A Study Access to 3D Object Detection Applied to features and Cars

Henry Schneiderman
Carnegie Mellon University, Pittsburgh, USA

Abstract

In this thesis, we describe a statistical method for 3D object detection. In this method, we decompose the 3D geometry of each object into a small number of viewpoints. For each viewpoint, we construct a decision rule that determines if the object is present at that specific orientation. Each decision rule uses the statistics of both object appearance and "non-object" visual appearance. We represent each set of statistics using a product of histograms. Each histogram represents the joint statistics of a subset of wavelet coefficients and their position on the object. Our approach is to use many such histograms representing a wide variety of visual attributes. Using this method, we have developed the first algorithm that can reliably detect faces that vary from frontal view to full profile view and the first algorithm that can reliably detect cars over a wide range of viewpoints.

1. Introduction

Object detection is a big part of our lives. We are constantly looking for and detecting objects: people, streets, buildings, hallways, tables, chairs, desks, sofas, beds, automobiles. Yet it remains a mystery how we perceive objects so accurately and with so little apparent effort. Comprehensive explanations have defied physiologists and psychologists for more than a century.

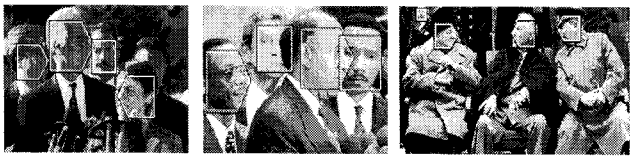


Figure 1. Examples of computer detection of human faces

Figure 2. Examples of computer detection of automobiles

In this thesis, our goal is not to understand how humans perceive, but to create computer methods for automatic object detection. Automated object detection could have many uses. The availability of large digital image collections has grown dramatically in recent years. Corbis estimates it currently has more than 67 million images in its current collection[1]. The Associated Press collects and archives an estimated 1,000 photographs a day[2]. The number of images on the World Wide Web is at least in the hundreds of millions. However, the usability of these collections is limited by a lack of effective retrieval methods. Currently, to find a specific image in such a collection, we have to search using text-based captions and low-level image features such as color and texture. Automatic object detection and recognition could be used to extract more information from these images and help automatically label and categorize them. By making these databases easier to search, they will become accessible to wider groups of users, such as television broadcasters, law enforcement agencies, medical practitioners, graphic and multimedia designers, book and magazine publishers, journalists, historians, artists, and hobbyists. Automatic object detection could also be useful in photography. As camera technology changes from film to digital capture, cameras will become part optics and part computer (giving true meaning to the term "computer vision"). Such a camera could automatically focus, color balance, and zoom on a specified object of interest, say, a human face. Also, specific object detectors, such as a face detectors and car detectors, have specialized uses. Face detectors are a necessary component in any system for automatic

face identification. Car detectors could be used for automatically monitoring traffic.

1.1. Challenges in Object Detection

Automatic object detection is a difficult undertaking. In 30 years of research in computer vision, little progress has been made. The main challenge is the amount of variation in visual appearance. For example, cars vary in size, shape, coloring, and in small details such as the headlights, grill, and tires. An object's orientation and distance from the camera affects its appearance. A more general difficulty is that visual information is ambiguous. Geometric ambiguity exists since the three dimensions of the world are projected on to two in the image. Also, a pixel's intensity depends on many dispersed factors in the environment. It depends on the light sources: their locations, their color, their intensity. It depends on the surrounding objects. Some objects may cast shadows on the object or reflect additional light on to the object. Pixel intensity also depends on the reflective properties of the viewed surfaces. A smooth surface will reflect light differently than a rough one.

