A New Method of Selecting Safe Neighbors for the Riemannian Manifold Learning Algorithm

Lucas P. Carlini, Gastão F. Miranda Junior, Gilson A. Giraldi and Carlos E. Thomaz

***Abstract*—Manifold learning (ML) comprehends a set of non-linear techniques for mining and representing high-dimensional data. In this work, we approach the well-known and successful ML technique called Riemannian Manifold Learning (RML). Firstly, we present a geometric interpretation of the main steps of selecting visible and safe neighborhoods to reconstruct geometry and topology in the original RML algorithm. Then, we describe and implement a new method of selecting safe neighbors for this algorithm. Our experimental results on synthetic and real data sets, using open source tools and a public face image database, have showed that the new method proposed shows similar results to the original one and reconstructions that favour local rather than holistic similarities described by the data. Additionally, since the new method proposed requires the specification of only one input parameter, its implementation is simpler and more intuitive than the original one.**

***Index Terms*—RML, manifold learning, safe neighbors.**

I. Introdução

Aprendizado de Variedades (em inglês, *Manifold Learning*) é um conjunto de técnicas não-lineares utilizado para mineração e análise de dados multidimensionais por meio de elementos de geometria [1], [2], [3], [4], [5]. Estas técnicas são aplicadas em áreas como visão computacional, reconhecimento de padrões, modelagem baseada em dados e análise de imagens [6], [1], [7], [8], [9], [10].

Nos últimos anos, diversas técnicas de aprendizado de variedades têm sido propostas com finalidades variadas, cada uma com a preocupação de preservação de alguma propriedade geométrica intrínseca dos dados. Propriedades como: preservar distâncias, em *Locally MultiDimensional Scaling* (LMDS) [11], Isomap [6] e *Riemannian Manifold Learning* (RML) [1]; preservar ângulos, em *Locally Linear Embedding* (LLE) [12] e *Local Tangent Space Alignment* (LTSA) [13]; ou preservar topologia (estrutura de vizinhos), em *t-Stochastic Neighbor Embedding* (t-SNE) [14], LargeVis [15] e *Uniform Manifold Approximation and Projection* (UMAP) [16]. Dentre os principais problemas abordados por técnicas de aprendizado de variedades constam: seleção de vizinhos, para construção de uma estrutura de vizinhança; redução de dimensionalidade, que visa encontrar uma representação dos dados utilizando uma quantidade de coordenadas (ou características) menor que a original; e síntese, que trata do processo de reconstrução e interpolação de dados que estão no espaço de dimensão reduzida para o espaço original dos dados.

L. P. Carlini, FEI, São Paulo, Brazil, lucaspcarlini10@gmail.com
G. F. Miranda Junior, UFS, Sergipe, Brazil, gastao@mat.ufs.br
G. A. Giraldi, LNCC, Rio de Janeiro, Brazil, gilson@lncc.br
C. E. Thomaz, FEI, São Paulo, Brazil, cet@fei.edu.br
Corresponding author: Carlos E. Thomaz.

Em particular, o algoritmo de aprendizado de variedades RML visa preservar a geometria intrínseca dos dados usando conceitos clássicos da Geometria Riemanniana [17], [18], [5]. Uma propriedade relevante do RML é o fato desta técnica implementar um método adaptativo de construção da variedade, determinando o tamanho da vizinhança local para cada dado (ou ponto), assim como a dimensão intrínseca local. Entretanto, o RML necessita da definição prévia de parâmetros de entrada pouco intuitivos, que são fortemente impactados pela densidade de amostragem da variedade. Na prática, a densidade de amostragem de uma variedade é desconhecida e precisa ser determinada experimentalmente, podendo, consequentemente, variar consideravelmente para conjuntos de dados de naturezas distintas.

Neste contexto, este trabalho revisita o RML com o objetivo de propor um novo método de seleção de vizinhos seguros para esse algoritmo de aprendizado de variedades. Inicialmente, como primeira contribuição da metodologia empregada aqui, descreve-se uma interpretação geométrica das etapas de seleção das vizinhanças visível e segura do algoritmo RML original. Depois, como segunda e principal contribuição, propõe-se e implementa-se, para o RML, um novo método de seleção de vizinhos seguros para navegação na variedade que contém os dados. O método proposto de seleção de vizinhos seguros está baseado na análise massivamente univariada [19] das diferenças locais estatisticamente significantes entre os dados. Mais especificamente, para construção da estrutura de vizinhança segura, o método proposto preserva a geometria intrínseca dos dados usando conceitos clássicos da Estatística, simplificando a parametrização do RML e tornando-a intuitiva. Visando avaliar comparativamente a seleção de vizinhos seguros do método original e do método proposto, a metodologia empregada contempla também a implementação de um algoritmo de navegação ponto-a-ponto na variedade que contém os dados.

Em experimentos realizados com dados sintéticos e imagens frontais de faces, como exemplos de dados reais multidimensionais, o método proposto mostra resultados gerais e principais de seleção de vizinhos seguros igualmente satisfatórios ao algoritmo original do RML. No entanto, para as imagens de faces, as navegações com o método proposto evidenciam priorização das similaridades locais (característica-por-característica), diferentemente do método original que tende a preservar características holísticas (do padrão como um todo) descritas pelos dados. A priorização das similaridades locais permite, como outros resultados principais destacáveis, navegações menos sensíveis ao tipo de variedade e mais suaves, sem interrupções precoces. Adicionalmente, como o novo método de seleção de vizinhos seguros requer a especificação

de um único parâmetro apenas, sua implementação é mais simples do que a do algoritmo original.

II. RML: Seleção de Vizinhos Visíveis e Seguros

Seja o conjunto de N dados (ou pontos) n-dimensionais de entrada $\mathcal{D} = (x_1, x_2, ..., x_N)$, isto é, $\mathcal{D} \subset \Re^n$. O algoritmo RML original de seleção de vizinhos visíveis e seguros é composto basicamente pelos seguintes três passos:

