Index spaces for 3D retrieval : Toward a better understand of their geometry and distribution

Paquet, E.; Viktor, H.L.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1117/12.837238
Proceedings of SPIE, 7526, 2010-01-21

Index spaces for 3D retrieval: toward a better understand of their geometry and distribution

E. Paquet [a, b] and H. L. Viktor [b]
[a] National Research Council, 1200 Montreal Road, Ottawa (Ontario) K1A 0R6 Canada;
[b] School of Information Technology and Engineering, University of Ottawa, 800 King Edward, Ottawa (Ontario) K1N 6N5 Canada

ABSTRACT

Distance is a fundamental concept when considering the information retrieval and cluster analysis of 3D information. That is, a large number of information retrieval descriptor comparison and cluster analysis algorithms are built around the very concept of the distance, such as the Mahalanobis or Manhattan distances, between points. Although not always explicitly stated, a significant proportion of these distances are, by nature, Euclidian. This implies that it is assumed that the data distribution, from a geometrical point of view, may be associated with a Euclidian flat space. In this paper, we draw attention to the fact that this association is, in many situations, not appropriate. Rather, the data should often be characterised by a Riemannian curved space. It is shown how to construct such a curved space and how to analyse its geometry from a topological point of view. The paper also illustrates how, in curved space, the distance between two points may be calculated. In addition, the consequences for information retrieval and cluster analysis algorithms are discussed.

Keywords: 3D, Clustering, curvature, curved space, distribution, Euclidian, geodesic distance, retrieval, metric, object, Riemannian.

1. INTRODUCTION

The majority of methods which calculate the similarity of information retrieval descriptors for 3D objects are based on the very concept of distance. That is, two descriptors are considered to be closely related if the distance between them is small. The same observation holds for cluster analysis algorithms. For example, partitioning methods, constrained-based methods and spectral methods all make use of a distance measure to construct the clusters [1, 2]. Depending on the nature of the application domain, one of a number of distances may be used in the implementation of such algorithms. These include the Euclidian, Manhattan, Minkovsky, Chebyshev, Mahalanobis and the angular distances [1, 2]. It follows that the selection of the distance measure requires careful consideration, as the success of these algorithms relies heavily on the choice thereof. Furthermore, the above-mentioned distances are Euclidian by nature, in the sense that they assume that the underlying space is flat. As we will see later, this is not always the case. In a flat space, the concept of distance corresponds to either the length of the straight line (the so called Euclidian distance in the literature) in between two points, or to another construction which is consistent with the former and the axioms of Euclidian geometry, such as the Manhattan, angular or Mahalanobis distances. That is, in order to determine how similar two data instances are, we in essence measure a straight line between them. Subsequently, to compare two vectors, one may thus transport one of the vectors so that their origins coincide. The key point here is that, when following this approach, the parallel transportation does not depend on the particular path. Rather, all the paths are considered to be equivalent. As we will see later, this is generally not the case in a curved (or Riemannian) manifold.

This paper is organized as follows. Section 2 presents some background on the general considerations of the curved manifold in the continuous case. When considering the instances used for constructing information retrieval descriptors and clusters, it follows that the data instances represents a discrete domain, or a so-called discrete distribution. To address this issue, in Section 3, we show how a Riemannian space may be associated with such a discrete distribution. Section 4 discusses the geometrical interpretation of this distribution, specifically focusing on constrained clustering, equivalence and impact of the curse of dimensionality for information retrieval. In Section 5, we show how the distance between points may be calculated in a curved space. Section 6 concludes the paper.

2. EUCLIDIAN AND RIEMANNIAN SPACES: GENERAL CONSIDERATIONS IN THE CONTINUOUS CASE

In this section, we present some notions related to Riemannian manifolds [3] that will facilitate, subsequently, our comprehension. Unless stated otherwise, all summations are to be taken from 1 to *n*: *n* being the number of dimensions associated with the distribution of *x* being an *n*-dimensional vector. In a Riemannian manifold, one may assume that it is possible to associate, with the manifold, a symmetrical tensor called the metric, which is defined by the quadratic form

$$ds^2 = \sum_{\mu\nu} g_{\mu\nu}\left(x\right) dx^{\mu} dx^{\nu} \tag{1}$$

In other words, it is assumed that the infinitesimal distance is a local invariant under general coordinate transformation. It follows that this is to be contrasted with Euclidian spaces, in which, the distance (not necessary infinitesimal) is invariant under orthogonal transformations, i.e. $SO\left(n\right)$. In Riemannian spaces, covariant and contravariant vectors may be defined depending of their behaviour under general coordinate transformation. They are given respectively by

