BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC QUY NHƠN

NGUYỄN NHẬT NAM

MỘT SỐ PHƯƠNG PHÁP PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC
Ngành: TOÁN ỨNG DỤNG

Người hướng dẫn
TS.LÂM THỊ THANH TÂM

Bình Định, 2023

MỞ ĐẦU

Nội dung khóa luận được trình bày trong 4 chương:

- **Chương 1. Một số kiến thức chuẩn bị**
- **Chương 2. Phân tích giá trị kì dị (SVD)**
- **Chương 3. Phân tích thành phần chính (PCA)**
- **Chương 4. Một số ứng dụng của SVD và PCA**

Chương 2. Phân tích giá trị kì dị (SVD)

- **Định lí về sự tồn tại của SVD**
- **Thuật toán SVD**

Định lý

Cho $\mathbf{A}$ là một ma trận có cấp $m \times n$. Khi đó, mọi giá trị riêng của ma trận $\mathbf{A}^T\mathbf{A}$ đều không âm.

Định lý

Mọi ma trận $\mathbf{A}$ cỡ $m \times n$ bất kì đều có phân tích SVD có dạng

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

Định lý

Cho $\mathbf{A}$ là ma trận cỡ $m \times n$ bất kì. Khi đó

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T.$$

trong đó σ_i là các giá trị kì dị dương của ma trận $\mathbf{A}$, $\mathbf{u}_i$ là các vector kì dị trái và $\mathbf{v}_i$ là các vector kì dị phải của ma trận $\mathbf{A}$.

Thuật toán SVD

- Bước 1: Tính ma trận $\mathbf{A}^T\mathbf{A}$ và giải phương trình $\det\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right) = 0$ từ đó suy ra các giá trị kì dị của $\mathbf{A}$ là $\sigma_i = \sqrt{\lambda_i}, i = \overline{1,n}$ và $\mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$.
- Bước 2: Với mỗi giá trị riêng λ_i, tìm vector riêng $\mathbf{v}_i$. Từ đó tìm được ma trận trực giao $\mathbf{V}$ cấp n chứa các vector kì dị phải của $\mathbf{A}$.
- Bước 3: Tìm ma trận trực giao $\mathbf{U}$ với

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i.$$

CHƯƠNG 3. PHÂN TÍCH THÀNH PHẦN CHÍNH

3.1 Ý Tưởng

Tìm một phép xoay trục toạ độ để được một hệ trục toạ độ mới sao cho trong hệ mới này, thông tin của dữ liệu chủ yếu tập trung ở một vài thành phần. Phần còn lại chứa ít thông tin hơn có thể được lược bỏ.

3.2 Phân tích thành phần chính

Với $\mathbf{X} \in \mathbb{R}^{m \times n}$ với mỗi giá trị đã được chuẩn hóa sao cho mỗi hàng có giá trị trung bình là 0. Thì PCA của $\mathbf{X}$ với r thành phần là tìm ma trận trực giao $\mathbf{A} \in \mathbb{R}^{m \times r}$ và ma trận $\mathbf{B} \in \mathbb{R}^{n \times r}$ sao cho $\mathbf{X}$ có thể biểu diễn được dưới dạng ma trận.

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T + \mathbf{E}.$$

trong đó $\|\mathbf{E}\|^2$ đạt giá trị nhỏ nhất.

Nhận xét

$\mathbf{A}\mathbf{B}^T$ là SVD "chặt cụt" của $\mathbf{X}$, nghĩa là nếu $\mathbf{U}_r\mathbf{D}_r\left(\mathbf{V}_r\right)^T$ là SVD "chặt cụt" của $\mathbf{X}$ thì $\mathbf{A} = \mathbf{U}_r$ và $\mathbf{B}^T = \mathbf{D}_r\left(\mathbf{V}_r\right)^T$.

