Semantic Visual Abstraction for Face Recognition

Yang Cai[1], David Kaufer[2], Emily Hart[3], and Elizabeth Solomon[4]

[1] Ambient Intelligence Lab, [2] English
[3] Electrical and Computer Engineering, [4] Art
Carnegie Mellon University
{ycai,kaufer,erhart,lizzeesolomon}@andrew.cmu.edu
www.cmu.edu/vis

Abstract. In contrast to the one-dimensional structure of natural language, images consist of two- or three-dimensional structures. This contrast in dimensionality causes the mapping between words and images to be a challenging, poorly understood and undertheorized task. In this paper, we present a general theoretical framework for semantic visual abstraction in massive image databases. Our framework applies specifically to facial identification and visual search for such recognition. It accommodates the by now commonplace observation that, through a graph-based visual abstraction, language allows humans to categorize objects and to provide verbal annotations to shapes. Our theoretical framework assumes a hidden layer between facial features and the referencing of expressive words. This hidden layer contains key points of correspondence that can be articulated mathematically, visually or verbally. A semantic visual abstraction network is designed for efficient facial recognition in massive visual datasets. In this paper, we demonstrate how a two-way mapping of words and facial shapes is feasible in facial information retrieval and reconstruction.

Keywords: semantic network, visual abstraction, visual search, human features, face, face recognition, information retrieval, video analytics.

1 Introduction

If a picture is worth 10,000 words [5], can a word be worth 10,000 images? The answer is *yes*. As visual abstractions, many linguistic referring expressions convey visual information with much greater efficiency than visual images. In our everyday life, we detect, recognize and retrieve images with words, which dramatically compress the representational information data space. For example, we often describe a traffic intersection with a letter 'T', or 'X', where we compress an image (e.g. 1 megabyte) to a letter (e.g. 1 byte). We also can retrieve images from our memory with words. This two-way transformation has been culturally supported since the invention of the alphabet. One central objective of this paper is to address how to improve visual search processes by the two-way mapping of images and words. This study will also demonstrate applications to real-world problems, such as the identification of faces

G. Allen et al. (Eds.): ICCS 2009, Part I, LNCS 5544, pp. 419–428, 2009.

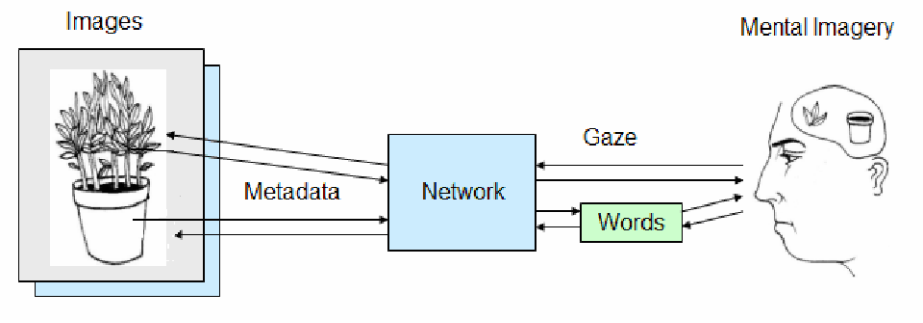

Fig. 1. Overview of the semantic visual abstraction in a network

from a massive video database. Given a method to detect gaze and objects, how do we encode our visual information in multiple resolutions to minimize the communication load and maximize the efficiency for information retrieving? Figure 1 illustrates the architecture of a visual abstraction network.

2 Image-Word Mapping

For many years, cognitive scientists have investigated visual abstraction from psychological experiments including visual search using *foveal vision* [17-27] and *mental rotation* [30]. Visual abstraction models have also been developed, notably Marr's cylinder model of human body structures [28] and the spring-mass graph model of facial structures [29]. Unfortunately, those visual abstractions don't address the image-word two-way mapping issues in particular. Recently, scientists have begun to model the relationship between words and images. CaMeRa [8], for example, is a computational model of multiple representations, including imagery, numbers and words. However, the mapping between the words and images in this system is linear and singular, lacking flexibility. An Artificial Neural Network model has been proposed to understand oil paintings [9], where Solso remarks that the hidden layers of the neural network enable the two-way mapping of words and visual features more effectively. With this method, Solso has argued that fewer neurons are required to represent more images. However, the content of the hidden layers of the neural network remains a mystery.