$$T'_{\mu}\left(x'\right) = \sum_{\nu} \frac{\partial x^{\nu}(x')}{\partial x'^{\mu}} T_{\nu}\left(x\left(x'\right)\right) \quad \text{and} \quad T'^{\mu}\left(x'\right) = \sum_{\nu} \frac{\partial x^{\mu}(x')}{\partial x^{\nu}} T^{\nu}\left(x\left(x'\right)\right) \tag{2}$$

depending if their transformation under general coordinate transformation is determined by the Jacobian matrix, or by its inverse. In curved space, one may transform a contravariant quantity into a covariant quantity by multiplying it with the metric; or the inverse metric in the converse case; note that the inverse and the contravariant metric are equal.

$$T^{\mu}\left(x\right) = \sum_{\nu} g^{\mu\nu}\left(x\right) T_{\nu}\left(x\right) \quad \text{where} \quad \sum_{\rho} g_{\mu\rho}\left(x\right) g^{\rho\nu}\left(x\right) = \delta^{\nu}_{\mu} \tag{3}$$

The derivative in Euclidian space is related to the concept of slope or, in other words, the difference in between two points located at two distinct positions. Such an operation is not problematical in standard Euclidian geometry since the space is flat, homogeneous and isotropic. For a Riemannian manifold, the space is not flat but curved and one cannot compare two points at two different locations because they "live"' or "reside", so to say, in two different tangential spaces, which are locally Euclidian. Consequently, in order to compare two vectors, or its generalisation, a tensor, one first has to transport one of the vectors "parallel" to itself to the first location and then compare the two vectors in the same Euclidian tangential space. One can demonstrate [3] that this may be obtained by introducing a covariant derivative defined as

$$\nabla_{\nu} T_{\mu}\left(x\right) \equiv \frac{\partial T_{\mu}\left(x\right)}{\partial x^{\nu}} - \sum_{\sigma} \Gamma^{\sigma}_{\mu\nu}\left(x\right) T_{\sigma}\left(x\right) \tag{4}$$

for which the correction associated with the parallel transport is provided by the affine connection which is related to the metric by

$$\Gamma^{\sigma}_{\mu\nu}\left(x\right) \equiv \sum_{\rho} \tfrac{1}{2} g^{\sigma\rho}(x) \left(\frac{\partial g_{\mu\rho}\left(x\right)}{\partial x^{\nu}} + \frac{\partial g_{\rho\nu}\left(x\right)}{\partial x^{\mu}} - \frac{\partial g_{\mu\nu}\left(x\right)}{\partial x^{\rho}} \right) \tag{5}$$

Riemannian manifolds are not conservative. If a vector is moved parallel to itself along a closed path, the resulting vector does not coincide, in general, with the original vector. For example, if one moves a vector on the surface of the earth from the North pole along a meridian, then along the equator and back to the North pole along another meridian, it follows that the direction of the vector is modified. One may characterise such behaviour by analysing the variation to which a vector is submitted if it is transported along an infinitesimal loop. For an infinitesimal closed path C_{ε}, one may demonstrate [3] that the variation is proportional to the Riemannian curvature tensor

$$\delta T_{\mu}\left(x\right)=-\oint_{C_{\varepsilon}}\sum_{\nu\sigma}\Gamma_{\mu\nu}^{\sigma}\left(x\right)T_{\sigma}\left(x\right)dx^{\nu}=\tfrac{1}{2}\sum_{\nu\sigma\tau}R_{\mu\sigma\tau}^{\nu}\left(x\right)T_{\nu}\left(x\right)A^{\sigma\tau}\left(C_{\varepsilon}\right) \tag{6}$$

where $A^{\sigma\tau}\left(C_{\varepsilon}\right)$ is the bivector associated to the surface bounded by C_{ε} (see also equation (16)) and where the Riemannian curvature tensor is equal to

$$R_{\mu\sigma\tau}^{\nu}\left(x\right)\equiv\frac{\partial\Gamma_{\mu\tau}^{\nu}\left(x\right)}{\partial x^{\sigma}}-\frac{\partial\Gamma_{\mu\sigma}^{\nu}\left(x\right)}{\partial x^{\tau}}+\sum_{\rho}\left(\Gamma_{\rho\sigma}^{\nu}\left(x\right)\Gamma_{\mu\tau}^{\rho}\left(x\right)-\Gamma_{\rho\tau}^{\nu}\left(x\right)\Gamma_{\mu\sigma}^{\rho}\left(x\right)\right) \tag{7}$$

