On the Generalization of Subspace Detection in Unordered Multidimensional Data*

Leandro A. F. Fernandes[1], **Manuel M. Oliveira**[1] **(Advisor)**

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{laffernandes,oliveira}@inf.ufrgs.br

Abstract. *We present a generalized closed-form framework for detecting data alignments in large unordered noisy multidimensional datasets. In our approach, the intended type of data alignment may be a geometric shape (*e.g.*, straight line, circle, conic section) or any other structure, with arbitrary dimensionality, that can be characterized by a linear subspace. We also present an extension of our detection scheme to data with Gaussian-distributed uncertainty. The proposed extension makes the framework more robust to the detection of spurious alignments. In contrast to existing solutions, the proposed approach is independent of the geometric properties of the alignments to be detected. Also, it is independent of the type of input data and automatically adapts to entries of arbitrary dimensionality. This allows application of the proposed framework (without changes) in a broad range of applications as a pattern detection tool.*

1. Introduction

A central component of many computer vision and data mining applications is to identify data alignments that emerge as well-defined structures or geometric patterns in datasets. The identification of data alignments is also key in scientific fields such as particle physics and astronomy, because data alignments define strong local coherence in data, and hence, important features to be analyzed. For this reason, automatic detectors have been developed and used both by computer scientists as well as by researcher in many different areas.

The development of automatic detectors has been explored and extended in many ways in order to produce techniques robust to the presence of noise and discontinuities in large datasets. But, traditionally, detectors have been designed for specific types of alignments in a given type of input data. Such a specialization prevents the development of generally applicable techniques and optimizations due to specificities in their formulations. Thus, improvements on existing solutions need to be done on a case-by-case basis.

This work introduces a more general approach for detecting data alignments in unordered noisy multidimensional data. It is focused on fulfill the lack of generality of existing solutions. The proposed approach is based on the observation that a wide class of alignments (*e.g.,* straight lines, planes, circles, spheres, conic sections, among others), as well as input 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

*L. A. F. Fernandes was supported by CNPq (process 142627/2007-0).

Keywords: subspace detection, parameterization, generalization, Hough transform, geometric algebra.

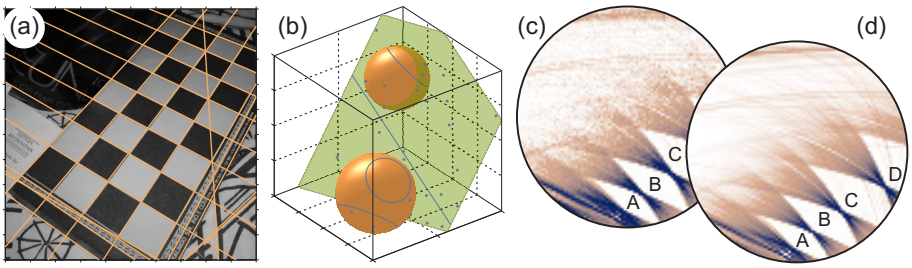

Figure 1. (a) Detection of the 22 lines that best fit the edge pixels of the image. (b) Concurrent detection of plane and spheres. The input dataset is comprised by 43 points, 1 straight line, and 3 circles. The proposed approach was used, without any changes, to automatically detect the structures shown in (a) and (b). (c)-(d) A portion of the accumulator arrays produced for (a) using the sampling-based and the error propagation-based voting schemes, respectively. The blue regions A-D represent peaks of votes (*i.e.,* detected straight lines).

datasets. The versatility of the framework is demonstrated in Figure 1, where it is applied both on a straight line detection case (Figure 1a) as well as on concurrent detection of multiple kinds of alignments with different geometric interpretations, in datasets containing multiple types of data (Figure 1b). Given its general nature, optimizations developed for the proposed framework immediately benefit all the detection cases.

The contributions of this work include: (i) a general framework for subspace detection in unordered multidimensional datasets; (ii) a parameterization scheme for subspaces based on the rotation of a canonical subspace with the same dimensionality; (iii) an algorithm that enumerates all instances of subspaces with a given dimensionality p that either contain or are contained by an input subspace of arbitrary dimensionality; (iv) a procedure that maps subspaces with Gaussian distributed uncertainty to the parameter space characterizing p-dimensional subspaces; (v) a number of experimental evidences supporting that the open affine covering of the Grassmannian (*i.e.,* the set of all p-dimensional linear subspaces of a vector space $\mathbb{R}^n$) can be used as an auxiliary space where the uncertainty of some classes of analytical geometric shapes can be handled in a unified fashion; and (vi) an algorithm that identifies local maxima in a multidimensional histogram.

