Visual Learning of Simple Semantics in ImageScape

Jean Marie Buijs and Michael S. Lew

Leiden Institute for Advanced Computer Science
Leiden University, Postbus 9512, 2300 RA Leiden, The Netherlands
{buijs,mlew}@cs.leidenuniv.nl

Abstract Learning visual concepts is an important tool for automatic annotation and visual querying of networked multimedia databases. It allows the user to express queries in his own vocabulary instead of the computer's vocabulary. This paper gives an overview of our current research directions in learning visual concepts for use in our online visual webcrawler, ImageScape. We discuss using the Kullback relative information for finding the most informative features in the case of human faces and generalize the method to other objects/concepts.

1 Introduction

In many content based retrieval systems, the user is asked to understand how the computer sees the world. An emerging trend is to try to have the computer understand how people see the world. However, understanding the world is a fundamental computer vision problem which has withstood decades of research. The critical aspect to these emerging methods is that they have modest ambitions. Petkovic[1997] has called this finding "simple semantics." From recent literature, this generally means finding computable image features which are correlated with visual concepts. The key distinction is that we are not trying to fully understand how human intelligence works. This would imply creating a general model for understanding all visual concepts. Instead, we are satisfied to find features which describe some small, but useful domains of visual concepts.

1.1 Visual Search Paradigms

Content based search researchers are constantly looking for methods which are usable by nonexperts. The typical method for this intuitive search is by finding a similar image. In this paradigm, the user clicks on an image, and then the search engine ranks the database images by similarity with respect to color, texture, and shape. In sketch search methods, the user draws a rough sketch of the goal image. The assumption is that the sketch corresponds to the object edges and contours. The database images which have the most similar shapes to the user sketch are returned. Sketches represent

Dionysius P. Huijsmans, Arnold W.M. Smeulders (Eds.): VISUAL'99, LNCS 1614, pp. 131-138, 1999.

an abstract level of reasoning of the image. Another abstract method uses icons to represent objects and/or concepts in the image. The user places the icons on a canvas in the position where they should appear in the goal image. In this context, the database images must have been preprocessed for the locations of the available objects/concepts. The database images which are most similar by the content of objects/concepts to the iconic user query are returned. For an overview, see Gudivada and Raghavan[1995] and especially, Flickner, et al. [1995].

We designed a system for searching networked multimedia databases called ImageScape. One of the principal query types in ImageScape is semantic icons. Semantic icons are essentially drag-and-drop icons which represent concepts from the user's vocabulary. These can be objects such as human faces, textures such as sand or wood, or even colors. The importance of the method is it does not require the user to learn how the computer understands the image. Instead, the computer learns how humans perceive images.

1.2 Imagescape System Overview

When an image is brought to the server, it is analyzed for features pertaining to the semantic icons (i.e. faces, sand, water, etc.) and for extraction of the computer sketches. Then a thumbnail of the image and the features are stored in a compressed database. When a user sends an image query from a WWW Java browser/client program, the query is sent to the server, and matched against the database. The user drawn sketch is compared to the computer sketches and the semantic icons are compared to the automatically extracted features. The best matches are then sent back to the WWW browser/client program to be displayed to the user.

In summary, the ImageScape system consists of the following modules:

- collection of text, images, audio, and video from the WWW
- compression of the image database
- semantic object detection in images
- computer sketching of images
- matching between the icons/sketches with the database images
- Java client connecting to host server for the visual query input and processing

There are other interesting WWW image search engines which have been described in the research literature. The WebSeek[Smith and Chang 1997] system from Columbia University finds similar images and performs automatic text based category classification. The Webseer[Frankel, Swain, and Athitsos 1996] system from the University of Chicago lets users search by the number of faces and by text queries. Taylor, Cascia, and Sclaroff[1997] designed the ImageRover system to primarily use relevance feedback for the search process, and in the PicToSeek system, Gevers and Smeulders[1997] search through the images using similar images and image features.

In a previous paper [Lew, et al. 1997], we introduced an early version of this system. In this work, our focus is entirely on the object/concept detection.

2 Learning Simple Semantics

In this paper we discuss learning simple semantics, or in another way, *visual learning of concepts*. This brings into mind the question raised by readers and referees, which is "What is visual learning?" Such a general term could refer to anything having to do with artificial or human intelligence in all its sophistication and complexity. Rather than a vague description, we seek to define it clearly at least within the boundaries of this paper as either (1) feature tuning; (2) feature selection; or (3) feature construction. Feature tuning refers to determining the parameters which optimize the use of the feature. This is often called parameter estimation. Feature selection means choosing one or more features from a given initial set of features. The chosen features typically optimize the discriminatory power regarding the ground truth which consists of positive and negative examples. Feature construction is defined as creating new features from a base set of atomic features and integration rules. In this paper, we focus on feature selection and in the section on future work, we reveal preliminary results toward feature construction.