From the inner product between the metric and the Riemannian curvature tensor, one may define the Ricci tensor and the Ricci scalar by contracting the Riemann tensor with the metric

$$R_{\nu\sigma}\left(x\right)\equiv\sum_{\mu\tau\rho}g^{\mu\rho}\left(x\right)g_{\mu\tau}\left(x\right)R_{\nu\rho\sigma}^{\tau}\left(x\right) \tag{8}$$

$$R\left(x\right)\equiv\sum_{\nu\sigma}g^{\nu\sigma}\left(x\right)R_{\nu\sigma}\left(x\right) \tag{9}$$

The Ricci tensor is symmetric [3]. The Riemann tensor and the Ricci tensor and scalar characterise the curvature of the space at a given point. The Riemann and the Ricci tensor are not invariant under general coordinate transformations. Rather, they transform as tensors. However, the Ricci scalar is an invariant since its value is the same, irrespectively of the general coordinate transformation applied. Such a feature is common to all scalars in a Riemannian space. At this stage, it is important to realise that the curvature is not associated with a particular coordinate system, but to the geometrical point itself. For instance, even if the Ricci scalar is an invariant, the coordinates are not.

Instead of defining the metric directly in the curved space, one may define the metric in the Euclidian tangential space at each point of the manifold. One then obtain

$$g_{\mu\nu}\left(x\right)=\sum_{ab}e_{\mu}^{a}\left(x\right)e_{\nu}^{b}\left(x\right)\delta_{ab} \tag{10}$$

where e_{μ}^{a} is the μ component of vector *a* spanning the tangential space; the Greek index being in curved space, while the Latin index being in flat or Euclidian space. This relation will be useful when we aim to establish a correspondence in between the continuous and the discrete cases, in order to apply this concept to information retrieval and cluster analysis spaces.

3. ASSOCIATION OF A RIEMANNIAN SPACE TO A DISCRETE DISTRIBUTION

In the previous section, we reviewed the basic concepts associated with Riemannian spaces. For all these concepts, it was assumed that the manifold is continuous. Very often, through, in practice, a distribution is by no mean a continuous space; on the contrary, it is constituted of a finite number of discrete points. For example, consider an algorithm which is used to create descriptors (or indexes) for a set of 3D objects, to be used for similarity search. Here, each of the descriptors is represented as a point in a discrete space. This implies that in order to proceed further; we must be able to transpose all our previously defined concepts on such a finite discrete set of data. In this section, we define our concepts in terms of geometric quantities.

Such a transposition may be achieved by tessellating the vertices forming the distribution. This is accomplished by finding, for each one of these vertices, the nearest neighbours and by constructing a set of connected simplices [3, 4]. In the following, we transpose a remarkable approach for quantum gravity developed in [4] to our specific problem. For instance, if the distribution of the data points is bidimensional, a simplex is represented by a triangle spanned by two vectors sharing a common origin, in three dimensions it is represented by a tetrahedron, and so on. The distribution is

represented by the ensemble of all the connected simplices associated to it. Each simplex is then considered as a local Euclidian flat space, while the various curvatures defined earlier are associated with the rotations, i.e. $SO(n)$, in between two or many simplices (to be defined later). With such a representation, the metric may be obtained with an approach similar to equation (10). The Euclidian simplex is assimilated to a local Euclidian flat space while the metric is obtained from the inner product in between the vectors spanning that space, i.e. the edges of the simplex. Consequently, the metric is given by

$$g_{ij}(s) = e_i \cdot e_j \tag{11}$$

where e_i is the vector that goes from the origin of the simplex to point (i) and the dot product is the Euclidian inner product in *n* dimensions. One may also define the square of the distance in between two points belonging to a given simplex

$$l_{i,j}^2 = \sum_{\mu\nu} \delta_{\mu\nu} \left(x(i) - x(j)\right)^{\mu} \left(x(i) - x(j)\right)^{\nu} = \left|e_i - e_j\right| \tag{12}$$

A point of notation here: in the following we will use *a*, *b* … to identify the vectors forming a given simplex; *i*, *j* … for the vertices associated with the data; μ, ν … for the coordinates and *n* for the number of dimensions. From the metric, it follows that the volume of the simplex may be obtained from

$$V_n(s) = \tfrac{1}{n!}\sqrt{\det \mathbf{g}(s)} = \frac{(-1)^{\frac{n+1}{2}}}{n!\,2^{n/2}} \det \begin{bmatrix} 0 & 1 & 1 & \cdots \\ 1 & 0 & l_{0,1}^2 & \cdots \\ 1 & l_{1,0}^2 & 0 & \cdots \\ 1 & l_{2,0}^2 & l_{2,1}^2 & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 & l_{n,0}^2 & l_{n,1}^2 & \cdots \end{bmatrix} \tag{13}$$

