Jurnal Matematika Integratif
Volume 12 No.1, April 2016, pp 27 – 34
ISSN 1412-6184

Pengenalan Huruf Arab Menggunakan Metode Reduksi *Two Dimensional Principal Component Analysis* (2DPCA) dan Klasifikasi *K-Nearest Neighbor* (K-NN)

Masitoh Majid, Arif F. Huda, Rini Cahyandari
Laboratorium Computational Inteligent, Jurusan Matematika,
Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Gunung Djati Bandung
Jln. A.H. Nasution No. 105 Cibiru, Bandung, Jawa Barat 40614 Indonesia
E-mail: masitohmajid@gmail.com, afhuda@gmail.com, rcahyandari@yahoo.com

ABSTRAK

Pengenalan huruf arab merupakan salah satu pengenalan pola gambar dengan mengetahui ciri-ciri utama dari gambar tersebut. Huruf arab dapat dikenali menggunakan metode reduksi *Two Dimensional Principal Component Analysis* (2DPCA) dan klasifikasi k-*Nearest Neighbor* (k-NN). 2DPCA menggunakan format data gambar input berupa matrik. Terdapat dua pendekatan dalam 2DPCA yaitu *Unilateral 2 Dimensional Principal Component Analysis* (U2DPCA) dan *Bilateral 2 Dimensional Principal Component Analysis* (B2DPCA). Dalam perhitungan 2DPCA, unilateral hanya menggunakan baris atau kolom dari matrik gambar. Sedangkan Bilateral menggunakan baris dan kolom secara bersamaan. Huruf arab yang digunakan 126 huruf yang terdiri dari 9 huruf hijaiyah yaitu alif, ba, ha, dal, sin, shad, tha, mim, dan Haa. Dengan masing-masing huruf digunakan 14 tipe penulisan yaitu *arial, corie new, microsoft san serif, microsoft ughur, sakhal majalla, sagoe UI, simplyfied arabic, simplyfied arabic fixed ,tahoma, times new roman, traditional arabic, arabic typeseting, arial unicode ms, dan andalus.* Berdasarkan percobaan, 9 tipe sebagai data latih dan 5 tipe sebagai data uji maka rata-rata akurasi pengenalan huruf arab menggunakan metode U2DPCA (baris) yaitu sebesar 70% dengan menggunakan 40 eigen vektor. Rata-rata akurasi pengenalan huruf arab menggunakan metode U2DPCA (kolom) yaitu sebesar 84% dengan menggunakan 50 eigen vektor. Sedangkan rata-rata akurasi pengenalan huruf arab menggunakan metode B2DPCA yaitu sebesar 95% dengan menggunakan 7 eigen vektor. Sehingga, pada penelitian ini metode reduksi yang paling baik untuk pengenalan huruf arab adalah B2DPCA.

Kata kunci: *Two Dimensional Principal Component Analysis* (2DPCA), U2DPCA, B2DPCA, k-*Nearest Neighbor* (k-NN), karakter huruf arab, matrik kovarian, nilai eigen, vektor eigen.

ABSTRACT

Recognition of the Arabic alphabet is one of the image recognition by determined main character of image. Arabic letters can be identified by using the reduction method*of Two Dimensional Principal Component Analysis* (2DPCA) and classification *k-Nearest Neighbor* (k-NN).2DPCA was using the input image in the form of a matrix. The approaches of 2DPCA are *Unilateral 2 Dimensional Principal Component Analysis* (U2DPCA) and*Bilateral 2 Dimensional Principal Component Analysis (B2DPCA).*In the calculation 2DPCA, unilateral just only using the rows or columns of an image matrix. While, Bilateral uses rows and columns simultaneously. The arabic alphabet are used in this research including 126 letters consisting of 9 letters of hijaiyah there are alif, ba, ha, dal, sin, shad, tha, mim, and Haa. With each lettersis used were come from 14 types of writing there is *arial, Corie new, microsoft san serif, microsoft ughur, sakhal Majalla, sagoe UI, simplyfied arabic, simplyfied arabic fixed, tahoma, Times New Roman, traditional arabic, arabic typeseting, arial unicode ms and andalus.* Based on the experiment, 9 types as training and 5 type as the testing, the average recognition accuracy of the Arabic alphabet using U2DPCA (rows) is 70% by using 40 eigen vectors. The average recognition accuracy of the Arabic alphabet using U2DPCA (columns) is 84% with 50 eigen vectors. While, the average recognition accuracy of the Arabic alphabet using B2DPCA methode

