Análise de Componentes Principais

José Francisco Pereira
George Darmiton da Cunha Cavalcanti

{jfp,gdcc}@cin.ufpe.br
CIn-UFPE

Roteiro

- Introdução
- Características da técnica
- Fundamentos matemáticos
- O algoritmo PCA
- *Toy Problem*
- Vantagens e desvantagens
- Aplicações

Introdução

- PCA é uma técnica de análise estatística útil para compressão, visualização e classificação de dados
- Proposta inicialmente em 1901 por Karl Pearson mas generalizada apenas em 1963 por Loève.
  - Também é conhecida como transformada de Karhunen-Loève
- O propósito é reduzir a dimensionalidade de um conjunto de dados.
- Para tanto, um novo conjunto de variáveis menor do que o conjunto original e que mantém a maioria da informação da amostra é calculado

Introdução

- Informação diz respeito à variação presente na base de dados. Em geral, suas variáveis são correlacionadas e possuem redundância

- As variáveis do novo conjunto produzido pela técnica são não-correlacionadas e guardam a maior parte da informação dos padrões

- Em geral, esta perda de informação é mais que compensada pela representação mais concisa e precisa dos dados

Características da Técnica

- Reduz a dimensionalidade eliminando a redundância dos dados
  - Variáveis que medem o mesmo evento
  - Que possuem dependência entre si
- A análise de redundância é feita pela análise da matriz de covariância destes dados
- Expressa os mesmos dados em um sistemas de eixos diferentes
- Cada eixo representa uma componente principal
- Em função do novo sistema de eixos ser ortogonal as variáveis são não-correlacionadas

Características da Técnica

- Os novos eixos são produzidos por *combinações lineares* dos eixos originais e são selecionados segundo sua variância (qtde de informação)

- Sobre o número de componentes principais e variáveis dos padrões
  - # de componentes = # de variáveis originais
  - Maior parte da informação concentra-se em poucos componentes
  - Obtém-se boa representação em baixa dimensão
  - Não há perda de informação. Os dados originais podem ser reconstruídos

Um pouco de matemática e estatística

- Média
  - Valor médio dos padrões da amostra

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Desvio Padrão
  - Medida da dispersão dos dados
  - Valor não-negativo e com valores na escala dos padrões da amostra

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}$$

Um pouco de matemática e estatística

- Variância e covariância
  - São medidas de dispersão estatística
  - Variância se aplica a uma variável enquanto covariância só se aplica a duas variáveis
  - Variância pode ser entendida como o quadrado do desvio padrão

$$var(\mathrm{X}) = \mathrm{s}^2 = \frac{\sum_{i=1}^{n}(\mathrm{X}_i - \bar{X})^2}{(n-1)}$$

$$cov(\mathrm{X},\mathrm{Y}) = \mathrm{s}^2 = \frac{\sum_{i=1}^{n}(\mathrm{X}_i - \bar{X})\,(\mathrm{Y}_i - \bar{Y})}{(n-1)}$$

Um pouco de matemática e estatística

- Matriz de covariância – (C)
  - Matriz de valores de covariância das variáveis (vetor) de um conjunto de dados
  - Matriz simétrica quadrada (*m x m)*. Sendo *m* o número de característica do padrão.

$$C = \frac{1}{M}\sum_{j=1}^{M}(X_j - \bar{X})^T(X_j - \bar{X})$$

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Um pouco de matemática e estatística

- Autovetores e autovalores
  - Representam os vetores base do novo espaço vetorial que melhor representa os dados originais
  - São vetores ortonormais
  - Autovalores representam a quantidade de informação presente em cada autovetor (dimensão do novo espaço)

O Algoritmo PCA

1. Calcula-se a média e normaliza-se todo o conjunto de dados

H	M	$(H_i - \bar{H})$	$(M_i - \bar{M})$
9	39	-4.92	-23.42
15	56	1.08	-6.42
25	93	11.08	30.58
14	61	0.08	-1.42
10	50	-3.92	-12.42
18	75	4.08	12.58
0	32	-13.92	-30.42
16	85	2.08	22.58
5	42	-8.92	-20.42
19	70	5.08	7.58
16	66	2.08	3.58
20	80	6.08	17.58