What is an object/concept? For our purposes, an object/concept is anything which we can apply a label or recognize visually. These could be clearly defined objects like faces or more difficult concepts such as textures. Most textures do not have corresponding labels in common language. Object/Concept detection is essential to the usage of the WWW image search engine because it gives the computer the ability to understand our notion of an object or concept. Instead of requiring all users to understand low level feature queries, we are asking the computer to understand the high level queries posed by humans. For instance, if we want to find an image with a beach under a blue sky, most systems require the user to translate the concept of beach to a particular color and texture. In our system, the user can pose the query visually as a beach under a blue sky using icons to represent beach and blue sky, respectively.

Giving a complete discussion of visual concept learning would not fit within the scope of a conference paper. In fact, it would require several books to do it justice. Furthermore, we suggest that what is necessary in the field now is a thorough survey on visual concept learning. For the scope of this paper, we give a brief overview of recent visual learning techniques in the research literature. We turn to an example of feature selection in the domain of human face detection, and then observe that it is straightforward to generalize the face detection method to other objects.

2.1 Background

Picard[1996] reported promising results in classifying blocks in an image into "at a glance" categories. What this means is that she investigated multiple model methods to classify an NxN block into categories which humans could classify without logically analyzing the content. Forsyth, et al. [1996] found objects from feature blobs. More recently, Vailaya, Jain, and Zhang [1998] have reported success in classifying images as city vs. landscape. They found that the edge direction features have sufficient discriminatory power for accurate classification of their test set. Buijs[1998] reported promising results in learning primary colors and textures using the Kullback relative information. The commonality between these methods was using multiple features for object/concept detection. Regarding object detection, the recent surge of research toward face recognition has motivated robust methods for face detection in complex scenery. Representative results have been reported by Sung and Poggio[1998], Rowley and Kanade[1998], Lew and Huijsmans [1996], and Lew and Huang[1996].

2.2 Feature Selection

We begin by describing a method of finding human faces in grayscale images with complex backgrounds and then show that the method is easily extensible to other objects/concepts. The Kullback relative information[1959] is generally regarded as one of the canonical methods of measuring discriminatory power. Specifically, we formulated the problem as discriminating between the classes of face and nonface, and used the Kullback relative information as a measurement of the class separation, which is the distance between the classes in feature space. As the class separation increases, the overlap between the classes decreases making the confidence in the class decision increase.

The detection algorithm can be stated concisely as follows:

(1) Create a ground truth set of positive (face) and negative(nonface) examples
(2) Compute the histograms from the ground truth set for the classes face and nonface.
(3) Find the N most informative features by maximizing the Kullback relative information combined with a Markov random field.
(4) Arrange the N most informative features in a vector, and apply a minimum distance classifier

and is shown in Figure 2.1.

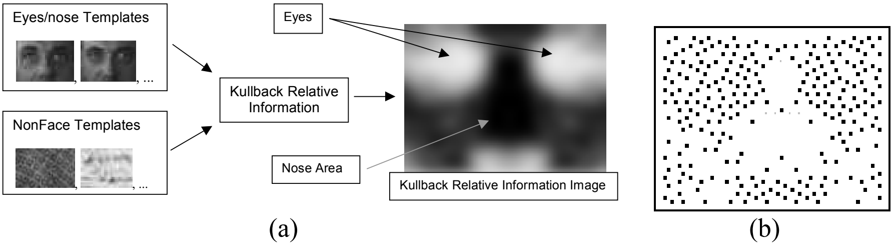

Figure 2.1. (a) Result of using Kullback relative information on face and nonface examples; and (b) the 256 pixels which have the greatest discriminatory power.