Figure 3. Objects of the same class can vary significantly in appearance

Figure 4. Variation due to the pose (the relationship between the orientation of the object and the position of the camera)

Figure 5. Variation due to lighting and shadowing

Object detection is also difficult because images contain a large amount of data. There may be hundreds or even thousands of input pixels, each of which may carry important information. To use this information to its fullest extent, we would have to build the detector as an enormous table with an entry for every possible input indicating its classification, object or non-object, such as Table 1. Such a representation would account for all the forms of variation we just mentioned. Unfortunately, such a table is infeasible. entries would be required for describing the classification of a 20x20 region. Computer power and memory limit us to using a classification rule that is hundreds of orders of magnitude smaller. However, there is hope that we can get by with such a representation. The physical world imposes constraints on the appearance of objects; that is, of all the possible images that could conceivably exist in the physical world, only a small subset actually do. Moreover, people and animals are living examples things that achieve successful perception within their own computational limits.

2. View-Based Detectors

Our overall goal is to be able to detect the object over a range of orientations, sizes, and positions in an image. We use a 2D view-based approach to accommodate variation in orientation and we use exhaustive search in position and scale to accommodate variation in size and position.

A view-based approach works as follows. For each object, we build several detectors where each one is specialized to specific orientation of the object and can accommodate small amounts of variation around this orientation. To be able to detect an object at any orientation we apply all these detectors to the image and merge their results such that they are spatially consistent. In Figure 7 we show such a face detection result using this approach. In this example, each detector detects all the faces corresponding to its orientation. The woman in front was initially detected by both the frontal and left profile detectors, because the orientation of her face is somewhat inbetween these orientations. However, when the algorithm spatially integrates these results and chooses the more confident detection which in this case is the frontal detection.

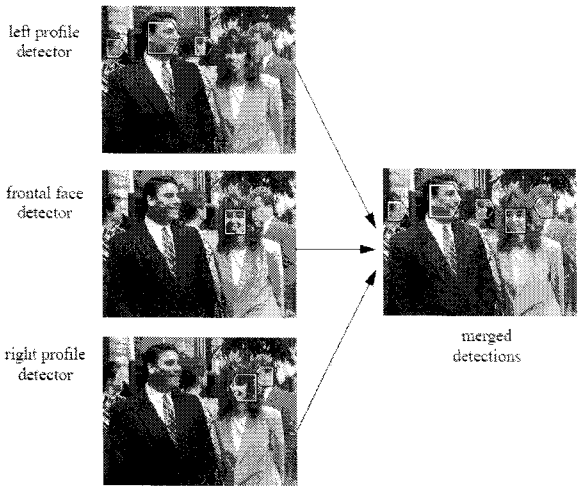


Figure 5.View-based detectors

It may seem counter-intuitive to use a 2D based model such as this to represent a 3D object, but there is an advantage to doing so. The problem with a 3D model is that we do not have explicit knowledge of the 3D geometry of the object. All our information is in the form of 2D images. To maintain a 3D representation we would have to rely on 3D recovery methods which are errorprone. By maintaining 2D models we avoid introducing such errors into our representation.

The question of how many and which viewpoints to use is an open question. One possible answer is to select viewpoints from aspect graphs if the object has well-defined surfaces. However, our approach was to simply determine the number of viewpoints through experimentation. For face detection, we found that three separate detectors was sufficient: left-profile, frontal views, and right-profile. In practice, we built only two detectors, right-profile and frontal, since we can detect left-profiles by applying the right profile detector to a mirror-reversed image. We show example training images for these in Figure 8. For automobile detection, we originally used