O Algoritmo PCA

2. Calcula-se a matriz de covariância

$$s = \frac{1}{M}\sum_{j=1}^{M}(X_j - \bar{X})^T(X_j - \bar{X})$$

4. Calculam-se os autovetores e os autovalores da matriz de covariância

O Algoritmo PCA

4. Escolhem-se os K autovetores com maior quantidade de informação associada
   a) Os autovalores associados expressam a quantidade de informação

5. Monta-se a matriz de projeção P baseado nos autovetores selecionados previamente

$$P = [e_1, e_2, \ldots, e_k]$$

O Algoritmo PCA

6. Projeta-se a imagem normalizada obtida na etapa 1 pela matriz de projeção produzida na etapa 5.

$$Z_j = (X_j - \bar{X}) \cdot P$$

7. Desta forma o novo vetor Z_j de dimensão K será a nova representação do padrão original X_j

8. Geralmente, $K < n$.

Exemplo de uso da técnica

- Elaborada uma base artificial de dados para estudo de caso

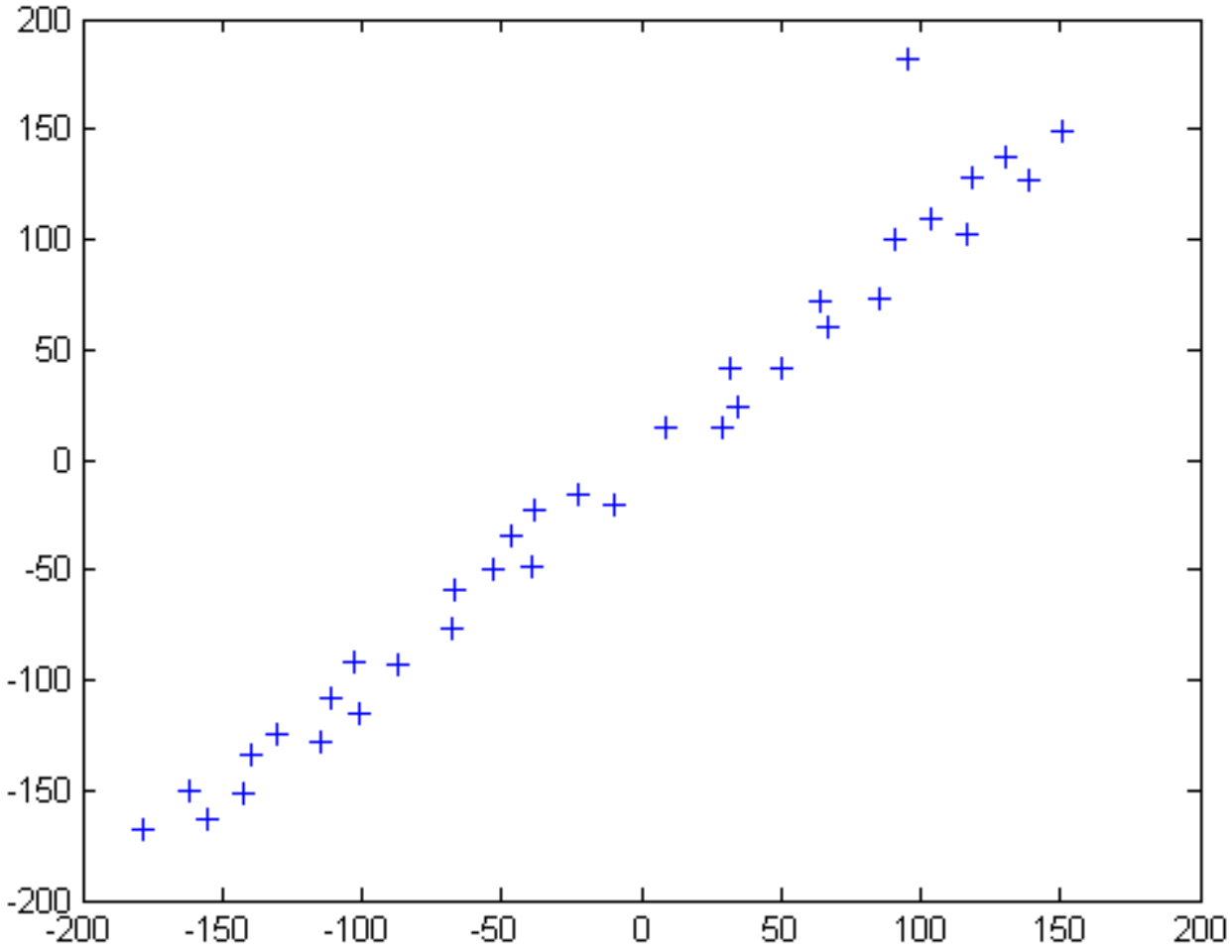
- "toy problem" simples visando:
  - Fazer um estudo de caso detalhado da técnica
  - Eliminar a redundância dos dados do problema
  - Analisar a distribuição de informação entre as componentes resultantes