is 95% by using 7 eigen vectors. Thus, in this study the most excellent reduction method for the introduction of the Arabic alphabet is B2DPCA.

Keywords: *Two Dimensional Principal Component Analysis (2DPCA), U2DPCA, B2DPCA, k-Nearest Neighbor* (k-NN), the character of the Arabic alphabet, covariance matrix, eigenvalues, eigenvectors.

1. Pendahuluan

Allah SWT menurunkan Al-Qur'an kepada Nabi Muhammad SAW melalui malaikat Jibril AS. Umat islam wajib memahami isi kandungan Al-Qur'an dengan terlebih dahulu mengetahui tulisan Al-Qur'an. Al-Qur'an ditulis dalam bahasa arab, dari kata qara'a - yaqra'u – qur'anan dengan makna bacaan atau sesuatu yang dibaca berulang-ulang. Namun, pengertian lebih rinci Al-Qur'an mencakup segala aspek kehidupan berupa *hudan* (petunjuk), *bayyinah* (penjelas) dan *furqan* (pembeda). Al-Qur'an surat Yusuf ayat 2 menerangkan tentang tulisan Al-Qur'an. *"Sesungguhnya kami menurunkannya berupa Qur'an berbahasa Arab, agar kamu mengerti."* (QS.Yusuf : 2). Huruf arab dalam Al-Qur'an disebut huruf hijaiyyah dengan jumlah 29 huruf. Huruf arab memiliki bentuk huruf yang berbeda-beda sesuai dengan posisi penulisannya. Huruf arab dapat ditulis bersambung atau berdiri sendiri yaitu pada awal, tengah dan akhir kalimat. Huruf arab dapat dipelajari langsung menggunakan metode baca Iqra tetapi memerlukan waktu yang lama. Huruf arab jarang digunakan dalam implementasi sehari-hari, karena memiliki pola yang unik sesuai posisi penulisan. Dengan perkembangan teknologi, komputer dapat membantu manusia untuk mengenali huruf arab dengan lebih mudah dan cepat. Komputer dapat mengenali huruf arab dengan dua metode, yaitu ekstraksi ciri dan klasifikasi.

Pada penelitian ini, metode ekstraksi ciri yang digunakan adalah metode reduksi *Two Dimensional Principal Component Analysis* (2DPCA). Setelah data direduksi, data akan diklasifikasikan menggunakan metode k-*Nearest Neighbor* (k-NN). 2DPCA merupakan pengembangan dari metode PCA. Dengan beberapa kelebihan yaitu matrik citra 2D langsung dihitung tanpa mengubah ke bentuk 1D (vektor). Perhitungan matrik kovarian pada 2DPCA lebih sederhana dan mudah untuk ekstraksi ciri [3].

