Available at https://jtsiskom.undip.ac.id (16 July 2019)
DOI: 10.14710/jtsiskom.7.2.2019.47-55

Metode Pengenalan Tempat Secara Visual Berbasis Fitur CNN untuk Navigasi Robot di Dalam Gedung

Visual Place Recognition Method Based-on CNN Features for Indoor Robot Navigation

Hadha Afrisal[*)]

Departemen Teknik Elektro, Fakultas Teknik, Universitas Diponegoro
Jl. Prof. Soedarto, SH, Kampus Undip Tembalang, Semarang, Indonesia 50275

Cara sitasi: H. Afrisal, "Metode Pengenalan Tempat Secara Visual Berbasis Fitur CNN untuk Navigasi Robot di Dalam Gedung," *Jurnal Teknologi dan Sistem Komputer*, vol. 7 no. 2, 2019. doi: 10.14710/jtsiskom.7.2.2019.47-55, [Online].

Abstract – Place recognition algorithm based-on visual sensor is crucial to be developed especially for an application of indoor robot navigation in which a Ground Positioning System (GPS) is not reliable to be utilized. This research compares the approach of place recognition of using learned-features from a model of Convolutional Neural Network (CNN) against conventional methods, such as Bag of Words (BoW) with SIFT features and Histogram of Oriented Uniform Patterns (HOUP) with its Local Binary Patterns (LBP). This research finding shows that the performance of our approach of using learned-features with transfer learning method from pre-trained CNN AlexNet is better than the conventional methods based-on handcrafted-features such as BoW and HOUP.

Keywords *– place recognition; convolutional neural network; visual navigation; mobile robot*

Abstrak – Algoritma pengenalan tempat berbasis sensor visual penting untuk dikembangkan, terutama untuk aplikasi navigasi robot di dalam gedung dimana Ground Positioning System (GPS) tidak reliabel untuk digunakan. Penelitian ini membandingkan antara pendekatan berbasiskan learned-features yang diperoleh dari model Convolutional Neural Network (CNN), terhadap metode konvensional berbasis handcrafted-features, seperti Bag of Words (BoW) dengan fitur SIFT dan Histogram of Oriented Uniform Patterns (HOUP) dengan Local Binary Patterns (LBP). Hasil pengujian menunjukkan bahwa performa pendekatan learned-features dengan metode transfer learning pada pre-trained CNN AlexNet memiliki performa yang lebih baik dibandingkan metode konvensional berbasis handcrafted features seperti BoW dan HOUP.

Kata kunci *– pengenalan tempat; convolutional neural network; navigasi visual; robot bergerak*

[*)] Penulis korespondensi (Hadha Afrisal)
Email: hadha.afrisal@elektro.undip.ac.id

I. PENDAHULUAN

Dalam beberapa dekade terakhir, aplikasi robot-bergerak (*mobile-robot*) semakin banyak dibutuhkan di dunia industri, salah satunya sebagai instrumen transportasi serta distribusi logistik secara otomatis. Dalam beberapa *setting* industri, robot-bergerak diharuskan untuk bernavigasi dari satu tempat ke tempat lain secara otomatis guna mengambil barang dan mendistribusikannya ke beberapa ruangan dalam satu gedung. Robot tersebut ditugaskan untuk mengenal dan mengingat ruang-ruang yang pernah disinggahi, sehingga proses distribusi logistik yang melibatkan proses memuat dan menurunkan barang dapat terlaksana secara tepat. Dalam skenario tersebut, robot-bergerak wajib dibekali dengan kemampuan pengenalan dan pengingat lokasi. Pada perkembangannya, metode pengenalan tempat untuk navigasi robot-bergerak menjadi dasar dalam pengembangan metode *Simultaneous Localisation and Mapping* (SLAM). Metode SLAM sampai saat ini pengembangannya masih dalam tahap awal dan membutuhkan inovasi algoritma untuk meningkatkan performa dan akurasinya [1].