Because of the two- or three-dimensional structure of images and the one-dimensional structure of language, the mapping between words and images is a challenging and still undertheorized task. Arnheim observed that, through abstraction, language categorizes objects. Yet language, through its richness, further permits humans to create categorizations and associations that extend beyond shape alone [2]. As a rich abstractive layer, language permits categorizations of textures, two- and three-dimensions, and sub-shapes. As an abstractive layer, natural language seems to be the only method we have to satisfactorily describe a human subject. To explore this insight further, Roy developed a computerized system known as Describer that learns to generate contextualized spoken descriptions of objects in visual scenes [10]. Describer illustrates how a description database could be useful when paired with images in constructing a composite image. However, Describer is limited in representing a simplified block world.

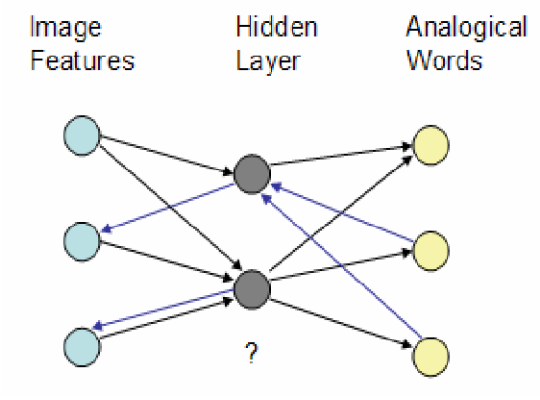

Fig. 2. The two-way mapping neural network model

3 Descriptions for Humans

Our own work and theoretical framework has focused on the mapping between words and images for human features. Why do we focus on human faces? Humans in general and human faces in particular provide among the richest vocabularies of visual imagery in any modern language. Imaginative literature is a well-known source of such descriptions, where human features are often described in detail. In addition, reference collections in the English language focused on visual imagery, such as description and pictorial dictionaries, never fail to have major sections devoted to descriptions of the human face. These sections are typically devoted to anatomical rather than social descriptions of faces based on cultural stereotypes and analogies. The mappings between images and faces we have been exploring are built upon cultural stereotype and analogical associations.

In the following sections, we briefly overview a variety of semantic visual description methods, including multiple resolution, semantic differentiation, symbol-number, and analogy. Then, we introduce the computational implementation of the human description through the interaction of words and images.

4 Multiple Resolution Descriptions

Human descriptions function as classifiers for shape, color, texture, proportion, size and dynamics in *multiple resolutions*. For example, one may start to describe a person's torso, then her hairstyle, face, eyes, nose, and mouth. Human feature descriptions have a common hierarchic structure [1]. For example, figure, head, face, eye, et al. Like a painter, verbal descriptions can be built in multiple resolutions. The words may start with a coarse global description and then 'zoom' into sub components and details. See Fig. 3 for a breakdown of a global description of a human head.

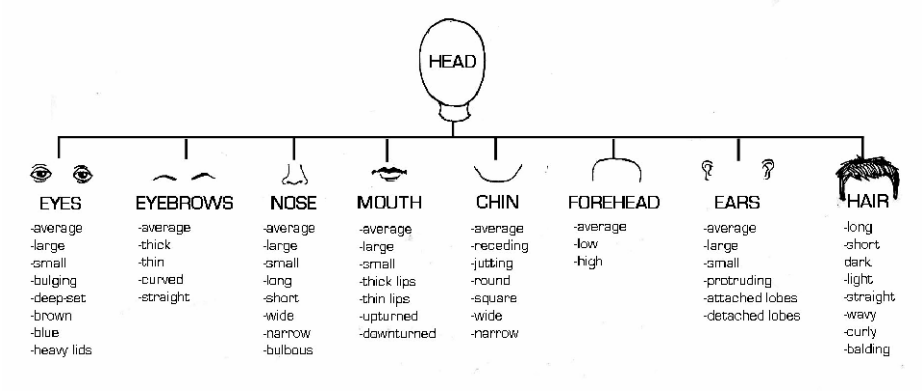

Fig. 3. Multi-resolution representation of a face

In our research to date, we have collected over 100 entries of multi-resolution descriptions from imaginative literature. Our collection ranges from global descriptions to components and details. Due to limitations of space, we will only enlist a few samples, where the underlined sections represent the global levels of description, the **bolded** show the component-based descriptions, and the *italicized* sections, the details:

- "For A **lean face** , pitted and scarred, very **thick black eyebrows** and **carbon-black eyes** with *deep grainy circles of black* under them. A *heavy five o'clock shadow*. But the skin under all was pale and unhealthy- looking. [11]"
- "Otto has a **face** like very ripe peach. **His hair** is fair and thick, growing low on his **forehead**. He has small sparkling **eyes**, full of naughtiness, and a wide, disarming **grin** which is too innocent to be true. When he grins, *two large dimples* appear in his peach **blossom cheeks**.[12]
- "Webb is the oldest man of their regular foursome, fifty and then some- a lean thoughtful gentleman in roofing and siding contracting and supply with a calming gravel voice, his **long face** broken into *longitudinal strips by creases* and **his hazel eyes** almost lost under an amber **tangle of eyebrows**.[13]"

5 Semantic Differential Representation

The Semantic Differential method measures perceptual and cognitive states in numbers or words arrayed on a linear scale. For example, the feeling of pain can be expressed with adjectives, ranging from weakest to strongest. Figure 4 shows a chart of visual, numerical and verbal expressions of pain in hospitals: No Hurt (0), Hurts Little Bit (2), Hurts Little More (4), Hurts Even More (6), Hurts Whole Lot (8) and Hurts Worst (10). This pictorial representations are very useful in patient communication where descriptions of pain type (e.g., pounding, burning) and intensity (e.g., little, a lot) lack a robust differentiated vocabulary.

The physical feeling can be quantified with mathematical models. When the change of stimulus (I) is very small, one won't detect the change. The minimal difference (ΔI) that is just noticeable is called perceptual threshold and it depends on the initial stimulus strength I. At a broad range, the normalized perceptual threshold is a constant, $\Delta I/I = K$. This is Weber's Law [16].

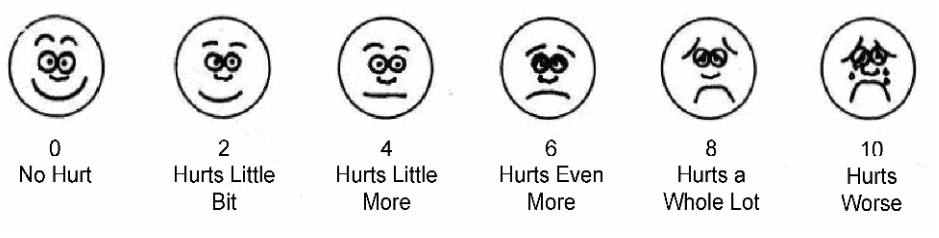

Fig. 4. Expressions of pain in pictures, numbers and words

Given the perceptual strength E, as the stimulus I changes by ΔI, the change of E is ΔE. We have the relationship $\Delta E = K*\Delta I/I$. Let ΔI be dI and ΔE be dE, thus we have the Weber-Fechner's Law:

$$E = K*ln(I) + C$$

where, C is constant and K is Weber Ratio, I is stimulus strength and E is the perceptual strength. Weber-Fechner's Law states that the relationship between our perceptual strength and stimulus strength is a logarithmic function. This perhaps explains why we are able to use limited words to describe a broad range of sensational experiences.

6 Symbol-Number Descriptions

In many cases, numbers can be added to provide even greater granularities. For example, the FBI's Facial Identification Handbook [14] comes with a class name such as bulging

