars for the word. These two models are used to handle cases in which a scene description does not have a valid referent in the scene e.g., the description *the red pair* for a scene that has no red objects. For handling word order and calculating a phrase conditional confidence for a group, a weighted pooling of all word-conditional probabilities is performed. The group with the highest phrase conditional confidence is returned as the answer. For handling cases where the initial processing does not return an answer, a backtracking procedure was implemented that loosens the semantic constraints on the search through the set of all candidate groups.

We evaluated and presented the results of the performance of our model on a visual referent identification task.

6.2 Future Work

Following are some directions for future work towards extending the ideas in this thesis:

- A more comprehensive model of gestalt grouping that will utilize all the laws of perceptual organization. This involves a two-fold challenge of implementing a detection function corresponding to each law and further Devising a framework for combining such functions, a first version of which is the weighted sum combination technique used in this thesis. We also plan to improve the method for learning weight preferences so as to take into account phenomenon such as the disproportionately large importance of proximity in grouping. As a step in that direction we wish to create a new experimental task in which a grouping can be unambiguously connected with the perceptual property used in the grouping.

- Our intention is to be able to build a program for visual parsing and description of natural images, specifically document images. For this purpose a direction for the

future is building a low level image processing module for segmenting scenes before forming gestalt grouping. The efficiency of the segmentation module will determine how complex a natural image can be handled.

- Currently our program performs natural language understanding. We wish to extend beyond this and use our visually grounded model for natural language description. One prospective step in this direction is to implement the techniques presented in [17] for adding the capability of forming scene descriptions.

6.3 Contributions

We have presented a computational model for grounding linguistic terms to gestalt perception. The model is a first step towards building a visual scene description understanding and generation system. The methods presented in this thesis can be extended to handle more complex images e.g., web pages, and other types of electronic text documents. As part of our model we also introduced a saliency measure for groups based on a hierarchical clustering framework, and a weighted distance function. We used adaptive weighting of visual properties in our model, where the probability of usage of a weight combination adapted to human judgement data collected from a visual group selection task. Our model was implemented as a program that can take a visual scene as input and identify the correct group referent. This functionality when transplanted into a visual description tool, such as a document describer for the visually impaired, would allow an interactive, intuitive and intelligent interface for accessing and navigating the document using natural language.

Appendix A

Lexicon

Word class	Word		
grouping	pair	triplet	group
	stuff	ones	one
	blocks	block	
spatial	middle	top	bottom
	left	right	rightmost
	topmost	leftmost	
color	red	green	blue
	purple	yellow	brown
	pink	grey	orange
	white	black	violet
size	large	small	tiny
	big		
other	on	the	in
	rectangles	of	to
	towards	at	

Appendix B

Description Phrases

Subject	Phrases	
Subject 1	the pair in the middle	the large green rectangle
	the blue pair in the top right	the stuff at the bottom
	the triplet to the left	The red pair
	the rightmost pair	the group of blue ones
	the pair of small rectangles	the small rectangle on the left
Subject 2	the pair in the middle	the blue rectangle
	the green pair on the right	the stuff at the top
	the triplet to the left	The red pair
	the topmost pair	the group of red ones
	the pair of tiny rectangles	the small rectangle towards the top
Subject 3	the pair in the middle	the pair at the top
	the small rectangle towards the top	the pair of tiny rectangles
	the pair at the bottom	the small pair on the left
	the red pair	the rightmost pair
	the group of red ones	the large block
Subject 4	the pair	the group of green rectangles
	the big block	the group at the bottom
	the red stuff to the left	the purple stuff in the middle
	the green one at the top	the brown pair
	the leftmost pair	the group of big rectangles

Subject	Phrases	
Subject 5	the stuff to the left	the pair towards the bottom
	the violet pair	the rightmost block
	the blue pair in the middle	the yellow one
	the group to the left	the topmost pair
	the small block	the left one on top
Subject 6	the large pair	the pair of small rectangles
	the tiny blocks at the top	the rightmost triplet
	the middle one	the pair in the top left
	the grey stuff	the group
	the triplet	the topmost pair on the right
Subject 7	the red stuff to the left	the stuff at the top
	the pair in the middle	the big block
	the large pair	the green one at the top
	the brown pair	the small pair on the left
	the group	the stuff at the bottom
Subject 8	the group at the bottom	the group of big rectangles
	the blue pair in the middle	the rightmost triplet
	the pair of tiny rectangles	the tiny blocks at the top
	the pair at the bottom	the pair at the top
	the grey stuff	the group of red rectangles
Subject 9	the middle one	the triplet to the left
	the pair in the top left	the large block
	the violet pair	the purple stuff in the middle
	the green pair on the right	the group of green rectangles
	the pair of small rectangles	the small block
Subject 10	the topmost pair on the right	the triplet
	the group of blue rectangles	the rightmost block
	the red pair	the pair towards the bottom
	the pair	the group to the left
	the yellow one	the small pair
Subject 11	the red one	the blue one
	the green one	the purple one
	the yellow one	the brown one
	the pink one	the grey one
	the orange one	the white one
	the black one	the violet one