Toy Problem

- Dados artificiais:
  - [151 149] [-38 -23] [85 73] [-101 -115]
    [130 137] [-47 -34] [64 72] [-111 -108]
    [139 127] [-39 -49] [67 60] [-115 -128]
    [118 128] [-53 -50] [50 41] [-130 -124]
    [117 102] [-67 -59] [32 41] [-140 -134]
    [104 109] [-68 -77] [35 24] [-142 -152]
    [91 100] [-87 -93] [29 15] [-155 -163]
    [95 182] [-103 -92] [9 15] [-162 -150]
    [-10 -20] [-178 -168] [-23 -16]

Toy Problem

- Representação gráfica dos dados

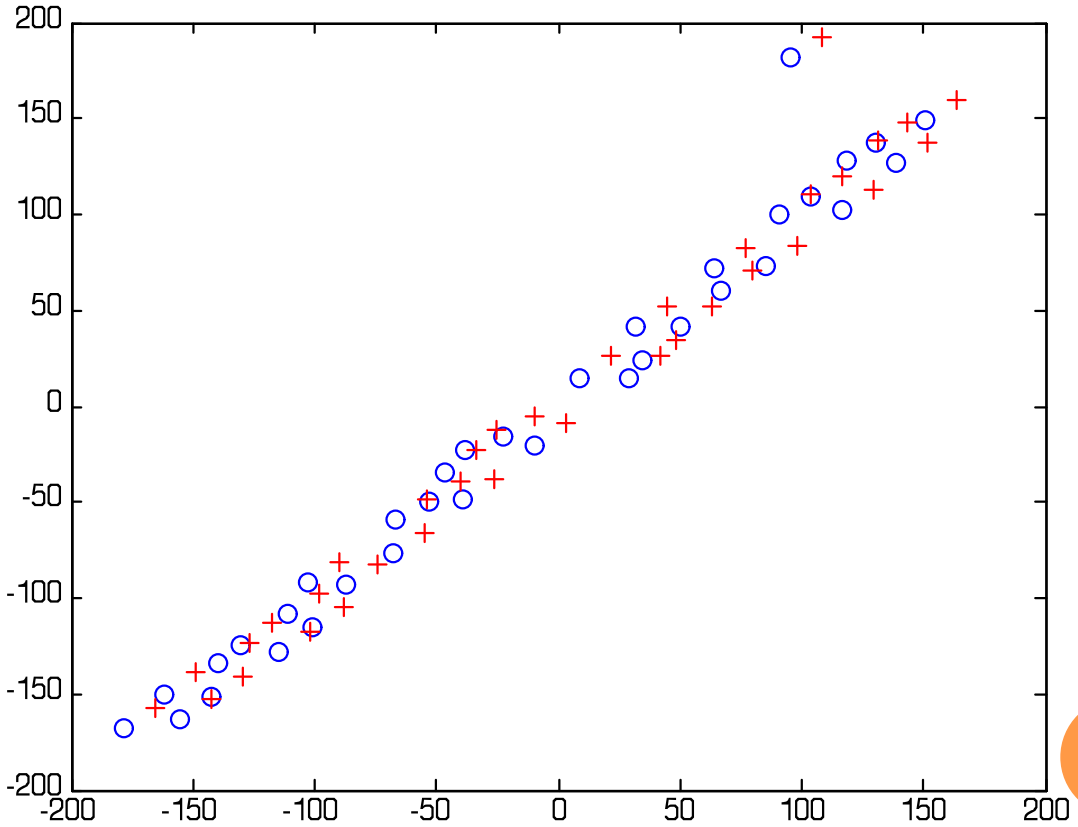


Toy Problem

- Média dos valores
  - [-12.9429 -10.8571]
