DENSE VISUAL WORD SPATIAL ARRANGEMENT DAN PENERAPANNYA PADA PENGENALAN GAMBAR SECARA OTOMATIS

Gama Wisnu Fajarianto[1)], dan Handayani Tjandrasa [2)]
[1, 2)] Jurusan Teknik Informatika ITS Surabaya
e-mail: gama.fajarianto12@mhs.if.its.ac.id[1)], handatj@its.ac.id[2)]

ABSTRAK

Bag of visual word (BoVW) merupakan metode yang menjelaskan isi dari gambar. Metode ini hanya menghitung banyaknya word dan tidak memberikan informasi spatial. Terdapat metode Visual word spatial arrangement (WSA) dimana metode ini memberikan informasi spatial tentang word tertentu pada gambar dengan menggunakan interest point sebagai detektor.

WSA kurang dapat memberikan informasi yang penting pada gambar dikarenakan interest point yang dihasilkan oleh detektor dapat memberikan titik-titik yang berpotensi tidak merupakan representasi yang penting dari gambar tersebut. Pada paper ini diusulkan metode dense visual word spatial arrangement (DVSA) yang merupakan modifikasi metode dari WSA. Metode ini tidak menggunakan detektor interest point untuk menghitung deskriptor lokal melainkan dengan menghitung deskriptor lokal pada bagian komponen piksel-piksel yang saling berdekatan.

Hasil pengujian pada 4485 gambar dengan 15 jenis kelas menggunakan 10-fold cross validation untuk 2 word metode yang diusulkan memberikan peningkatan performa sebesar 12.68 % dari akurasi BoVW sedangkan akurasi WSA lebih baik 15.62 % dari BoVW. Untuk 4 word metode yang diusulkan memberikan peningkatan performa akurasi sebesar 30.99 % dari akurasi BoVW dan peningkatan performa 18.16 % dari WSA. Sedangkan untuk 6 word metode yang diusulkan memberikan peningkatan performa sebesar 29.98 % dari akurasi BoVW dan peningkatan performa 18.75 % dari WSA. Peningkatan performa akurasi sebesar 36.2 % didapatkan oleh metode yang diusulkan dengan 6 word terhadap BoVW dengan 2 word. Peningkatan performa sampai 18.75 % yang dihasilkan DVSA dibandingkan WSA dan peningkatan performa sampai 30.99 % dibandingkan BoVW dengan jumlah word yang sama menunjukkan metode yang diusulkan kompetitif untuk mengenali jenis gambar.

***Kata Kunci**: deskriptor lokal, bag of visual word, klasifikasi, ekstraksi fitur*

ABSTRACT

Bag of visual word (BoVW) is a method that describes the contents of an image. This method simply counts the number of words, but it doesn't provide spatial information. Besides there is a method that provides spatial information about particular words in the image by using an interest point as a detector. The method is Visual word spatial arrangement (WSA).

WSA can provide less important information on the image generated due to the interest point doesn't represent the main aspects of the image. In this paper, Dense visual word spatial arrangement (DVSA) method which is proposed is a modification of the WSA method. The proposed method doesn't use an interest point detector to compute local descriptor but it uses a local descriptor that computes at the component pixels adjacent to each other.

The test result on 4485 images with 15 types of classes is computed using 10 fold cross validation for 2 words of the proposed method that provides an improved performance by 12.68% of accuracy BoVW, while WSA has better accuracy by 15.62% from BoVW. For 4 words, the proposed method provides an improved performance by 30.99% from the accuracy of BoVW, and an improved performance by 18.16% from WSA. While for 6 words, the proposed method provides an improved performance by 29.98% from the accuracy of BoVW, and an improved performance by 18.75% from WSA. The improved performance of the accuracy by 36.20% is obtained by the proposed method with 6 words than BoVW with 2 words. From the result can be concluded that the proposed method or DVSA method is more competitive to recognize images.

Keywords: local descriptor,bag of visual word, classification, fitur extraction

I. PENDAHULUAN

INFORMASI visual telah menjadi informasi yang banyak ditemui dalam kehidupan sehari-hari dan tidak kalah pentingnya dengan informasi tekstual. Semisal pada saat mengendarai sepeda motor pengendara melihat informasi visual lampu merah yang artinya pengendara sepeda motor harus berhenti. Dengan mengerti isi dari gambar akan membantu mengetahui informasi visual pada gambar tersebut. Salah satu pendekatan populer dalam

menjelaskan isi dari gambar adalah dengan pendekatan BoVW [4][5][7][9][13]. Selain populer pendekatan ini juga efektif dalam menjelaskan isi dari gambar [3][8][9].

Pendekatan BoVW dapat dijelaskan secara umum dengan 3 langkah yaitu pertama *local image descriptor* diekstraksi dari gambar, kemudian *visual dictionary* didapat dari sekumpulan fitur vektor *local image descriptor* yang bisa diperoleh dengan menggunakan *clustering k-means*, langkah kedua adalah fitur *encoding* yaitu mengaktifkan *visual word* dengan memetakan fitur deskriptor ke *visual dictionary*, dan langkah ketiga yaitu *pooling* yang menjadikan hasil dari *encoding* fitur deskriptor dengan *visual dictionary* menjadi satu fitur vektor [1][5].

Aplikasi yang menggunakan BoVW dapat ditemui pada aplikasi pembeda tulisan tangan dengan tulisan teks mesin print pada suatu dokumen [12]. Kemudian aplikasi *scene categorization*, contohnya *keyword suggestion* yaitu menawarkan beberapa label yang berhubungan dengan isi gambar, dan *retrieval* yaitu melakukan penyaringan gambar pada internet berdasarkan *keyword* [6]. Aplikasi lainnya adalah pada bidang rekayasa biomedis atau *biomedical engineering* yaitu otomatisasi analisa dari *time series* sinyal biomedis *electroencephalogram* (EEG) dan *electrocardiographic* (ECG) dimana BoVW digunakan untuk merepresentasikan *biomedical time series* [11].

