Slimane LARABI

Computer Vision

From Bidimensional Images to Three Dimensional Scene

January 02, 2025

Springer Nature

Chapter 1

Object Recognition from Visual Appearance

Abstract The objective of chapter is to study object recognition of 3D objects from their visual appearance in 2D images. This is referred as appearance matching. There are two basic approaches for objects representation: Shape: the explicit representation of the 3D geometry of the object. Appearance: depends on the pose, illumination. An image set is then obtained represented the appearance of the object. The reduction of the dimensionality is necessary in order to achieve the matching. Using PCA, we lower dimensional subspace in which we can represent the original image set by simply projecting the image set. We end up for each object , a compact parametric representation of the object. We terminate with a pipeline for appearance matching.

1.1 Introduction

The objective of chapter is to study object recognition of 3D objects from their visual appearance in 2D images. This is referred as appearance matching. We begin this chapter by introducing the concepts and motivations behind appearance-based analysis. Following this, we examine the relationship between shape and appearance, highlighting their respective contributions to visual understanding and recognition tasks. The process of Learning Appearance is then explored, focusing on how appearance features can be captured and encoded effectively.

An image set is then obtained represented the appearance of the object. The reduction of the dimensionality is necessary in order to achieve the matching. Using PCA, we lower dimensional subspace in which we can represent the original image set by simply projecting the image set.

Building on these foundational concepts, the chapter discusses Parametric Appearance Representation, which encapsulates appearance information using mathematical models, facilitating efficient and robust analysis. Finally, we address Appearance Matching, a process crucial for tasks such as object recognition, image retrieval, and alignment, where learned appearance models are utilized to compare and identify visual data.

1.2 Object recognition based on Shape Representation

The 3D representation learning approaches presented so far are based on voxel occupancy. We can cite [20, 18, 19, 21]. Authors are also payed attention on point clouds [24], [22] and explicit shape parametrization [23].

Figure 1.1 shows three explicit representations, such as triangle meshes are exceedingly popular in the graphics community, Voxel useful in computer graphics (see figures 1.1, 1.2) are defined on fixed regular grids making them exceptionally well suited for learning applications, in particular convolutional approaches[13]. Point clouds are also commonly used to describe the shape of 3D objects [25].

Many approaches have been proposed: Planes, spheres, complex splines, super quadric (see figure 1.3).

The superquadric shape have been considered as an extension of spherical or ellipsoidal particles and used for modeling of spheres, ellipsoids, cylinder like and box(dice) like particles just varying shape parameters [15].

The function $f(x)$ uses five parameters to describe super quadric, and is given by the equation 1.1.

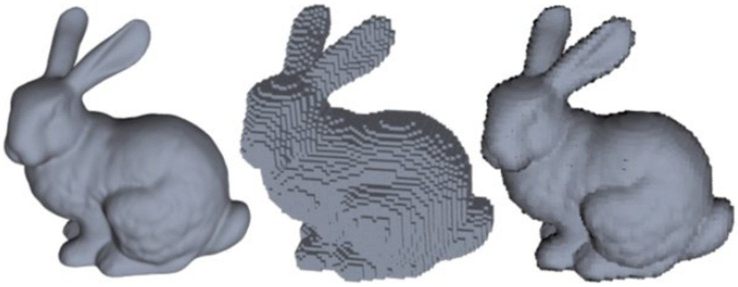
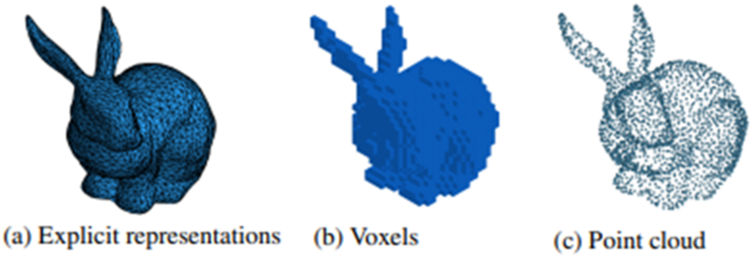


Fig. 1.1 Left: the original 3D shape. Middle: the voxelized 3D shape. Right: the octree representation with normal sampled at the finest leaf octants [13].



(a) Explicit representations (b) Voxels (c) Point cloud

Fig. 1.2 Common representations of 3D shape.[14]

$$f(\mathbf{x}) = (|\frac{x}{a}|^{n_2} + |\frac{y}{b}|^{n_2})^{n_1/n_2} + |\frac{z}{c}|^{n_1} - 1 = 0 \tag{1.1}$$

where $\mathbf{x} = (x, y, z)^T$.