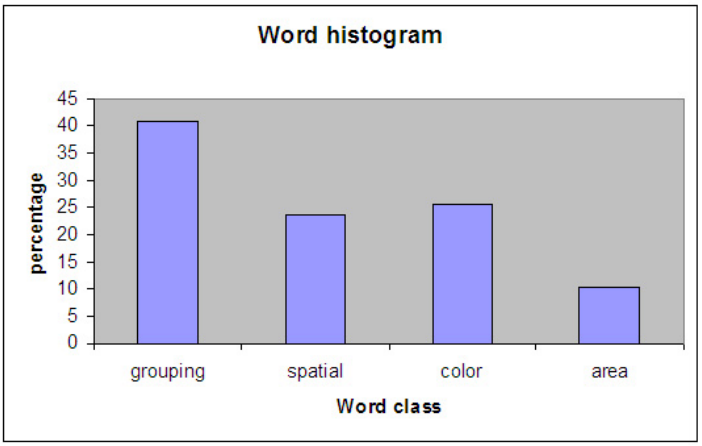


Figure B-1: Word histogram

Appendix C

r, g, b to CIEL*a*b* conversion

C.1 r,g,b to XYZ

CIE XYZ color space is the cone-shaped space formed by the tri-stimulus values that when applied to the CIE primaries, match any visible color.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \begin{bmatrix} r \\ g \\ b \end{bmatrix}$$

C.2 Conversion form XYZ to L*a*b*

CIE 1976 L*a*b* linearizes the perceptibility of color differences. The distance between colors in this system is intended to mimic the logarithmic response of the eye. Coloring information is with respect to the color of the white point of the system denoted by subscript *n*. The conversion formula [2] is as follows,

$$\begin{aligned} L* &= 116(Y/Y_n)^{1/3} - 16 \qquad for Y/Y_n > 0.008856 \\ L* &= 903.3 * Y/Y_n \qquad else \end{aligned}$$

$$
\begin{aligned}
a* &= 500 * (f(X/X_n) - f(Y/Y_n)) \\
b* &= 200 * (f(Y/Y_n) - f(Z/Z_n)) \qquad \textit{where,} \\
f(t) &= t^{1/3} \qquad if\ t > 0.008856 \\
f(t) &= 7.787 * t + 16/116 \qquad if\ t \leq 0.008856
\end{aligned}
$$

Here X_n, Y_n and Z_n are the tri-stimulus values of the reference white.

Bibliography

[1] Arnon Amir and Michael Lindenbaum. A generic grouping algorithm and its quantitative analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(2):168–185, 1998.

[2] R. S. Bern, R. J. Motta, and Gorzynski. Crt colorimetry: Part 1 theory and practice, part 2 metrology. *Color Research and Application*, 18, 1993.

[3] E. Brunswik and J. Kamiya. Ecological validity of proximity and other gestalt factors. *American Journal of Psychology*, pages 20–32, 1953.

[4] J. M. Buhmann, J. Malik, and P. Perona. Image recognition: Visual grouping, recognition and learning. *Proc. of National Academy of Sciences*, 96(25):14203–14204, 1999.

[5] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*. Springer, 1999.

[6] R. Cattoni, T. Coianiz, S. Messelodi, and C. Modena. Geometric layout analysis techniques for document image understanding: a review, 1998.

[7] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* Wiley, 2001.

[8] J. Elder and R. M. Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):324:353, 2002.

[9] Attneave F. Some informational aspects of visual perception. *Psychology Reviews*, 61:183–193, 1954.

[10] P. Gorniak and D. K. Roy. Grounded compositional semantics for referring noun phrases. *forthcoming*, 2003.

[11] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[12] Gerd Herzog and Peter Wazinski. VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions. *Artificial Intelligence Review*, 8:175–187, 1994.

[13] B. Landau and R. Jackendoff. What and where in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–265, 1993.

[14] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer, June 1985.

[15] George Miller and Philip Johnson-Laird. *Language and Perception*. Harvard University Press, 1976.

[16] T. Regier and L. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130:273–298, 2001.

[17] D. K. Roy. Learning words and syntax for a visual description task. *Computer Speech and Language*, 16(3), 2002.

[18] Simone Santini and Ramesh Jain. Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[19] E. Saund. Finding perceptually closed paths in sketches and drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.25(4):475– 491, 2003.

[20] A. Sha'ashua and S. Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *Proceedings of the International Conference on Computer Vision*, pages 321–327, 1988.

[21] R. N. Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–398, 1980.

[22] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[23] R. K. Srihari. Computational models for integrating linguistic and visual information: A survey. *Artificial Intelligence Review, special issue on Integrating Language and Vision*, 8:349–369, 1995.

[24] L. Talmy. How language structures space. In H. L. Pick Jr. and L. P. Acredolo, editors, *Spatial orientation: Theory, research and application*, pages 225–282. Plenum, NY, 1983.

[25] K. R. Thorisson. Simulated perceptual grouping: An application to human-computer interaction. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 876–881, Atlanta, GA, 1994.

[26] A. Treisman. Features and objects in visual processing. *Scientific American*, 254(11):114–125, 1986.

[27] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[28] B. Tversky and P. U. Lee. How space structures language. In C. Freksa, C. Habel, and K. F. Wender, editors, *Spatial Cognition*. Springer, Berlin, 1998.

[29] D. Waltz. On the interdependance of language and perception. In *Proceedings of the theoretical issues in natural language processing-2*, pages 149–156, Urbana-Champaign, IL, 1978.

[30] M. Wertheimer. Laws of organization in perceptual forms. In S. Yantis, editor, *Visual Perception: Essential Readings (Key Readings in Cognition)*, pages 216–264. Psychology Press, Philadelphia, 2000.

[31] R. A. Wilson and F. Keil, editors. *MIT Encyclopedia of Cognitive Sciences*. The MIT Press, 1999.

[32] G. Wyszecki and W. S. Stiles. *Colour Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley & Sons, New York, 1982.

[33] A. L. Zobrist and W. B. Thompson. Building a distance funtion for gestalt grouping. *IEEE Transactions on Computers*, C-24(7):718–728, 1975.