3.2 Phân tích thành phần chính

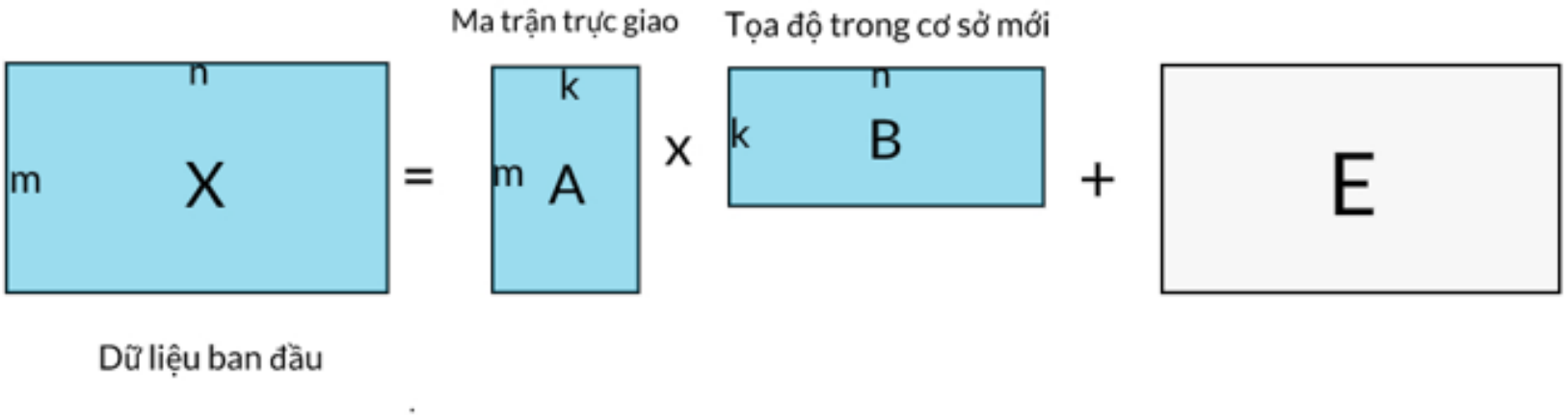


Figure: Ảnh minh họa PCA

3.3 Tính duy nhất nghiệm của PCA

Định lí

Nếu $(\mathbf{A}, \mathbf{B})$ là một nghiệm của mô hình PCA thì $(\mathbf{AQ}, \mathbf{BQ})$ cũng là một nghiệm của mô hình PCA, với $\mathbf{Q}$ là ma trận trực giao cấp r.
Lúc này, $\mathbf{Q}$ được gọi là phép quay trực giao.

3.4 Thuật toán PCA

- Bước 1: Chuẩn hóa dữ liệu.
- Bước 2: Tìm SVD "chặt cụt" của ma trận $\mathbf{X}$, ta được $\mathbf{X} = \mathbf{U}_r\mathbf{D}_r\left(\mathbf{V}_r\right)^T$ với $r \leq n$.
- Bước 3: Tìm ma trận $\mathbf{A}$ và $\mathbf{B}$.
- Bước 4: Nếu nghiệm $(\mathbf{A}, \mathbf{B})$ chưa tốt thì chọn phép quay $\mathbf{Q}$, với $\mathbf{Q}$ là ma trận trực giao cấp r.

CHƯƠNG 4. MỘT SỐ ỨNG DỤNG CỦA SVD VÀ PCA

4.1 Ứng dụng của SVD trong xấp xỉ hạng thấp tốt nhất của ma trận

Định lí Eckart- Young, 1936

Với mọi ma trận $\mathbf{B}$ cỡ $m \times n$ và $\text{rank}(\mathbf{B}) \leqslant k$, ta có

$$\|\mathbf{A} - \mathbf{B}\|_F \geqslant \|\mathbf{A} - \mathbf{A}_k\|_F .$$

4.1 Ứng dụng của SVD trong xấp xỉ hạng thấp tốt nhất của ma trận

Định lí

Giả sử $\mathbf{A}$ là ma trận cỡ $m \times n$ bất kì, với $\text{rank}(A) = r$ và $\mathbf{A}$ có khai triển kì dị SVD là

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathcal{V}_i^T.$$

Khi đó, với mọi ma trận $\mathbf{B}$ có cỡ $m \times n$ bất kì và $\text{rank}(\mathbf{B}) \leqslant k$, ta có

$$\|\mathbf{A} - \mathbf{B}\|_2 \geqslant \sigma_{k+1}.$$

Dấu "=" xảy ra khi $\mathbf{B} = \mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathcal{V}_i^T$.