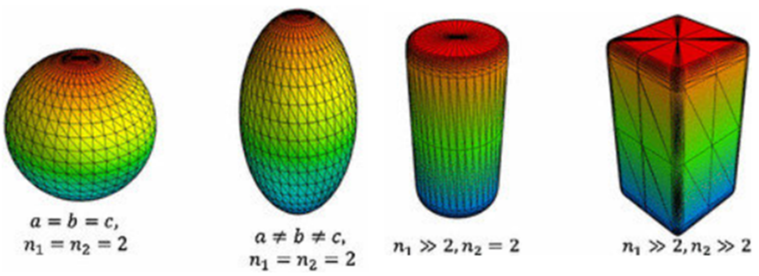


Fig. 1.3 Four examples of superquadric particle shape composed by the five shape parameters (a, b, c, n_1, n_2) [15].

For representation based on shape description using geometrical functions, the $3D$ Shape matching requires:

- Creating the database of object shapes (offline) Recognition (online),
- Computationally expensive for large databases, require hardware (scanners).

To overcame these constraints, new representation have been proposed, based on learning appearance.

1.3 Object Recognition Based on Learning Appearance

1.3.1 Basic Principle

When we see an object for the first time, we need to see it from different points of views so as it will be possible to recognize it if we see it again.

The first step of object recognition based on their appearance is to acquire a dataset of images of each object, associate them a useful representation and store all information.

For this step, some factors are important because they characterize the set of acquired images and define their visual appearance. We can cite (see figures 1.4, 1.5):

- Intrinsic factors: shape, reflectance,,
- Extrinsic factors: pose, illumination) .

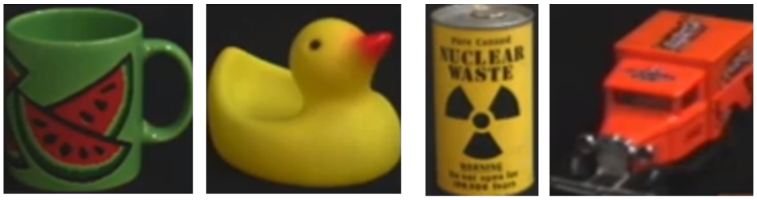


Fig. 1.4 Factors such as shape, reflectance, pose, illumination define the visual appearance.

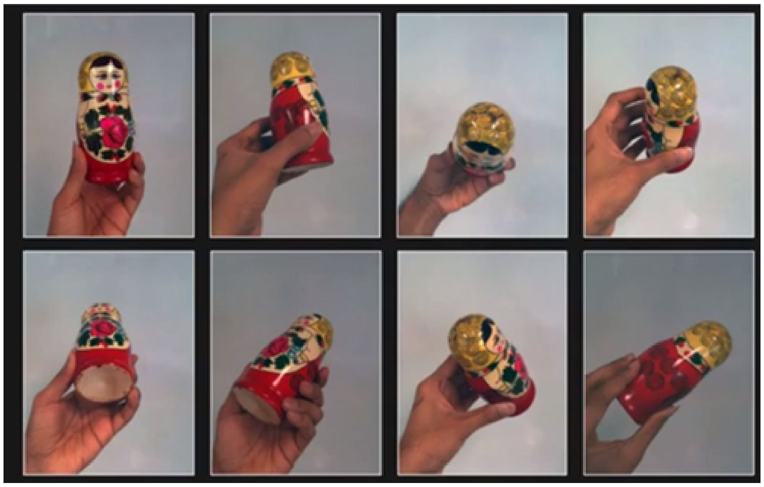


Fig. 1.5 Importance of the object pose for the recognition by visual appearance.

We need then to capture images under all poses and lighting directions as indicated by figure 1.6. We collect then a set of images (database) , segmented and without occlusion. Figure 1.7 shows examples of such datasets built by S.K.Nayar et al. [17].

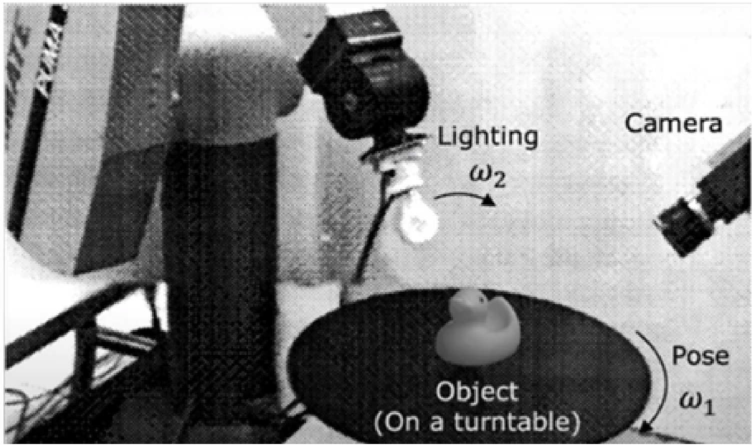


Fig. 1.6 Procedure of image model capture.