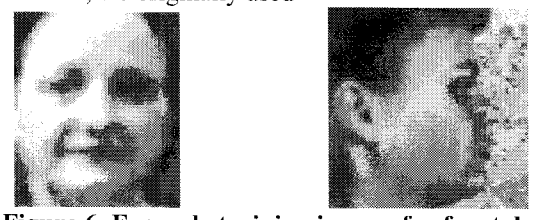


Figure 6. Example training images for frontal and right profile face views

three detectors, left-side, front, and right-side, but found it was necessary to use more. There are several explanations for this. Automobile photographs tend to be taken from a wider variety of vantages, from road level to views from a higher vantage point. In comparison, we usually photograph faces at eye level, except in surveillance cameras. Also, the shape of an automobile is rectilinear. Small changes in angle will produce bigger changes in appearance than they do for a sphere. Overall, we used 15 decision rules corresponding to the following orientations: one frontal viewpoint and 14 side viewpoints. Here again, we only had to train 8 detectors (7 right side detectors and one frontal detector), since 7 viewpoints are mirror reflections of each other. In Figure 9 we show example training images for each of the viewpoint we trained on . We do not detect back views of cars. Also for both object we do not represent in-plane rotations. Both faces and cars tend to appear as upright objects.

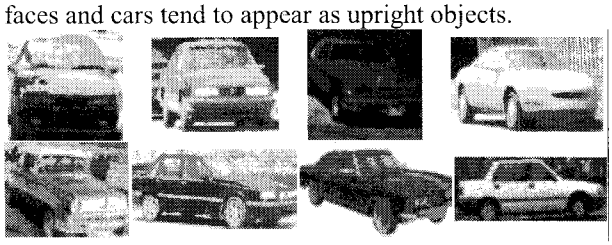


Figure 7. Example training images for each car viewpoint

In addition to detecting the object over variation in orientation, we also have to detect it over variation in size and position within the image. Our approach to detecting the object under these variations is to use exhaustive search. We train each view-based detector to only find the object when it is normalized in size and centered in a given rectangular image window. (We design each such detector to accommodate small variation about this size and alignment.) To then detect the object at position in an image, we have to re-apply each detector at all possible positions of the rectangular window. Then to detect the object at any size, we have to repeat this process over magnified and contracted versions of the original image.

3. Functional Form of Detector: Statistical Representation Using Histograms

In this chapter we derive the functional form of our detector. We use a statistical representation to model variation in visual appearance. We model both the statistics of appearance of the object and the statistics of the rest of the world. The difficulty in modeling these distributions is that we do not know their true characteristics. We do not know if they are Gaussian, Poisson, multimodal, etc. These characteristics are unknown since it is not tractable to analyze the joint statistics of large numbers of pixels. Therefore, we sought statistical models that avoid making strong

assumptions about distributional structure while still retaining good properties for estimation and retrieval. As we will explain, the best compromise we found was histograms.

Histograms, however, have one fundamental limitation. A histogram can only use a discrete number of values to describe appearance. More importantly, because of limited computer memory and finite training data, a histogram can only use a relatively small number of discrete values. To overcome this limitation we will describe how we use multiple histograms where each histogram represents the statistical behavior of a different group of quantized wavelet coefficients. With this representation, each histogram represents a different attribute of appearance in terms of spatial extent, frequency range, orientation. Our approach is to use many such histograms to make up for the limited scope and resolution of each individual one.

In this approach, by modeling groups of wavelet coefficients, we capture the statistics of appearance over limited spatial extents. However, we would also like to capture the overall geometric configuration of the object. Therefore, as we will explain, in each histogram, we represent the joint statistics of appearance and position, where we measure position with respect to a local coordinate frame affixed to the object. This representation implicitly captures each part's relative position with respect to all the others.

4. Functional Form of Detector: Re-derivation from an Ideal Form

We can also view the functional form of our detector as representing our best attempt at approximating to an ideal functional form within our computational constraints. In this chapter, we re-derive our functional form through a series of approximations to an ideal functional form. By deriving our decision rule this way we get a clear picture of the functional form's representational capacity and its deficiencies. In particular, we have a complete record of all the transformations and simplifications that limit its representational power. Also, through this analysis, we gain a better understanding of several issues that were not apparent in the first derivation in Chapter 3. This derivation reveals why it is useful to select non-object samples by bootstrapping and why we should train the detector to explicitly reduce the classification error on the training set. Also, we see why we divide object probability by non-object probability (in section 3.1) -- a consequence of Bayes' decision rule. Similarly, we see how multiplying the probabilities from different attributes (Section 3.6) corresponds to an assumption of statistical independence of the observations.