4.2 Phân tích SVD trong xử lí ảnh

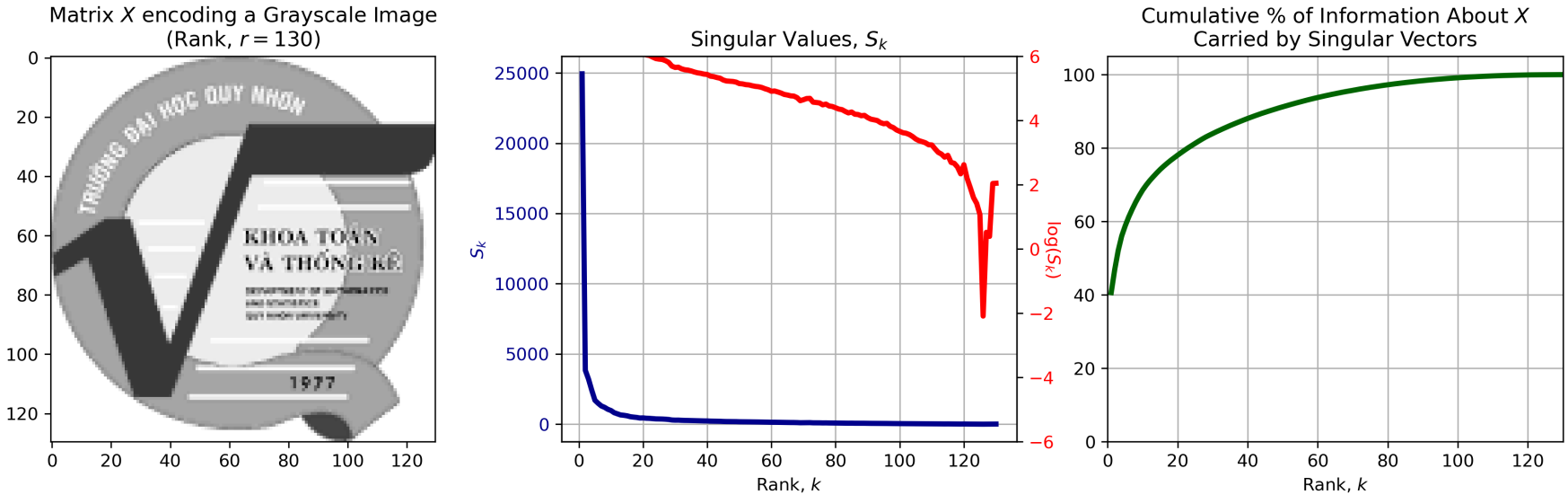


Figure: Tính quan trọng của các thành phần chính trong ma trận ảnh và mức độ giữ lại thông tin khi ta giảm số lượng các thành phần chính