An n-simplex is bounded by (n-1) dimensional faces. A normal vector can be associated to each one of them

$$\Upsilon_{\mu}(f) = \sum_{\nu_1 \cdots \nu_{n-1}} \varepsilon_{\mu\nu_1 \ldots \nu_{n-1}} e_{(1)}^{\nu_1} \cdots e_{(n-1)}^{\nu_{n-1}} \tag{14}$$

where $\varepsilon_{\mu\nu_1 \ldots \nu_{n-1}}$ is the totally antisymmetric tensor with $\varepsilon_{1\ldots n} = 1$ and where the $\left\{e_{(1)}^{\nu_1} \cdots e_{(n-1)}^{\nu_{n-1}}\right\}$ are the vectors spanning the face. The $n-1$ dimensional volume of the face may be calculated from this vector

$$V_{n-1}(f) = \left(\sum_{\mu=1}^{n} \Upsilon_{\mu}^2(f)\right)^{1/2} \tag{15}$$

Simplices meet on a common volume called a hinge. For instance, a hinge is a point in a bidimensional space (and also a vertex), a line in three dimensions, a triangle in four dimensions, and so on. A bivector (antisymmetric matrix) may be associated with each hinge. The later is defined as

$$\Upsilon_{\mu\nu}(h) = \sum_{\lambda_1 \cdots \lambda_{n-2}} \varepsilon_{\mu\nu\lambda_1 \ldots \lambda_{n-2}} e_{(1)}^{\lambda_1} \cdots e_{(n-2)}^{\lambda_{n-2}} \tag{16}$$

where $\varepsilon_{\mu\nu\lambda_1\ldots\lambda_{n-2}}$ is again the totally antisymmetric tensor and the $\left\{e_{(1)}^{\lambda_1}\cdots e_{(n-2)}^{\lambda_{n-2}}\right\}$ span the hinge. This bivector can be normalized

$$\sum_{\mu\nu}\hat{\Upsilon}_{\mu\nu}\left(h\right)\hat{\Upsilon}^{\mu\nu}\left(h\right)=2 \tag{17}$$

For a given hinge, it is possible to construct a curve that entangles that hinge and that goes through the barycentres of all the simplices meeting at that particular hinge. Mathematically this is called the dual curve and it shall prove itself a useful concept later on. The normalized bivector associated with a hinge and the bivector associated with the surface delimited by the dual curve are related by equation (18) where $A_{C_h}^{\mu\nu}$ is the bivector associated with the dual curve and A_{C_h} is the area associated with the later. Figure 1 shows an example containing five (5) data points in a bidimensional space.

$$\hat{\Upsilon}_{\mu\nu}\left(h\right)A_{C_h}^{\mu\nu}=2A_{C_h} \tag{18}$$

The $n-2$ dimensional volume associated with a hinge can be related to its bivector by the following expression

$$V_{n-2}\left(h\right)=\tfrac{1}{(n-2)!}\left(\sum_{\mu<\nu}\Upsilon_{\mu\nu}^{2}\left(h\right)\right)^{1/2} \tag{19}$$

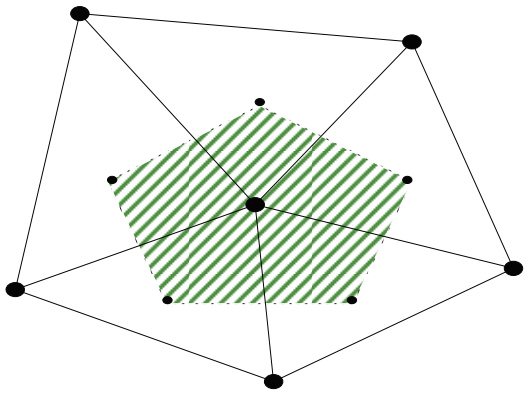

Figure 1. Calculating the area associated with the dual curve in a bidimensional space.