1. Utilizando-se o algoritmo *nearest-neighbour* [20], calcula-se o conjunto $KN(x_i)$ formado pelos k vizinhos mais próximos de x_i, segundo a métrica Euclideana. A escolha do valor de k deve ser suficientemente grande para abranger todos os dados ao redor de x_i, tal que $k \leq (N-1)$;
2. Determina-se o conjunto $VN(x_i)$, que representa a vizinhança visível de x_i composta por v pontos, tal que $v \leq k$, por meio da expressão

$$VN(x_i) = \{y \in KN(x_i) \mid \langle x_i - z, y - z\rangle \geq 0, \forall z \in KN(x_i)\}, \quad (1)$$

   em que $\langle \cdot, \cdot \rangle$ representa o produto interno usual. Em outras palavras, um ponto y é dito um vizinho visível de x_i se não houver nenhum outro ponto z intermediário que separe y de x_i, ou seja, o produto interno entre os vetores $(x_i - z)$ e $(y - z)$ é maior ou igual a zero para todo z pertencente a $KN(x_i)$;
3. Determina-se o conjunto $SN(x_i)$, que representa a vizinhança segura de x_i, com o intuito de preservar a estrutura de vizinhos local dos dados e remover os pontos denominados de curto-circuito [1], [4] do conjunto de v visíveis. Para isso, primeiramente, ordena-se ascendentemente os comprimentos dos vetores $e_j = x_j - x_i$, tal que $x_j \in VN(x_i)$. Cria-se então a lista ordenada $\{e_1, e_2, ..., e_v\}$, tal que $|e_1| \leq |e_2| \leq \ldots \leq |e_v|$. Estima-se a dimensionalidade intrínseca local em x_i computando-se o número d_j de componentes principais [21], cujos autovalores são não-nulos e maiores do que um determinado parâmetro T, dos primeiros j elementos de $\{e_1, e_2, ..., e_v\}$, tal que $1 \leq j \leq v$. Se $d_j > d_{j-1}$, então compara-se o valor do salto $|e_j| - |e_{j-1}|$ com outro determinado parâmetro φ. Se esse salto for maior que φ, então o dado j e os seguintes são considerados como curto-circuito e, portanto, removidos. Caso contrário, todos os pontos são considerados como pertencentes a vizinhança segura de x_i.

Para execução do algoritmo RML original faz-se necessária, conforme descrito, a especificação no passo 1 do paramêtro k (número de vizinhos mais próximos), e dos parâmetros T (valor limiar de autovalores) e φ (valor limiar de salto) no passo 3. Esses dois últimos parâmetros T e φ dependem fortemente da densidade de amostragem dos dados e são pouco intuitivos na prática, pois se referem a limiares específicos para projeção dos mesmos.

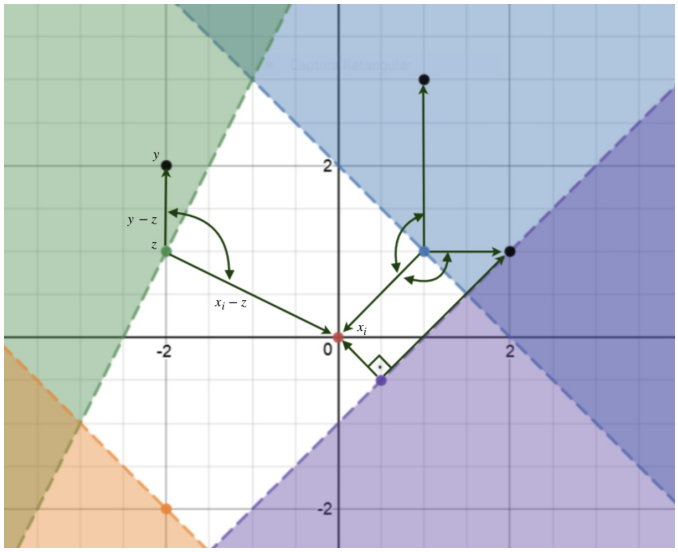


Fig. 1. RML: Interpretação Geométrica (passo 2 original).

A. Interpretação Geométrica: Seleção de Vizinhos Visíveis (passo 2)

A Figura 1 ilustra um exemplo hipotético para interpretação geométrica da obtenção da vizinhança visível de um ponto x_i em $(0, 0)$ (ponto vermelho), denotada por $VN(x_i)$, dados os $k = 7$ vizinhos mais próximos existentes (3 pontos pretos, 1 verde, 1 azul, 1 roxo e 1 laranja).

Começando pelo ponto verde z localizado em $(-2, 1)$, deseja-se saber se o ponto preto y localizado em $(-2, 2)$ é visível em relação ao ponto vermelho x_i em $(0, 0)$. Como os vetores $(x_i - z)$ e $(y - z)$, ilustrados na figura, formam um ângulo obtuso, ou seja, $\langle x_i - z, y - z\rangle < 0$, então o ponto preto y é dito não visível a x_i em razão da existência do ponto verde z intermediário. Na verdade, todo e qualquer ponto que esteja à esquerda nesse caso da reta[1] normal ao vetor $(x_i - z)$ que passa por z, na região verde colorida da figura, não é visível a x_i. O contrário, ou seja, avaliando-se a partir do ponto preto $(-2, 2)$, verifica-se que o ponto verde continuaria visível ao ponto vermelho, pois os vetores $(x_i - y)$ e $(z - y)$ formariam um ângulo agudo. Analogamente, os demais pontos delimitam as suas respectivas zonas de exclusão, evidenciando a área branca como região de vizinhança visível para x_i, tal que os pontos visíveis são somente os 4 pontos coloridos de verde, azul, roxo e laranja na figura.

B. Interpretação Geométrica: Seleção de Vizinhos Seguros (passo 3)

Partindo da Figura 1 anterior, sejam agora o ponto x_i em $(0, 0)$ e seus 4 vizinhos visíveis já determinados, tal que $VN(x_i) = \{(0.5, -0.5), (1, 1), (-2, 1), (-2, -2)\}$, descritos na Figura 2.

Para verificar se esses 4 vizinhos visíveis são seguros, deve-se, conforme passo 3 do algoritmo original RML descrito anteriormente, inicialmente calcular e ordenar ascendentemente

[1]Reta no caso planar exemplificado, hiperplano para o caso geral.

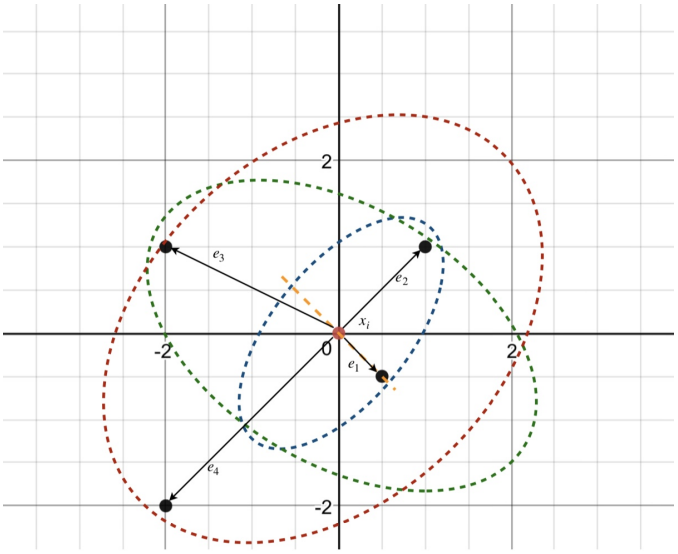


Fig. 2. RML: Interpretação Geométrica (passo 3 original).

