Pattern Recognition 45 (2012) 3566–3579

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# A general framework for subspace detection in unordered multidimensional data

Leandro A.F. Fernandes [a,*], Manuel M. Oliveira [b]

[a] Instituto de Computação, Universidade Federal Fluminense (UFF), CEP 24210-240 Niterói, RJ, Brazil
[b] Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), CP 15064 CEP 91501-970, Porto Alegre, RS, Brazil

A R T I C L E  I N F O

*Article history:*
Received 14 August 2011
Received in revised form
30 December 2011
Accepted 19 February 2012
Available online 5 March 2012

*Keywords:*
Hough transform
Geometric algebra
Parameter space
Subspace detection
Shape detection
Blade
Grassmannian
Coordinate chart
Line
Circle
Plane
Sphere
Conic section
Flat
Round
Quadric

A B S T R A C T

The analysis of large volumes of unordered multidimensional data is a problem confronted by scientists and data analysts every day. Often, it involves searching for data alignments that emerge as well-defined structures or geometric patterns in datasets. For example, straight lines, circles, and ellipses represent meaningful structures in data collected from electron backscatter diffraction, particle accelerators, and clonogenic assays. Also, customers with similar behavior describe linear correlations in e-commerce databases. We describe a general approach for detecting data alignments in large unordered noisy multidimensional datasets. In contrast to classical techniques such as the Hough transforms, which are designed for detecting a specific type of alignment on a given type of input, our approach is independent of the geometric properties of the alignments to be detected, as well as independent of the type of input data. Thus, it allows concurrent detection of multiple kinds of data alignments, in datasets containing multiple types of data. Given its general nature, optimizations developed for our technique immediately benefit all its applications, regardless the type of input data.


## 1. Introduction

Data analysis is a fundamental element in scientific discovery and data mining. In many scientific fields, visual inspection of experimental datasets is often performed in order to identify strong local coherence in the data. Such coherence results from data alignments (in some multidimensional space), and usually emerges as geometric shapes and patterns. For instance, straight lines and circles appear as well-defined structures in the analysis of electron backscatter diffraction (Fig. 1a) and clonogenic essays (Fig. 1c), respectively. However, when large volumes of data need to be analyzed, visual inspection becomes impractical. For this reason, automatic detectors for specific types of data alignments have been broadly applied by scientists in many different areas, such as particle physics [1,2], astronomy [3,4], microbiology [5,6], crystallography [7,8], and medicine [9,10]. Such detectors are also a central component of many computer vision and image processing applications [11–13]. The goal of automatic detectors is to identify certain kinds of alignments that best fit a given unordered dataset, even in the presence of noise and discontinuities.

We describe a general approach for detecting data alignments in unordered noisy multidimensional data. Our approach is based on the observation that a wide class of alignments, and also input data entries, can be represented as linear subspaces. Thus, instead of defining a different detector for each specific case and input data type, it is possible to design a unifying framework to detect the occurrences of emerging subspaces in multidimensional datasets. In our framework, these datasets may be heterogeneous and contain entries with different dimensionalities (Fig. 2).

Our approach has a broad range of applications as a pattern detection tool. For instance, it can be applied, without any changes, to all kinds of data alignments that can be represented

* Corresponding author. Tel.: +55 21 2629 5665; fax: +55 21 2629 5669.
*E-mail addresses:* laffernandes@ic.uff.br (L.A.F. Fernandes), oliveira@inf.ufrgs.br (M.M. Oliveira).
*URLS:* http://www.ic.uff.br/~laffernandes (L.A.F. Fernandes), http://www.inf.ufrgs.br/~oliveira (M.M. Oliveira).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.02.033