A quantity of particular interest is the angle in between two faces meeting at a particular hinge. This is given by the Euclidian inner product in between the two normal vectors associated with these faces.

$$\cos\theta\left(f,f'\right)=\frac{\Upsilon_{n-1}\left(f\right)\cdot\Upsilon_{n-1}\left(f'\right)}{V_{n-1}\left(f\right)V_{n-1}\left(f'\right)} \tag{20}$$

The scalar product may also be calculated directly from the vectors spanning the two faces from the following determinant

$$\Upsilon_n \cdot \Upsilon_n{}' = \frac{1}{(n!)^2} \det\left[e_i \cdot e_j{}' \right] \tag{21}$$

Perhaps the most important quantity is the deficit angle, in the sense that, as we will show, it is directly related to the space curvature. For a given hinge, one is able to calculate all the angles in between each pair of faces associated to the simplices meeting at that particular hinge. The complement of the sum of these angles is called the deficit angle and, as we will show later, is related to the local curvature.

$$\psi(h) = 2\pi - \sum_{s \cap h \neq \varnothing} \theta(s,h) \tag{22}$$

The deficit angle may be visualized in two dimensions, as follows. Let us consider the object in Figure 2. The vertices, which are also the hinges, for which the deficit angles are null, have no curvature and consequently the space is locally flat. The boundary, though, is characterised by non-zero deficit angles and, due to this reason, presents a curvature that varies from hinge to hinge.

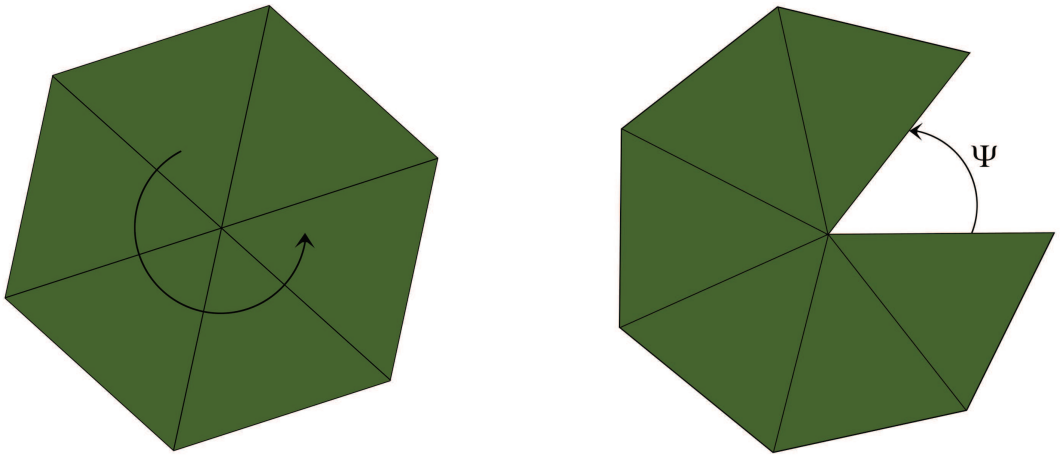

Figure 2. Deficit angle in the case of a bidimensional distribution.

When one adds more dimensions, non-zero deficit angles appear, not only on the boundary, but also inside the manifold.

4. CHARACTERIZATION OF RIEMANNIAN SPACES

In this section, we transpose the notions of the previous section in the case of a discrete distribution, such as an information retrieval space or a cluster. Recall, as we mentioned earlier, that each simplex represents a locally flat Euclidian space. Given two simplices, it is possible to relate their local reference frames by a set of rotations (orthogonal matrices)

$$e_a(s_2) = \sum_b \mathbf{R}_a^b(s_2, s_1)\, e_b(s_1) \tag{23}$$

where $\mathbf{R}_a^b(s_2, s_1) = \left[\left[\mathbf{R}_a^b(s_2, s_1) \right]_\mu^\nu \right]$ is the rotation matrix in between the vector $e_b(s_1)$ associated with the first simplex and the vector $e_a(s_2)$ associated with the second simplex. Then, the rotation associated to a closed path can be written as

$$\mathbf{R}\left(C\right) = \mathbf{R}\left(s_1, s_m\right) \cdots \mathbf{R}\left(s_3, s_2\right) \mathbf{R}\left(s_2, s_1\right) \tag{24}$$

Now, we saw earlier that the variation of a vector along a closed path is given by equation (6). That is, the variation is given by the product of the Riemann tensor, the bivector associated with the closed path and the vector itself. In terms of rotation matrices, the same variation may be written as

$$\begin{aligned} \delta T^{\mu} &= \tfrac{1}{2} \sum_{\nu\lambda\sigma} R^{\mu}_{\nu\lambda\sigma} \; A^{\lambda\sigma}\left(C_{\varepsilon}\right) T^{\nu} = \sum_{\nu} \left[\mathbf{R}\left(C_{\varepsilon}\right) - \mathbf{1}\right]^{\mu}_{\nu} \; T^{\nu} \\ &\Rightarrow \tfrac{1}{2} \sum_{\lambda\sigma} R^{\mu}_{\nu\lambda\sigma} \; A^{\lambda\sigma}\left(C_{\varepsilon}\right) = \left[\mathbf{R}\left(C_{\varepsilon}\right) - \mathbf{1}\right]^{\mu}_{\nu} \end{aligned} \tag{25}$$