The second step is to develop an algorithm to retrieve the identity of the object from a new query image using the stored data.

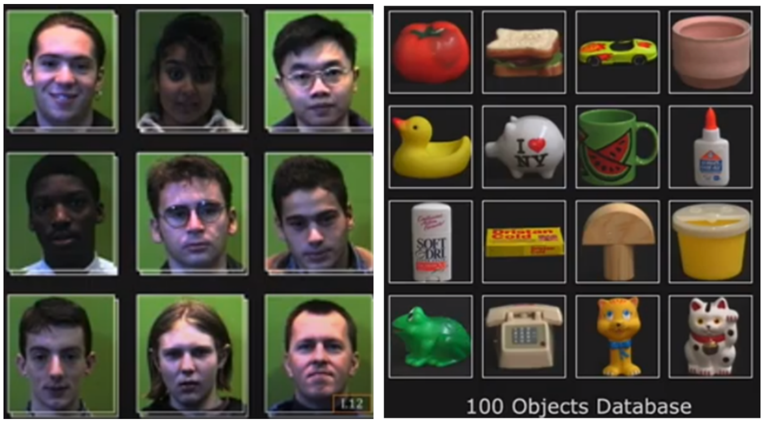


Fig. 1.7 Two examples of datasets of objects [17] .

To perform recognition, we many approaches are possible.

1.3.2 Object Recognition using Template Matching

A naive solution is to apply template matching to compare a query template with all image models of the dataset using one of the known similarity measures: SAD (Sum Absolute Distances), or SSD (Sum Squared Differences) or NCC (Normalized Cross Correlation).

However, this technic is not a practical given the number of images that we need to deal with, in addition, it is then very expensive and time consuming.

In addition, for a given object image set, there is a similarity between two consecutive images and high redundancy between images as depicted by figure 1.9. We can exploit this redundancy between images to reduce the dimensionality of the image set in order to achieve a compact appearance representation that makes matching efficient.

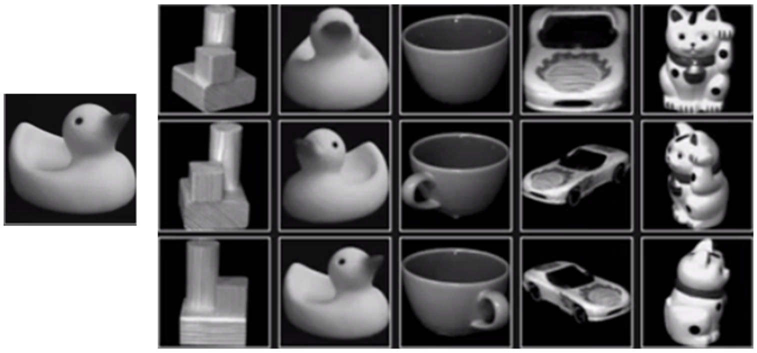


Fig. 1.8 Template matching technic; (Left) the query image, (Right) the dataset of image models.

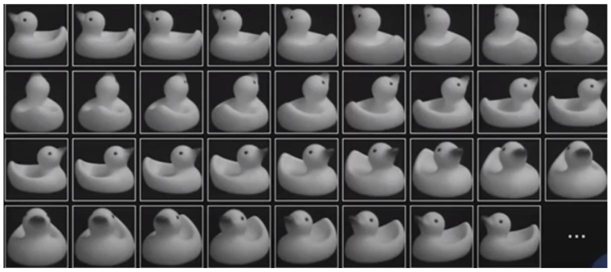


Fig. 1.9 Redundancy between images of the dataset of object image models.

1.3.3 Dimensionality reductions using Principal Component Analysis (PCA)

We want to transform the images into different space where matching one image with another is more efficient. We assume that image est represented by a vector (concatenated rows).

We can construct an N-dimensional space. Each one of the dimensions space represents the brightness at corresponding pixel.

The N pixels are represented using N vectors units. The image is simply a point in that space.

We treat $i_1, i_2, .., i_N$ as an orthonormal basis. Example: $i3 = (0, 0, 1, 0, ..., 0)$

The Correlation in image space (SSD), is computed as the sum of squared difference between pixels (p, q) of two images (model I_1 and query I_2) as given by equation 1.2 (see figure 1.10). For example, the distance SSD between two elements: $(123, 0, 0)$, $(120, 20, 0)$, the distance $SSD = (123 - 120)^2 + (20 - 0)^2 = 409$.

$$SSD = \sum_p \sum_q (I_1(p,q) - I_2(p,q))^2 \tag{1.2}$$

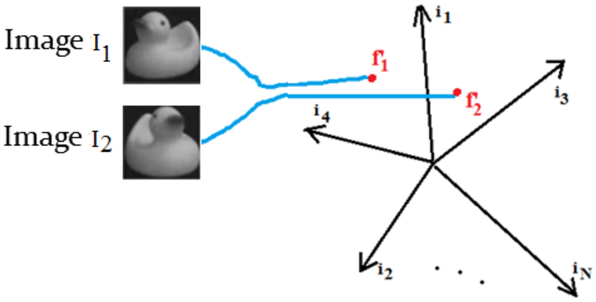


Fig. 1.10 SSD computation between query and model image.