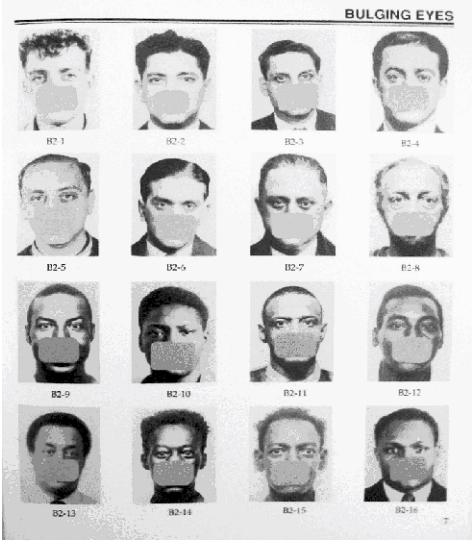

Fig. 5. Bulging Eyes from FBI Facial Identification Catalog

eyes and then a number to give specific levels and types. The FBI has already created a manual for witnesses, victims, and other suspect observers to use in identifying possible suspect features. The catalog presents several images per page under a category such as "bulging eyes." Each image in such a category has bulging eyes as a feature, and the respondent is asked to identify which image has bulging eyes most closely resembling the suspect. See Figure 5 for an example. This book is an extremely efficient and effective tool for both forensic sketch artists and police detectives. It is most commonly used as a tool in helping a witness or victim convey the features of the suspect to the sketch artist in order to render an accurate composite sketch.

7 Analogical Descriptions

From the multi-resolution point of view, an analogy describes in a coarse way in contrast to symbolic-number descriptions. Instead of describing features directly, people often find it more intuitive to refer a feature to a stereotype, for example, a movie star's face. The analogical mapping includes structural mapping (e.g. face to face), or component mapping (e.g. Lincoln's ear and Washington's nose). Children often use familiar objects to describe a person, for example using 'cookie' as an analogical reference for a round face.

Analogies are culture-based. In the Western world, several nose stereotypes are named according to historical figures. Many analogies are from animal noses or plants. Fig. 6 illustrates examples of the nose profiles as described above. We use a simple line drawing to render the visual presentation.

Analogies are a trigger of experience, which involves not only images, but also dynamics. The nose at the far right in Fig. 6 shows a 'volcano nose', which triggers a reader's physical experience such as pain, eruption, and explosion. In this case, readers not only experience it but also predict the consequence. Therefore, it is an analogy of a novel physical process that remains under the visible surface.

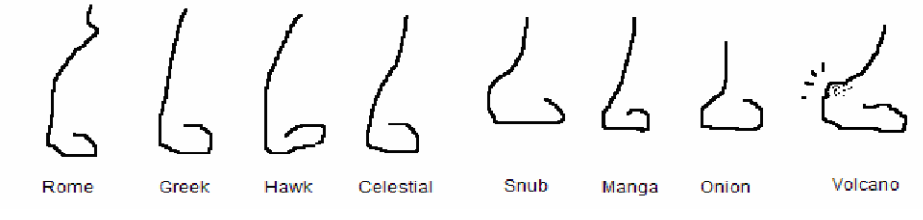

Fig. 6. Analogical description of noses

Given a verbal description of the nose, how do we visually reconstruct the nose profile with minimal elements? In this study, we use a set of 5 to 9 'control points' to draw a profile. By adjusting the relative positions of the control points, we can reconstruct many stereotypes of the profiles and many others in between. To smooth the profile contour, we apply the Spline [15] curve fitting model. See Fig. 7.

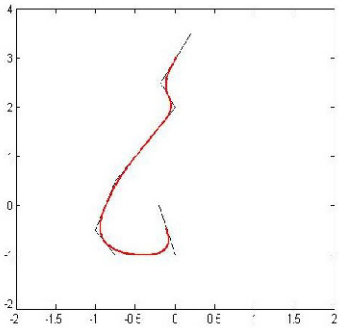

Fig. 7. Reconstructing a nose profile with points (black) and Spline curve (red)

8 The Verbal Description Database for Human Features

In this study, we have also collected over 100 verbal descriptions of human faces from several thesauri and descriptive dictionaries. The structure of the database is as follows: 1) the entity, 2) the side of the body, 3) the region of the body, 4) the part of the body, and 5) subtypes. The database is organized in terms of the resolution based on a hierarchy of human features reduced to each final descriptor. The database is intended to list all possible measurable descriptors of human features including face, body, and movement.