4.2 Phân tích SVD trong xử lí ảnh

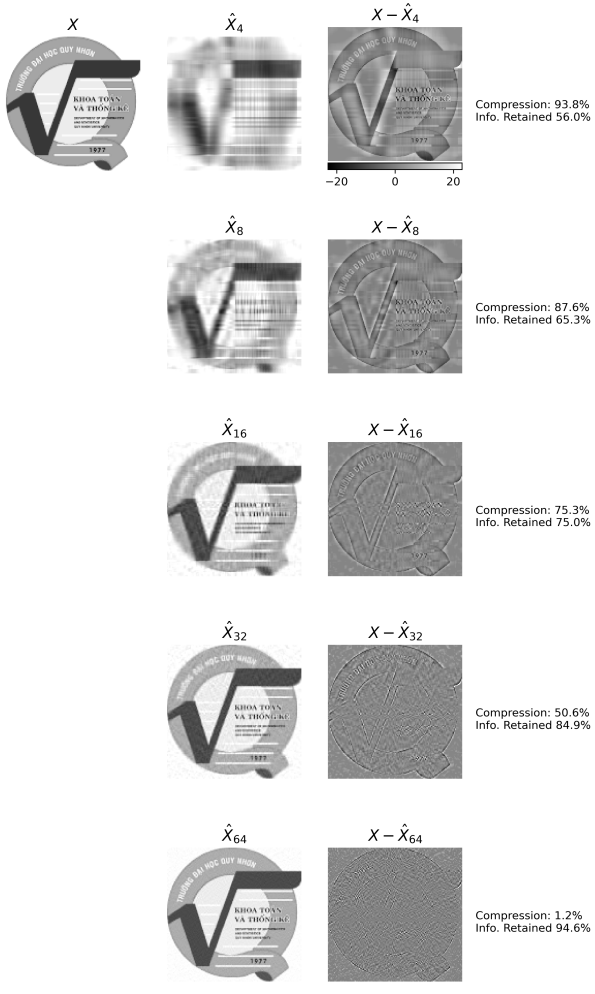
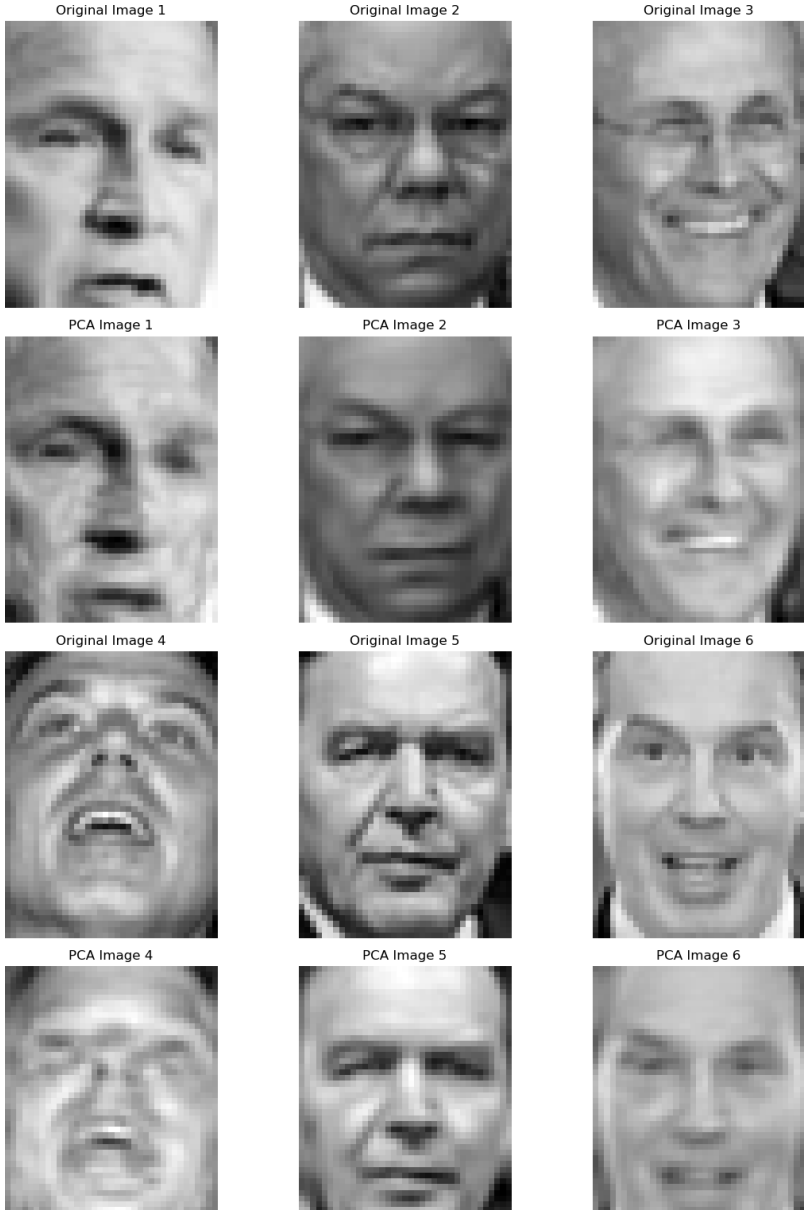


Figure: Ứng với từng k ta quan sát được mức độ tối ưu trong việc lưu trữ thông tin cũng như là lượng thông tin được giữ lại.