The basic principle of dimensionality reduction may be explained by the the example of a distribution of 3D points that lie on a 2D plane as depicted by figure 1.11. It is redundant to represent each point with 3 coordinates (see figure). If we use a new coordinates system (e1, e2) that lies on the plane, each point can represented with just 2 coordinates.

We consider M images, each one is represented in the N dimensional space, where is the number of pixels of the image (see figure 1.12). The appearance distribution provides insights into whether there is a correlation between images. In this context, the distribution of feature points is highly structured and often lies within a low-dimensional subspace.

In order to express the feature point distribution in a lower dimension space ($k < N$) and the calculate the new basis $(e_1, \ldots, e_k)$, we perform the following steps:

- step 1: **Subtracting the mean**

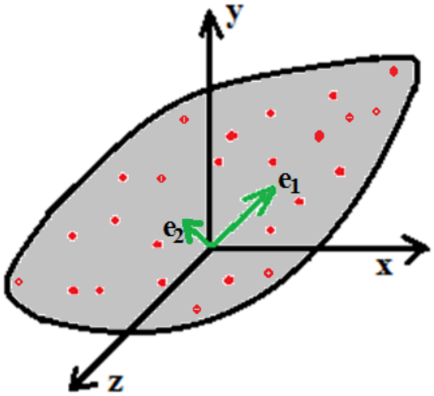


Fig. 1.11 Reduction the representation of points from 3D to 2D.

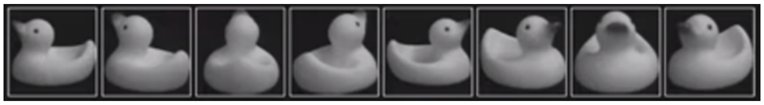


Fig. 1.12 Object Image set where there is a redundancy between successive images).

Given M images $(f'_1, f'_2, \ldots, f'_M)$ of an object, the Mean Image is:

$$c = \frac{1}{M}\sum_{m=1}^{M} f'_m \tag{1.3}$$

Step 2: Subtract the mean from the object image set so as to move the origin of the new basis to the centroid of the distribution.

$f_m = f'_m - c$

- Step 3: The first Principal component (e_1) corresponds to the direction of maximum variance in the image set.

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{1.4}$$

Knowing e_1, the image is represented by projecting it onto the principal component e_1. Image is then represented by a single number p, where $p = e_1.f$ (dot product) (see figure 1.13).

Step 3: The second principal component e_2 is the direction of the second maximum variance in the image set such that: $(e_1 \perp e_2)$. Image is the represented by a two numbers p_1, p_2. $p^T = (p_1 p_2)^T = (e_1 e_2)^T f$

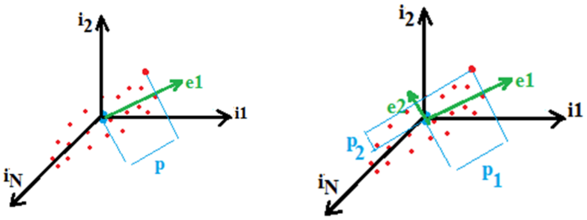


Fig. 1.13 New basis.

Step 4: The k^{th} principal component e_k is the direction of the k^{th} maximum variance in the image set such that: $e_1 \perp e_2 \perp e_3 .... \perp e_k$. Image is represented by a k numbers $p_1, p_2, .., p_k$ (in $k \leq N$ dimensions): The forward projection is written as follow:

$$p^T = (p_1 p_2 ... p_k)^T = (e_1 e_2 e_3 ... e_k)^T f \tag{1.5}$$

The Back projection is given by:

$$f = \sum_{i=1}^{k} p_i e_i \tag{1.6}$$

$(e_1, e_2, \ldots, e_k)$ is referred to as **Linear Subspace**.

1.3.4 Finding Principal Components

Given the Mean-Subtracted Image Set $(f_1, f_2, ..., f_M)$, f_i is $(N \times 1$ vector.

Find the orthogonal basis $(e_1, e_2, \ldots, e_k)$ where e_i is $(N \times 1$ vector.