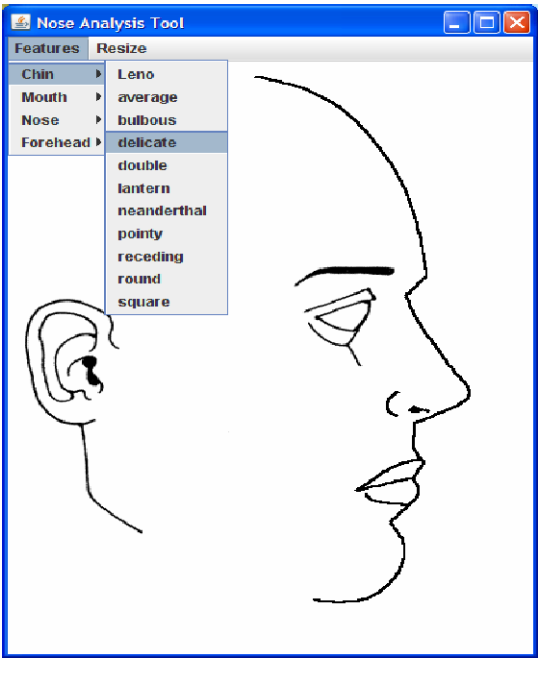

Fig. 8. Interactive facial profile reconstruction based on line-drawing. The code is written in Java so that it is possible to run on the Internet. These descriptions can then be rendered and distributed on the network: http://www.cmu.edu/vis/project9.html

9 Interactive Facial Reconstruction

We developed a computationally working prototype of the interactive system for facial reconstruction. In the system, a user selects the feature keywords in a hierarchical structure. The computer responds to the selected keyword with a pool of candidate features that are coded with labels and numbers. Once a candidate is selected, the computer will superimpose the components together and reconstruct the face. See Fig. 8 and Fig. 9.

As we know, composite sketches of a suspect are typically done by highly-trained professionals. Our system enables inexperienced users to reconstruct a face using only a simple menu driven interaction. In addition, this reconstruction process is reversible. We have designed it for use not only in facial description studies, but also in studies for robotic vision and professional training.

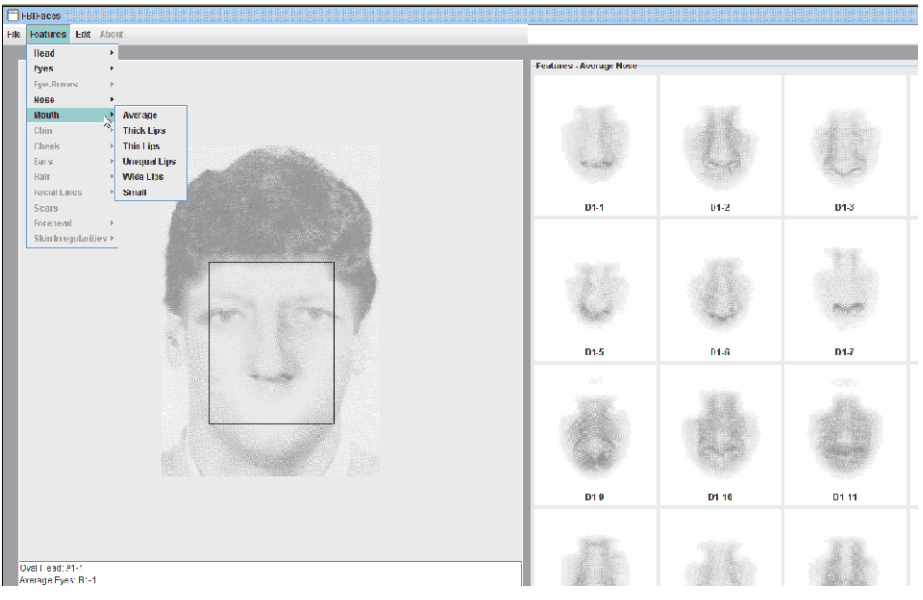

Fig. 9. Interactive front facial reconstruction based on image components

10 Conclusions

In this study, we assume a hidden layer between the human perception of facial features and referential words that contain 'control points' that can be articulated mathematically, visually or verbally. Our framework of a semantic network associating verbal and visual information remains in its early stages. Nevertheless, we countenance its long-term promise for understanding how we meaningfully and effortlessly map between visual and verbal information in the successful interpersonal communication about faces. At this moment, we only have profile and frontal facial reconstruction models. In the future, we plan to develop both whole head and body

models with far more control points and referencing expressions indexed into those points.