Permasalahan pada representasi gambar dengan BoVW yaitu tidak memberikan informasi geometris dari gambar [3][8][9][13]. Dengan tidak adanya informasi geometris jenis gambar yang berbeda dapat memiliki histogram BoVW yang mirip. Oleh karena itu pada [8] mengajukan WSA dimana ia memberikan informasi geometris *visual word* dari gambar. Pada gambar jenis tertentu misalnya, WSA akan memberikan informasi *visual word* tertentu memiliki kecenderungan berada pada bagian tertentu, semisal bagian kanan atas.

Pada WSA [8] *local image descriptor* diekstraksi pada *interest point*. Terdapat permasalahan yang muncul apabila menggunakan *interest point*. Permasalahan tersebut adalah *interest point* memiliki potensi menghasilkan *local image descriptor* yang kurang handal [10]. Pada hasil uji coba WSA, WSA kurang dapat membedakan gambar jenis MITopencountry dengan gambar jenis MITcoast. Secara umum, visual *interest point* yang dihasilkannya berkumpul pada objek-objek dan bukan kepada langit cerah [8]. Kurangnya *local image descriptor* pada bagian lain dapat membuat performa yang kurang baik apabila bagian lain tersebut merupakan bagian yang penting dalam mendeskripsikan gambar. Pada kedua kelas ini yaitu gambar jenis MITopencountry dan gambar jenis MITcoast sering dikenali kelasnya tertukar. MITopencountry dikenali dengan MITcoast dan MITcoast dikenali dengan MITopencountry. Selain itu WSA juga kurang dapat membedakan dengan benar gambar MITtallbuilding dengan gambar industrial dimana banyak gambar struktur bangunan tinggi [8]. Dengan kata lain pada kedua kelas tersebut juga sering dikenali kelasnya tertukar.

Pada tulisan ini diajukan metode modifikasi dari WSA yaitu modifikasi pada bagian ekstraksi fitur deskriptor lokal. Metode yang diajukan pada tulisan ini memodifikasi WSA dengan tidak menggunakan *interest point* karena ekstraksi fitur deskriptor lokal dengan *interest point* berpotensi menghasilkan sekumpulan titik yang tidak dapat diandalkan. *Interest point* dapat tidak mengikut sertakan daerah yang penting untuk mengenali citra tersebut. Oleh karena itu pada tulisan ini ekstraksi fitur deskriptor lokal WSA dengan *interest point* diganti dengan ekstraksi fitur deskriptor lokal yang dihitung pada bagian komponen yang saling berdekatan atau *densely* pada keseluruhan bagian gambar.

II. DENSE VISUAL WORD SPATIAL ARRANGEMENT

Pada tulisan ini diusulkan metode DVSA untuk klasifikasi citra. Metode ini menghitung fitur lokal pada citra tidak dengan menggunakan *interest point* melainkan fitur dihitung pada semua daerah secara berurutan pada citra dengan jarak kerapatan piksel tertentu. Proses selanjutnya dilakukan WSA dengan menggunakan fitur lokal yang telah dihitung sebelumnya.

Ide dari metode DVSA adalah menghitung fitur lokal dengan jarak piksel tertentu yang dilakukan pada keseluruan piksel dari citra. Terdapat dua parameter yaitu *step* dan *size*. *Step* digunakan untuk setiap berapa *step* piksel pusat perhitungan deskriptor ini berada, sedangkan *size* digunakan sebagai ukuran radius perhitungan deskriptor. Langkah pertama yang dilakukan metode ini adalah melakukan perhitungan fitur lokal pada setiap titik dengan jarak dan ukuran tertentu. Kemudian dilakukan proses *coding* dan *pooling*.

$$H = \begin{array}{c} \\ c_1 \\ \\ c_m \\ \\ c_M \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & & x_j & & x_N \end{array} \\ \begin{bmatrix} \alpha_{1,1} & \cdots & \alpha_{1,j} & \cdots & \alpha_{1,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m,1} & \cdots & \alpha_{m,j} & \cdots & \alpha_{m,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{M,1} & \cdots & \alpha_{M,j} & \cdots & \alpha_{M,N} \end{bmatrix} \end{array} \Rightarrow g: pooling$$

$$\Downarrow$$

$$f: coding$$

Gambar 1 Matrix coding dan pooling

Matrix untuk proses *coding* dan *pooling* ditunjukkan pada Gambar 1. Proses *coding* ini dilakukan dengan mengisi nilai α. Pengisian nilai α menggunakan persamaan (1). Pada proses *coding* bagian perhitungan jarak dapat dilakukan dengan perhitungan jarak semisal *Euclidean distance, Manhattan distance* atau perhitungan jarak lainnya. Sedangkan pada proses *pooling* dimana menjadikan hasil *coding* menjadi satu fitur vektor. Dilakukan dengan menghitung nilai α pada setiap c. Proses *pooling* dapat dihitung dengan menjumlahkan, melakukan rata-rata, mengambil nilai minimum atau nilai maksimum dari nilai α. Pada tulisan ini proses *coding* menggunakan jarak dengan *Euclidean distance* sedangkan untuk proses *pooling* nya yang digunakan pada BoVW menggunakan jumlahan untuk membentuk satu fitur vektor.

$$\alpha_{m,j} = 1 iff m = \arg\min_{k \in \{1,\ldots M\}} \| x_j - c_k \|_2^2 . \quad (1)$$

Dengan c_1, c_m, c_M adalah *word* ke 1,*word* ke m, dan *word* ke M. x_1, x_j, dan x_N adalah *instance* ke 1, *instance* ke j dan *instance* ke N.