Due to space limitation, this paper does not present a detailed description of the proposed algorithms or results achieved. The full dissertation [Fernandes 2010] and the list of related publications, courses, pending submissions and implementations [Fernandes and Oliveira 2008, Fernandes and Oliveira 2009, Fernandes and Oliveira 2010, Fernandes and Oliveira 2011] are available in *http://www.inf.ufrgs.br/∼laffernandes/ctd2011*.

2. Geometric Algebra

We have formulated the subspace detector using Geometric Algebra (GA). GA is a powerful mathematical system encompassing many mathematical concepts (*e.g.,* complex numbers, quaternions, and Plücker coordinates) under the same formalism. In GA, subspaces are treated as primitives for computation. As such, it is an appropriate tool for modeling the subspace-detection problem. Also, GA has been proven to be capable of representing many types of geometry. It is because GAs can be constructed over any type of quadratic space, which includes real-valued vector spaces, and also more sophisticated

Hilbert spaces, such as finite Fourier basis, finite random-variable spaces, and basis of orthogonal polynomials, among others. In all cases, the concepts of subspaces, intersections and combinations of subspaces are still valid and independent of the underlying metric space, even though they may not have the same geometric meaning. We explore the generality of these concepts while defining our subspace detection scheme.

By assuming a model of geometry (MOG), one defines the space where data will be encoded and provides a practical (geometric) interpretation to subspaces as input data entries or resulting data alignments. Examples of MOGs successfully encoded by GA include Euclidean, Projective, Spherical, Hyperbolic, and Conic spaces. These MOGs provide practical applications for the proposed technique as a detector of emerging geometric shapes on datasets like, but not limited to, images, volumetric datasets, and point clouds.

Only in the past few years GA became accessible to the computer science researchers through specialized literature. With the aim of disseminating GA within the Brazilian computer science community, the experience obtained from this work has been used in the preparation of courses [Fernandes and Oliveira 2009, Fernandes and Oliveira 2010] for providing an introduction to the fundamental concepts of GA and discussing its great potential as a tool for representing and solving problems in computer graphics, computer vision, and image processing.

3. Related Work

Most of the techniques for detecting data alignments are derived from the Hough Transform (HT), Random Sample Consensus (RANSAC), or Tensor Voting (TV) paradigms. In order to use the HT or RANSAC approaches, one needs to assume a mathematical model for the intended type of data alignment (*e.g.,* the normal equation of the line; the center-radius parameterization of circles) with respect to the expected type of input data (usually points). Although there are partial generalizations of the HT and RANSAC designed to some classes of analytic shapes and HTs for non-analytic shapes, such approaches are still restrictive regarding the assumed input or intended output data. The TV, on the other hand, follows a generalized definition. The TV, however, returns all possible features (with any dimensionality) at the same time. Such a behavior prevents the efficient detection of predefined types of alignments, because it requires a subsequent filtering step.

We propose a voting-based framework for detecting the occurrences of emerging linear subspaces (with a given dimensionality) in multidimensional datasets. The proposed approach is based on the representation of subspaces as primitives in GA (Section 2). By assuming a MOG, such subspaces can be geometrically interpreted as some shape (*e.g.,* straight lines, circles, planes, spheres, among others) or other data alignments (*e.g.,* customer behaviors may emerge as linearly correlated data in e-commerce data). Unlike in conventional HTs, the parameterization used by the proposed approach is independent of the geometric properties of the structure to be detected. When applied to the detection of geometric shapes the proposed framework can be seen as the generalization of the HTs for analytic shapes that can be represented by some linear subspace.

4. Overview of the General Subspace Detection Framework

The proposed subspace detection scheme takes as input a set $\mathcal{X}$ of subspaces (*i.e.,* the input dataset encoded into a MOG), the dimensionality p of subspaces interpreted as the

intended data alignment in the same MOG, and the dimensionality n of the whole underlying space imposed by the MOG. The algorithm outputs the p-dimensional subspaces that best fit the input set $\mathcal{X}$. The detection is performed using a three-step process: (i) create an accumulator array as a discrete representation of the parameter space characterizing p-dimensional subspaces; (ii) perform a voting procedure where the input dataset is mapped to the accumulator array; and (iii) search for the peaks of votes in the accumulator, as they correspond to the p-dimensional subspaces that best fit the input dataset. For the case of uncertain input data, extended mapping and voting procedures are performed in step (ii).