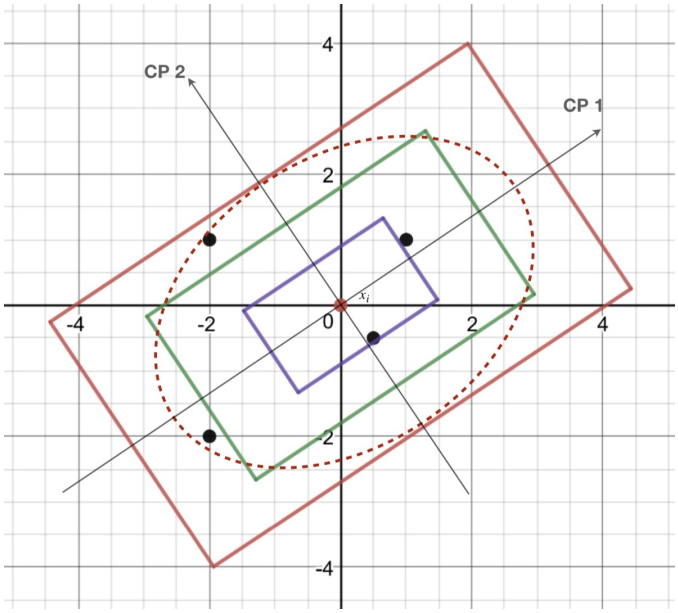


Fig. 3. RML: Interpretação Geométrica (passo 3 novo).

os comprimentos dos vetores $e_j = x_j - x_i$, como ilustrado na Figura 2, tal que $1 \leq j \leq 4$ neste caso.

A seguir, estima-se a dimensionalidade intrínseca d_1 em x_i, calculando o número de componentes principais para a amostra de dados centralizada[2] em x_i e composta apenas por x_i e e_1. Neste caso, a direção de maior variância é dada pela dimensão que liga esses dois pontos, conforme reta pontilhada em cor laranja na Figura 2. Essa única dimensão tem autovalor não-nulo e, assumindo que esse seja maior do que T (valor limiar de autovalores), tem-se então $d_1 = 1$. Como há diferença entre os números de componentes principais, ou seja, $d_1 > d_0$ pois, por definição, $d_0 = 0$, calcula-se o salto $|e_1| - |e_0|$. Pontos considerados como vizinhos seguros devem atender a condição de que tal salto seja menor ou igual a φ (valor limiar de salto), ou seja, $|e_1| - |e_0| \leq \varphi$, que equivale a $|e_1| \leq \varphi$, pois $|e_0| = 0$, por definição. Assumindo que a condição fora satisfeita, então o ponto $(0.5, -0.5)$ referente a e_1 é considerado vizinho seguro de x_i.

Posteriormente, calcula-se então o número d_2 de componentes principais para a amostra de dados centralizada em x_i e composta agora por x_i, e_1 e e_2. Têm-se como componentes principais os eixos da elipse ilustrada em pontilhado azul na mesma Figura 2. Essas duas componentes principais têm autovalores não-nulos e maiores do que T, então tem-se $d_2 = 2$ e testa-se a condição de salto $|e_2| - |e_1| \leq \varphi$, pois $d_2 > d_1$. Se essa condição for satisfeita então o ponto $(1, 1)$ referente a e_2 é considerado também vizinho seguro de x_i. Analogamente, os pontos restantes $(-2, 1)$ e $(-2, -2)$ também são considerados vizinhos seguros, pois não haverá mais condições de salto para teste devido ao fato das dimensões intrínsecas de d_3 e d_4 serem também iguais a 2, descritas pelos eixos das elipses ilustradas, respectivamente, em pontilhado verde e vermelho com autovalores não-nulos e maiores do que T.

Portanto, se a condição de salto $|e_2| - |e_1| \leq \varphi$ for satisfeita, todos os vizinhos visíveis ilustrados na Figura 2 são considerados também seguros, pois estão sobre o mesmo plano geométrico. Caso contrário, o ponto $(0.5, -0.5)$ é o único vizinho seguro de x_i, assumindo-se que $|e_1| \leq \varphi$.

Experimentalmente, tem-se considerado [1], [4] o valor limiar de autovalores igual a aproximadamente zero ($T \approx 0$) e o valor limiar de salto diretamente proporcional ao tamanho médio dos comprimentos dos vetores e_j, ou seja,

$$\varphi \propto \frac{1}{v}\sum_{j=1}^{v}|e_j| = \rho \times (\frac{1}{v}\sum_{j=1}^{v}|e_j|), \tag{2}$$

tal que ρ é uma constante multiplicativa e v o número de vizinhos visíveis. No caso da Figura 2, $v = 4$.

III. Novo Método de Seleção de Vizinhos Seguros

O método proposto de seleção de vizinhos seguros visa preservar a geometria intrínseca dos dados usando conceitos clássicos da Estatística. Para isso, mantem-se a seleção da vizinhança local descrita pelo algoritmo RML original, definida pelos passos 1 e 2 já descritos na seção anterior, e altera-se exclusivamente o passo 3 deste algoritmo redefinindo o conjunto $SN(x_i)$ que representa a vizinhança segura de x_i.

A ideia principal deste novo método é a construção de uma vizinhança ao redor de x_i que seja considerada estatisticamente similar e, portanto, segura, assumindo que a dispersão dos dados pertencentes a $VN(x_i)$ possa ser modelada por uma distribuição normal. A representação geométrica desta vizinhança considerada estatisticamente segura encontra-se na Figura 3, para os mesmos pontos exemplificados nas figuras anteriores, para três níveis de similaridade estatística: roxo, verde e vermelho. Os detalhes sobre os cálculos desses níveis são explicados a seguir.