Pada setiap deskriptor lokal citra dihitung berapa banyak *word* sesuai deskriptor lokal tersebut pada setiap kuadrant. Kemudian setelah selesai dihitung untuk semua lokal deskriptor, dilakukan penggabungan untuk menghasilkan fitur vektor. Fitur *vector* akhir yang dihasilkan sebanyak 4 dikali W dengan W adalah jumlah *word* yang digunakan. Penjelasan mengenai pembuatan *dictionary* dan masing-masing metode yaitu BoVW, WSA dan DVSA dijelaskan pada point A dan point B pada bagian selanjutnya dari tulisan ini.

A. Membuat Dictionary

Dictionary dibuat dari pengambilan sampel secara acak pada 4485 gambar *training*. Langkah pertama diambil sampel dengan jumlah masing-masing 30 gambar untuk setiap kelas. Gambar yang diambil terebih dahulu di *resize* dengan ukuran ½ kali citra asli. Kemudian pada setiap gambar dilakukan proses detektor untuk mendapatkan *interest point*. Langkah detektor ini dilakukan pada BoVW dan WSA. Detektor yang digunakan adalah SIFT. Setelah didapat *interest point* maka pada *interest point* tersebut dilakukan perhitungan deskriptor. Perhitungan deskriptor pada *interest point* ini dilakukan untuk BoVW dan WSA. Sedangkan untuk DVSA tidak dilakukan deteksi *interest point* melainkan dengan cara mendefinisikan per piksel jarak dan radius tertentu dari piksel tersebut. Kemudian dari nilai per piksel jarak dan radius tersebut langsung dihitung deskriptornya. Deskriptor yang digunakan pada BoVW, WSA, dan DVSA ini adalah SIFT.

Setelah didapat semua deskriptor dari gambar sampel langkah selanjutnya adalah pengambilan secara acak deskriptor lokal. Jumlah yang diambil berdasarkan pada berapa *word* yang akan digunakan. Apabila menggunakan 2 *word* maka yang diambil adalah 2 deskriptor lokal secara acak. Deskriptor lokal yang diambil untuk *dictionary* ini jumlahnya sama dengan *word* yang akan digunakan. Pada tulisan ini menggunakan 2, 4 dan 6 *word*. Jadi diambil 2, 4 dan 6 deskriptor lokal secara acak.

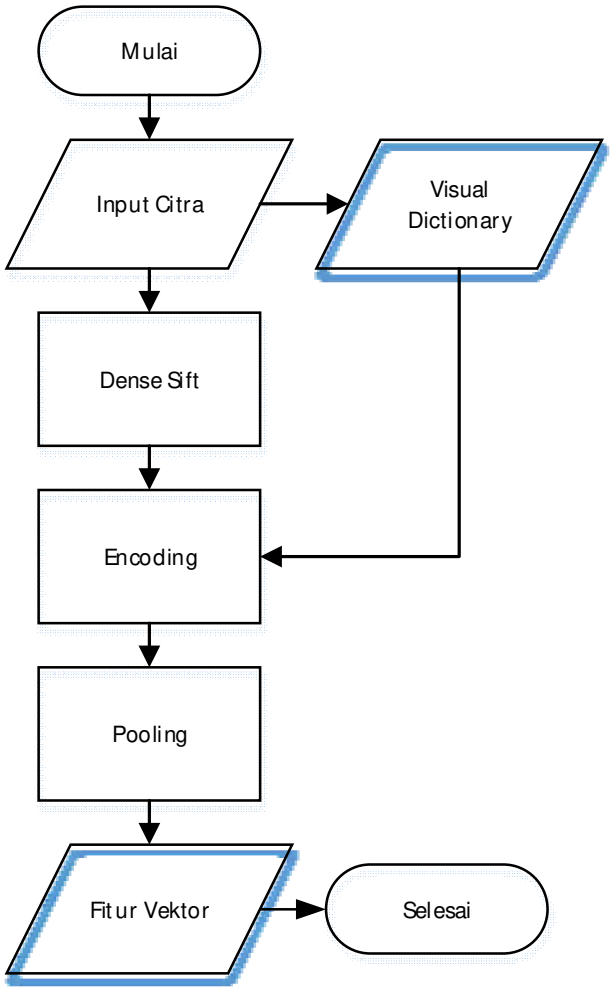
B. BoVW, WSA dan DVSA

Setelah didapat *dictionary* langkah selanjutnya adalah dihitung kembali detektor pada masing-masing gambar dari 4485 gambar data *training*. Gambar yang diambil terlebih dahulu di *resize* dengan ukuran ½ kali citra asli. Pada BoVW dan WSA dilakukan perhitungan detektor yang dilanjutkan dengan perhitungan deskriptor. Detektor dan deskriptor yang digunakan adalah SIFT. Dengan parameter *default* SIFT [2].

Berbeda dengan metode yang diajukan yaitu DVSA. DVSA tidak menghitung *interest point* dengan detektor melainkan dengan cara per piksel jarak dan radius tertentu dari piksel tersebut langsung dihitung deskriptornya.