Equation (25) establishes a correspondence in between the rotation matrix along a closed path and the Riemann tensor. Let us now concentrate on the rotation around a hinge. We saw earlier, in Section 3, that the amplitude of such a rotation is given by the deficit angle. The generator of this rotation is given by the normalized bivector associated with the hinge, which means that the rotation matrix may be written as

$$\left[\mathbf{R}\right]^{\mu}_{\nu}\left(C_h\right) = \exp\left(\psi\left(h\right) \hat{\Upsilon}^{\mu}_{\nu}\left(h\right)\right) \tag{26}$$

From equations (18), (25) and (26) the following correspondence may be established in between the Riemann tensor and the deficit angle, the area of the hinge and its normalized bivector

$$R_{\mu\nu\lambda\sigma}\left(h\right) = \frac{\psi\left(h\right)}{A_C\left(h\right)} \hat{\Upsilon}_{\mu\nu}\left(h\right) \hat{\Upsilon}_{\lambda\sigma}\left(h\right) \tag{27}$$

Contracting the Riemann tensor, one obtains the Ricci scalar

$$R\left(h\right) = 2\frac{\psi\left(h\right)}{A_C\left(h\right)} \tag{28}$$

As a result, we have managed to express the metric, the Riemann tensor and the Ricci scalar in terms of the geometrical quantities associated with the representation of the data distribution in terms of simplices. As we saw earlier, even in two dimensions, bounded distributions are in general curved. This implies that it is not mathematically acceptable to express the distance in terms of Euclidian distance, except in the regions in which the space is flat, e.g. inside the boundary for a bidimensional distribution. In Section 5, we will thus address the issue of calculating the distance and similarity in a curved space, as applicable for example for similarity search algorithms.

4.1 Discussion

We conclude this section with some remarks on constraints, equivalence and the curse of dimensionality, all from a geometrical perspective. Constrained clustering refers to clustering data analysis methods that use background knowledge encoded as constraints [5]. That is, in constrained clustering, information that the data instances should obey (must-link or cannot-link) is maintained to guide the cluster analysis algorithm. Clustering is importance, for instance, if one wants to group similar objects in various classes. In our work, the construction in terms of simplices must obey certain inner geometrical constraints. Since the simplices are built from triangles, it means that each triangle much obeys the triangle inequality

$$\begin{cases} l_{i,j} + l_{j,k} > l_{i,k} \\ l_{j,k} + l_{k,i} > l_{j,i} \\ l_{k,i} + l_{i,j} > l_{k,j} \end{cases} \tag{29}$$

and that simplicial spaces have build-in constraints. Despite its simplistic appearance, this constraint has far-reaching consequences. For example, each time a calculation involving a variation (e.g. a derivative) is performed this constraint must be enforced at each stage of the calculation. This implies that we are not free, in order to perform a given calculation, to move the points (or data instances) as we please. If we do so, the final result might be inconsistent in the sense that it does not satisfy equation (29) which is a necessary condition. Consequently, the end results of our calculation may be incorrect, since equation (29) is not satisfied.

Having this constraint in mind, we would like to address the question of equivalence. For a given data distribution it is possible to generate two distributions that are equivalent from a mathematical point of view, but for which the vertices are not the same. For example, this is relevant when determining whether two groups of descriptors (or index distributions) are similar. If the two distributions generate the same local volumes and the same curvatures e.g. Riemann and Ricci, then the two simplicial spaces associated with the two distributions are equivalent. To understand better what this implies, let us consider a given vertex and its neighbourhood. Geometrically, equivalent configurations are obtained by moving the vertex within the perimeter defined by the dual neighbourhood (the barycentres of the surrounding simplices). A distinct configuration is generated if the vertex is displaced out of this perimeter. The representation of a distribution in terms of simplices has the advantage of giving a clear geometrical interpretation of equivalent data distributions. This may be demonstrated with equation (13) by applying infitesimal variations to the distances.