4.3 Ứng dụng của PCA trong nhận dạng khuôn mặt



4.3 Ứng dụng của PCA trong nhận dạng khuôn mặt

	precision	recall	f1-score	support
Ariel Sharon	0.62	0.77	0.69	13
Colin Powell	0.77	0.83	0.80	60
Donald Rumsfeld	0.59	0.63	0.61	27
George W Bush	0.88	0.84	0.86	146
Gerhard Schroeder	0.83	0.80	0.82	25
Hugo Chavez	0.75	0.60	0.67	15
Tony Blair	0.73	0.75	0.74	36
accuracy			0.80	322
macro avg	0.74	0.75	0.74	322
weighted avg	0.80	0.80	0.80	322

Figure: Báo cáo phân loại sau khi huấn luyện bộ phân loại SVM.

4.3 Ứng dụng của PCA trong nhận dạng khuôn mặt

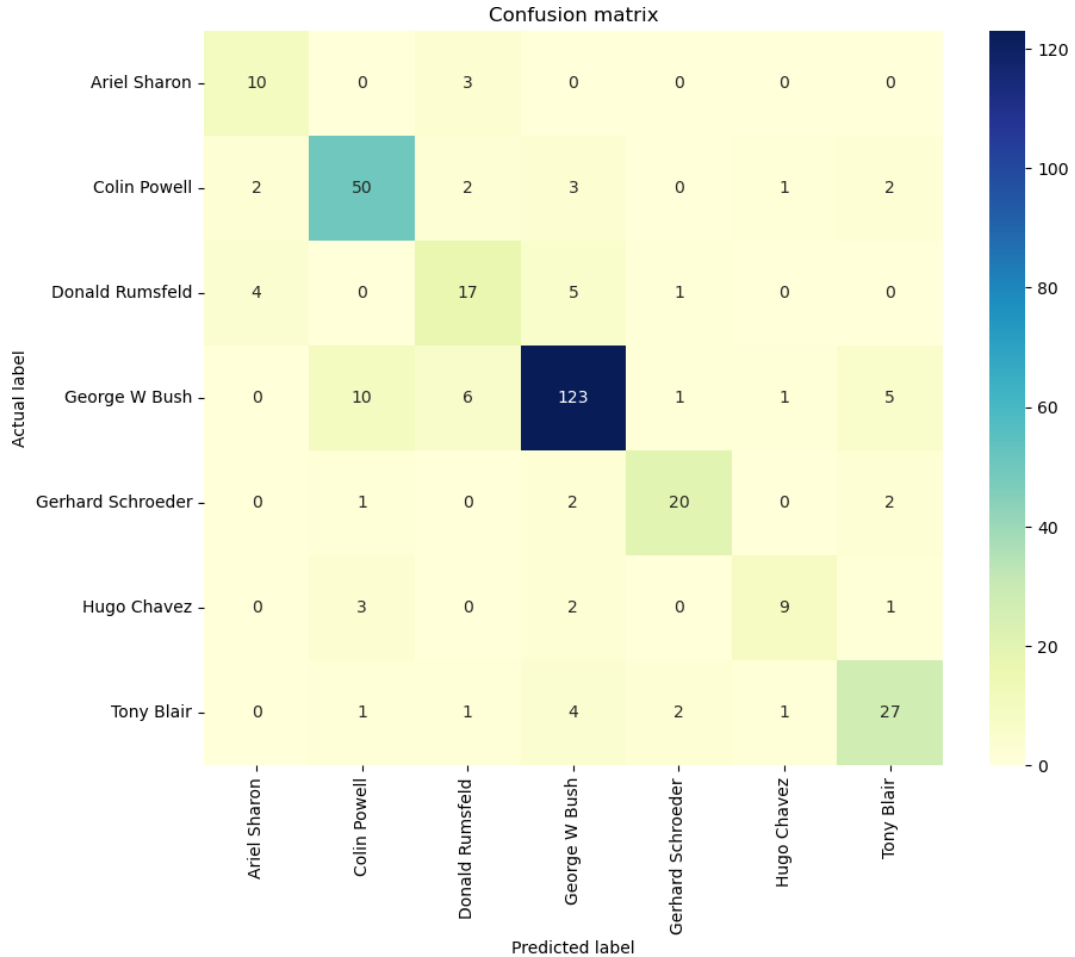


Figure: Ma trận lỗi(confusion matrix)