Such that, : for each $m = 1..k$, we have:

$$f_m = \frac{1}{n} \sum_{i=1}^{K} p_i^m e_i \tag{1.7}$$

where : $p_i^m = e_i^T f_m$ is the projection of the image f along the i^{th} principal component.

In order to compute the first principal component, we search e that maximizes Variance of the $f.e$:

$$Var(f.e) = E(e^T f f^T e) \tag{1.8}$$

For this, we need to maximizes $E(e^T f f^T e)$ such that $e^T e = 1$. We note :$R = f f^T$, this is equivalent to maximize the objective function $L(e, \lambda)$ such that:

$$L(e, \lambda) = e^T Re - \lambda(e^T e - 1) \tag{1.9}$$

Taking derivatives of $L(e, \lambda)$ with respect to e and equating to zero: $Re - \lambda e = 0$, then $Re = \lambda e$: The first principal Component is the eigenvector corresponding to the maximum eigenvalue.

The algorithm

- Data Matrix: $F = [f_1, f_2, .., f_M]$

Covariance Matrix $R = FF^T$

$X_i = X_j = fe$

$Cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))]$

Solve EigenValue problem:

Eigenvalues: $(\lambda_1, \lambda_2, \ldots, \lambda_k)$

Eigenvectors: $(e_1, e_2, \ldots, e_k)$

1.3.5 Parametric Appearance Representation

From an object image set, we get K Eigenvectors (see example below)

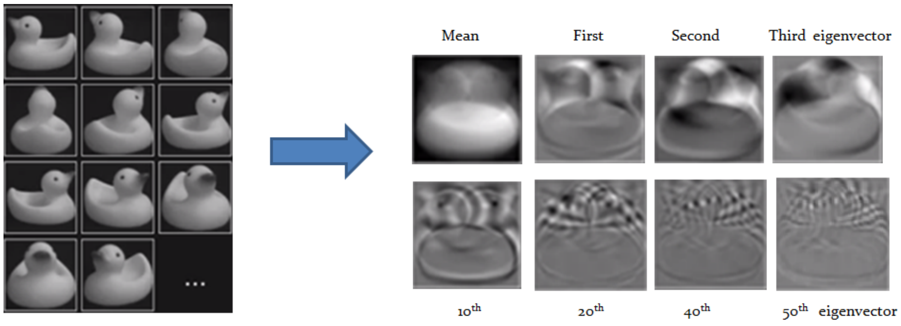


Fig. 1.14 (Left) Initial dataset, (Right) The new basis defined by K eigen vectors.

How many principal components (K) are sufficient? If we want to capture 95% of variations in the data set, we examine the cumulative explained variance ratio.

Each eigenvalue corresponds to a principal component and represents the amount of variance captured by that component. We normalize the eigenvalues to compute the explained variance ratio for each component as follow:

Explained Variance Ratio=(Eigenvalue of Component / (Sum of All Eigenvalues Explained Variance Ratio)

After this, we compute Cumulative Explained Variance as the sum of the explained variance ratios sequentially of the first K components:

At the end, we select K such that the cumulative explained variance is at least 95% as shown by figure 1.15.

1.3.6 Appearance Matching

Algorithm 1: Dataset representation

Given M learning images $I_1^{(q)}, I_2^{(q)}, \ldots, I_M^{(q))}$ for each object $q(= 1..Q)$ of Q training objects

1- Normalize all images to remove brightness variations: $I_m^{'(q)} = \frac{I_m^{'(q)}}{||I_m^{'(q)}||}$

2- Convert image $I_m^{'(q)}$ to a vector $f_m^{'(q)}$

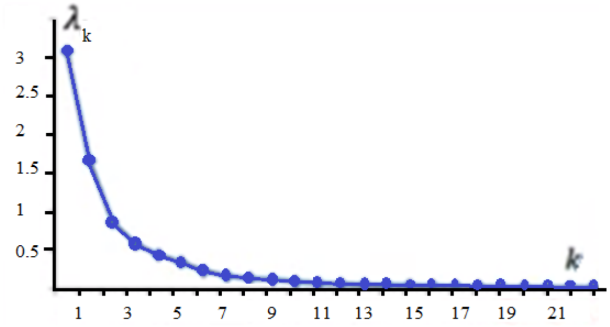


Fig. 1.15 Computation of the Cumulative Explained Variance.

3- Compute the mean vector c^q of each object q.