Today, we have an overabundance of data but not nearly enough attention or bandwidth. Image and video collections grow at an explosive rate that exceeds the capacity of network and human attention. In real-time surveillance systems, over a terabyte per hour are transmitted for only a small number of platforms and sensors. We believe that the visual abstraction network described in this paper is one of the feasible solutions that can and should be more thoroughly developed.

Acknowledgement

We would like to thank Army Research Office (ARO) and National Science Foundation (NSF) for their sponsorship. We are also in debt to Brian Zeleznik for his comments and editing.

References

1. Cai, Y.: How Many Pixels Do We Need to See Things? In: Ganter, B., de Moor, A., Lex, W. (eds.) ICCS 2003. LNCS, vol. 2746. Springer, Heidelberg (2003)
2. Arnheim, R.: Visual Thinking. University of California Press (1969)
3. Allport, A.: Visual Attention. MIT Press, Cambridge (1993)
4. Yarbus, A.L.: Eye Movements during Perception of Complex Objects. Plenum Press, New York (1967)
5. Larkin, J.H., Simon, H.A.: Why a diagram is (sometimes) worth 10,000 words. Cognitive Science 11, 65–100 (1987)
6. Geisler, W.S., Perry, J.S.: Real-time foveated multiresolution system for low-bandwidth video communication. In: Proceedings of Human Vision and Electronic Imaging. SPIE, Bellingham (1998)
7. Shell, J.S., Selker, T., Vertegaal, R.: Interacting with groups of computers. Communications of the ACM 46, 40–46 (2003)
8. Tabachneck-Schijf, H.J.M., Leonardo, A.M., Simon, H.A.: CaMeRa: A computational model of multiple representations. Cognitive Science 21, 305–350 (1997)
9. Solso, R.L.: Cognition and the Visual Arts. The MIT Press, Cambridge (1993)
10. Roy, D.: Learning from Sights and Sounds: A Computational Model. Ph.D. In: Media Arts and Sciences, MIT (1999)
11. Doctorow, E.L.: Loon Lake. Random House, New York (1980)
12. Isherwood, C.: Goodbye to Berlin. Signet. (1952)
13. Updike, J.: The Rabbit is Rich. Ballantine Books (1996)
14. FBI Facial Identification Catalog (November 1988)
15. Spline (2007), http://en.wikipedia.org/wiki/Spline_(mathematics)
16. Li, Q., Rosa, M.D., Daniela, R.: Distributed Algorithms for Guiding Navigation across a Sensor Network. Dartmouth Department of Computer Science (2003)
17. Wolfe, J.M.: Visual Search. In: Pashler, H. (ed.) Attention, East Sussex. Psychology Press, UK (1998)
18. Theeuwes, J.: Perceptual selectivity for color and form. Perception & Psychophysics 51, 599–606 (1992)

19. Treisman, A., Gelade, G.: A feature integration theory of attention. Cognitive Psychology 12, 97–136 (1980)
20. Verghese, P.: Visual search and attention: A signal detection theory approach. Neuron 31(13), 523–535 (2001)
21. Visual Seach (2008), http://en.wikipedia.org/wiki/Visual_search
22. Yarbus, A.L.: Eye Movements during Perception of Complex Objects. Plenum Press, New York (1967)
23. Larkin, J.H., Simon, H.A.: Why a diagram is (sometimes) worth 10,000 words. Cognitive Science 11, 65–100 (1987)
24. Duchowski, A.T., et al.: Gaze-Contingent Displays: A Review. Cyber-Psychology and Behavior 7(6) (2004)
25. Kortum, P., Geisler, W.: Implementation of a foveated image coding system for image bandwidth reduction. In: SPIE Proceedings, vol. 2657, pp. 350–360 (1996)
26. Geisler, W.S., Perry, J.S.: Real-time foveated multiresolution system for low-bandwidth video communication. In: Proceedings of Human Vision and Electronic Imaging. SPIE, Bellingham (1998)
27. Majaranta, P., Raiha, K.J.: Twenty years of eye typing: systems and design issues. In: Eye Tracking Research and Applications (ETRA) Symposium. ACM Press, New Orleans (2002)
28. Marr, D.: Vision. W.H. Freeman, New York (1982)
29. Ballard, D.H., Brown, C.M.: Computer Vision. Prentice-Hall Inc., New Jersey (1982)
30. Mental Rotation (2007), http://en.wikipedia.org/wiki/Mental_rotation