O novo passo 3 do algoritmo RML modificado pode ser assim descrito:

3*. Determina-se o conjunto $SN(x_i)$, que representa a vizinhança segura de x_i, com o intuito de preservar a

[2]A partir dos vizinhos visíveis de x_i, obtemos as componentes principais tomando x_i como a origem do sistema de coordenadas local.

similaridade estatística dos dados e remover os pontos de curto-circuito do conjunto de v visíveis, pois são mais distantes do que os demais pertencentes a $VN(x_i)$. Para isso, primeiramente, calcula-se o número total p de componentes principais cujos autovalores são não-nulos para os vetores $e_j = x_j - x_i$, centralizados em x_i, tal que $x_j \in VN(x_i)$ e $p = \min(v, n)$. Projetam-se então todos os v pontos visíveis nas p componentes principais e calcula-se a variância σ_j^2 desses pontos projetados em cada j-ésima componente principal. O novo conjunto $SN^\star(x_i)$ é dado por

$$SN^\star(x_i) = \{y \in VN(x_i) \mid (x_{ij} - c \cdot \sigma_j) \leq y_j \leq (x_{ij} + c \cdot \sigma_j)\} \quad (3)$$

$\forall j$, tal que $j = 1, 2, \ldots, p$, onde x_{ij} e y_j são, respectivamente, a j-ésima coordenada dos pontos x_i e y. O parâmetro c é uma constante multiplicativa para os desvios padrões σ_j.

A execução do novo passo 3 do algoritmo RML modificado requer apenas a especificação do parâmetro c. Esse parâmetro é intuitivo, pois permite avaliar, por exemplo, mais de um desvio padrão para definir a vizinhança estatisticamente segura, como ilustrado na Figura 3, onde as vizinhanças delimitadas pelos retângulos roxo, verde e vermelho correspondem ao uso na equação (3) do valor de c igual a 1, 2 e 3, respectivamente. Portanto, para um desvio (ou seja, considerando $c = 1$), apenas o ponto $(0.5, -0.5)$ é vizinho seguro de x_i. Se forem considerados dois desvios ($c = 2$), o ponto $(1, 1)$ também é considerado seguro. Para três desvios ($c = 3$), todos os quatro pontos visíveis são também considerados seguros estatisticamente, como mostra o retângulo vermelho da figura.

IV. Algoritmo de Navegação Ponto-a-Ponto na Variedade

Visando avaliar comparativamente a seleção de vizinhos seguros do método original e do método proposto, implementou-se o Algoritmo 1 para navegação ponto-a-ponto na variedade que contém esses dados. Define-se aqui vizinho seguro navegável como todo e qualquer ponto calculado como vizinho seguro que ainda não tenha sido visitado pelo algoritmo.

O processo de navegação inicia-se, basicamente, com a escolha de um ponto arbitrário inicial x_i da base de dados. Esse ponto x_i é adicionado à lista de pontos visitados, denominada L no Algoritmo 1. Calculam-se os vizinhos mais próximos $KN(x_i)$, visíveis $VN(x_i)$ e seguros $SN(x_i)$. A quantidade de vizinhos seguros de x_i, isto é, a cardinalidade de $SN(x_i)$, é adicionada à lista Q_1, assim como a quantidade de vizinhos seguros navegáveis, isto é, a cardinalidade de $SN^*(x_i)$, é adicionada à lista Q_2. Esses passos se repetem para cada ponto considerado como o vizinho seguro mais próximo[3] ao ponto da vez. A navegação se encerra quando o conjunto $SN(x_i)$ de vizinhos seguros for vazio, ou seja, quando não houver mais vizinhos seguros para navegação, ou quando o conjunto $SN^*(x_i)$ de vizinhos seguros navegáveis for vazio, ou seja, todos os pontos pertencentes a $SN(x_i)$ também pertencerem a lista de navegação L, evitando que o algoritmo entre em *loop* e revisite pontos já navegados.

[3]Se houver pontos equidistantes, escolhe-se aleatoriamente qualquer um desses pontos que ainda não tenha sido visitado pelo algoritmo de navegação.

Algorithm 1: Retorna L, Q_1 e Q_2.

```
Defina k, número de vizinhos mais próximos
Defina x_i, ponto arbitrário inicial
Inicialize L, lista de pontos visitados
Inicialize Q_1, lista de quantidades de vizinhos seguros
Inicialize Q_2, lista de quantidades de vizinhos seguros navegáveis
flag = 0
do
    Adicione x_i em L
    Calcule KN(x_i), VN(x_i) e SN(x_i)
    Calcule a cardinalidade de SN(x_i) e adicione em Q_1
    if SN(x_i) = {} then
        Mensagem ('Conjunto SN(x_i) vazio.')
        flag = 1
    else
        SN*(x_i) = SN(x_i) − L
        if SN*(x_i) = {} then
            Mensagem ('Todo conjunto SN(x_i) já visitado.')
            flag = 1
        else
            Calcule a cardinalidade de SN*(x_i) e adicione em Q_2
            Calcule x_{i+1}, vizinho mais próximo a x_i, tal que x_{i+1} ∈ SN*(x_i)
            x_i = x_{i+1}
while flag = 0
```

V. Experimentos e Resultados

Os experimentos foram conduzidos com o objetivo de comparar os resultados do método original e do método proposto de vizinhos seguros referentes ao processo de navegação em variedades contendo dados sintéticos e reais.

Para os experimentos com dados sintéticos, foram utilizadas curvas amplamente investigadas na literatura afim de aprendizado de variedades [1], que representam dispersões de pontos no plano Cartesiano em formatos S e espiral, e geradas com auxílio da ferramenta *scikit-learn* [22] de código aberto.

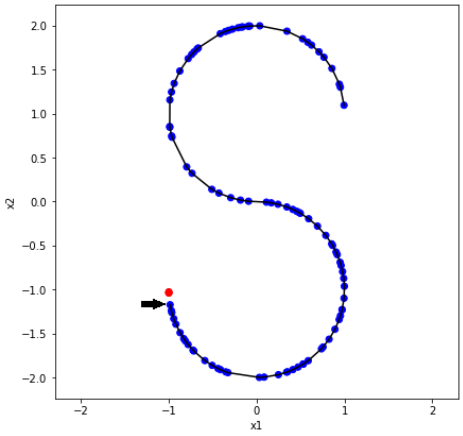
Para os experimentos com dados reais, foram utilizadas imagens frontais em tons de cinza do banco de faces da FEI [23]. Ao todo, foram 400 imagens de faces de 200 pessoas distintas (100 homens e 100 mulheres) capturadas com expressão facial neutra e sorrindo. Todas as imagens têm resolução de 300×250 pixels e, portanto, foram representadas aqui como vetores de dimensão $n = 75000$.