Dengan nilai radius *size* yang digunakan adalah 10 dan jarak antar piksel *step* yang digunakan adalah 20. Setelah didapat deskriptor lokal pada setiap gambar maka masing-masing deskriptor lokal tersebut dihitung jaraknya dengan *dictionary*. Jarak yang digunakan adalah *Euclidean distance*. Untuk BoVW maka dihitung dengan *dictionary* BoVW, untuk WSA maka dihitung dengan *dictionary* WSA, dan begitu juga dengan DVSA dihitung dengan *dictionary* DVSA yang dihasilkan dari proses sebelumnya. Hasil perhitungan deskriptor lokal dengan masing-masing *word* pada *dictionary* tersebut diambil jarak yang terdekat. Apabila deskriptor paling dekat dengan *word* yang pertama maka deskriptor lokal tersebut termasuk *word* pertama. Begitu seterusnya. Hasilnya adalah didapat *word* untuk masing-masing citra.

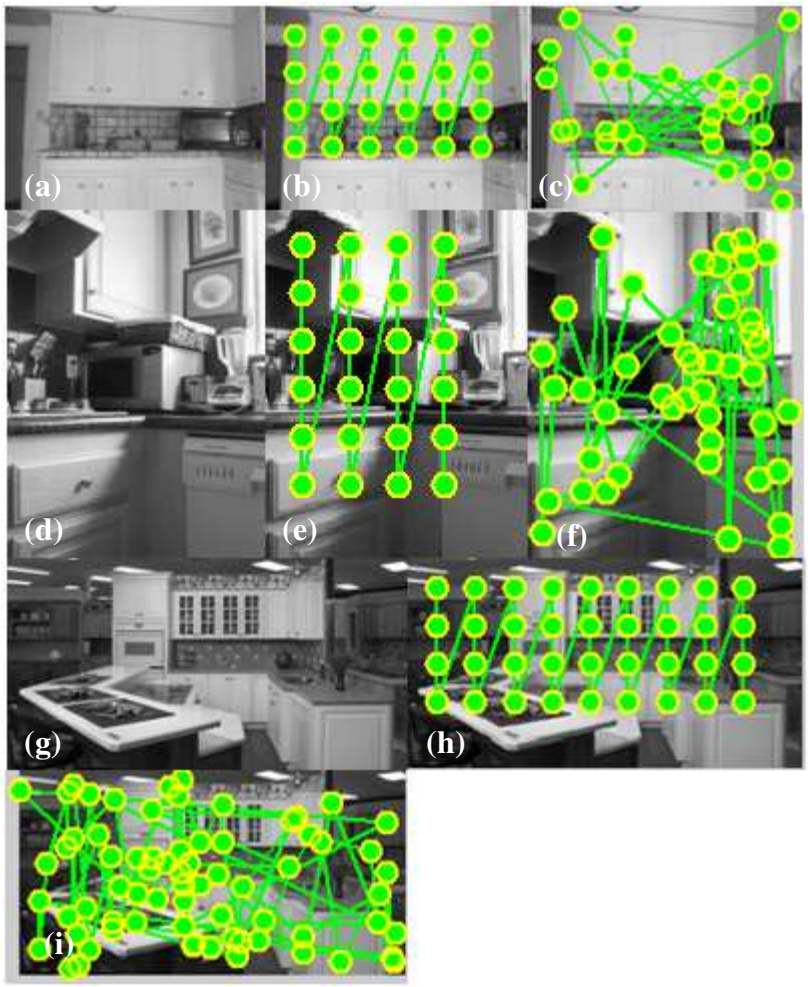
Pada Gambar 2 menunjukkan diagram tahapan untuk metode yang diajukan. Ditunjukkan pada Gambar 2, pertama dimulai dengan input citra kemudian pada input citra tersebut dilakukan proses *dense SIFT* yaitu perhitungan *SIFT* pada *point* yang merepresentasikan keseluruhan bagian permukaan citra. *Visual dictionary* juga dihasilkan dari input citra. Hasil proses *dense SIFT* lalu diproses kembali dengan proses *encoding*. Pada proses *encoding* ini membutuhkan inputan *visual dictionary*. Proses selanjutnya adalah *pooling* dimana proses ini menjadikan hasil *encoding* menjadi satu fitur vektor.



Gambar 2 Diagram alir tahapan metode yang diajukan

Pada BoVW untuk setiap gambar dihitung jumlah word dari deskriptor lokal. Berapa *word* dari deskriptor lokal yang termasuk *word* pertama, berapa *word* dari deskriptor lokal yang termasuk *word* kedua dan seterusnya. Jumlah *word* ini merupakan fitur vektor.

Dapat dilihat pada Gambar 3 adalah visualisasi hasil interest point dari masing-masing metode. Untuk huruf a, d dan g pada Gambar 3 tersebut adalah citra asli. Citra dengan huruf b, e, dan h adalah citra visualiasi *point* menggunakan metode yang diusulkan. Sedangkan citra dengan huruf c, f dan I adalah citra visualisasi *interest point* dari metode BoVW dan WSA. *Point-point* atau *interest point* tersebut digambarkan dengan bulatan. Visualisasi garis yang menghubungkan antar bulatan menunjukkan urutan *deskriptor lokal* pada citra tersebut.