Let us conclude this section by a short digression about the curse of dimensionality, which refers to the scalability problems incurred when the number of dimensions is large [6]. This problem is highly relevant, since a high dimensional information retrieval descriptor (or index) with too many dimensions may lose, or reduce, its descriptive power. Let us take into account the geometrical consequences associated with distributions of high dimensionality. Consider a uniform distribution for which the distance in between nearest neighbour is given by $l_{ij}^{(0)2}$. Let us apply a small perturbation to the distance and look what happens to the volume of a given simplex when the number of dimensions is very high. The distance is now given by

$$l_{ij}^2 = l_{ij}^{(0)2} + \varepsilon_{ij} \tag{30}$$

and the volume associated to the simplex is given by equation (13) which in limit of large n becomes

$$V_n(s) \underset{n\to\infty}{\sim} \frac{\sqrt{n}}{n!2^{n/2}} \left[1 - \frac{1}{2}\varepsilon_{12}^2 + \ldots + \frac{1}{n}\left(\varepsilon_{12} + \cdots + \varepsilon_{12}\varepsilon_{13} + \cdots\right) + O\left(n^{-2}\right)\right] \tag{31}$$

Equation (31) shows that the volume of the simplex is at most quadratic in the perturbation. Concretely, it means that when the number of dimensions is high, the sensitivity of the details is lost; the later being associated with terms of cubic and higher order. This may lead to a low precision and recall, when e.g. creating a descriptor to retrieve similar images from a high dimensional image repository.

5. CALCULATION OF THE DISTANCE IN BETWEEN TWO POINTS

The concept of distance is quite different in flat and curved spaces. Recall that, in a flat space, the distance between two points corresponds to the length of the straight line in between these points, or another construction compatible with the axioms of Euclidian geometry. Naturally, the situation differs in a curved space; one must again find the shortest distance in between the points. However, in this case, the displacements are constrained by the underlying geometry; this is called the geodesic distance. For curved Riemannian geometry [3], the geodesic distance is given in the continuous case by

$$d\left(x,y;\mathbf{g}\right)=\min_{\gamma}\int_{\tau(x)}^{\tau(y)}d\tau\,\sqrt{\sum_{\mu\nu}g_{\mu\nu}\left(\gamma\right)\,\frac{d\gamma^{\mu}}{d\tau}\frac{d\gamma^{\nu}}{d\tau}} \tag{32}$$

The geodesic distance in between two points in the discrete case is relatively easy to determine: given all connected paths in between two points, the geodesic distance corresponds to the path of shortest length. Of course, if the space is flat, the Euclidian distance and the geodesic distance are almost the same; the difference coming from the fact that the path is a perfect straight line in the Euclidian case, while it is a slightly broken line (as a consequence of following the edges) in the geodesic case. Various heuristic methods may be used to calculate the geodesic distance. For instance, the shooting and path rectifying methods have been proposed, amongst others [7].

It is also possible to estimate the geodesic distance with the Laplacian [8] associated with the simplicial construction if the distance is large compare to a certain scale μ

$$d\left(i,j\right)\underset{d\left(i,j\right)\gg\mu}{\sim}-\tfrac{1}{\mu}\ln G_{ij}\left(l^{2}\right) \tag{33}$$

The quantity $\left[G_{ij}\left(l^{2}\right)\right]$ is the inverse of the Laplacian and it is given by

$$G_{ij}\left(l^{2}\right)=\left[\frac{1}{-\boldsymbol{\Delta}\left(l^{2}\right)+\mu^{2}}\right]_{ij}=\tfrac{1}{\mu^{2}}\left[\sum_{m=0}^{\infty}\left(\tfrac{1}{\mu^{2}}\,\boldsymbol{\Delta}\left(l^{2}\right)\right)^{m}\right]_{ij} \tag{34}$$

where the parameter μ is required in order to regularise the expression; otherwise the later would be divergent because the Laplacian has a zero Eigen value. The fact that this parameter is to some extend arbitrary might explain, in part, why it is so difficult to define a similarity measure in a totally unambiguous way. This parameter is constrained by the minimum distance in between two vertices and the correlation length associated with the data distribution. Here, the Laplacian is inverted with the Ornstein-Zernike formula. In terms of simplices, the Laplacian may be defined in various ways like the weighted tangent and cotangent representations [8]. In its simplest form, the Laplacian may be defined as

$$\boldsymbol{\Delta}=\mathbf{1}-\mathbf{D}^{-1}\mathbf{A}\qquad\left[\mathbf{A}\right]_{ij}=\begin{cases}1 & j\in\Xi\left(i\right)\\ 0 & j\notin\Xi\left(i\right)\end{cases}\qquad\mathbf{D}=\operatorname{diag}\left[\aleph\left(\Xi(i)\right)\right] \tag{35}$$

where $\mathbf{A}$ is the adjacency (connectivity) matrix, $\Xi\left(i\right)$ is the neighbourhood of vertex *i* (the closest vertices associated with it) and $\aleph$ is the cardinal of the neighbourhood.