4-Subtract the mean feature vector c^q for object q

$f_m^{(q)} = f_m^{'(q)} - c^q$ 5- Construct the data matrix and covariance matrix:

$F^q = |f_1^{(q)} f_2^{(q)} f_3^{(q)} .. f_M^{(q)}|$

$R^q = F^q F^{q^T}$ 6- Compute the K eigenvectors

$e_1^{(q)}, e_2^{(q)}, e_3^{(q)}, ..., e_K^{(q)}$ of $R^{(q)}$ 7- Project feature vector to eigenvectors for object q:

$p_m^{(q)} = [e_1^{(q)}, e_2^{(q)}, e_3^{(q)}, ..., e_K^{(q)}]^T \times f_m^{(q)}$

Algorithm 2: Object recognition

Given input image (I) for object recognition

1- Normalize the image to remove brightness variations: $I' = I/||I||$

2- Convert image I' to a vector f'

For each object q in the database, perform steps $3 - 6$:

3- Compute the mean vector $c^{(q)}$ of each object.

4-Subtract the mean feature vector $c^{(q)}$ for object q: $f_q = f' - c^{(q)}$

5- Project feature vector to eigenspace for object q:

$p_m^{(q)} = [e_1^{(q)}, e_2^{(q)}, e_3^{(q)}, ..., e_K^{(q)}]^T \times f^{(q)}$

6- In the eigenspace of object q find the closest point to projected point, compute the distance $d^{(q)}$.

7- Find the object for which $d^{(q)}$ is minimum.

1.4 Example of face recognition

Figure 1.16 shows sample faces of the dataset.

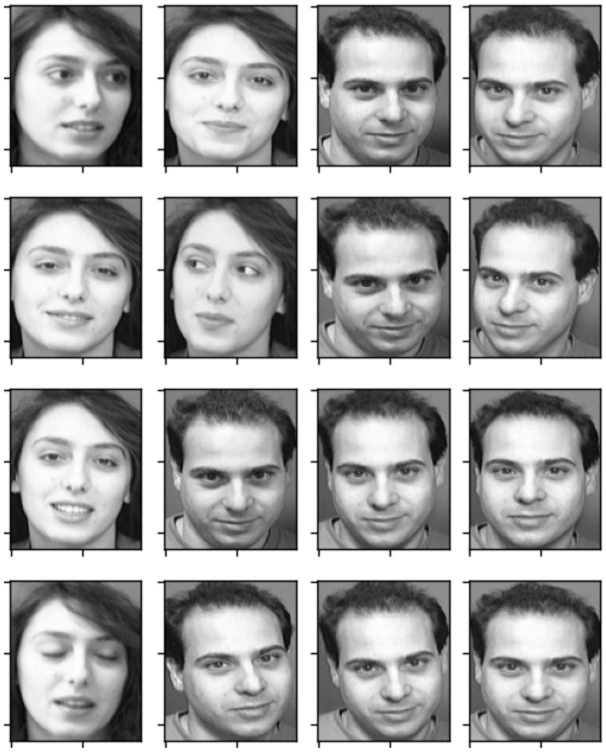


Fig. 1.16 A sample of images of the dataset of faces.

Figure 1.17 illustrates the explained variance ratio. The the first 16 eigenfaces are presented by figure 1.18.

The test on out-of-sample image of existing class, we obtain the result depicted by figure 1.19.

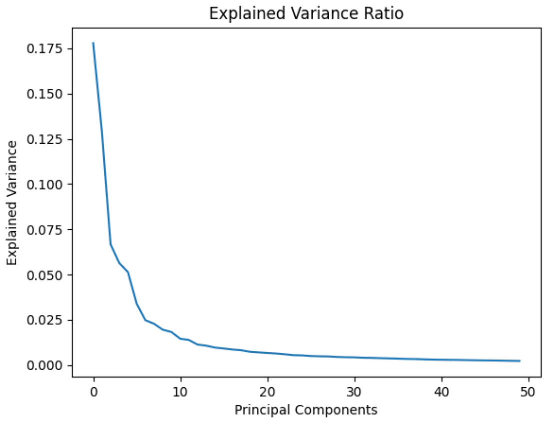


Fig. 1.17 The explained variance ratio.

The test on out-of-sample image of new class, we obtain the result depicted by figure 1.20.

Using the first 16 eigen vectors, the reconstructed face corresponding to the face illustrated by figure (left) is presented at the right of the same figure 1.21.

1.5 Conclusion

In this chapter, we presented the fundamental aspects of object recognition from visual appearance. We examined the two primary paradigms of object recognition: shape representation and learning-based appearance recognition.

The shape-based approach highlighted the utility of geometric features in identifying objects, providing a robust framework for applications where structural consistency is key. On the other hand, the learning-based appearance methods showcased the adaptability and effectiveness of modern techniques that leverage statistical and machine learning tools.