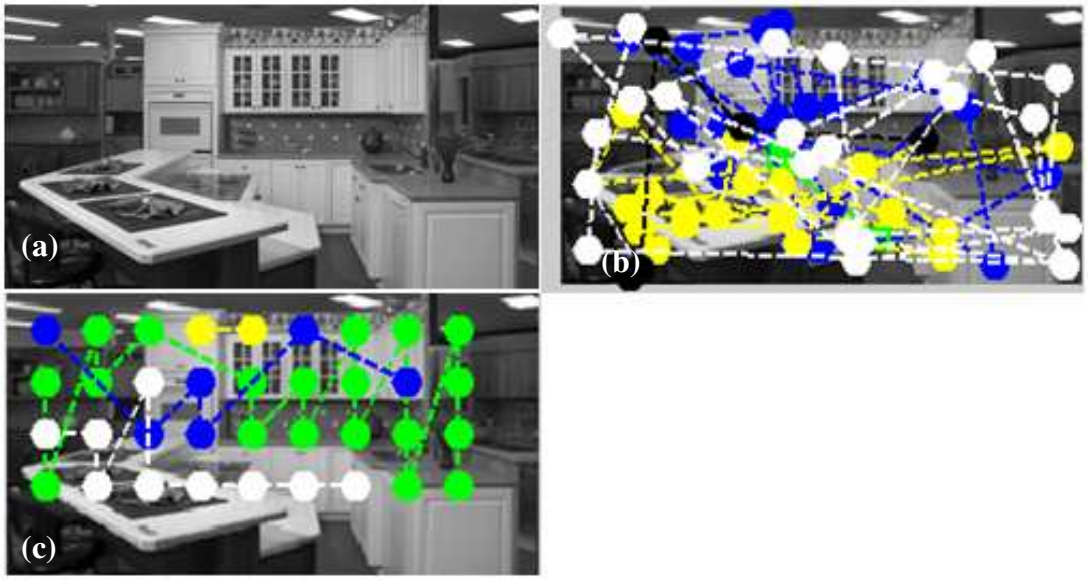


Gambar 3 Perbandingan visual citra hasil interest point dari BoVW dan WSA ditunjukkan pada bulatan warna hijau (c),(f) dan (g). Dengan DVSA yang ditunjukkan pada bulatan warna hijau (b),(e) dan (h). Citra (a),(d) dan (g) adalah citra asli. Garis yang menghubungkan antar bulatan menunjukkan urutan interest point.

Dapat dilihat pada Gambar 4 merupakan contoh visualisasi *word* dari masing-masing metode. Untuk citra a merupakan citra asli. Citra b merupakan visualisasi *word* dari BoVW dan WSA sedangkan citra c merupakan visualisasi word dari metode yang diusulkan. Warna bulatan yang berbeda menunjukkan pada *point* tersebut memiliki *word* yang berbeda. Garis putus-putus pada setiap bulatan menunjukkan urutan *point* pada setiap word yang sama.

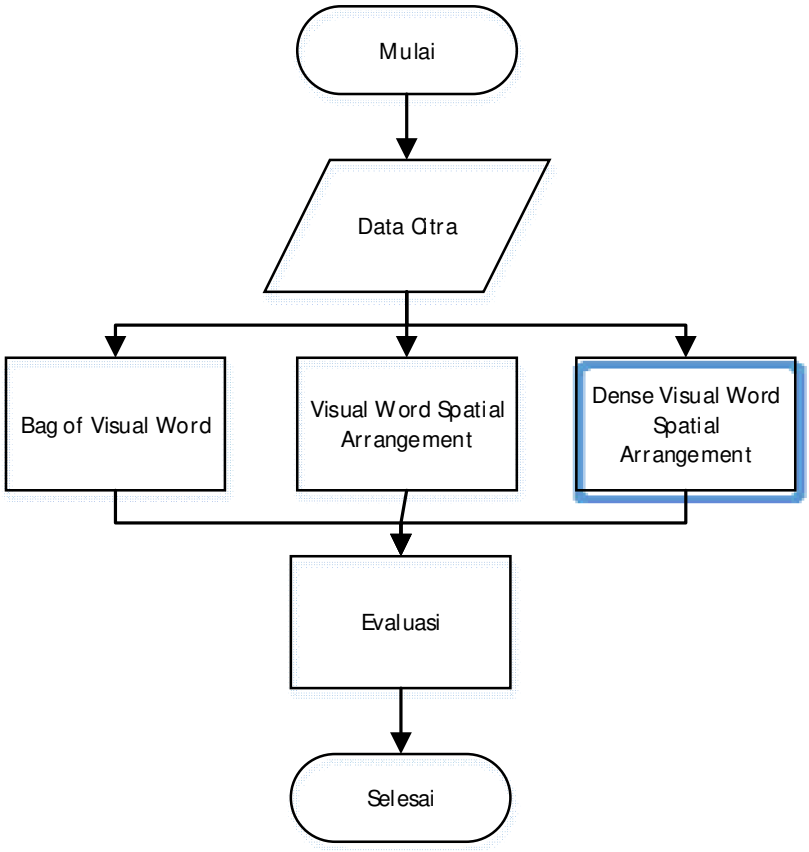
Berbeda dengan BoVW yang menjumlahkan *word* dari deskriptor lokalnya. Pada WSA dan DVSA selain dihitung deskriptor lokal tersebut termasuk pada word berapa, juga dihitung posisi deskriptor lokal tersebut. Posisi deskriptor lokal ini yang digunakan untuk proses pembentukan fitur vektor. Setelah didapat deskriptor lokal pada citra termasuk *word* berapa, maka untuk masing-masing deskriptor lokal ditarik garis horizontal dan vertical. Dari penarikan garis ini didapat 4 kuadran pada keseluruhan piksel dari citra yaitu daerah kiri atas kuadran 1, kanan atas kuadran 2, kiri bawah kuadran 3 dan kanan bawah kuadran 4.

Pada masing-masing kuadran dilakukan pengecekan apakah ada deskriptor lokal yang termasuk *word* pertama jika ada maka dihitung berapa jumlahnya. Begitu juga untuk *word* kedua dan seterusnya. Setiap berganti deskriptor lokal dari pembentukan kuadrant, hasil jumlahan sebelumnya ditambahkan ke hasil selanjutnya. Hal ini dilakukan sejumlah deskriptor lokal yang ada pada gambar. Jadi WSA dan DVSA menghasilkan fitur vektor 4 x jumlah *word*. Hasil penjumlahan pada proses pembentukan kuadran untuk masing-masing deskriptor sampai yang terakhir adalah merupakan fitur vektor yang digunakan untuk proses klasifikasi.