5. Training Detectors

So far we have only chosen the form of the decision rule; that is, we have specified the number of histograms, the size of each histogram and the variables over which we compute each histogram. We have not specified the actual values within each histogram, Pk(patternk(x,y), i(x), j(y) | object) and Pk(patternk(x,y), i(x), j(y) | non-object) that are used in the decision rule. We compute these statistical values from various sets of images. This process of gathering statistics is usually referred to as training. Specifically, we use a set of images of the object to generate samples for training Pk(patternk(x,y), i(x), j(y) | object) and we use images that do not contain the object to train Pk(patternk(x,y), i(x), j(y) | non-object). In this chapter, we begin with a discussion of our training images for faces and cars in Section 5.1. In Section 5.2 we discuss the training images we use for the non-object class. Then in section 5.3, we describe a basic training algorithm in which we estimate each histogram separately then in Section 5.4 we describe an alternative training procedure which minimizes the classification error on the training set using the AdaBoost algorithm.

6. Implementation of the Detectors

In this chapter we describe how we implement our detectors. Our main concern is speed of execution. We would like detection to be as fast as possible. Our strategy is to re-use multi-resolution information wherever possible and to use a coarse-to-fine search strategy and various other heuristics to prune out unpromising object candidates.

6.1. Exhaustive Search

As explained in Chapter 2, each detector is specialized for a specific orientation, size, alignment, and intensity of the object. However, an object can occur at any position, size, orientation, and intensity in the image. Our approach is to use an exhaustive search along all these dimensions to find objects in the image. First, to be able to detect the object at any position in the image, we have to re-apply all the detectors at regularly spaced intervals in the image. At each of these sampling sites we evaluate the candidate at five different intensity corrections and select the one that gives the best response. Then, to detect the object at any size, we have to repeat this process for magnified and contracted versions of the original image. We

search at scales of magnification that decrease by a multiplicative factor of $2\sqrt{4} = 1.189$

As we explain below, we chose an integer root of 2 so we could reuse information at each octave in this search through scale. We then combine the results of running all these detectors. If there are multiple detections at the same or adjacent locations and/or scales, the algorithm chooses the strongest detection.

Since it will be very time-consuming to evaluate the image in such an exhaustive fashion, we experimented with several methods for decreasing computation time, as we will describe later in this chapter.

7. Face Detection Performance

In this chapter we describe our results in face detection in Section 7.1, provide analysis of how the different parts of the face influence detection in Section 7.2, and assess statistical dependency across the extent of the face in Section 7.3.

7.1. Results in Face Detection

The distinguishing characteristic of our face detector is that it works for both frontal and out of- plane rotational views. To date, several researchers [14] have had success developing algorithms that work for frontal views of faces, but none, to our knowledge, have had success with profile (side) views except (below we will compare our performance with).

We believe there are several reasons why profile view faces are more difficult to detect than frontal views. First, the salient features on the face (eyes, nose, and mouth) are not as prominent when viewed from the side as they are when they are viewed frontally. Also, for frontal views these features are interior to the object, whereas on a profile one of the strongest features is the silhouette with the background. Since the background can be almost any visual pattern, a profile detector must accommodate much more variation in the silhouette's appearance than a frontal detector does for interior features.