2.3 Generalizing to Multiple Models

In the previous explanation covering face detection, we used the 256 pixels which had the greatest discriminatory power. For lack of a better word, we define this set of features as a *discriminatory model.* This discriminatory model has the advantage that for N features in the model, it minimizes the misdetection rate. The question arises then of how to generalize the method to more features such as color, texture, and shape. For each object we wish to detect, a large set of positive and negative examples is collected. We measured a variety of texture, color, and shape features, and for each one we calculate the Kullback discriminant. The candidate features for our system included the color, gradient, Laplacian, and texture information from every pixel as shown in Figure 2.2. For the texture models, we used Trigrams[Huijsmans, Poles, and Lew 1996], LBP [Wang and He 1990], LBP/C, Grad. Mag., XDiff, YDiff, [Ojala, Pietikainen, and Harwood 1996]. For shape comparison, we used the features derived from elastic contours[Del Bimbo and Pala 1997], invariant moments[Hu 1962], and Fourier Descriptors [Gonzalez and Woods 1993].

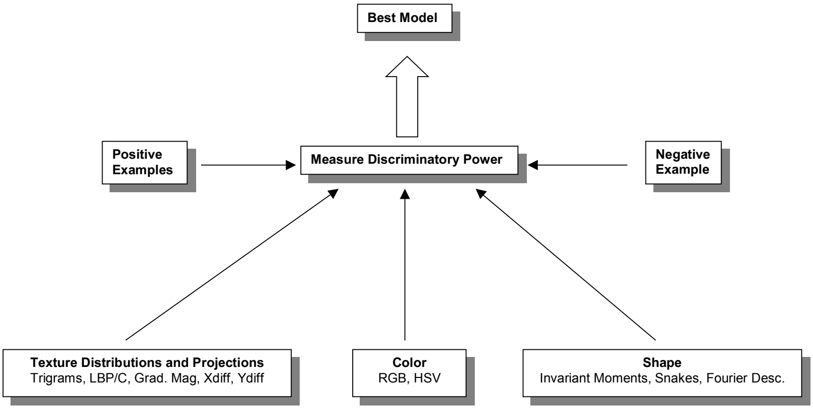
Best Model
Positive Examples
Measure Discriminatory Power
Negative Example
Texture Distributions and Projections
Trigrams, LBP/C, Grad. Mag, Xdiff, Ydiff
Color
RGB, HSV
Shape
Invariant Moments, Snakes, Fourier Desc.

Figure 2.2. Selecting the best discriminatory model of N features from texture, color, and shape features.

3 Feature Construction

In the previous discussion, we proposed using the Kullback relative information for feature selection. The next logical step was to consider methods for feature construction. In recent work[Buijs 1998], we presented a rule based method for combining the atomic features of color, texture, and shape toward representing more sophisticated visual concepts.

Recall that in feature construction there are atomic features and rules for integrating them. Our atomic features were instances of color, texture, and shape. The atomic colors were red, yellow, purple, green, blue, brown, orange, white, gray, and black. From the Kullback relative information, the color model with the greatest discriminatory power was the HSV.

The atomic textural features were coarse, semi-coarse, semi fine, fine, nonlinear, semi-linear, linear, and texture features based on examples: marble, wood, water, herringbone, etc. LBP/C had the greatest discriminatory power regarding the Kullback relative information.

For the atomic shape features, we created a basic set of geometric primitives: circular, elliptical square, triangular, rectangular, and pentagonal. We also tuned a variety of shape examples. Shape features were detected using template matching and active contour energy [del Bimbo and Pala 1997].

Simple concepts were represented as AND conjoined boolean expressions:

```
If (color is orange)
    AND (texture is coarse)
    AND (shape is circular)
Then object is an orange
```

More complex concepts were represented using AND/OR expressions such as:

```
If ((color is yellow OR color is white)
    AND (texture is fine)
    AND (texture is nonlinear))
Then object is sand
```

Rules were automatically generated from positive and negative example training sets using decision trees. We ranked the features used in the decision trees by the Kullback relative information, and created the tree using the features with greater discriminatory power first. Results for five outdoor categories are shown in Table 1.

Table 1. Probability Of Misdetection

	forest	mountain	sand	sky	water
Misdetection	0.23	0.14	0.27	0.09	0.15

4 Conclusions

Visual concept learning has the potential of bringing intuitive searching for visual media to the general public. Regarding the World Wide Web, we can bring image and video search to anyone with a Web browser if these visual learning technologies mature. At Leiden University, we are currently creating a large library of visual feature training databases and detectors. In this paper, we gave an overview of the visual concept learning methods being used for the ImageScape project. From the perspective of visual learning as feature tuning, feature selection, and/or feature construction, we have shown a progression of techniques for learning simple concept domains. Regarding future work, we think that the methods for combining the results of multiple classifiers[Kittler 1998] have the most versatility and potential for improvement of simple semantic detection.