Gambar 4 Perbandingan visual hasil word dari BoVW dan WSA yang ditunjukkan dengan bulatan warna (b) dengan hasil word dari DVSA (c). Bulatan dengan warna yang sama merupakan word dengan jenis yang sama dengan kata lain bulatan dengan warna yang berbeda merupakan word dengan jenis yang berbeda Garis putus – putus antar bulatan menunjukkan urutan pada setiap jenis word yang sama.

Gambar 5 merupakan tahapan diagram tahapan evaluasi. Dimulai dari inputan berupa data citra kemudian di proses dengan masing-masing metode yaitu *bag of visual word*, *visual word spatial arrangement* dan *dense visual word spatial arrangement*. Hasil masing-masing metode kemudian dievaluasi.

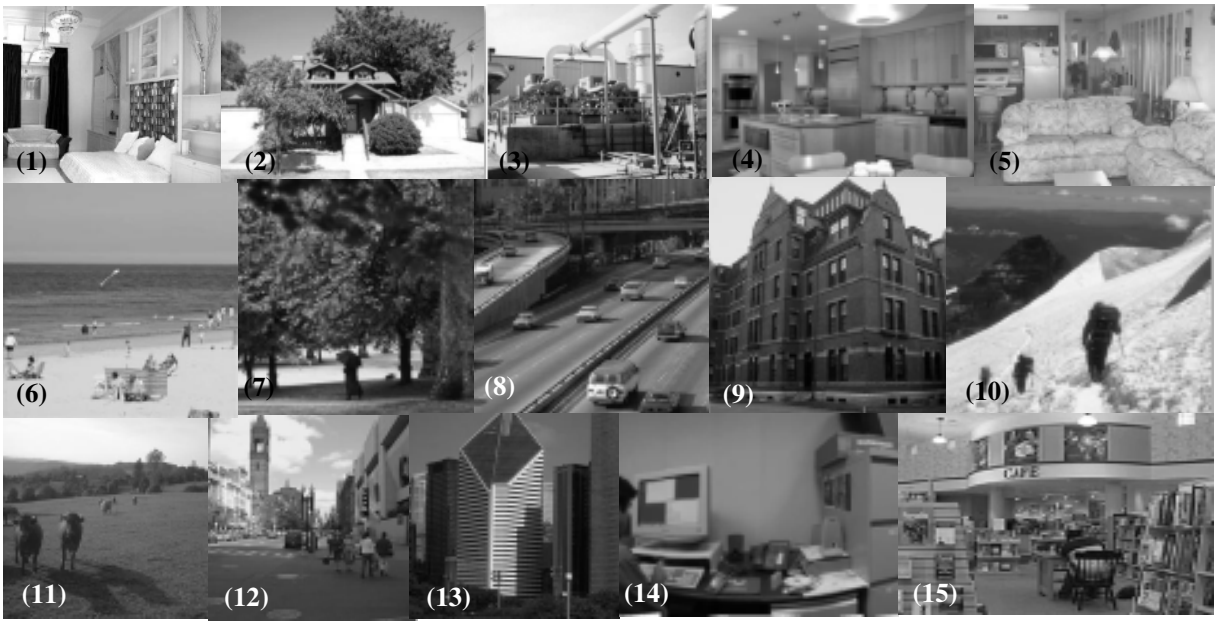


Gambar 5 Diagram tahapan evaluasi

III. HASIL UJI COBA

Dataset yang digunakan adalah dataset 15-*scenes* terdiri dari 4485 citra, yang dibagi menjadi 15 kategori. Contoh citra dataset dapat dilihat pada Gambar 6. Secara berurutan dari citra nomer 1 sampai citra nomer 15 adalah kelas

dengan kategori *bedroom*, *CALsuburb*, *industrial*, *kitchen*, *livingroom*, *MITcoast*, *MITforest*, *MIThighway*, *MITinsidecity*, *MITmountain*, *MITopencountry*, *MITstreet*, *MITtallbuilding*, *PARoffice* dan *store*. Jumlah citra untuk masing-masing kategori secara berurutan dari kelas *bedroom*, *CALsuburb*, *industrial* sampai kelas *store* adalah 216, 241, 311, 210, 289, 360, 328, 260, 308, 374, 410, 292, 356, 215 dan 315. Total terdapat 4485 citra. Dataset 15-*scenes* ini juga memiliki ukuran panjang dan lebar piksel yang beragam. Ada yang 200 x 267 piksel, 220 x 411 piksel, 256 x 256 piksel sampai ukuran 552 x 220 piksel. Ragam ukuran dataset yang digunakan ini memiliki ukuran terkecil dimensi pertama yaitu 200 piksel dan ukuran terbesarnya 552 piksel. Sedangkan untuk dimensi kedua memiliki ukuran terkecil 220 piksel dan ukuran terbesar 411 piksel.