A. Resultados com Dados Sintéticos

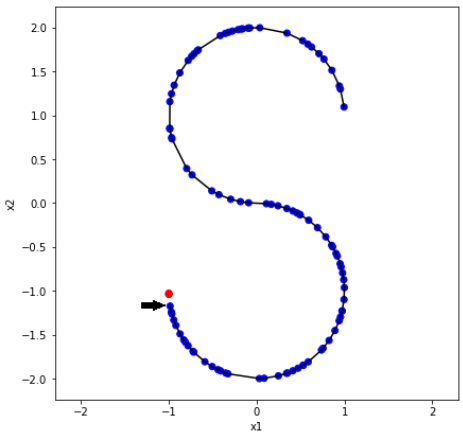
As Figuras 4 e 5 mostram navegações realizadas pelo método original e pelo método proposto na variedade descrita

por uma curva em formato S (100 amostras), usando $k = 30$ para os vizinhos mais próximos. Os pontos coloridos em vermelho representam a geometria original enquanto que os pontos em azul descrevem os caminhos gerados pelo processo de navegação. Em ambas figuras, o ponto arbitrário inicial escolhido foi $(-1.0, -1.17)$, destacado pela seta preta.

Na Figura 4, utilizando os parâmetros iniciais sugeridos na literatura para o método original, ou seja, $T = 0.08$ e $\rho = 1.2$ [1], [4], e $c = 3$ para o método proposto, ambos métodos apresentaram o mesmo resultado de navegação na variedade, reconstruindo toda a geometria da curva. No entanto, na Figura 5, com o parâmetro ρ reduzido para $\rho = 0.9$, pode-se verificar que a navegação pelo método original foi interrompida precocemente. Nesta situação, com o parâmetro c reduzido para $c = 1$, o método proposto descreveu melhor a estrutura de dados em questão, reconstruindo parcialmente a variedade.



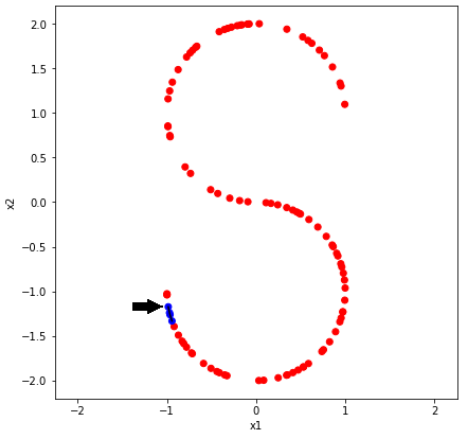
(a) Método Original



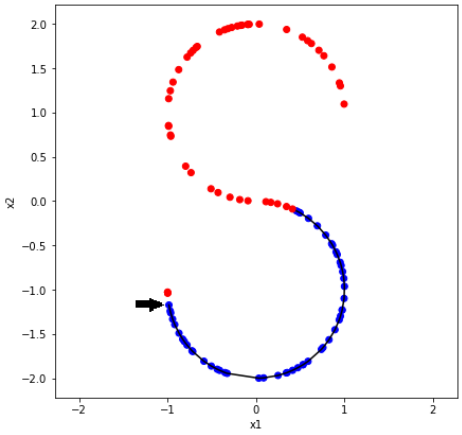
(b) Método Proposto

Fig. 4. Navegação em variedade de formato S (100 amostras): (a) Método Original ($T = 0.08$ e $\rho = 1.2$) e (b) Método Proposto ($c = 3$).

Há situações em que os parâmetros iniciais sugeridos na literatura podem ser inadequados. Por exemplo, as Figuras 6 e 7 mostram navegações na variedade descrita por uma curva em formato espiral (100 amostras), usando $k = 30$ para encontrar os vizinhos mais próximos. Em ambas figuras, o ponto arbitrário inicial escolhido foi $(-9.0, -0.56)$, destacado



(a) Método Original



(b) Método Proposto

Fig. 5. Navegação em variedade de formato S (100 amostras): (a) Método Original ($T = 0.08$ e $\rho = 0.9$) e (b) Método Proposto ($c = 1$).

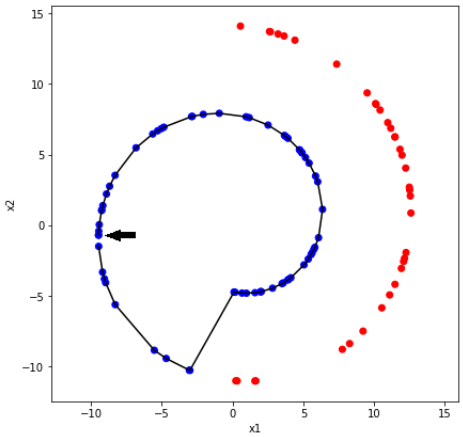
pela seta preta.

Na Figura 6, utilizando novamente os parâmetros iniciais sugeridos para os métodos original e proposto, ambos apresentaram o mesmo resultado de navegação, com ocorrência inapropriada de curto-circuito. Reduzindo os respectivos parâmetros para $\rho = 0.9$ e $c = 1$, a navegação de ambos métodos foi idêntica e sem curto-circuito, como mostrado na Figura 7. Faz-se importante destacar, no entanto, que quando utilizado $\rho = 0.95$ ao invés de $\rho = 0.9$ para o método original, também houve ocorrência de curto-circuito. Este resultado sugere que o método original apresenta maior sensibilidade ao tipo de variedade e, portanto, para um mesmo ρ e dependendo da forma da variedade pode-se ter navegações interrompidas precocemente, como mostrado na Figura 5a, ou parcial, como ilustra a Figura 7a.

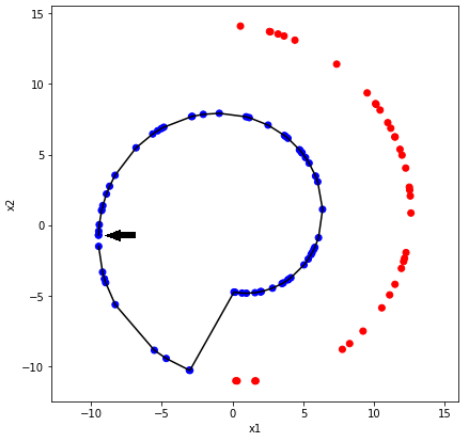
B. Resultados com Dados Reais

Em todos os resultados com dados reais, foram usados como parâmetros $k = 30$ para determinar os vizinhos mais próximos, $T = 0.08$ e $\rho = 1.2$ para o método original, como sugerido em [1], [4], e $c = 3$ para o método proposto.