We compared the performance of our detectors with that reported by Rowley and Kanade on a test set of profile views selected from a set of proprietary images Kodak provided to Carnegie Mellon University. These images consists of typical amateur photographs with some of the typical problems of such images, including poor lighting, contrast and focus. This test set consisted of 17 images with 46 faces, of which 36 are in profile view (between 3/4 view and full profile view):

Table 8: Face Detection results on Kodak data set

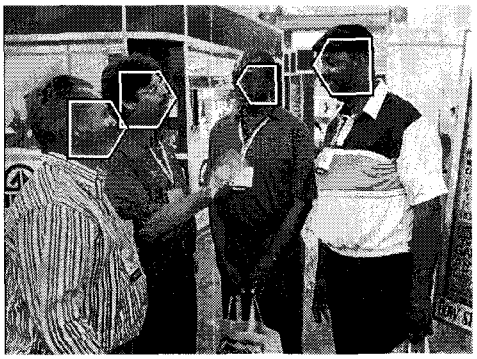
Rowley & Kanade [87]		Schneiderman and Kanade (using AdaBoost)			
Detection	False Detections	γ	Detection (all faces)	Detection (profiles only)	False Detections
58.7%	1347	0.5	80.4%	86.1%	105
41.3%	617	1.0	70.0%	69.4%	7
32.6%	136	1.5	63.0%	61.1%	1

We also collected a larger test set for benchmarking face detection performance for out-of plane rotation. This test set consists of 208 images with 441 faces that vary in pose from full frontal to side view. Of these images approximately 347 are profile view (between 3/4 view and full profile view). We gathered these images from a variety of sites on the World Wide Web, mainly news sites such as Yahoo and the New York Times. Most of these images were taken by professional photographers and of better quality than the Kodak images in terms of composition, contrast, and focus. Otherwise, they are unconstrained in terms of content, background scenery, and lighting. Below in Table 9 we show the performance at different values of the threshold γ controlling the sensitivity of the detectors. By changing γ we linearly scale the detection thresholds of both the profile and frontal detectors. We also compare the performance of the detectors trained with AdaBoost and without AdaBoost. Below in Figure 42 we show some typical results on this image set evaluated at $\gamma = 1.0$ using detectors trained with AdaBoost.

Table 9: Face Detection Results on Schneiderman & Kanade Test Set

	With AdaBoost			Without AdaBoost	
γ	Detections (all faces)	Detection (profiles)	False Detections	Detection (all faces)	False Detections
0.0	92.7%	92.8%	700	82%	137
1.5	85.5%	86.4%	91	74%	27
2.5	75.2%	78.6%	12	60%	3

In terms of frontal face detection, the accuracy of our detector compares favorably with those



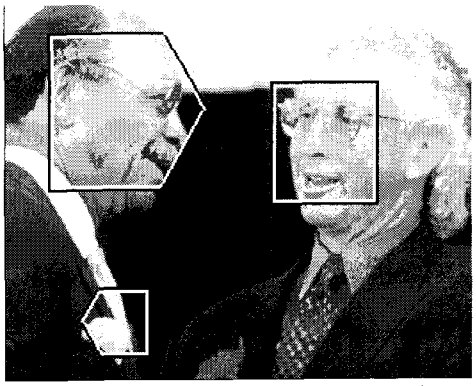


Figure 42. Face detection results

of other researchers. In these experiments, we also noticed some differences in performance between the detector described in this thesis and an improved version of the detector we described in. Both of these detectors use similar probabilistic structures but differ mainly in that the detector in uses visual attributes based on localized eigenvectors rather than wavelet coefficients. The wavelet based detector described in this thesis performs much better for profile view faces. However, the eigenvector based detector seems to be perform slightly better on frontal faces. Below in Table 10 we compare our face detectors (wavelet-based and eigenvector-based) with those results reported by others on the combined frontal face test set combining the test images from Sung and Poggio and Rowley, Baluja, and Kanade [14].

8. Review of Other Statistical Detection Methods

In this chapter in Section 9.1 we describe some of the major theoretical differences between our method of object detection and other methods of object detection and in Sections 9.2 and 9.3 we summarize several methods that have been applied to face and car detection. In this discussion we emphasize the particular modeling choices in each of these methods.

