**Laporan Tugas Penamabangan Data untuk Klasifikasi Gambar menggunakan K Nearest Neighbor**



Febrian Nathan 2016730020

Muhammad Ravi 2016730041

Gerry Fernando Lesmana 2016730050

## Latar belakang

Data adalah kumpulan-kumpulan informasi yang dapat berwujud berbagai macam, seperti gambar, text,suara, dll. Dan pada era sekarang ini karena data mudah didapatkan, tetapi data tersebut belum tentu bermakna. Untuk mendapatkan data yang lebih bermakna maka data harus diolah terlebih dahulu dan proses tersebut disebut Data mining.

Data mining adalah proses menemukan pola dalam kumpulan data besar yang melibatkan metode di persimpangan pembelajaran mesin, statistik, dan sistem basis data. Tujuan dari data mining adalah ekstraksi pola dan pengetahuan dari sejumlah besar data.

## Hal yang akan dibahas

Dalam laporan ini, akan dibahas teknik yang digunakan, masalah yang ditemukan, data yang diolah, dan masalah yang belum terselesaikan.

## Deskripsi Topik

Dalam tugas ini kami membuat sebuah aplikasi python yang akan memprediksi jenis gambar berdasarkan warna dari gambar. Sebelum aplikasi memprediksi, aplikasi akan melakukan training terlebih dahulu dan proses labelling akan dilakukan secara manual terlebih dahulu oleh manusia.

Aplikasi menggunakan library opencv2 dan beberapa library untuk clustering dan classification. Untuk clustering kami menggunakan k-means sedangkan classification menggunakan k-nearest neighbor.

## Data yang akan diolah

Data terdiri dari 4 kategori yaitu pantai,gedung,salju, dan hutan. Data akan dibagi menjadi dua yaitu untuk prediksi data untuk train ada 44-47 gambar untuk masing-masing kategori sedangkan gambar untuk test ada 5 untuk tiap-tiap kategori adapun gambar untuk test bisa dilihat pada tabel berikut.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gambar** | **1** | **2** | **3** | **4** | **5** |
| Pantai | D:\kuliah\Semester8\datmin\image-classification\test\test1.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test2.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test3.jpg |  | D:\kuliah\Semester8\datmin\image-classification\test\test5.jpg |
| Gedung | D:\kuliah\Semester8\datmin\image-classification\test\test6.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test7.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test8.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test9.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test10.jpg |
| Salju | D:\kuliah\Semester8\datmin\image-classification\test\test11.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test12.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test13.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test14.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test15.jpg |
| Hutan | D:\kuliah\Semester8\datmin\image-classification\test\test16.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test17.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test18.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test19.jpg | D:\kuliah\Semester8\datmin\image-classification\test\test20.jpg |

## Metode experimen

Eksperimen terdiri dari dua tahap yaitu; Pemilihan fitur dan klasifikasi.

### 1. Pemilihan Fitur

Fitur yang akan diambil adalah potongan-potongan area pada gambar. Pemilihan fitur terdiri dari beberapa langkah yaitu; mengkonversi warna, menghitung histogram untuk tiap warna, melakukan clustering untuk menemukan warna dominan.

1. **Konversi Warna pada Gambar**

Secara *default,* opencv mengkonversi warna pada gambar dengan format BGR(Blue, Green, Yellow). Jarak antara pixel pada BGR kurang akurat sehingga apa yang dianggap warna berdekatan oleh manusia belum tentu dianggap sama oleh komputer karena itu warna dikonversi menjadi format yang lebih uniform secara perceptual, CIEl\*a\*b. Berikut hasil gambar setelah dikonversi:

|  |  |
| --- | --- |
| Gambar 1 Image BGR | Gambar 2 Image CIEL\*a\*b |

1. **Menghitung histogram**

Histogram dihitung pada keseluruhan gambar tetapi pada bagian-bagian gambar. Gambar akan dibagi menjadi 16 bagian dan masing-masing akan kita tentukan histogramnya.

|  |
| --- |
| Gambar 3 contoh bagian histogram 1 |
| Gambar 4 contoh bagian histogram 2 |

Seperti kedua gambar di atas untuk menghitung histogram pada bagian-bagian gambar perlu dilakukan proses masking yaitu menutup bagian gambar yang tidak diperlukan. Proses ini akan terus diiterasi sebanyak 16 kali.