Os processos de navegação iniciaram-se sempre a partir de uma mesma imagem de referência escolhida aleatoriamente,

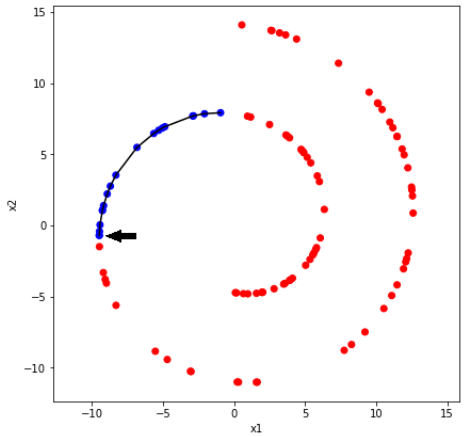


(a) Método Original

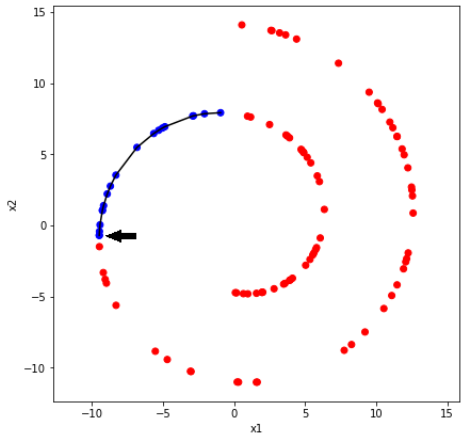


(b) Método Proposto

Fig. 6. Navegação em variedade de formato espiral (100 amostras), com presença do problema de curto-circuito: (a) Método Original ($T = 0.08$ e $\rho = 1.2$) e (b) Método Proposto ($c = 3$).



(a) Método Original



(b) Método Proposto

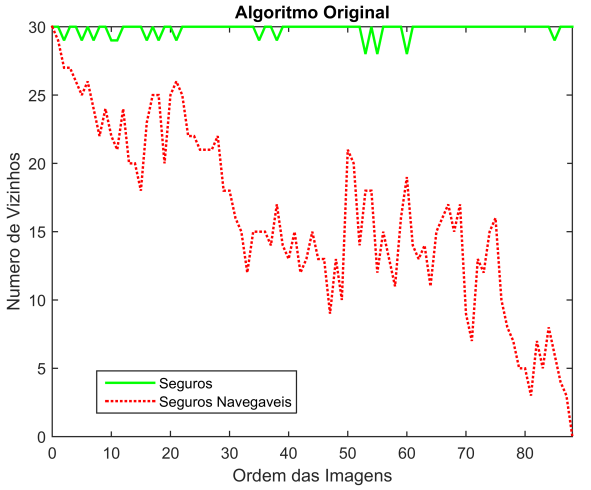
Fig. 7. Navegação em variedade de formato espiral (100 amostras): (a) Método Original ($T = 0.08$ e $\rho = 0.9$) e (b) Método Proposto ($c = 1$).

que descreve um rosto do gênero masculino com expressão neutra. No geral, o método original apresenta uma navegação contendo 88 imagens de faces e o método proposto 161 imagens. Ambas navegações são parciais, pois potencialmente poderiam conter 400 imagens de faces, mas, de novo, agora para dados reais, o método original parece ter interrompido a sua navegação mais precocemente que o método proposto.

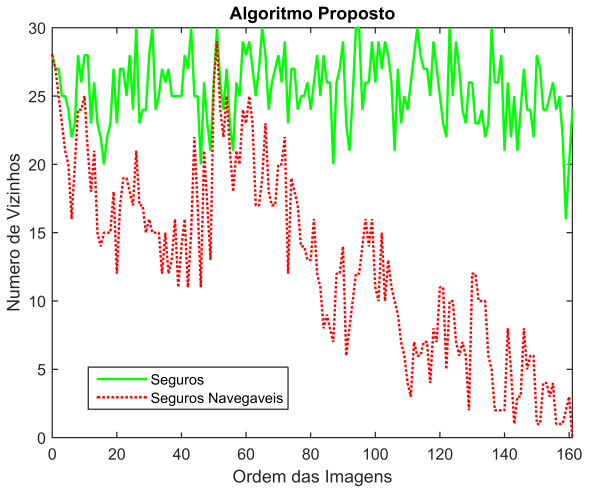
A Figura 8 apresenta as quantidades de vizinhos seguros (Q_1) e seguros navegáveis (Q_2) para os métodos original e proposto. Pode-se verificar que o método original tende a obter a quantidade máxima de vizinhos seguros para cada ponto da vez, ou seja, igual a quantidade $k = 30$ de vizinhos mais próximos. Com o decorrer da navegação, o método original apresenta diminuição na quantidade desses vizinhos seguros que são navegáveis. Já o método proposto apresenta-se mais restritivo na obtenção dos vizinhos seguros, oscilando entre quantidades que variam de 16 a 30 pontos, mas, de maneira similar ao método original, também há decréscimo na quantidade dos vizinhos seguros que são navegáveis. Ambos os métodos se encerram pelo mesmo motivo, ou seja, quando não existe mais nenhum vizinho seguro válido para navegação. No entanto, dadas as quantidades de imagens visitadas iguais a 88 e 161 para os métodos original e proposto, respectivamente, constata-se uma diminuição mais expressiva na quantidade de vizinhos seguros navegáveis por parte do método original que do método proposto.

Em detalhes, na Figura 9, pode-se analisar as características faciais de gênero (Figura 9a) e expressão facial (Figura 9b) preservadas por ambos os métodos. Verifica-se, na Figura 9a, que o método original navega majoritariamente entre imagens de face de homens, preservando a característica de gênero masculino da face inicial de referência. Por outro lado, o método proposto transita entre gêneros, apresentando, após as primeiras 60 imagens, tendência maior em navegar entre imagens de face de mulheres exclusivamente. No entanto, como mostrado na Figura 9b, nenhum dos métodos tende a preservar as características de expressão facial, oscilando entre expressões faciais neutra e sorrindo do início ao fim da navegação.