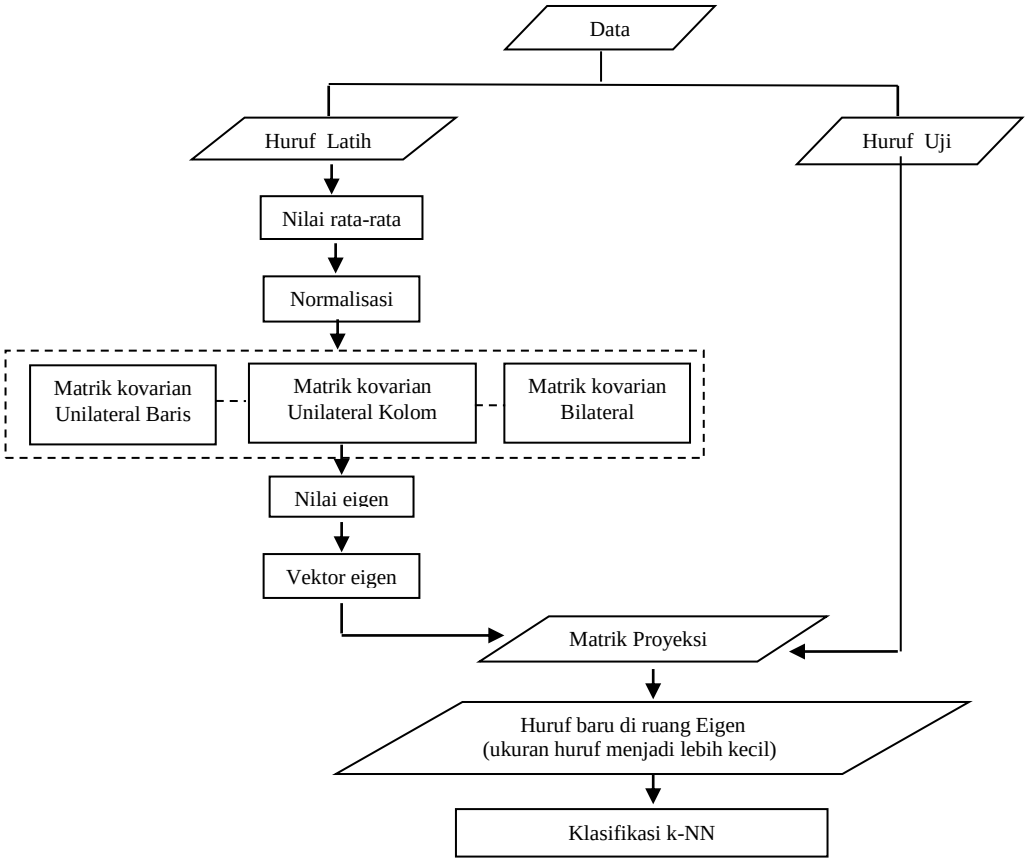
Metode 2DPCA telah dikembangkan oleh beberapa peneliti. Diantaranya, Dhiraj K, 2012, yang berjudul *"Comparative Analysis of PCA and 2DPCA in Face Recognition"*. Penelitian Dhiraj yaitu perbandingan metode PCA dan 2DPCA dalam pengenalan data wajah manusia yang berbeda. Dapat disimpulkan bahwa metode 2DPCA lebih baik dari pada metode PCA. Metode 2DPCA dikembangkan juga oleh Kong et.al, 2005, yang berjudul *"Generalized 2D principal component analysis for face image representation andrecognition"*. Penelitan Kong et.al yaitu menggunakan metode 2DPCA dengan pendekatan *Unilateral* 2DPCA dan *Bilateral* 2DPCA. Didapatkan hasil penelitian Kong et.al bahwa metode 2DPCA dengan pendekatan *Bilateral* 2DPCA lebih baik dari pada menggunakan pendekatan *Unilateral* 2DPCA [2][4]. Berdasarkan latar belakang diatas, maka akan diteliti pengenalan huruf arab menggunakan metode reduksi 2DPCA dan klasifikasi k-NN.

2. Metode Penelitian

Penelitian ini menggunakan metode reduksi 2DPCA yaitu mengekstraksi ciri dengan cara mengambil ciri-ciri yang penting. Metode reduksi 2DPCA pertama kali dikembangkan oleh Kong et.al, 2005 yaitu menggunakan data input matrik gambar 2D yang tidak diubah ke bentuk 1D [4]. Ukuran kemiripan yang digunakan dalam klasifikasi k-NN adalah *Euclidean Distance* [5].