- Normaliza-se a base de dados

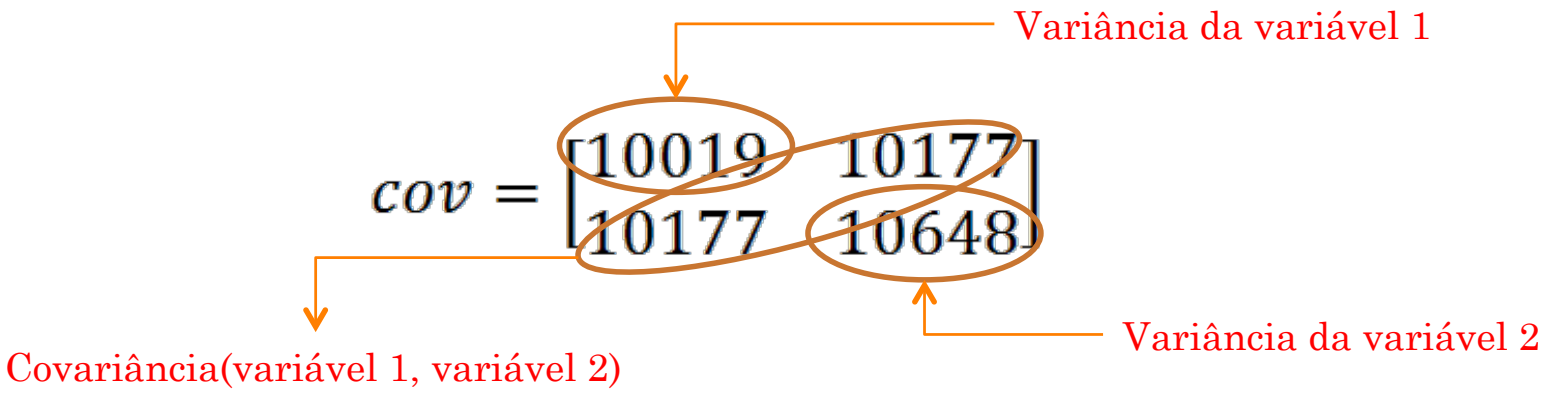
‘o’ Originais

‘+’ Normalizados



Toy Problem

- Matriz de covariância dos dados



- Percebe-se um elevado valor de covariância entre as duas variáveis dos dados (diagonal secundária)

Toy Problem

- Extrai-se os autovetores e autovalores da matriz de covariância
  - Autovetor 1

$$e_1 = \begin{bmatrix} 0{,}6961 \\ 0{,}7179 \end{bmatrix} \rightarrow \lambda_1 = 20516$$

  - Autovetor 2

$$e_2 = \begin{bmatrix} -0{,}7179 \\ 0{,}6961 \end{bmatrix} \rightarrow \lambda_2 = 151$$