Em mais detalhes, na Figura 10, são descritas transições visualmente relevantes nas imagens durante a seleção das vizinhanças seguras de ambos os métodos. Pode-se verificar igualdade entre os métodos nas 5 primeiras imagens das navegações correspondentes. No entanto, após essas primeiras imagens descritas nas primeiras linhas das Figuras 10a e 10b,

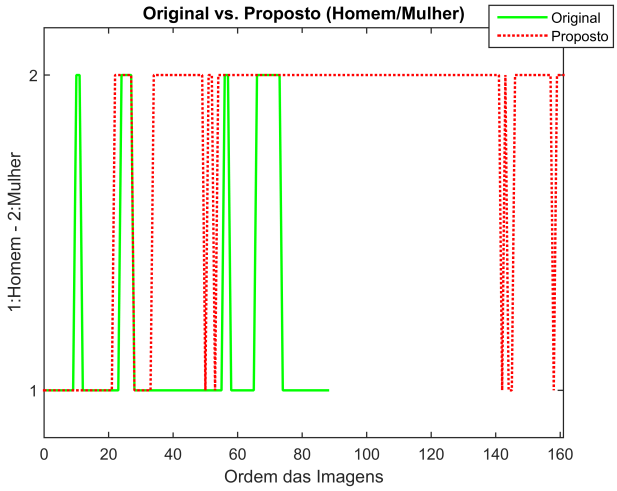


(a) Método Original

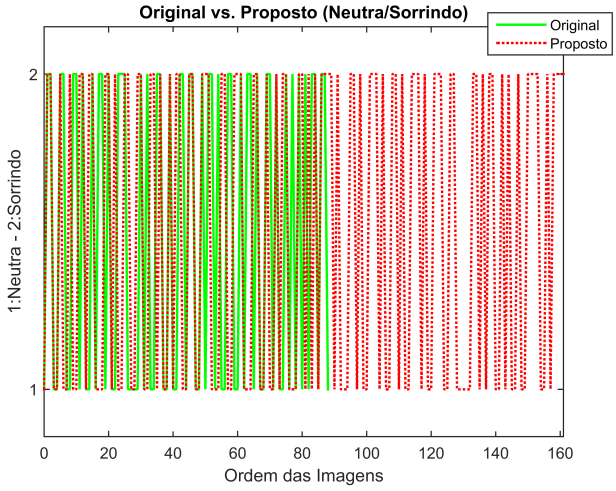


(b) Método Proposto

Fig. 8. Quantidades de vizinhos seguros (Q_1) e seguros navegáveis (Q_2): (a) Método Original ($T = 0.08$ e $\rho = 1.2$) e (b) Método Proposto ($c = 3$).



(a) Gênero



(b) Expressão Facial

Fig. 9. Detalhes da navegação pelos métodos original e proposto para diferenças de (a) gênero e (b) expressão facial.

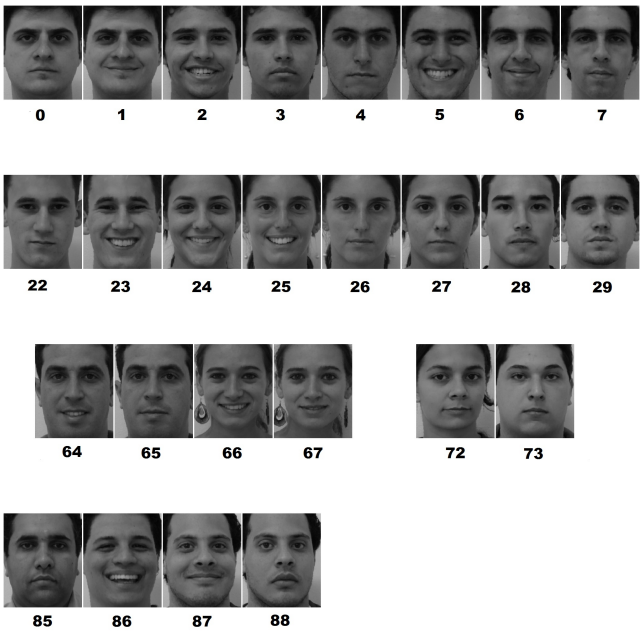
observam-se diferenças claras entre os métodos de seleção. Na segunda linha das Figuras 10a e 10b, é mostrado a primeira transição de seleção de somente faces de homens para, posteriormente, faces de homens e mulheres. Essa seleção intercalada se mantém em menor proporção até o final da navegação para o método proposto (Figura 10b). Já a navegação pelo método original volta a apresentar somente imagens de homens em seu final (Figura 10a). Faz-se importante notar também a transição mostrada na quarta linha da Figura 10b. Nessa sequência de imagens, é possível verificar que a navegação na variedade pelo método proposto transita de imagens de faces com cabelos curtos (terceira linha) para faces largas (quarta linha), e depois para cabelos longos (última linha) até o final da navegação. Esse comportamento não acontece na navegação do método original (Figura 10a), em que a característica referente ao cabelo curto mantém-se em toda navegação.

Resumindo-se, para os resultados com dados reais, o método proposto tende a navegar suavemente entre as imagens preservando as características locais de cada face. Por outro lado, o algoritmo original tende a preservar as características globais das imagens em relação à imagem de referência.

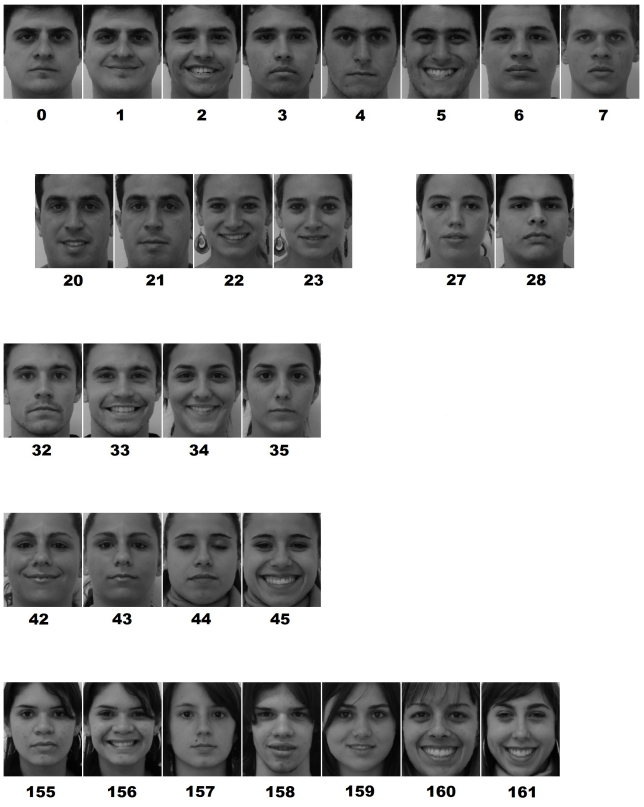
VI. Conclusão

Este artigo propõe um novo método de seleção de vizinhos seguros para o algoritmo RML. Esse novo método visa preservar a geometria intrínseca dos dados usando conceitos clássicos da Estatística. A ideia principal é a construção de uma vizinhança de navegação na variedade de dados multidimensionais que seja considerada estatisticamente similar.