Metode pengenalan tempat berbasiskan citra digital menggunakan sensor kamera pada robot-bergerak merupakan salah satu solusi untuk pengembangan SLAM, terutama pada aplikasi navigasi *indoor*. Pada umumnya, navigasi robot (yang melibatkan *positioning* dan *localization*) berpedoman pada sensor GPS saja. Namun, pada lokasi *indoor* pembacaan nilai GPS cenderung tidak dapat diandalkan karena memiliki kepresisian dan akurasi yang rendah [2]. Penggunaan sensor visual, seperti kamera, menjadi alternatif yang tepat untuk skenario ini. Kamera sebagai sensor utama pada robot-bergerak juga dinilai lebih menguntungkan karena instalasi dan perawatannya yang cukup mudah dan harganya yang terjangkau jika dibandingkan dengan sensor-sensor lain, seperti *laser range finder*, *ultrasound*, *infra-red*, dan sejenisnya.

Pada teknik visual SLAM, beberapa pendekatan berbasiskan *handcrafted features* atau fitur buatan telah banyak dikembangkan, antara lain SIFT [3] dan SURF

Submitted: 21 February 2019; Revised: 18 April 2019; Accepted: 29 April 2019; Published: 30 April 2019

[4] yang berbasiskan fitur lokal berbentuk vektor, BRISK [5] dan ORB [6] yang berbasiskan fitur lokal berbentuk biner, serta fitur dan deskriptor lain yang dikenali oleh penglihatan manusia, seperti garis [7] maupun sudut dan tepi [8].

Salah satu metode pengenalan tempat yang cukup banyak digunakan saat ini adalah *Bag-of-Words* (BoW) atau *Bag-of-Features* (BoF). Metode tersebut secara umum terdiri dari tahapan: (1) ekstraksi fitur lokal pada citra, (2) proses *encoding* fitur lokal sebagai deskriptor, dan (3) klasifikasi deskriptor citra [9]. Salah satu teknik pengenalan tempat berbasis BoW telah berhasil didemonstrasikan dengan algoritma Dorian [10], yakni dengan menggunakan teknik *encoding* biner untuk mendeskripsikan fitur-fitur yang telah diekstraksi dari citra digital. Metode tersebut cukup efektif dan efisien untuk pengenalan tempat. Namun, metode tersebut memiliki akurasi yang cukup rendah jika digunakan pada citra-citra yang memiliki bentuk rotasi dan skala perspektif yang bervariasi. Selain itu, tantangan lain dalam pengembangan teknik BoW yang paling banyak ditemukan adalah algoritma tersebut tidak dapat digunakan pada citra-citra yang memilki struktur atau pola bentuk yang repetitif dalam citra yang sama. Padahal pada umumnya penggunaan robot-bergerak di lokasi indoor banyak menemui pola-pola visual repetitif tersebut, seperti misalnya pada ubin-ubin, pola pada plafon, pintu, jendela, dan lain sebagainya [11]. Selain itu, metode-metode konvensional berbasiskan BoW pada umumnya memiliki akurasi yang rendah jika digunakan pada kondisi cahaya yang berubah-ubah serta dengan sudut pandang kamera yang beragam [12].

Perkembangan teknik *machine learning* (pembelajaran-mesin) dan *deep learning* (pembelajaran-mendalam) memberikan alternatif terobosan baru dalam pengembangan metode pengenalan tempat menggunakan citra digital yakni dengan menggunakan *Convolutional Neural Network* (CNN) [13]. CNN dengan *learned-features* (fitur hasil-pembelajaran) yang dihasilkan dapat digunakan untuk menyelesaikan permasalahan-permasalahan yang ada pada proses pengenalan objek dan pemulihan citra, misalnya pada metode pengenalan dan kategorisasi tempat [14].

Implementasi metode tersebut untuk aplikasi robot-bergerak masih menemui banyak tantangan dan keterbatasan, terutama terkait dengan dinamika serta tingginya tingkat ketidakpastian di dunia nyata. Jumlah citra yang digunakan dalam proses training haruslah mencukupi dan representatif (tidak *overfitting* maupun *underfitting*). Selain itu, kemungkinan jumlah citra dalam proses *training* akan menjadi jauh lebih besar karena proses tersebut berlangsung secara terus menerus dan dalam waktu cukup lama [15]. Sebagai solusi atas tantangan serta keterbatasan yang ada pada metode-metode terdahulu, penelitian ini bertujuan untuk mengembangkan metode alternatif untuk pengenalan tempat dan kategorisasi citra lokasi *indoor* dengan menggunakan metode *transfer learning* CNN yang berbasis *learned-features*.