4.1. Parameterization of Subspaces Interpreted as the Intended Data Alignment

In [Fernandes 2010] we show that a p-dimensional subspace in a n-dimensional space can be characterized by a set of $m = p\,(n - p)$ rotations applied to a canonical subspace used as reference, where the values of n and p are related to the MOG where data has been encoded and the type of data alignment one wants to detect, respectively.

By assuming that each one of the m rotation angles (θ_t) related to the sequence of rotation operations are in the $[-\pi/2, \pi/2)$ range, we ensure that such angles define a parameter space for p-dimensional subspaces:

$$\mathbb{P}^m = \{(\theta_1, \theta_2, \cdots, \theta_m) \mid \theta_t \in [-\pi/2, \pi/2)\}, \tag{1}$$

where each parameter vector $(\theta_1, \theta_2, \cdots, \theta_m) \in \mathbb{P}^m$ characterizes an instance of a p-dimensional subspace in the underlying n-dimensional space. The proposed parameterization guarantees the use of the smallest set of parameters in the representation of the intended subspaces. Thus, when applied as a shape detector (Figures 1a and 1b), the proposed approach always represents the intended shapes in the most compact way.

In its first step, the proposed subspace detection framework discretizes $\mathbb{P}^m$, for which an accumulator array is built to receive "votes", and initially set its bins to zero.

4.2. Voting Process for Input Subspaces

The second step maps the input dataset to parameter space. Essentially, the mapping procedure takes each r-dimensional subspace $\mathbf{X}_{\langle r \rangle}$ in the input dataset $\mathcal{X}$ and identifies the parameters (coordinates in $\mathbb{P}^m$, Equation 1) of all p-dimensional subspaces related to it. When $r \leq p$, the mapping procedure identifies in $\mathbb{P}^m$ all p-dimensional subspaces containing $\mathbf{X}_{\langle r \rangle}$. If $r \geq p$, the procedure identifies in $\mathbb{P}^m$ all p-dimensional subspaces contained in $\mathbf{X}_{\langle r \rangle}$. As the input entries are mapped, the bins of the accumulator related to such a mapping are incremented by some importance value of the entry.

In conventional voting-based approaches, such as the HTs, the input data type is known *a priori*. Thus, conventional mapping procedures predefine which parameters of the related parameter vectors must be arbitrated and which ones must be computed. The proposed approach, on the other hand, does not have prior information about input data. It decides at runtime how to treat each parameter. Such a behavior is key for the generality of the proposed detection framework, providing a closed-form solution for the detection of subspaces of a given dimensionality p on datasets that may be heterogeneous and contain elements (*i.e.,* subspaces) with arbitrary dimensionalities ($0 \leq r \leq n$). Such a feature is illustrated by Figure 1b, where the input dataset is comprised by subspaces geometrically interpreted as points, straight line, and circles.

4.3. Voting Process for Input Subspaces with Uncertainty

Experimental (real) data often contain some uncertainty due to imprecision in the instruments used to collect them. Such an uncertainty can be taken into account while performing subspace detection by supersampling input entries according to their distribution of uncertainty and, in turn, by processing each sample with the technique described in Section 4.2. The quality of sampling-based approaches, however, depends on the number of samples, and the computational load increases as more samples are used.

In order to avoid the brute force sampling approach, we propose an extended mapping and voting procedures for input data with Gaussian-distributed uncertainty. The extended mapping procedure is based on first-order error propagation analysis. It transports the uncertainty of each input element throughout the computations into an auxiliary parameter space where the uncertainty is described by a multivariate Gaussian distribution. In turn, such a distribution is mapped to the actual parameter space, leading to non-Gaussian distributions of votes in the accumulator array.

Figures 1c and 1d present a comparison between a portion of the accumulator arrays produced for Figure 1a with a sampling-based voting using the technique described in Section 4.2 and the technique described in the current section, respectively. Notice that error propagation produces smoother distributions of votes than the sampling-based approach. As a result, the latter is less prone to the detection of spurious subspaces.