Experimentos foram realizados para comparar os resultados do método proposto com o método original em variedades de dados sintéticos e reais. Como dados sintéticos, foram utilizadas curvas amplamente investigadas na literatura afim de aprendizado de variedades e, como dados reais, imagens frontais de faces de um banco de dados disponível publicamente. No geral, os resultados comparativos mostram que o método original apresenta maior sensibilidade ao tipo de variedade, podendo apresentar navegações que são interrompidas preco-



(a) Método Original



(b) Método Proposto

Fig. 10. Detalhes das imagens durante a navegação na variedade descrita pelas faces frontais (400 amostras): (a) Método Original ($T = 0.08$ e $\rho = 1.2$) e (b) Método Proposto ($c = 3$).

cemente pois procuram preservar as características globais dos dados. Já o método proposto tende a navegar suavemente entre os dados preservando as características locais das vizinhanças.

Acredita-se que o método proposto seja uma alternativa para o método original quando objetiva-se priorizar as similaridades de característica-por-característica e não do padrão como um todo descritas pelos dados. Em outras palavras, acredita-se que o método proposto seja mais eficiente do que o método original em mineração e análise de dados multidimensionais que contenham artefatos ou ruídos, desde que esses artefatos ou ruídos não descaracterizem por completo o objeto de interesse.

Como trabalhos futuros, sugere-se o estudo e possível extensão do método proposto de seleção de vizinhos seguros para outras técnicas de aprendizado de variedades que abordem o problema de construção de uma estrutura de vizinhança, tal como o LMDS [11], por exemplo.

AGRADECIMENTOS

Os autores deste trabalho gostariam de agradecer o apoio da FEI por meio da concessão de bolsa de Iniciação Científica (116/17) para Lucas P. Carlini por um período de 12 meses.

REFERÊNCIAS

[1] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 796–809, 2008.

[2] A. A. Jamshidi, M. J. Kirby, and D. S. Broomhead, "Geometric manifold learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 69–76, 2011.

[3] H. Qiao, P. Zhang, D. Wang, and B. Zhang, "An explicit nonlinear mapping for manifold learning," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 51–63, 2013.

[4] G. F. M. Junior, "Composicao de coordenadas normais de riemann locais e geometria poliedral em aprendizado de variedades com aplicacoes de teoria de folheacoes," Ph.D. dissertation, LNCC, 2015.

[5] G. F. Miranda, C. E. Thomaz, and G. A. Giraldi, "Geometric data analysis based on manifold learning with applications for image understanding," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*. IEEE, 2017, pp. 42–62.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 22, pp. 2319–2322, 2000.

[7] C. Chen, J. Zhang, and R. Fleischer, "Distance approximating dimension reduction of riemannian manifolds," *IEEE Transactions on Systems, Man, and Cybernetics: Part B*, vol. 40, no. 1, pp. 208–217, 2010.

[8] J. Zhang, H. Huang, and J. Wang, "Manifold learning for visualizing and analyzing high-dimensional data," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 54–61, 2010.

[9] Y. Huang, G. Kou, and Y. Peng, "Nonlinear manifold learning for early warnings in financial markets," *European Journal of Operational Research*, vol. 258, no. 2, pp. 692–702, 2017.

[10] S. Kadoury, "Manifold learning in medical imaging," in *Manifolds II - Theory and Applications*. IntechOpen, 2018.

[11] L. Yang, "Alignment of overlapping locally scaled patches for multidimensional scaling and dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 438–450, 2008.

[12] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[13] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, pp. 313–338, 2002.

[14] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[15] J. Tang, J. Liu, M. Zhang, and Q. Mei, "Visualizing large-scale and high-dimensional data," in *Proceedings of the 25th International Conference on World Wide Web (WWW)*. ACM Press, 2016.

[16] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.
[17] A. Brun, C.-F. Westin, M. Herberthson, and H. Knutsson, "Fast manifold learning based on riemannian normal coordinates," in *Scandinavian Conference on Image Analysis*. Springer, 2005, pp. 920–929.
[18] M. P. Carmo, *Geometria Riemanniana*. Projeto Euclides, Instituto de Matematica Pura e Aplicada, 2015.
[19] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall, 1998.
[20] P. A. Devijver and J. Kittler, *Pattern Classification: A Statistical Approach*. Prentice-Hall, 1982.
[21] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1990.
[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[23] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image and Vision Computing*, vol. 28, no. 6, pp. 902 – 913, 2010.

Lucas Carlini is currently working towards his B.Sc. degree in automation and control engineering at FEI University Center, São Paulo, Brazil. Since 2018, he has received an undergraduate research fellowship from FEI to develop research activities on manifold learning and pattern recognition.

Gastão F. Miranda Junior received his B.Sc. degree in mathematics from Federal University of Alagoas, Alagoas, Brazil, M.Sc. in mathematics from State University of Campinas, Sao Paulo, Brazil, and Ph.D. degree in computational modeling from the National Laboratory for Scientific Computing, Petropolis, Brazil, in 2000, 2002 and 2015, respectively. Since 2006 he has worked in the Department of Mathematics at the Federal University of Sergipe, Sergipe, Brazil. His main research interests are in the areas of computer graphics and computer vision.

Gilson Giraldi received his B.Sc. degree in mathematics from Pontifical Catholic University of Campinas, Sao Paulo, Brazil, M.Sc. in applied mathematics from State University of Campinas, Sao Paulo, Brazil, and Ph.D. degree in computer graphics from Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, in 1986, 1993 and 2000, respectively. Since 2000 he has been with the National Laboratory for Scientific Computing, Petropolis, Brazil, where he is responsible for academic research projects in the area of image segmentation, data analysis, machine learning and physics-based animation.

Carlos Thomaz received in 1993 the B.Sc. degree in electronic engineering from Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil. After working for six years in industry, he obtained the M.Sc. degree in electrical engineering from PUC-Rio in 1999. In October 2000, he joined the Department of Computing at Imperial College London where he obtained the Ph.D. degree in statistical pattern recognition in 2004. He joined the Department of Electrical Engineering, FEI University Center, São Paulo, Brazil, in 2005, as an Associate Professor, where he has been, since 2006, head of the Image Processing Laboratory. Since 2014 he has been Professor of Statistical Pattern Recognition at FEI. His research interests include pattern recognition, cognitive perception and machine learning. From 2015 to 2018, Professor Thomaz was awarded a Newton Advanced Fellowship from the Royal Society, UK.