Gambar 6 Contoh dataset 15-scenes

Evaluasi pengujian dilakukan menggunakan software bantu WEKA dengan 10-fold cross validation dan klasifikasi random forest. Hasil evaluasi pada 4485 gambar dengan 15 jenis kelas dapat dilihat pada Tabel 1

TABEL 1
HASIL PENGUJIAN PADA METODE YANG DIUSULKAN

Jumlah Word	Metode BoVW			
	Akurasi	Presisi	Recall	F-Measure
2	30.82 %	22.10 %	30.80 %	23.20 %
4	36.89 %	35.20 %	36.90 %	35.60 %
6	37.03 %	34.00 %	37.00 %	34.90 %
Jumlah Word	Metode WSA			
	Akurasi	Presisi	Recall	F-Measure
2	**46.45 %**	**44.00 %**	**46.50 %**	**44.10 %**
4	49.72 %	47.80 %	49.70 %	47.20 %
6	48.27 %	46.60 %	48.30 %	45.60 %
Jumlah Word	Metode yang diusulkan (DVSA)			
	Akurasi	Presisi	Recall	F-Measure
2	43.51 %	41.10 %	43.50 %	41.40 %
4	**67.88 %**	**66.90 %**	**67.90 %**	**66.90 %**
6	**67.02 %**	**66.30 %**	**67.00 %**	**66.00 %**

Tabel 1. Hasil pengujian DVSA pada 4485 gambar dengan 10-fold cross validation menggunakan 2 word, 4 word dan 6 word. Angka yang tebal menunjukkan performa terbaik yang dihasilkan oleh metode pada setiap word.

Dapat dilihat pada Tabel 1 metode DVSA unggul dalam semua word yang diuji terhadap metode dasar BoVW baik menggunakan 2 *word,* 4 *word* maupun 6 *word*. Sedangkan WSA unggul terhadap DVSA ketika menggunakan word yang sangat sedikit yaitu 2 word.

Untuk hasil peningkatan performa dapat dilihat pada Tabel 2. Peningkatan performa akurasi diperoleh oleh DVSA terhadap BoVW meningkat sampai 30.99 %. Sedangkan peningkatan performa DVSA terhadap WSA sampai 18.75%.

TABEL 2
HASIL PENINGKATAN PERFORMA METODE UNTUK SETIAP WORD

Metode	Jumlah Word	Peningkatan Performa Terhadap BoVW			
		Akurasi	Presisi	Recall	F-Measure
WSA	2	15.62 %	21,90 %	15,70 %	20,90 %
WSA	4	12.82 %	12,60 %	12,80 %	11,60 %
WSA	6	11.23 %	12,60 %	11,30 %	10,70 %
DVSA	2	12.68 %	19,00%	12,70%	18,20 %
DVSA	4	30.99 %	31,70 %	31,00 %	31,30 %
DVSA	6	29.98 %	32,30 %	30,00 %	31,10 %
Metode	Jumlah Word	Peningkatan Performa Terhadap WSA			
		Akurasi	Presisi	Recall	F-Measure
DVSA	2	-2.94 %	-2.90 %	-3.00 %	-2.70%
DVSA	4	18.16 %	19.10 %	18.20 %	19.70 %
DVSA	6	18.75 %	19.70 %	18.70 %	20.40 %

Dari Tabel 1 dan Tabel 2 dapat dilihat metode yang diusulkan unggul pada semua uji coba terhadap metode dasar. Metode WSA juga unggul pada semua uji coba terhadap metode dasar. Unggul baik dalam hal akurasi, presisi, *recall* maupun *f-measures*. Peningkatan performa akurasi diperoleh sampai 36.20% dari metode yang diusulkan terhadap metode dasar. Dengan metode dasar 2 *word* dan metode yang diusulkan 6 *word*.

Untuk jumlah word yang sama pada metode yang diusulkan terdapat peningkatan performa akurasi sebesar 30.99% terhadap metode dasar dengan sama-sama menggunakan 4 *word*. Metode *spatial* lain sebagai pembanding atau WSA unggul terhadap semua metode pada jumlah *word* yaitu 2 *word*. Selain itu metode yang diusulkan unggul baik terhadap metode dasar maupun metode WSA. Didapat peningkatan akurasi sebesar 18.75% dari metode yang diusulkan terhadap metode WSA. Peningkatan ini ditemui pada semua evaluasi yaitu akurasi, presisi, *recall* dan *f-measures*.

Hasil akurasi dari masing-masing metode memiliki nilai akurasi tertinggi 67.88% didapat oleh metode DVSA. Nilai akurasi ini merupakan hasil dari 10 *fold cross validation* dimana dievaluasi pada 4485 gambar. Jadi secara bergantian dilakukan pengambilan data *training* dan *testing*. Dengan data yang diambil tidak diambil lagi atau *sampling without replacement*. 4485 gambar tersebut juga sangat beragam walaupun berada pada kelas yang sama. Dari sinilah diambil analisa nilai akurasi yang rata-rata dari metode. Baik metode yang diusulkan maupun metode lain sebagai pembanding.

Metode spatial lain yang menjadi pembanding atau WSA, pada semua uji coba selalu unggul saat *word* berjumlah dua. Dengan kata lain ketika hanya sedikit sekali word yang digunakan maka metode tersebut unggul terhadap semua metode untuk semua uji coba. Dapat diberikan penjelasan dikarenakan metode WSA menggunakan *interest point* yang dapat fokus sehingga titik-titiknya akan banyak yang saling berdekatan. Jadi ketika menggunakan dua *word* akan menjadi sederhana dan lebih mengelompokkan anggota mana yang menjadi *word* pertama dan mana yang menjadi *word* kedua, dengan menggunakan *interest point* ini. Berbeda ketika seperti metode yang diajukan atau DVSA dimana titik-titik akan lebih melebar luas seperti merata pada keseluruhan gambar. Jadi untuk dua *word* yang memiliki pembeda ciri yang sangat sedikit sekali ini, ketika diberikan secara merata pada titik-titiknya maka akan kurang dapat membedakan ciri dari gambar tersebut dibandingkan dengan titik-titik yang mengelompok.