For object recognition, we explained the requirement of dimensionality reduction using Principal Component Analysis (PCA) for efficient data representation. High-

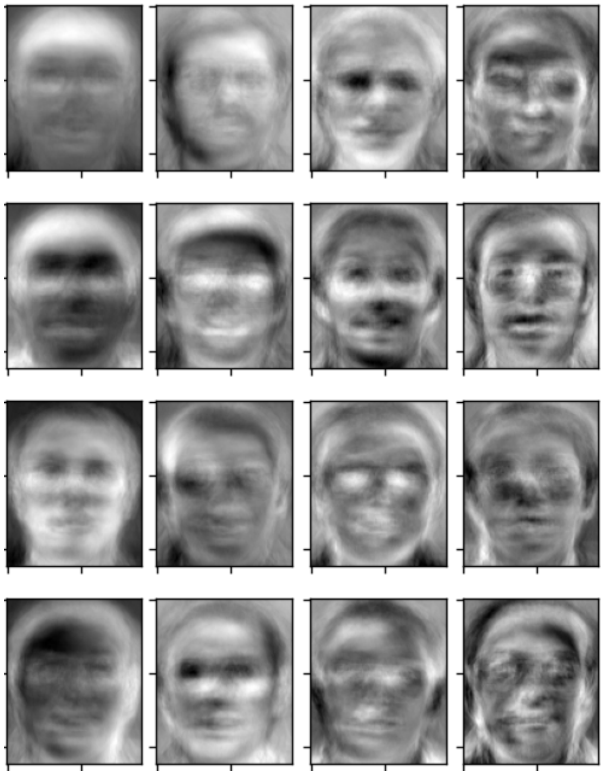


Fig. 1.18 The the first 16 Eigenfaces.

dimensional data can be transformed into a compact yet descriptive form and are used to retrieve query images.

The end of this chapter is devoted for illustrating how to retrieve a query image of face using a dataset of faces.

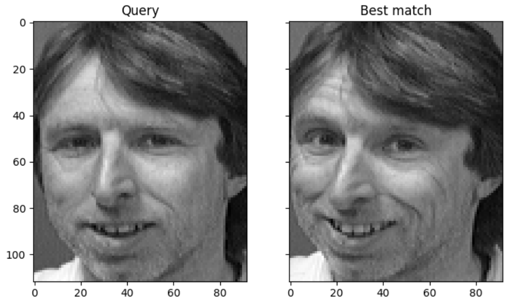


Fig. 1.19 Result (right) of recognition of the query face (left).

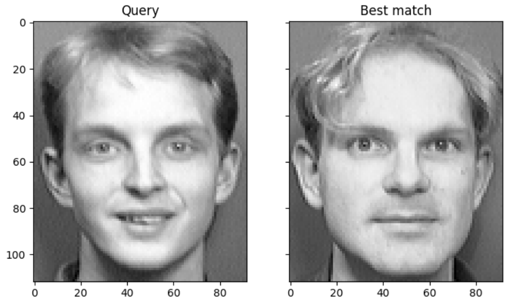


Fig. 1.20 Result (right) of recognition of the query face (left) of a new class.

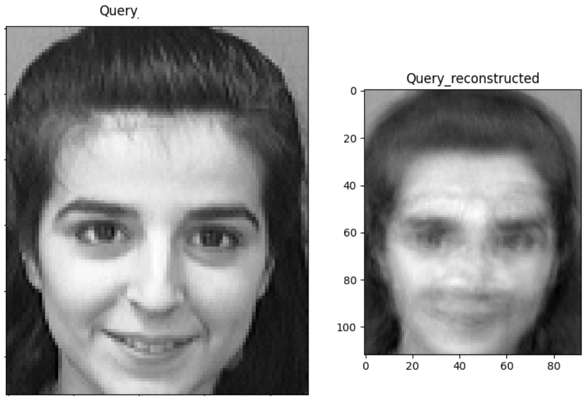


Fig. 1.21 Result (right) of recognition of the query face (left) of a new class.

References

1. Seymourt Papert. The Summer Vision Project. Artificial Intelligence Group, MIT, 1966. https://dspace.mit.edu/handle/1721.1/11589
2. Bob Sumner. Augmented Creativity, TEDxZurich, 19th of January, 2015. https://www.youtube.com/watch?v=AJJOWemfOYI
3. Sachiko Iwase and Hideo Saito. Tracking soccer players based on homography among multiple views”, Proc. SPIE 5150, Visual Communications and Image Processing 2003, (23 June 2003); https://doi.org/10.1117/12.502967
4. Martin A. Fischler, Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, Volume 24, Issue 6, pp 381–395, https://doi.org/10.1145/358669.358692
5. ”The Eye of a Robot: Studies in Machine Vision at MIT” and ”TX-O Computer”. Courtesy of MIT Museum.
6. Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, 2000.
7. D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), volume 1, pages 195-202, Madison, WI, June 2003.
8. Zhengyou Zhang. A Flexible New Technique for Camera Calibration. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 11, NOVEMBER 2000
9. A computer algorithm for reconstructing a scene from two projections. Nature volume 293, pages 133–135 (1981)
10. OD Faugeras, QT Luong, SJ Maybank. Camera self-calibration: Theory and experiments. Second European Conference on Computer Vision ECCV’92, 1992
11. Olsson, Carl; Enqvist, Olof, Stable Structure from Motion for Unordered Image Collections Scandinavian Conference on Image Analysis, 2011.
12. Marc Pollefeys and Luc Van Gool. 3-D Modeling from Images. COMMUNICATIONS OF THE ACM July 2002/Vol. 45, No. 7.
13. Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Xin ; O-CNN: Octree-based Convolutional Neural Networks for 3D Shape Analysis , July 2017. ACM Transactions on Graphics, 36(4):1-11
14. M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, A. Eriksson. Deep Level Sets: Implicit Surface Representations for 3D Shape Inference. 2019. arXiv:1901.06802 [cs.CV]. https://arxiv.org/pdf/1901.06802.pdf