Acknowledgements

This research was funded by the Dutch National Science Foundation and the Leiden Institute for Advanced Computer Science

References

Buijs, J. M., “Toward Semantic Based Multimedia Search,” Master’s Thesis, , Leiden Institute for Advanced Computer Science, August 13, 1998.

Del Bimbo, A., and P. Pala, “Visual Image Retrieval by Elastic Matching of User Sketches,” IEEE Trans. Pattern Analysis and Machine Intelligence, February, pp. 121-132, 1997.

Flickner, M., H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, “Query by Image and Video Content: The QBIC System,” Computer, IEEE Computer Society, pp. 23-32, Sept. 1995.

Forsyth, D., J. Malik, M. Fleck, T. Leung, C. Bregler, C. Carson, and H. Greenspan, "Finding Pictures of Objects in Large Collections of Images," Proceedings, International Workshop on Object Recognition, Cambridge, April 1996

Frankel, C., M. Swain and V. Athitsos, “WebSeer: An Image Search Engine for the World Wide Web,” Technical Report 96-14, University of Chicago, August 1996.

Gevers, T. and A. Smeulders, “PicToSeek: A Content-Based Image Search System for the World Wide Web,” VISUAL’97, San Diego, December, pp. 93-100.

Gonzalez, R. and R. E. Woods, “Digital Image Processing”, Addison Wesley, 1993.

Gudivada, V. N., and V. V. Raghavan, “Finding the Right Image, Content-Based Image Retrieval Systems,” Computer, IEEE Computer Society, pp. 18-62, Sept. 1995.

Hu, M., “Visual Pattern Recognition by Moment Invariants”, IRA Trans. on Information Theory, vol. 17-8, no. 2, pp. 179-187, Feb. 1962.

Huijsmans, D. P., M. Lew, and D. Denteneer, “Quality Measures for Interactive Image Retrieval with a Performance Evaluation of Two 3x3 Texel-based Methods,” International Conference on Image Analysis and Processing, Florence, Italy, September, 1997.

Kittler, J., M. Hatef, R. Duin, and J. Matas, “On Combining Classifiers,” IEEE Trans. Patt. Anal. and Mach. Intel., vol. 20, no. 3, March 1998.

Kullback, S. “Information Theory and Statistics,” Wiley, New York, 1959.

Lew, M., K. Lempinen, and N. Huijsmans, "Webcrawling Using Sketches," VISUAL'97, San Diego, December, 1997, pp. 77-84.

Lew, M. and N. Huijsmans, “Information Theory and Face Detection,” Proceedings of the International Conference on Pattern Recognition, Vienna, Austria, August 25-30, 1996, pp.601-605.

Lew, M. and T. Huang, “Optimal Supports for Image Matching,” Proc. of the IEEE Digital Signal Processing Workshop, Loen, Norway, Sept. 1-4, 1996, pp. 251-254.

Ojala, T., M. Pietikainen and D. Harwood, “A Comparative Study of Texture Measures with Classification Based on Feature Distributions,” vol. 29, no. 1, pp. 51-59, 1996.

Petkovic, D., "Challenges and Opportunities for Pattern Recognition and Computer Vision Research in Year 2000 and Beyond, "Proc. of the Int. Conf. on Image Analysis and Processing, September, Florence, vol. 2, pp. 1-5, 1997.

Picard, R.. "A Society of Models for Video and Image Libraries." IBM Systems Journal. 1996.

Rowley, H, and T. Kanade, Neural Network Based Face Detection, IEEE Trans. Patt. Anal. and Mach. Intell., vol. 20, no. 1, pp. 23-38, 1998.

Smith, J. R. and S.F. Chang, “Visually Searching the Web for Content,” IEEE Multimedia, 1997, pg. 12-20.

Sung, K. K., and T. Poggio, Example-Based Learning for View-Based Human Face Detection, IEEE Trans. on Patt. Anal. and Mach. Intell, vol. 20, no. 1, pp. 39-51, 1998.

Taycher, L., M Cascia, and S. Sclaroff, “Image Digestion and Relevance Feedback in the ImageRover WWW Search Engine,” VISUAL’97, December, San Diego, pp. 85-91.

Tekalp, A. M., **Digital Video Processing**, Prentice Hall, New Jersey, 1995.

Vailaya, A., A. Jain and H. Zhang, "On Image Classification: City vs. Landscape," IEEE Workshop on Content-Based Access of Image and Video Libraries, Santa Barbara, June 21, 1998.

Wang, L. and D. C. He, “Texture Classification Using Texture Spectrum,” Pattern Recognition 23, pp. 905-910, 1990.