8.1. Comparison of Our Approach to Previous Detection / Recognition Methods

Below we summarize the main difference between our approach and other previous approaches to object detection / recognition

8.1.1. Local Appearance Versus Global ppearance

Much work in object recognition treats the appearance of the object in terms of full-sized rigid templates including the work of [5],[6]. These methods represent the appearance of the entire object as one entity rather than decomposing the object into smaller parts. There are several disadvantages to this type of model. First, the global methods that involve dimensionality reduction [5],[6] will end up emphasizing the coarse attributes of object appearance rather the distinctive nature of the smaller parts such as the eyes, nose, and mouth on a face. Second the matching of large template is known to be sensitive to small differences in scale, position, and orientation. Finally the matching of large regions can also be strongly influenced by "irrelevant" pixels. On many objects, such as a car, there will be large indistinctive areas such as the hood and windshield that are punctuated by relatively smaller areas of distinctive detailing such as the grill and headlights. In matching a large region, the majority of the pixels will come from the untextured parts and dominate selection of the match (using any norm that weighs each pixel equally such as L1 or L2).

9. Conclusion

In this thesis we have advanced the state of the art in 3D object detection in the following ways. We have developed the first algorithm that can reliably detect faces that vary in viewpoint from frontal to side view. Previously only frontal face detection had been demonstrated reliably. We have also demonstrated the first method for car detection that

works robustly over a range of view points.

Several concepts contribute to the effectiveness of these methods:

• Joint statistics of appearance and position - Much research in "parts-based" approaches to object recognition overlook the importance of representing the geometric arrangement of the parts. In our experiments, we have found performance improves drastically when we model the statistics of appearance and position jointly.
• Powerful representation of appearance - In our experiments we have observed that increased representation power improves the accuracy of the object detector. For example, we originally developed a weaker representation based on a subset of localized eigenvectors. Although this representation worked well for frontal face detection, it was not fully satisfactory for profile detection. We also noticed that when we reconstructed profile images from this representation, many of the small features that form of the silhouette of the face were lost. We then redesigned our representation using the wavelet-based representation described in this thesis. With this new representation, our visual representation of these features was better leading to improved detection performance for profile views of faces.
• Representation of the non-object - We also observed that performance depended on how we represented the non-object class. We noticed that having some model was an improvement over having no model, even if our model was based on randomly sampled non-object images. Performance improved further by using bootstrapping to select non-object samples and improved still further by using AdaBoost to weight them.
• Visual cues based on local relationships - We have shown that by using a combination of visual cues with selective localization in space, frequency and orientation we can achieve accurate detection of faces and cars.
• Coarse-to-fine heuristics - By using coarse to fine heuristics, we have demonstrated that we can use a large model with many visual cues in a computationally feasible way. There are several research areas we see as a natural continuation of this work:
• Representation - Representation remains the most important issue in object detection. Perhaps, some day researchers will develop specific statistic models for visual appearance in the same way Gaussian and Poisson models were developed to model specific physical phenomena. Of course, to do so, we would need a way to cope with the high dimensional nature of images. One approach would be to look at the pair-wise statistics among wavelet coefficients as we have suggested in Chapters 7 and 8. Another approach would be to use our knowledge of the physics of image formation. Good models for the effects of illumination, reflection, geometry, and material type on appearance exist. These models have been used to synthesize images that look fairly realistic. It may be possible to use such models to characterize the statistics of appearance. In particular, it may be easier to characterize the statistical variation of the "input" to the image formation process -- the illumination, the geometry of the scene, and the surface characteristics -- than to characterize the image variation directly. We could then analyze how these imaging models transform these stochastic inputs and thereby indirectly arrive at a statistical characterization of appearance.
• Intensity/Lighting correction - In our work, performance improved when we were able to correct for differences in illumination. Most of the existing methods for intensity correction use simple methods such as histogram equalization or transforming the intensities to have zero mean and unit variance. We believe that better performance can be achieved by explicitly accounting for the appearance of the object we are trying to enhance. One approach would be to use a probabilistic model of the object's appearance to choose the correction. Let us assume we have some lighting correction model:

$$\text{image}' = C(\text{image}, \theta) \tag{65}$$

where the parameter, θ, controls the lighting correction. Given that we have models for $P(image \mid \text{object})$ and $P(image \mid \text{non-object})$, we could choose the value of θ that gives the highest response and therefore is corrected so it most looks like a member of its class:

$$\hat{\theta} = \arg max_{\theta}\left(\frac{P(C(image,\theta)|\text{object})}{P(C(image,\theta)|\text{non-object})}\right) \tag{66}$$

We have conducted some preliminary

experimentation with this approach but have not had much success.