III. METODE PENELITIAN

Penelitian ini bertujuan untuk menguji metode *transfer learning* CNN yang berbasiskan *learned-features* dan membandingkan performanya terhadap metode konvensional yang berbasiskan *handcrafted-features*. Eksperimen yang dilakukan menggunakan pendekatan baru, yakni memanfaatkan *learned-features* dari teknik *transfer learning* pada arsitektur CNN AlexNet [16]. Metode konvensional *Bag of Words* (BoW) [10] dan *Histogram of Oriented Uniform Pattern* (HOUP) [17] digunakan sebagai pembanding terhadap metode yang dikembangkan tersebut. Pengujian dilakukan secara *offline* dengan menggunakan software Matlab pada komputer dengan CPU Intel Core i5 2,8 GHz dan kapasitas RAM 4 GB.

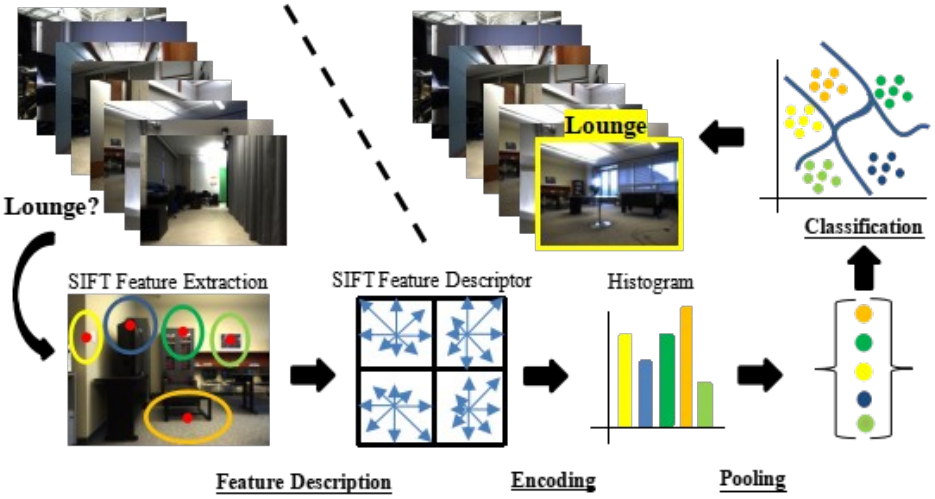
A. Dataset Citra *Indoor*

Untuk membandingkan performa metode yang dikembangkan terhadap metode konvensional yang sudah ada, secara fair eksperimen ini akan memanfaatkan 2 jenis dataset, yaitu *Indoor Scene Recognition (ISR) CVPR MIT dataset* [18] dan *York University dataset* [17]. Dataset ISR-CVPR MIT dipilih karena memiliki jumlah kategori citra yang relatif banyak dan dengan sudut pandang serta variasi fitur yang cukup untuk menguji akurasi algoritma pada kategori kelas yang cukup besar. Dataset York University ini dipilih sebagai data uji karena citra direkam menggunakan 2 jenis robot-bergerak yang berbeda dan dalam kondisi iluminasi yang berbeda, sehingga cocok untuk menguji kepresisian dan *repeatability* dari algoritma yang dikembangkan dalam situasi dan *setting* yang variatif.

Dataset ISR-CVPR MIT terdiri dari 67 kategori tempat *indoor* dengan total citra sejumlah 15.620 yang mana terdapat sekurang-kurangnya 100 citra pada tiap kategorinya. Citra dataset berformat JPG dengan klasifikasi jenis tempat yang beragam dari mulai *store, home, public spaces, leisure,* dan *working places*. Dataset York University terdiri dari 11 dan 17 kategori ruangan dengan jumlah total citra 29.917 berformat PNG dengan resolusi 640 x 480 piksel. Citra-citra ruangan tersebut diambil dengan menggunakan sensor kamera *Point Grey Bumblebee* dari 2 robot yang berbeda, yakni robot Pioneer (tinggi 88 cm) dan Virtual Me (tinggi 117 cm). Dataset diambil pada 2 kondisi pencahayaan yang berbeda, yakni pada siang dan malam hari.