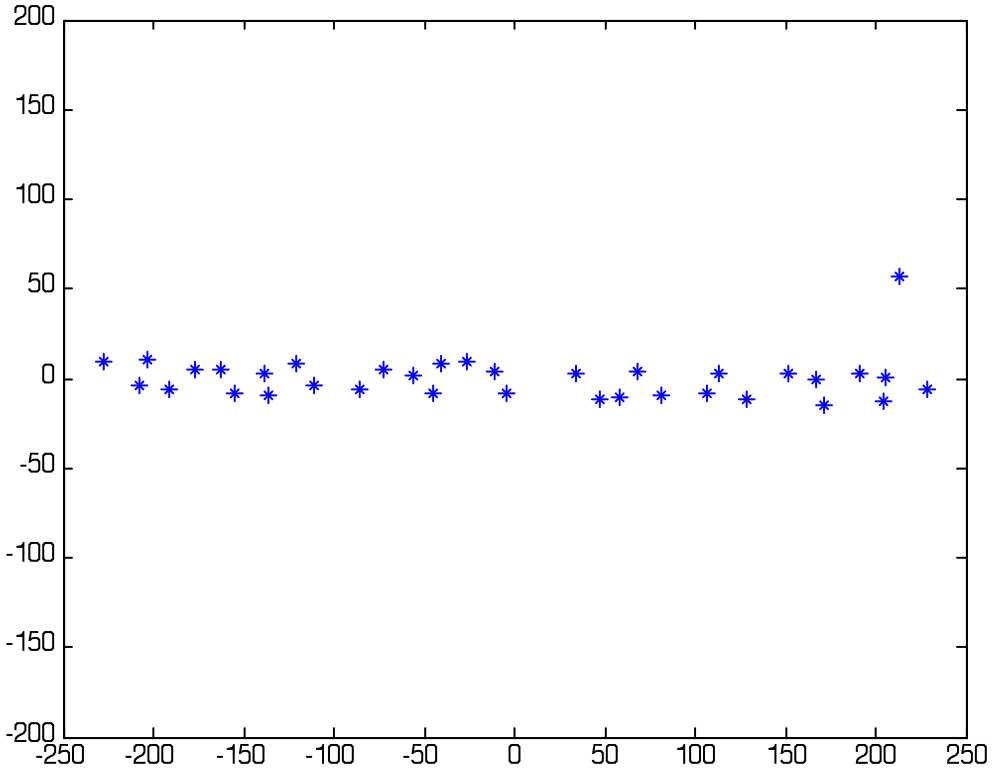
Toy Problem

- Distribuição de informação pelos autovetores
  - Vetor 1: $\lambda 1 = 20516 \approx 99{,}27\%$
  - Vetor 2: $\lambda 1 = \quad 151 \approx \quad 0{,}73\%$

- Fazendo uso de apenas uma componente principal (Vetor 1) obtém-se:
  - Redução da dimensionalidade pela metade
  - Preservação de aprox. 99,3% da informação dos dados
  - Troca do sistema de eixos que define o espaço vetorial. Do original {(1, 0), (0, 1)} para o novo sistema de eixos {(0.6961, 0.7179)}