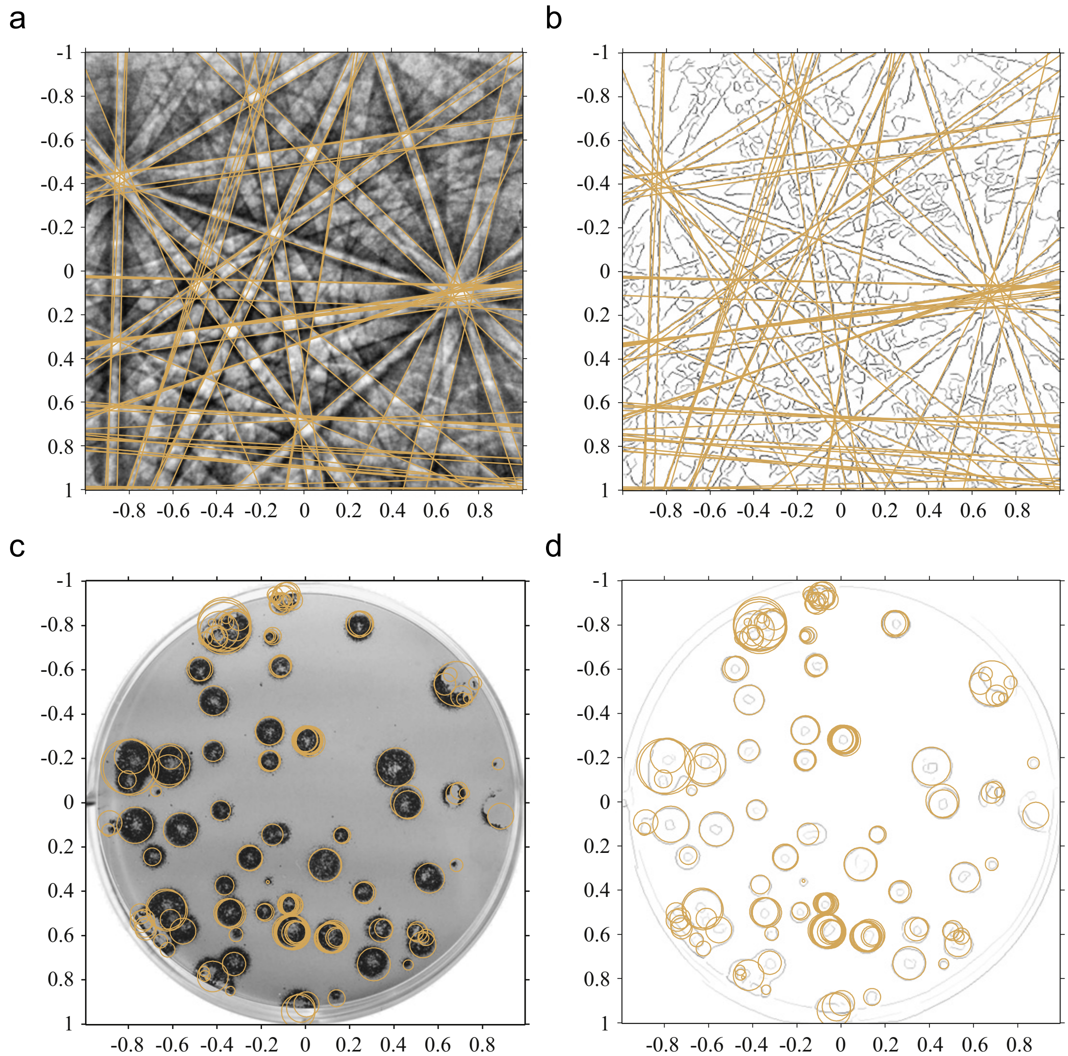


**Fig. 1.** (a) Electron backscatter diffraction image taken from a particle of wulfenite. The detection of straight lines is key for the identification of the particle's crystalline phase. (c) Gray image of infection with H1N1 in MDCK-SIAT1 cells. The detection of circles is important for automated counting process in clonogenic assays. Our approach was used, without any changes, to automatically detect the straight lines and circles shown in (a) and (c) from the edge information shown in (b) and (d), respectively. (a) $445 \times 445$ pixels. (b) $445 \times 445$ pixels. (c) $529 \times 534$ pixels. (d) $529 \times 534$ pixels.

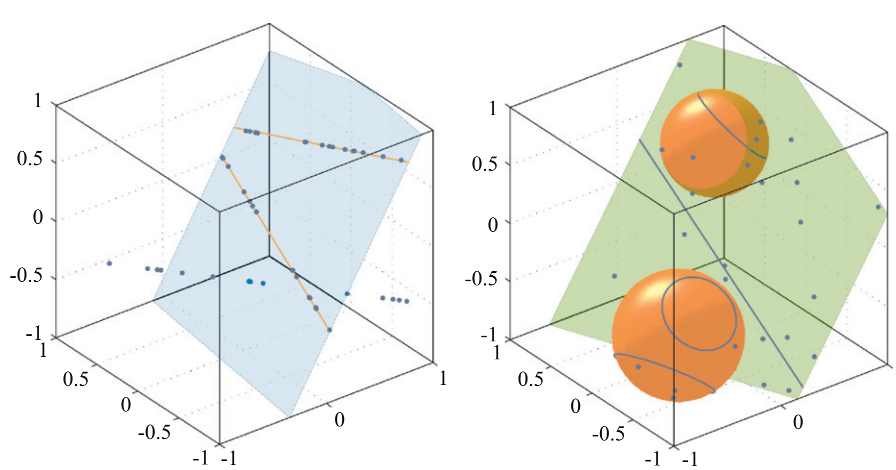


**Fig. 2.** Shape detection on heterogeneous synthetic datasets using our approach. (left) Detection of lines on the input plane that best fit subsets of input points. (right) Concurrent detection of plane and spheres by a single application of the proposed approach. The input dataset is composed by points, circles, and straight line.

as linear subspaces in any complete metric spaces (see Section 3). Examples include, but are not limited to, data alignments decomposed into real- or complex-valued vector spaces, orthogonal polynomials, wavelets, and spherical harmonics. For the purpose of illustration, however, we restrict the examples shown in the paper to the important problem of detecting analytic geometric