B. Deskriptor Fitur Citra

Deskriptor citra merepresentasikan informasi unik yang dikandung oleh sebuah citra dalam bentuk *keypoints*[19]. Deskriptor citra diklasifikasikan ke dalam 2 jenis, yaitu *handcrafted-features descriptor* dan *learned-features descriptor* [13].



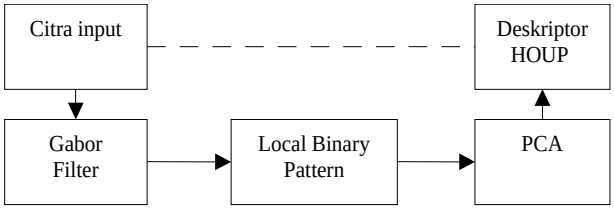
Gambar 1. *Pipeline* pembentukan deskriptor BoW dengan fitur SIFT

Handcrafted-Features Descriptor: BoW dan HOUP

Handcrafted-features atau fitur buatan-tangan, diperoleh dari teknik ekstraksi fitur lokal menggunakan ekstraktor yang telah ditentukan polanya, misalnya *Scale Invariant Feature Transform* (SIFT) [3] dan *Speeded-up Robust Features* (SURF) [4]. Pada penelitian ini, 2 buah *handcrafted-features descriptor* yang digunakan dan dibandingkan performanya adalah *Bag of Words* (BoW) dan *Histogram of Oriented Uniform Patterns* (HOUP).

Metode *Bag of Words* (BoW) digunakan dalam penelitian sebagai deskriptor citra untuk proses klasifikasi dan pengenalan objek dalam bidang *computer vision* dan disebut sebagai *Bag of Visual Words*, sebagai representasi dari kelompok fitur pada citra yang disandikan melalui deskriptor. Pada penelitian ini, fitur yang digunakan dalam BoW adalah SIFT. *Pipeline* deskriptor BoW dengan fitur SIFT ditunjukkan pada Gambar 1, yakni terdiri dari 4 tahapan utama, yaitu: (1) ektraksi fitur menggunakan SIFT, (2) proses encoding menggunakan histogram, (3) proses *max pooling* untuk masing-masing wilayah spasial pada fitur, dan (4) proses klasifikasi. Pada metode BoW yang dilakukan di penelitian ini, proses klasifikasi menggunakan *Support Vector Machine* (SVM).

Fitur yang diperoleh dengan SIFT sebagai *keypoints*, berbentuk lingkaran yang menandai sebuah wilayah citra, lengkap dengan orientasi gradien (Gambar 1). Deskriptor SIFT pada sebuah *keypoint* merupakan bentuk histogram spasial 3-D dari gradien citra. Gradien pada masing-masing piksel dianggap sebagai sampel dari dasar fitur vektor 3 dimensi, yang didefinisikan dari lokasi piksel dan juga orientasi gradien. Sampel-sampel tersebut ditimbang dan dinormalisasi terhadap histogram 3 dimensi secara total. Standar SIFT menggunakan fungsi pembobotan model Gaussian untuk mengeliminasi nilai-nilai gradien yang terlalu jauh dari titik pusat *keypoint*. Pada eksperimen yang dilakukan, metode SIFT yang digunakan mengacu pada standar SIFT pada VLFeat [20] yang berbeda dengan



Gambar 2. *Pipeline* pembentukan deskriptor HOUP

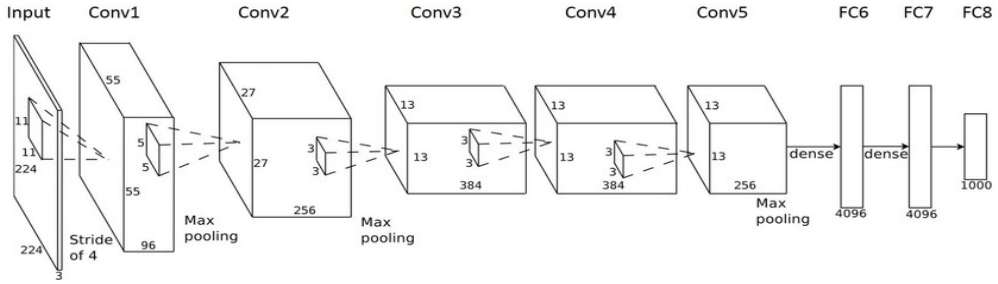
metode SIFT original yang dikembangkan [3], yakni menggunakan y-axis ke arah bawah dan konvensi orientasi searah putaran jarum jam.