4.4. Peak Detection

The last step of the subspace detection framework is performed after the voting procedure has been applied to all input data entries $\mathbf{X}_{\langle r \rangle} \in \mathcal{X}$. It consists in identifying the bins that correspond to local maxima in the accumulator array. For this step we propose a sweep-hyperplane-based peak detection scheme developed for accumulator arrays having arbitrary dimensionality. The proposed approach is an extension of the peak detection technique described in [Fernandes and Oliveira 2008] for 2-dimensional accumulator arrays. The technique returns a list with all detected vote peaks, sorted according to their importance (*i.e.,* number of votes). The coordinates of such bins (*i.e.,* parameter vectors) are used to retrieve the most significant p-dimensional subspaces.

5. Results

The proposed approach has been demonstrated by proof of concept implementations of the described algorithms. We have used our own GA library (*i.e.,* Geometric Algebra Template Library, GATL) in such implementations. We intend to make all C++ and MATLAB® code publicly available after the publication of pending submissions.

The implementations have been validated by applying the subspace detection framework to real and synthetic datasets. As a closed-form solution, the same implementation of the proposed framework allows the detection of subspaces that best fit an input set of subspaces with different dimensionalities and different geometric interpretations (*e.g.,* points, straight line and circles – Figure 1b). Also, it allows the concurrent detection of subspaces with different geometric interpretations but with the same dimensionality in a given MOG (*e.g.,* plane and spheres – Figure 1b). Our results have shown that the proposed approach can identify subspaces even in the presence of noise and outliers.

In [Fernandes 2010] we show that an approximation of the dth-order Voronoi diagram of a set of points in $\mathbb{R}^d$ can be retrieved as byproduct of the detection of subspaces geometrically interpreted as circles, spheres, and their higher-dimensional counterparts.

6. Conclusions

We presented a framework for detecting emerging data alignments in unordered noisy multidimensional data. The proposed subspace detector is based on a voting strategy, and it is formulated with GA. By doing so, the technique takes advantage of the conceptual simplicity of the voting paradigm for pattern recognition, while exploring the superior modeling capability of computational primitives and operations in GA.

The time complexity of our approach is the same as of conventional HTs[1] multiplied by p^2, for $r \geq p$: $\mathcal{O}(p^2\,(m-k)\,s^k\,N)$. A naive implementation of our approach suffers from the same drawbacks as HTs: large memory requirement and computational cost. However, as any HT, it is robust to the presence of outliers and is suitable for implementation on massively parallel architectures. Moreover, the generality of our technique should enable new and exciting applications in many different areas, because it avoids tailoring a different solution for each specific case of detection. As a result, we believe it will stimulate research on new optimization approaches for subspace detection. We also hope it will contribute to the popularization of GA among the computer science community.

We have demonstrated the application of the proposed approach on datasets chosen because of their visually-compelling structures. However, one should note that, given its generality, our framework is not restricted to the detection of geometric shapes. It can be applied to any domain in which a problem can be cast as a subspace detection one. For example, the subspace clustering problem in data mining applications. Also, the proposed general parameterization for data alignments may be useful while defining machine learning techniques. Since our approach is independent of the metric properties of the underlying space where data resides, it can be used, without any change, for the detection of subspaces having different interpretations (*e.g.,* different MOGs), *including some that may be defined in the future*.

References

Fernandes, L. A. F. (2010). *On the generalization of subspace detection in unordered multidimensional data*. PhD thesis, PPGC-UFRGS, Porto Alegre, Brazil.

Fernandes, L. A. F. and Oliveira, M. M. (2008). Real-time line detection through an improved Hough transform voting scheme. *Pattern Recognit.*, 41(1):299–314.

Fernandes, L. A. F. and Oliveira, M. M. (2009). Geometric algebra: a powerful tool for solving geometric problems in visual computing. In *Tutorials of Sibgrapi*, pages 17–30.

Fernandes, L. A. F. and Oliveira, M. M. (2011). A general framework for subspace detection in unordered multidimensional data. *IEEE Trans. PAMI*. (submitted).

Fernandes, L. A. F. and Oliveira, M. M. (Jan. 11-22, 2010). Introduction to geometric algebra. Lecture Series of the VISGRAF Laboratory at IMPA, Summer School in CG.

[1]The time complexity of conventional HTs is $\mathcal{O}((m-k)\,s^k\,N)$, where m is the number of parameters, k is the number of arbitrated parameters, s is the number of samples along one dimension of the accumulator array, and N is the number of input entries. It is usually presented as $\mathcal{O}(s^{m-1}\,N)$ in the literature because it is assumed that only one parameter is not arbitrated, *i.e.,* $k = m-1$.