IV. KESIMPULAN

Pada tulisan ini diusulkan metode DVSA yang merupakan metode modifikasi dari metode WSA. Idenya adalah menghitung deskriptor lokal tidak dengan *interest point* tetapi dengan skala point yang merupakan representasi secara keseluruhan daerah citra. Hal ini dilakukan karena perhitungan deskriptor dengan *interest point* dapat tidak merepresentasikan bagian penting dari citra dikarenakan lokasi *interest point* yang tidak mengenai daerah yang merupakan representasi penting dari citra tersebut.

Pengujian menggunakan 4485 citra yang memiliki 15 kelas dengan evaluasi 10-*fold cross validation*. Hasilnya untuk *akurasi, presisi, recall* dan f-measure sebagaimana dituliskan pada bagian pembahasan menunjukkan bahwa metode yang diusulkan lebih baik daripada metode dasar yaitu BoVW dengan menggunakan 2, 4 atau 6 *word*. Metode DVSA ini juga lebih baik daripada metode WSA untuk 4 dan 6 *word*. Untuk 2 *word* metode yang diusulkan lebih baik akurasinya sebesar 12.68 % dari akurasi BoVW sedangkan akurasi WSA lebih baik 15.62 % dari BoVW. Untuk 4 *word* metode yang diusulkan lebih baik akurasinya sebesar 30.99 % dari akurasi BoVW dan lebih baik 18.16 % dari WSA. Untuk 6 *word* metode yang diusulkan lebih baik akurasinya sebesar 29.98 % dari akurasi BoVW dan lebih baik 18.75 % dari WSA. Peningkatan akurasi sebesar 36.20 % didapatkan oleh metode yang

diusulkan dengan 6 word terhadap metode dasar 2 word dari BoVW. Hasil ini menunjukkan bahwa metode yang diusulkan kompetitif dalam mengenali jenis gambar.

Untuk saran atau *future work*, saran yang pertama adalah melakukan perhitungan informasi *spatial* lain seperti membentuk kuadran yang tidak dibentuk dari garis horisontal dan garis vertikal. Semisal garis horisontal dan garis vertikal tersebut diputar beberapa derajat. Atau menambahkan kuadran lain seperti daerah yang dibentuk oleh *spatial pyramid*. Dari kuadran baru inilah informasi *spatial* didapatkan. Saran lain yaitu perhitungan deskriptor bisa lebih dinamis semisal dengan ekstraksi fitur yang populer. Seperti ekstraksi fitur tekstur *local binary pattern* dan ekstraksi fitur warna. Dari ekstraksi fitur yang populer ini kemudian dibentuk deskriptor lokal yang digunakan untuk menghasilkan word. Saran selanjutnya adalah melakukan penggabungan hasil informasi *spatial* dan hasil penjumlahan dari *word*. Jadi tidak hanya menggunakan informasi *spatial* saja atau informasi penjumlahan dari word saja. Penggabungan inilah yang merupakan fitur vektor akhir dimana masing-masing dari hasil informasi *spatial* dan hasil penjumlahan dari *word* didapat dari proses *pooling*. Pada penggabungan sebagaimana saran sebelumnya proses *pooling* dapat dilakukan dengan *average pooling* maupun *max pooling*. Untuk metode *word* yang menjumlahkan dari *word*, maupun untuk metode yang *spatial*.

DAFTAR PUSTAKA

[1] Avila, S., Thome, N., Cord, M., Valle, E., de A. Araújo, A., 2013. Pooling in image representation: The visual codeword point of view. Computer Vision and Image Understanding 117, 453–465.

[2] A. Vedaldi and B.Fulkerson, 2008. (VLFeat): An Open and Portable Library of Computer Vision Algorithms

[3] Bolovinou, A., Pratikakis, I., Perantonis, S., 2013. Bag of spatio-visual words for context inference in scene classification. Pattern Recognition 46, 1039–1053.

[4] Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S., 2014. Fast and efficient visual codebook construction for multi-label annotation using predictive clustering trees. Pattern Recognition Letters 38, 38–45.

[5] Koniusz, P., Yan, F., Mikolajczyk, K., 2013. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. Computer Vision and Image Understanding 117, 479–492.

[6] Li, Z., Yap, K.-H., 2013. An efficient approach for scene categorization based on discriminative codebook learning in bag-of-words framework. Image and Vision Computing 31, 748–755.

[7] López-Sastre, R.J., García-Fuertes, A., Redondo-Cabrera, C., Acevedo-Rodríguez, F.J., Maldonado-Bascón, S., 2013. Evaluating 3D spatial pyramids for classifying 3D shapes. Computers & Graphics 37, 473–483.

[8] Penatti, O.A.B., Silva, F.B., Valle, E., Gouet-Brunet, V., Torres, R. da S., 2014. Visual word spatial arrangement for image retrieval and classification. Pattern Recognition 47, 705–720.

[9] Sánchez, J., Perronnin, F., de Campos, T., 2012. Modeling the spatial layout of images beyond spatial pyramids. Pattern Recognition Letters 33, 2216–2223.

[10] Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A., 2013. Fisher vector faces in the wild, in: British Machine Vision Conference. p. 7.

[11] Wang, J., Liu, P., She, M.F.H., Nahavandi, S., Kouzani, A., 2013. Bag-of-words representation for biomedical time series classification. Biomedical Signal Processing and Control 8, 634–644.

[12] Zagoris, K., Pratikakis, I., Antonacopoulos, A., Gatos, B., Papamarkos, N., 2014. Distinction between handwritten and machine-printed text based on the bag of visual word model. Pattern Recognition 47, 1051–1062.

[13] Zhang, C., Wang, S., Huang, Q., Liu, J., Liang, C., Tian, Q., 2013. Image classification using spatial pyramid robust sparse coding. Pattern Recognition Letters 34, 1046–1052.