Berdasarkan teoritis diatas, skema metode penelitian pengenalan huruf arab menggunakan metode reduksi 2DPCA dan klasifikasi k-NN sebagai berikut:

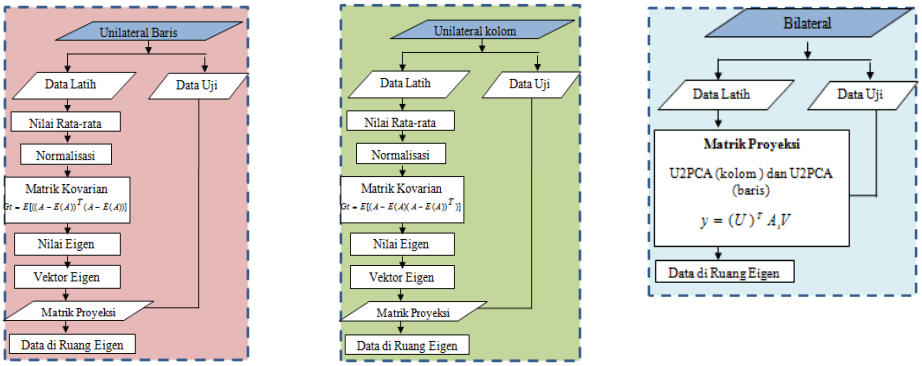
 ISSN 1412-6184



Gambar 2.1 Skema metode 2DPCA

2.1 Metode Reduksi Two Dimensional Principal Component Analysis (2DPCA)

Metode reduksi 2DPCA yaitu prsoses ekstraksi ciri yang dapat menurunkan dimensi gambar menjadi lebih kecil. Sehingga, gambar akan mudah digunakan untuk proses selanjutanya. Metode 2DPCA terbagi beberapa pendekatan yaitu *Unilateral Two Dimensional Principal Component Analysis* (U2DPCA) dan Bilateral *Two Dimensional Principal Component Analysis* (B2DPCA). U2DPCA mempunyai dua solusi yaitu U2DPCA baris dan U2DPCA kolom [3][4]. Skema metode 2DPCA sebagai berikut:



Gambar 2.2 Skema pendekatan 2DPCA

2.2 Algoritma Two Dimensional Principal Component Analysis (2DPCA)

Berikut ini dijelaskan langkah-langkah penentuan matriks kovarian gambar huruf arab menggunakan metode 2DPCA [3]:

1. Terdapat himpunan sebanyak M gambar huruf A_j, dimana $A_j = A_1, A_2, \dots, A_M$ dimana $j = 1, 2, \dots, M$ dengan dimensi gambar (50x40) yang diproyeksikan ke dalam matriks dua dimensi (*Y*).

$$\mathrm{Y} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{M1} \\ x_{12} & x_{22} & \dots & x_{M2} \\ \dots & \dots & \dots & \dots \\ x_{1N} & x_{2N} & \dots & x_{MN} \end{bmatrix} \quad (1)$$

2. Perhitungan rata-rata dari total matriks latih $\bar{A}$:

$$\bar{\mathrm{A}} = \frac{1}{M} \sum_{j=1}^{M} Y_j \quad (2)$$

3. Perhitungan matriks selisih dari setiap citra A_j dengan $\bar{A}$:

$$B = A_j - A \quad (3)$$

4. Perhitungan matriks kovarian dari himpunan citra latih Gt :

$$Gt = \frac{1}{M} \sum_{j=1}^{M} \left(A_j - \bar{A}\right)^T \left(A_j - \bar{A}\right) \quad (4)$$

Gt berupa matriks square.

5. Perhitungan nilai eigen dan vektor eigen dari matriks kovarian. Secara matematis dapat diekspresikan sebagai berikut

$$Ax = \lambda x \quad (5)$$

dimana :
A : *square* matrik (*mxn*)
x : vektor eigen
λ : skalar / nilai eigen
Nilai eigen selalu berkorespondensi dengan perubahan eigen vektor. Kemudian, eigen vektor diproyeksikan sesuai nilai eigen dari yang terbesar $\lambda_1 > \lambda_1 > \cdots > \lambda_n$.