Histogram of Oriented Uniform Patterns (HOUP) sebagai bentuk deskriptor citra digunakan untuk pengenalan tempat dan kategorisasi citra [21]. Metode ini menggunakan Gabor filter untuk memfiltrasi sub-blok pada citra dengan orientasi yang berbeda-beda (Gambar 2). *Output* dari Gabor filter tersebut digunakan untuk menghitung *Local Binary Patterns* (LBP). *Principal Component Analysis* (PCA) digunakan untuk mengurangi dimensionalitas dari jumlah fitur yang diperoleh,

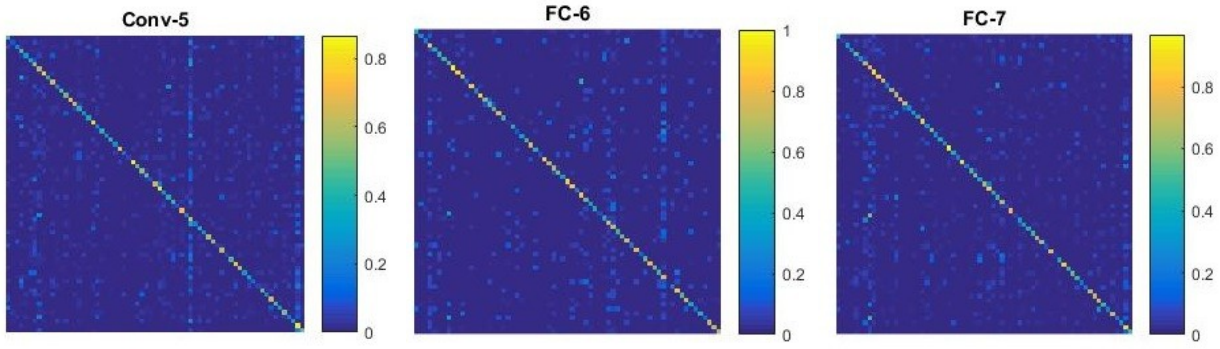
Learned-Features Descriptor: TL-AlexNet

Learned-descriptor sebagai bentuk deskripsi citra menggunakan fitur-fitur yang dihasilkan dari pembelajaran-mesin (*machine learning*) [13]. Pada penelitian ini, *learned-features* diperoleh dari metode *transfer learning* dari arsitektur *Convolutional Neural Network* (CNN) AlexNet [16]. Pendekatan *transfer learning* untuk pembelajaran mesin menggunakan sebuah model yang sudah dilatih pada dataset tertentu untuk kemudian digunakan kembali (*re-purposed*) pada dataset dengan *setting* dan keperluan yang lain.

Pre-trained AlexNet digunakan untuk tujuan baru yaitu sebagai metode pengenalan tempat dengan pendekatan *transfer learning* ini. Model tersebut sudah terbukti memiliki performa yang sangat baik pada *ImageNet Large-Scale Visual Recognition Challenge*



Gambar 3. Model *Convolutional Neural Network* (CNN) AlexNet [16]



Gambar 4. Visualisasi *confusion matrix* dari pengujian terhadap dataset ISR-CVPR (67 kategori) dengan metode *transfer learning* AlexNet pada layer Conv-5, FC-6, dan FC-7, dimana sumbu Y merepresentasikan kategori citra sebenarnya (*actual category*) dan sumbu X merepresentasikan kategori citra hasil prediksi (*predicted category*)

(ILSVRC-2012), yakni meraih peringkat top-5 dengan tingkat galat cukup kecil yakni 15,3%[22]. Hal ini sesuai dengan intuisi otak manusia, yaitu bahwa sebuah tempat biasanya terdapat beberapa objek yang familiar, misalnya peralatan-peralatan dapur di dapur, komputer dan meja kerja di kantor, atau ubin dan plafon atau pintu-pintu dalam jumlah banyak di lorong. Fitur-fitur yang telah ada pada pre-trained model AlexNet digunakan kembali untuk mengenali tempat. Model pre-trained AlexNet digunakan dengan MatConvNet [23] pada Matlab. Citra-citra pada dataset sebelumnya dilakukan preproses dengan melakukan skala ulang menjadi ukuran 227 x 227 piksel agar sesuai dengan spesifikasi pada model CNN AlexNet.