6. CONCLUSIONS

In this paper, we have demonstrated that, for a given distribution, one may construct an associated simplicial Riemannian space. This curved space is characterised by various curvatures including the Riemannian tensor, the Ricci tensor and the Ricci scalar [3]. We saw that even bounded bidimensional spaces are curved on their boundary. When more dimensions are added, the curvature becomes an integral part of the inner geometry. Such a state of affairs has many repercussions, of which the calculation of distance is perhaps the most important. Many retrieval, comparison and clustering algorithms rely on the concept of distance. Furthermore, most of the time, for numerical data, an implicit assumption of flatness is made which, as we have shown, is not necessarily valid. One of the main consequences is that distances must be computed in terms of geodesic distance and that all calculation must be performed keeping in mind that the space is curved; for instance scalar products must be calculated with the local metric. The effect of curvature is even more dramatic for non-convex distributions.

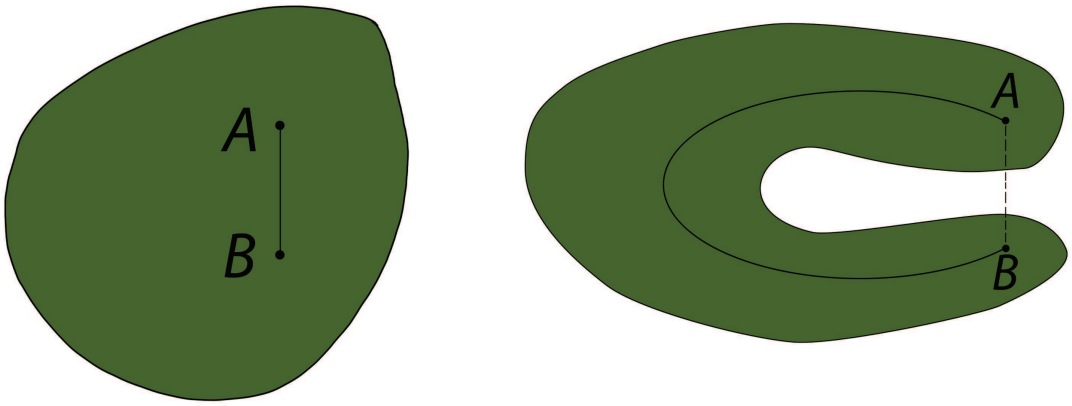

Figure 3. Convex and non-convex distributions and their impact on the determination of distance.

Consider for example Figure 3, which shows two bidimensional distributions; one which is non-convex ("U" shaped). For both distributions, the only curvature is the one associated with the boundaries. As shown by the left illustration, the Euclidian distance is a very good approximation for almost any pair of points. The situation is quite different in the non-convex case (right illustration). Consider e.g. points A and B. It is not possible to calculate the Euclidian distance (dotted line) while remaining inside the distribution. However, the distance may be defined in an unambiguous way with the geodesic distance (full line). This example illustrates that two points that, in appearance, seem to be relatively close are in fact distant. The situation becomes even more complex in higher dimensional spaces. Consequently, it is to be expected that the curvature effect is even more pronounced in non convex spaces. Future work will involve an evaluation of the overall effect of curvature on various information retrieval descriptor algorithms and cluster analysis approaches.

REFERENCES

[1] Luxburg, U., "Tutorial on Spectral Clustering", Statistics and Computing 17 (4), 395-416 (2007).
[2] Jain, A. K., Murty, M. N., and Flynn, P. J., "Data Clustering: A Review", ACM Computing Surveys 31 (3), 264-323 (1999).
[3] do Carmo, M. P., Riemannian Geometry, Birkhäuser, Boston, (1992).
[4] Hamber, H. W., Williams, R. M., "Simplicial Quantum Gravity with Higher Derivative Terms: Formalism and Numerical Results in Four Dimensions", Nuclear Physics B269, 712-743 (1986).
[5] Basu, S., Davidson, I., Wagstaff, K. L. (eds.), Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapman & Hall / CRC, New York, (2009).
[6] Agrawal, R. et al., "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", ACM SIGMOD Record 27 (2), 94-105 (1998).
[7] Samir C., Srivastava A., Daoudi M., Klassen E., "An Intrinsic Framework for Analysis of Facial Surfaces", International Journal on Computer Vision 82 (1), 80-95 (2009).
[8] Sorkine, O., "Differential Representations for Mesh Processing", Computer Graphics Forum 25 (4), 789-807 (2006).