Pre-processing merupakan proses awal setelah pengumpulan data gambar huruf. Pada proses ini huruf latih dan uji melalui *Pre-processing*. Dimulai dengan proses konversi citra RGB menjadi *grayscale*. Tujuan dari proses ini adalah untuk memudahkan proses selanjutnya [5]. karena citra *grayscale* mengandung nilai yang lebih sedikit yaitu 8 bit warna daripada citra RGB dengan 24 bit warna.

1. Pada penelitian ini menggunakan data huruf arab yaitu 9 huruf Hijaiyah yang terdiri dari 9 huruf hijaiyyah tunggal. Pada setiap hurufnya memilki 14 tipe penulisan dengan ukuran 50x40 piksel.

ا ب ح د س ص ط م ه

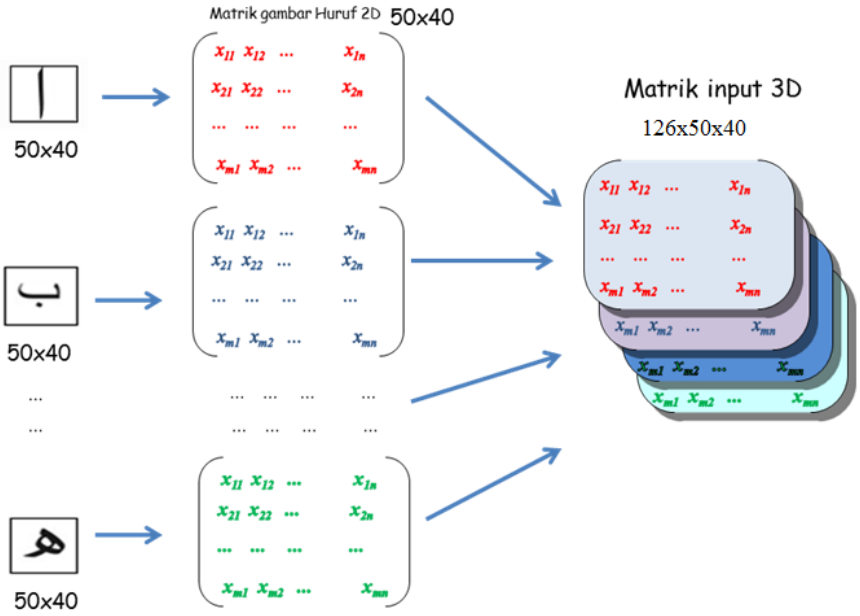
 ISSN 1412-6184

Dari 14 tipe dibagi menjadi 2 bagian yaitu sebagai huruf latih dan uji. Huruf arab dibuat dari *'paint'* kemudian di save dengan format PNG.

Gambar 2.3 Ukuran huruf

2. 126 huruf arab dengan format PNG diubah menjadi citra keabu-abuan *(grayscale)* untuk memudahkan proses komputasi.
3. Membentuk nama file citra huruf yang sudah di grayscale dalam masing-masing folder huruf latih dan uji. Sehingga citra dapat diketahui ketika Matlab sedang menjalankan proses.

Pada tahapan ini setiap gambar huruf arab dapat direpresentasikan dalam sebuah matriks 2D, dengan ukuran 50x40. Matriks input untuk reduksi gambar dan klasifikasi berupa matriks 3D. Data yang digunakan yaitu 9 gambar huruf arab dengan ukuran 50x40 piksel, masing-masing hurufnya memiliki 14 tipe penulisan. Sehingga jumlah huruf arab yaitu 126 huruf dan dapat direpresentasikan dalam matriks 3D, 126x50x40. Dari 126 huruf tersebut dibagi menjadi dua bagian yaitu sebagai data latih dan uji. Penentuan banyaknya huruf latih dan uji ditentukan dengan beberapa percobaan. Pada percobaan pertama matriks latih diambil dari tipe penulisan ke-1 sampai ke-9 dan terbentuk matriks 3D, 81x50x40. Matriks uji diambil dari tipe ke-10 sampai ke-14 [6].