Eksperimen dilakukan dengan cara membandingkan performa metode berbasis *learned-features* terhadap metode konvensional yang berbasiskan *handcrafted-features*, yaitu meliputi tahapan berikut (Gambar 3):

1. Menguji performa metode Transfer Learning (TL) menggunakan pre-trained CNN dengan arsitektur AlexNet, dimana fitur-fitur yang diekstraksi dan diolah adalah dari layer ke-5 (convolutional 5), ke-6 (fully-connected layer 6), dan ke-7 (fully-connected layer 7). Performa transfer learning dengan accuracy, precision, recall, dan F1 score terbaik kemudian digunakan sebagai pembanding terhadap metode konvensional berbasis handcrafted-features.
2. Metode berbasiskan learned-features dengan performa terbaik kemudian dibandingkan dengan metode konvensional yang paling banyak digunakan serta dengan performa yang relatif paling baik dalam metode pengenalan objek, yakni (a) BoW berfitur SIFT dengan pengklasifikasi SVM, dan (b) HOUP dengan LBP yang sudah direduksi dimensionalitasnya menggunakan metode PCA.

Perhitungan nilai *accuracy*, *precision*, *recall*, dan *F1 score* menggunakan Persamaan 1 sampai Persamaan 4. TP menyatakan *True Positive*, TN *True Negative*, FP *False Positive*, dan FN *False Negative*.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

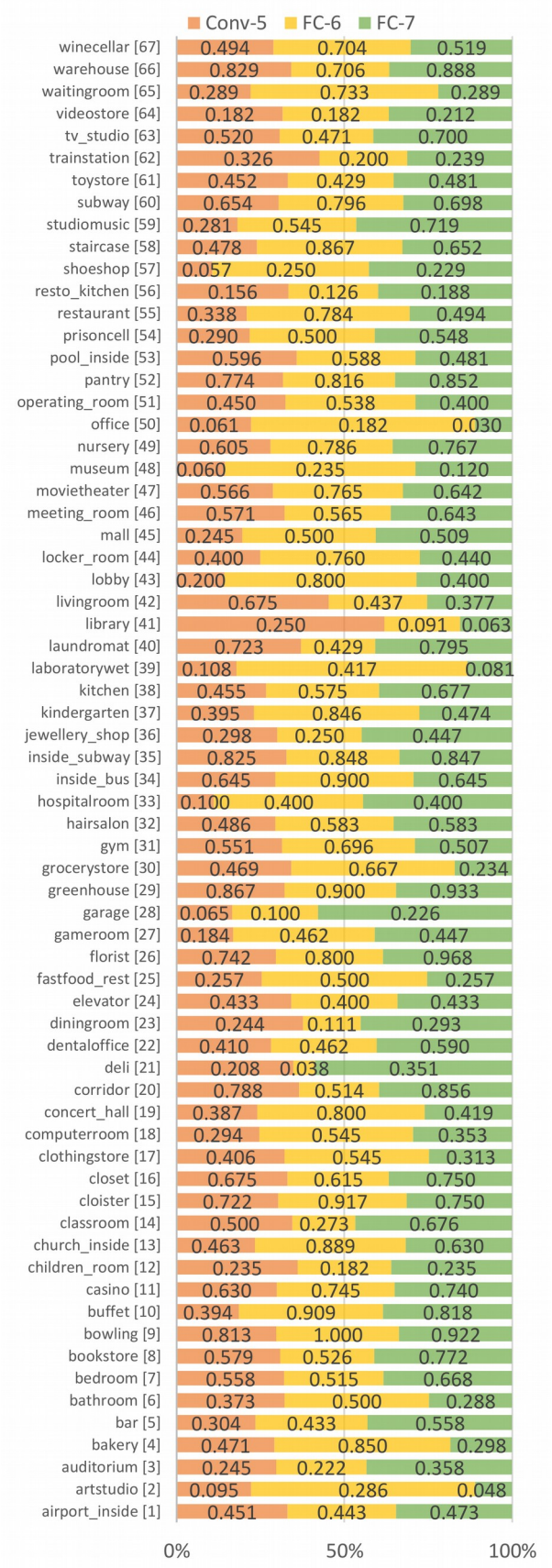
$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1_{score} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Pengujian awal berusaha membandingkan performa metode *transfer learning* pada layer 5, 6, dan 7, terhadap 2 buah dataset, yaitu: ISR-CVPR dan York Univ, dimana masing-masing dataset memiliki jumlah kategori dan jenis citra yang berbeda. *Confusion matrix* yang dihasilkan dari proses klasifikasi 67 kategori citra tempat *indoor* menggunakan dataset ISR-CVPR (Gambar 4). Matriks tersebut menunjukkan bahwa