15. A. Podlozhnyuk, S. Pirker, C. Kloss. Efficient implementation of superquadric particles in Discrete Element Method within an open-source framework. Computational Particle Mechanics 4(1), 2016. DOI: 10.1007/s40571-016-0131-6
16. M.A. Turk et al. Face recognition using eigenfaces,CVPR 1991.
17. S.K.Nayar et al., Real-Time 100 Object Recognition System, ICRA, 1996
18. C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3dr2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision, pages 628–644. Springer, 2016
19. R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In European Conference on Computer Vision, pages 484–499. Springer, 2016
20. D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In Advances in Neural Information Processing Systems, pages 4996–5004, 2016
21. S. R. Richter and S. Roth. Matryoshka networks: Predicting 3d geometry via nested shape layers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1936–1944, 2018
22. C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., 1(2):4, 2017.
23. Y. Liao, S. Donne, and A. Geiger. Deep marching cubes: ´ Learning explicit surface representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2916–2925, 2018
24. H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., volume 2, page 6, 2017
25. Mateusz Michalkiewicz et al. Deep Level Sets: Implicit Surface Representations for 3D Shape Inference. arXiv:1901.06802v1 [cs.CV], 21 Jan 2019.
26. Roberts, Lawrence G. 1963. Machine perception of three-dimensional solids. Outstanding PhD dissertations in the computer sciences. Garland Publishing, New York. 1963.
27. Adolfo Guzman-Arénas, Computer Recognition of Three-Dimensional Objects In a Visual Scene. PhD Dissertations in the computer sciences, MIT 1968.
28. M.B. Clowes. On seeing things. Artificial Intelligence, Volume 2, Issue 1, Spring 1971, Pages 79-116.
29. David L. Waltz. GENERATING SEMANTIC DESCRIPTIONS FROM DRAWINGS OF SCENES WITH SHADOWS. MIT Artificial Intelligence Laboratory. Technical Report 271, November 1972.

30. Kemp, M. Julesz's joyfulness. Nature 396, 419 (1998). https://doi.org/10.1038/24753
31. Julesz B. Foundations of Cyclopean Perception. Chicago University Press; Chicago, IL, USA: 1971.
32. David Marr. Vision, A Computational Investigation into the Human Representation and Processing of Visual Information. Originally published: San Francisco : W. H. Freeman, c1982.
33. Arthur Coste. Affine Transformation, Landmarks registration, Non linear Warping. https://www.sci.utah.edu/ãcoste/uou/Image/project3/ArthurCOSTE_Project3.pdf October 2012.
34. Sid Yingze Bao, Mohit Bagra, Yu-Wei Chao, Silvio Savarese. Semantic Structure From Motion with Points, Regions, and Objects. CVPR 2012.
35. Marco Crocco, Cosimo Rubino, Alessio Del Bue. Structure from Motion with Objects, CVPR 2016.
36. D.P. Robertson and R. Cipolla. Structure from Motion. In Varga, M., editors, Practical Image Processing and Computer Vision, John Wiley, 2009.
37. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. International Journal of Computer Vision, 9(2):137–154, 1992.
38. P. F. Sturm andW. Triggs. A factorization based algorithm for multi-image projective structure and motion. In European Conference on Computer Vision (ECCV'96), pages 709–720, 1996.
39. F. Schaffalitzky, A. Zisserman, R. I. Hartley, and P. H. S. Torr. A six point solution for structure and motion. In European Conference on Computer Vision (ECCV'00), pages 632–648, 2000.
40. W. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment: A modern synthesis. In W. Triggs, A. Zisserman, and R Szeliski, editors, Vision Algorithms: Theory and Practice, LNCS, pages 298–375. Springer Verlag, 2000.
41. D. C. Brown. The bundle adjustment - progress and prospects. International Archives of Photogrammetry, 21(3), 1976.