Gambar 2.4 Representasi gambar ke matriks

Tahap selanjutnya yaitu membentuk matriks rata-rata dari matriks huruf latih. Entri-entri atau nilai elemen pada setiap matriks huruf latih, dihitung nilai rata-ratanya sebanyak gambar huruf latih. Sehingga terbentuk sebuah matriks rata-rata, dengan ukuran 50x40. Dengan rumus sebagai berikut [3][4]:

$$\bar{A} = \frac{1}{M}\sum_{j=1}^{M} Y_j \tag{2}$$

Matriks normalisasi terbentuk dari nilai-nilai setiap matriks huruf latih (50x40), dikurangi dengan nilai-nilai matriks rata-rata (50x40). Sehingga terbentuk matriks normalisasi Data, sebanyak jumlah matriks latih (50x40). Dengan rumus sebagai berikut[3]:

$$B = A_j - A \tag{3}$$

Pada tahapan ini matriks normalisasi data sejumlah matriks latih (50x40) akan ditranspose. Sehingga terbentuk matriks normalisasi yang sudah ditanspose (40x50), sejumlah matriks latih. Metode transpose yaitu nilai-nilai baris dalam matriks menjadi nilai-nilai kolom dalam matriks dan sebaliknya.

Pada tahapan selanjutnya, matriks kovarian dapat dibentuk dari hasil variasi perkalian matriks normalisasi, dengan matriks normalisasi yang sudah ditranspose. Sehingga beberapa variasi pembentukan matriks kovarian sebagai berikut:

Matriks kovarian unilateral baris yaitu hasil dari jumlah perkalian matriks normalisasi yang sudah ditranspose (40x50), dengan matriks normalisasi (50x40) sebanyak matriks latih. Sehingga terbentuk sebuah matriks kovarian unilateral baris dengan ukuran (40x40). Dengan rumus sebagai berikut [1][4]:

$$Gt = \frac{1}{M}\sum_{j=1}^{M}\left(A_j - \bar{A}\right)^T\left(A_j - \bar{A}\right) \tag{4}$$

Matriks kovarian unilateral kolom yaitu hasil dari jumlah perkalian matriks normalisasi (50x40), dengan matriks normalisasi yang sudah ditranspose(40x50) sebanyak matriks latih . Sehingga terbentuk sebuah matriks kovarian unilateral kolom dengan ukuran (50x50). Dengan rumus sebagai berikut [4]:

$$Gt = \frac{1}{M}\sum_{j=1}^{M}\left(A_j - \bar{A}\right)\left(A_j - \bar{A}\right)^T \tag{6}$$

Pada tahapan Bilateral yaitu langsung mendapatkan matriks final data, untuk proses pengujian *(testing)* huruf uji ke huruf latih. Perhitungan matriks final data yaitu hasil perkalian matriks proyeksi yang sudah ditranspose, dari metode unilateral kolom. Dengan matriks latih dan matriks proyeksi dari metode unilateral baris. Metode bilateral dapat dibentuk dengan rumus sebagai berikut [4]:

$$P_{i,j} = \left(W_{opt\ kolom}\right)^T \times A_{i,j} \times \left(W_{opt\ baris}\right) \tag{7}$$

Proses selanjutnya perhitungan mendapatkan nilai eigen dan vektor eigen, dari matriks kovarian yang sudah dibentuk pada metode unilateral baris maupun unilateral kolom. Dimana nilai eigen yang didapatkan akan diurutkan dari nilai yang terbesar, dan vektor eigen akan terurut sesuai nilai eigen. Adapun nilai eigen dan vektor eigen pada persamaan berikut [1] :

$$det\ (\lambda I - A) = 0 \tag{8}$$

$$Ax = \lambda x \tag{5}$$

dimana λ merupakan nilai karakteristik dari suatu matriks *A,* disebut dengan nilai eigen, x merupakan vektor eigen yang bersesuaian dengan λ, I merupakan matriks identitas dari matriks *A,* dan *A* merupakan matriks varian kovarian.