4.3 Ứng dụng của PCA trong nhận dạng khuôn mặt

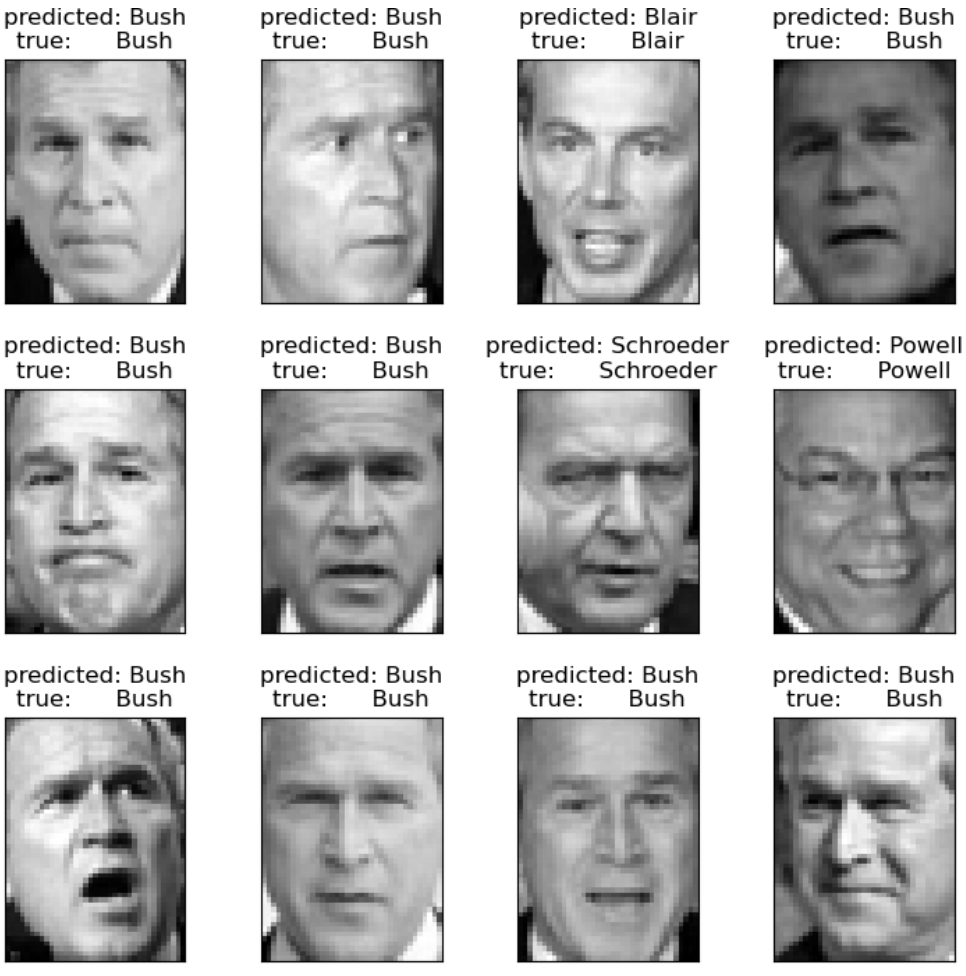


Figure: kết quả suy luận của bộ phân loại trên tập kiểm thử

4.3 Nghiên cứu về ứng dụng SVD trong kiến trúc Transformer

Kiến trúc mô hình	Độ phức tạp thời gian	Số phép toán tuần tự
Recurrent	$O(n)$	$O(n)$
Transformer	$O\left(n^2\right)$	$O(1)$
Sparse Tansformer	$O(n\sqrt{n})$	$O(1)$
Reformer	$O(n\log(n))$	$O(\log(n))$
Linformer	$O(n)$	$O(1)$

KẾT THÚC BÁO CÁO

TRÂN TRỌNG CẢM ƠN!