Frekuensi kemunculan pixel pada bagian tertentu akan diubah ke dalam bentuk bins jadi pada pixel pertama warna L ada berapa, warna a, ada berapa, dan warna b ada berapa. Berikut adalah contoh dari bins yang telah dinormalisasi dari bins 1 sampai 256:

|  |  |  |  |
| --- | --- | --- | --- |
| Bin | L | a | b |
| 1 | 0.06468699872493744 | 0.0 | 0.0 |
| .. | ... | ... | .. |
| 256 | 0.10609951615333557 | 0.0 | 0.0 |

Bins akan dinormalisasi sebelum dilakukan clustering dengan cara membagi nilai tiap bin dengan panjang bin. Setelah itu akan dilakukan clustering untuk mengelompokkan bins menjadi beberapa bagian.

**C. Clustering**

Clustering dilakukan dengan K Means dan jarak ditentukan dengan jarak euclidean. Kmeans akan mengelompokan bins menjadi 20 cluster. Cluster dengan anggota terbanyak akan diambil centroidnya untuk dijadikan fitur. Fitur akhir akan berupa centroid dari warna dominan pada tiap bagian gambar.

### 2. Klasifikasi

Klasifikasi dilakukan dengan K nearest Neighbor (KNN). Dilakukan training terhadap model dengan 47 gambar hutan, 44 gambar salju, 44 gambar gedung, dan 50 gambar pantai. Pada klasifikasi juga dilakukan iterasi untuk menemukan jumlah tetangga (k) yang optimum dari (1-16). Gambar yang dites ada sebanyak 20 buah. Berikut hasil perhitungan akurasi KNN:



Dari 16 kali percobaan jumlah tetangga terdekat (k) yang memberi nilai akurasi tertinggi dan konsisten terdapat pada range k=7 sampai k=11. Maka range tersebut akan dipilih untuk dianalisis hasilnya dan akan dijadikan threshold.

* Akurasi k=7 adalah 0.65 atau 65% artinya dari 20 gambar ada 13 gambar yang ditebak secara benar.
* Akurasi k=8 adalah 0.7 atau 70% artinya dari 20 gambar ada 14 gambar yang ditebak secara benar
* Akurasi k=9 adalah 0.65 atau 65% artinya dari 20 gambar ada 13 gambar yang ditebak secara benar
* Akurasi k=10 adalah 0.75 atau 75% artinya dari 20 gambar ada 15 gambar yang ditebak secara benar
* Akurasi k=11 adalah 0.7 atau 70% artinya dari 20 gambar ada 14 gambar yang ditebak secara benar

Dengan demikian, rata-rata tingkat akurasi klasifikasi dari kNN adalah 65-75%

## Hasil



Dari 20 *Test case* ada 6 kesalahan yang konsisten dalam klasifikasi yaitu pantai dianggap sebagai hutan, gedung dianggap sebagai hutan, hutan dianggap sebagai pantai, dan hutan dianggap sebagai gedung. Berikut adalah gambar-gambar yang di klasifikasikan secara salah:

|  |  |  |
| --- | --- | --- |
| **No** | **Gambar** | **Deskripsi** |
| 1 | D:\kuliah\Semester8\datmin\image-classification\test\test1.jpg | Pantai ini dianggap hutan karena tone dan komposisi warna senada dengan hutan ditambah banyak pepohonan, pulau, dan warna air laut yang kehijauan. |
| 9 | D:\kuliah\Semester8\datmin\image-classification\test\test9.jpg | Gedung dianggap hutan |
| 14 | D:\kuliah\Semester8\datmin\image-classification\test\test14.jpg | Salju dianggap pantai karena tekstur salju sama dengan pantai dan bebatuan dianggap sebagai semak atau pohon-pohon sekitar pantai |
| 18 | D:\kuliah\Semester8\datmin\image-classification\test\test18.jpg |  |
| 20 | D:\kuliah\Semester8\datmin\image-classification\test\test20.jpg | Hutan dianggap pantai karena air. Namun bisa dideteksi oleh KNN dengan k=10 |

## Masalah yang belum terpecahkan atau bisa ditambahkan

# Klasifikasi masih dapat tertipu ketika gambarnya mirip dengan dataset yang lain, Hal ini mungkin dapat diperbaiki dengan metode pairing.

**VII. Lampiran Kode**