• Sample selection and weighting - In our experiments, performance improved when we used bootstrapping to select samples and boosting to weight samples. Perhaps there is a principled way of combining these two methods to achieve better performance.
• Coding of appearance -We have used scalar quantization to discretize each wavelet coefficient separately. Methods of vector quantization whereby a whole group of coefficients is quantized together may improve performance. We noticed such an improvement in an earlier method based on eigenvector responses.

There are also several research problems that we see as a natural continuation of this work:

• Detection of other rigid objects - We would like to test the generality of this algorithm by applying it to other rigid objects such as boats, airplanes, animals, pedestrians, etc.
• Detection of more challenging objects - There are more challenging objects we would like to detect such as buildings, trees, and text in video. These objects have some structural regularity but less so than faces or cars. We believe it is possible to detect such objects accurately with current computing power, but new representations will have to be developed to do so.
• Other classifications - There are many other classification problems that are probably solvable such as discriminating between indoor and outdoor scenes, urban and rural scenes, etc. It should also be possible to classify people based on activity (talking, smiling, walking) and their characteristics (age, gender, hair color, facial hair, glasses, etc.). It may even be possible to robustly identify people by computer. Many research efforts are making progress in this area.

10. References

[1]. www.salon.com. "Bill Gates' Other CEO." 2/7/2000.
[2] www.ap.org
[3] D. P. Huttenlocher, S. Ullman. "Recognizing Solid Objects by Alignment with an Image." IJCV. 5:2, pp. 195-212, 1990.
[4] A. Pentland, B. Moghaddam, T. Starner. "View-Based and Modular Eigenspaces for Face Recognition." CVPR 1994.
[5] M. Turk, A. Pentland. "Eigenfaces for Recognition." Journal of Cognitive Neuroscience, 3:1, pp. 71-86. 1991
[6] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman."Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. PAMI. 19:7 pp. 711-720. July, 1997.
[7]. K. S. Arun, T. S. Huang, S. D. Blostein. "Least-Squares fitting of two 3-D point sets." IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI). vol. 9. pp. 698 - 700. Sept., 1987.
[8] T. Niblett. "Constructing decision trees in noisy domains." Proceedings of the second Europena working session on learning. pp 67-78. Bled, Yugoslavia.
[9]B. Schiele, J. L. Crowley. "Recognition without Correspondence using Multidimensional Receptive Field Histograms." MIT Media Lab. Tech Report 453.
[10]B. Schiele, A. Pentland. "Probabilistic Object Recognition and Localization." ICCV '99.
[11] Martial Hebert, Jean Ponce, Terrance Boult, and Ari Gross. "Report on the 1995 Workshop on 3-D Object Recognition in Computer Vision." Object Recogntion in Computer Vision. International NSF-ARPA Workshop, Dec., '94. Lecture Notes in Computer Science, 994. pp. 1 - 18. Springer, 1995.
[12]. Baback Moghaddam and Alex Pentland. "Probabilistic Visual Learning for Object Detection." 5th International Conference on Computer Vision (ICCV '95) (also M.I.T. Media Laboratory Perceptual Computing Section, Technology Report No. 326).
[13]. Jae S. Lim. Two-Dimensional Signal and Image Processing. Prentice-Hall. 1990.
[14]. Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. "Neural Network-Based Face Detection." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, January, 1998, pp. 23-38.
[15] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag. 1995.