Toy Problem

- A nova representação dos dados utilizando um componente principal



"+" dados originais normalizados

"*" dados rotacionados após o uso do PCA

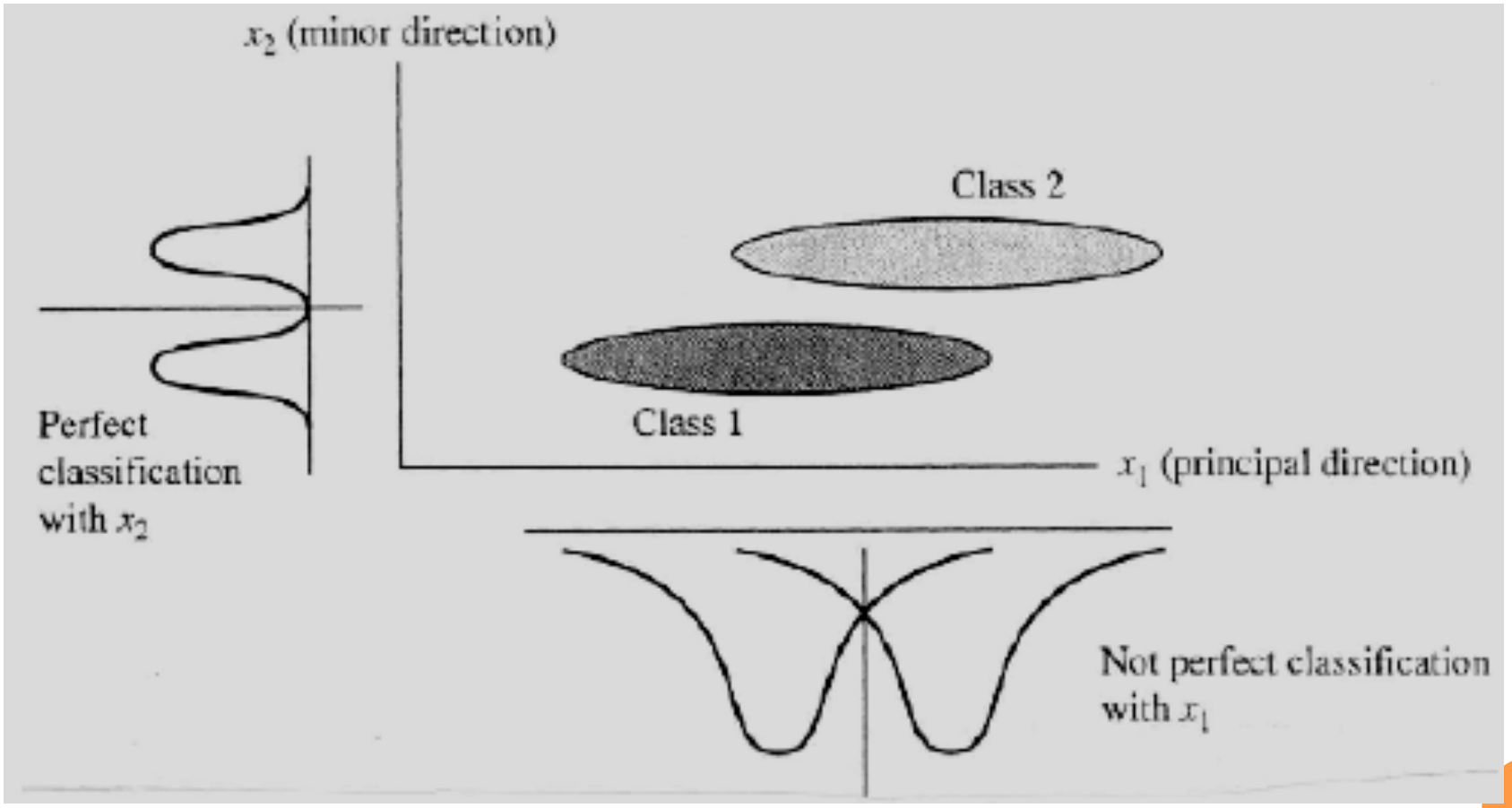
Vantagens e Desvantagens

- Vantagens
  - Alto poder de representação
  - Técnica puramente estatística
  - Robusta e largamente utilizada
  - Possui muitas adaptações
  - Reduz custo de armazenamento e posterior classificação
  - Fácil implementação

Vantagens e Desvantagens

- Desvantagens
  - Limitação na distribuição dos dados
  - Número *X* dimensionalidade dos protótipos
  - Não consideram as classes dos padrões envolvidos
  - Não é uma técnica de redução de dimensionalidade ótima para classificação

PCA e classificação



Aplicações

- Reconhecimento de faces
- Detecção de faces
- Reconstrução de imagens
- Compressão de dados
- Visualização de dados multidimensionais

Referências

- Jonathon Shlens. *A Tutorial on Principal Component Analysis* (v. 2), University of California, San Diego.

- M. Turk and A. Pentland. *Eigenfaces for Recognition*. Journal of Cognitive Neuroscience. 3(1). pp. 71-86, 1991.

- Zhao, Chellappa, Rosenfeld and Phillips. *Face recognition: A literature survey*. UMD CAR Technical Report CAR-TR-948, 2000.