Pada penelitian ini, melakukan beberapa percobaan pengambilan vektor eigen yang optimal sesuai nilai eigen. Vektor eigen yang optimal disebut dengan matriks proyeksi (sistem). Matriks proyeksi akan digunakan untuk pembentukan matriks final data, yaitu sebagai proses metode pengenalan untuk matriks data uji.

2.3 Klasifikasi k-Nearest Neighbor (k-NN)

Klasifikasi k-*Nearest Neighbor* (k-NN) merupakan metode klasifikasi yang menggunakan huruf latih, untuk menentukan kelas dari huruf uji. Metode klasifikasi membandingkan jarak huruf uji dengan sejumlah *k* huruf latih yang paling dekat. Jika huruf latih *D* dan huruf uji *d* disajikan. Maka akan dihitung jarak antara huruf latih *D* dengan huruf uji *d,* dinamakan perhitungan *Euclidean Distance.* Kemudian *k* buah data dalam *D,* yang memilki jarak terdekat (nilai min) diambil. Kelas *d* ditentukan berdasarkan mayoritas kelas dalam himpunan *k* tetangga terdekat [5]. Himpunan *k* merupakan *k-Nearest neighbor*. Rumus euclidean distance sebagai berikut:

$$D(a,b) = \sqrt{\sum_k (a_k - b_k)^2} \tag{9}$$

Dimana *a* adalah nilai-nilai dari hasil perkalian matriks kovarian dengan matriks proyeksi. *b* adalah data uji atau data baru yang berada di ruang eigen. Pada penelitian ini, *a*

 ISSN 1412-6184

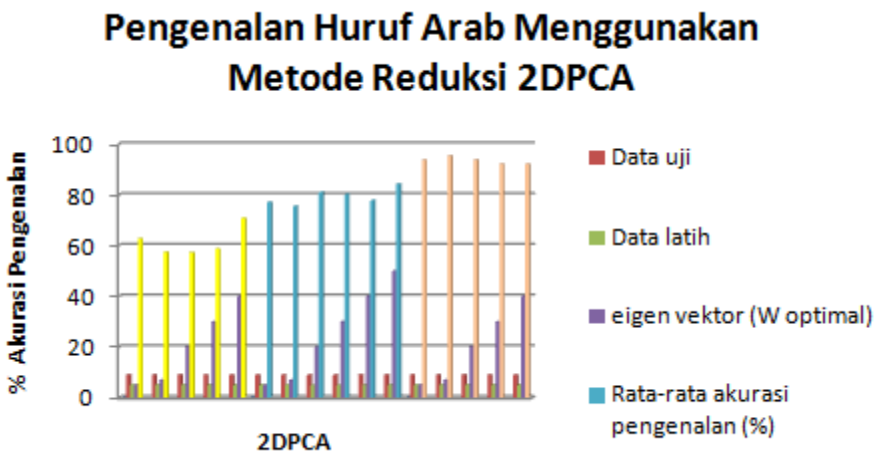
merupakan hasil perkalian matriks latih dengan matriks proyeksi disebut final data. Sedangkan *b* merupakan data uji yang diproyeksikan dengan matriks proyeksi disebut data baru diruang eigen[1][2].

3. Hasil dan Pembahasan

Pada penelitan ini pengenalan huruf arab menggunakan metode reduksi 2DPCA dengan pendekatan U2DPCA baris, U2DPCA kolom, dan B2DPCA. Bertujuan untuk mengetahui berapa persentase kebenaran rata-rata pengenalan huruf arab. Pada setiap percobaan menggunakan 9 tipe huruf sebagai data latih dan 5 tipe huruf sebagai data uji, dengan pengunaan data latih dan data uji secara *rooling*. Diantaranya, percobaan pertama digunakan data latih dan uji sebagai berikut: Data latih : Tipe huruf ke-1 sampai tipe huruf ke-9 ; Data uji : Tipe huruf ke-10 sampai tipe huruf ke-14. Percobaan kedua digunakan data latih dan uji sebagai berikut: Data latih : Tipe huruf ke-2 sampai tipe huruf ke-10 ; Data uji : Tipe huruf ke-1 dan ke-11 sampai tipe huruf ke-14. Digunakan beberapa pengambilan eigen vektor yaitu 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40 dan 50 eigen vektor. Hasil percobaan dapat disajikan dalam tabel dan grafik sebagai berikut:

Tabel 3.5 Perbandingan akurasi pengenalan huruf arab menggunakan metode reduksi 2DPCA

2DPCA	Data uji	Data latih	eigen vektor (W optimal)	Rata-rata akurasi pengenalan (%)
U2DPCA(baris)	9	5	5	62,70042857
	9	5	7	57,30414286
	9	5	20	57,14368571
	9	5	30	58,5737
	9	5	40	70,63714286
U2DPCA(kolom)	9	5	5	76,985
	9	5	7	75,39785714
	9	5	20	81,0
	9	5	30	80,15928571
	9	5	40	77,77785714
	9	5	50	84,12785714
B2DPCA	9	5	5	93,65142857
	9	5	7	95,23857143
	9	5	20	93,65142857
	9	5	30	92,06428571
	9	5	40	92,06428571



Gambar 3.4 Grafik akurasi pengenalan huruf arab

Pada percobaan ini pengenalan huruf arab menggunakan U2DPCA baris telah diperoleh persentase rata-rata kebenaran paling tinggi yaitu sebesar 70,6 % dengan menggunakan 40 eigen vektor. U2DPCA kolom telah diperoleh persentase rata-rata kebenaran paling tinggi yaitu sebesar 84,1 % dengan menggunakan 50 eigen vektor.B2DPCA telah diperoleh persentase rata-rata kebenaran paling tinggi yaitu sebesar 95,2 % dengan menggunakan 7 eigen vektor.

4. Simpulan

Pada penelitian ini telah berhasil dilakukan pengenalan huruf arab dengan menggunakan metode reduksi 2DPCA dan metode klasifikasi *k*-NN. Berdasarkan 3 pendekatan 2DPCA yaitu U2DPCA baris, U2DPCA kolom dan B2DPCA, dengan digunakan 9 tipe data latih dan 5 tipe data uji. Didapatkan hasil rata-rata akurasi pengenalan tertinggi diperoleh pada B2DPCA sebesar 95,2 % pada pengambilan 7 eigen vektor, sedangkan U2DPCA baris sebesar 70,6 % pada pengambilan 40 eigen vektor, dan U2DPCA kolom sebesar 84,1 % pada pengambilan 50 eigen vektor. Sehingga, dapat dinyatakan bahwa B2DPCA adalah metode reduksi yang paling baik untuk pengenalan Huruf Arab pada penelitian ini.

Daftar Pustaka

[1] Anton, H. 1987. *Aljabar Linear Elementer*. Penerbit Erlangga. Jakarta.

[2] Dhiraj, K. 2012. *Comparative Analysis of PCA and 2DPCA in Face Recognition*. Jurnal, Nepal Engineering College. ISSN: 2250-2459, Volume 2, Issue 1, January 2012.

[3] Fatchul Huda, A.2008. *Pengenalan wajah 3D pada gambar berkualitas rendah menggunakan Kernel Discriminant Analysis*. Tesis. Universitas Indonesia. Depok.

[4] Kong Hui, dkk. 2005.*Generalized 2D principal component analysis for face image representation and recognition*. Jurnal, Nanyang technological University, Singapore. ISSN: 585-594.

[5] Nuraiman, D. 2011. *Principal component analysis (PCA) untuk pengenalan huruf arab tulisan tangan menggunakan metode klasifikasi backpropagation*. Skripsi, UIN SGD. Bandung.

[6] Wang dong, dkk. 2011.*Two dimensional principal component of natural images and its application*. Jurnal, Dalian University of technologi , China. ISSN: 2745-2753.