Gambar 5. Perbandingan nilai akurasi dalam rentang 0 – 1 (0 – 100%) pada masing-masing kategori citra dataset ISR-CVPR menggunakan metode pengujian TL-AlexNet pada FC-7, FC-6, dan Conv-5

Tabel 1. Nilai *accuracy*, *precision*, *recall*, dan *F1 score* untuk masing-masing proses klasifikasi dataset ISR-CVPR (67 kategori) menggunakan metode TL dengan mengambil fitur CNN AlexNet pada layer 5 (Conv-5), 6 (FC-6), dan 7 (FC-7)

CNN Layer	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Conv-5	0,427	0,427	0,507	0,464
FC-6	0,542	0,542	0,612	0,575
FC-7	0,503	0,503	0,578	0,538

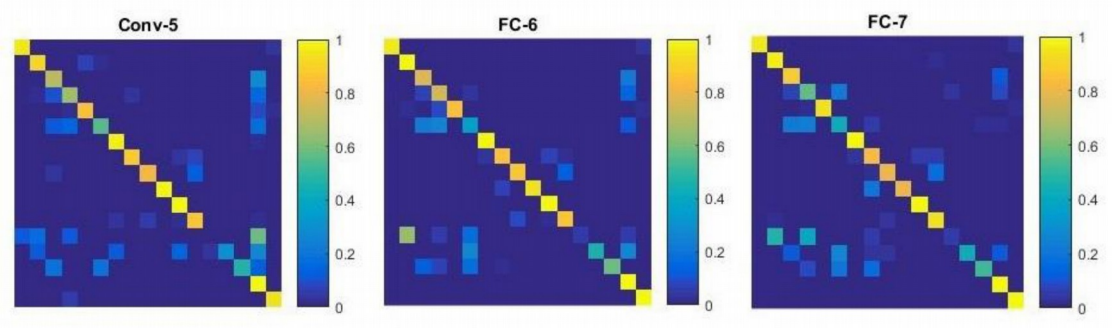
model *transfer learning* pada layer 5, 6, dan 7 AlexNet untuk dataset ISR-CVPR dengan jumlah 67 kategori citra masih memiliki akurasi cukup baik karena secara visual nilai TP pada diagonal matriks masih bernilai cukup tinggi.

Secara lebih detail, nilai sebaran akurasi pada masing-masing kategori citra ditampilkan pada Gambar 5. Dari pengujian yang dilakukan, terlihat bahwa secara umum nilai akurasi pada masing-masing kategori untuk metode TL-AlexNet FC-6 lebih dominan mengungguli Conv-5 dan FC-7. Nilai akurasi Conv-5 dan FC-7 hanya unggul pada beberapa kategori saja, sedangkan FC-6 unggul pada lebih dari 50% jumlah kategori yang ada.

Perbandingan performa secara keseluruhan dari masing-masing metode TL-AlexNet, nilai accuracy, precision, recall, dan F1 score disajikan pada Tabel 1 sesuai dengan perhitungan menggunakan Persamaan 1-4. Performa keseluruhan dari metode TL-AlexNet pada FC-6 mengungguli dua layer lain (Conv-5 dan FC-7), yakni dengan accuracy 54,2%, precision 54,2%, recall 61,2%, dan F1 score 57,5%. Dengan keunggulan tersebut, maka dapat disimpulkan bahwa metode TL-AlexNet paling optimal berada pada layer ke-6, yaitu Fully-Connected Layer 6, untuk dataset ISR-CVPR dengan 67 kategori citra.

Selanjutnya, ketiga buah metode TL-AlexNet (Conv-5, FC-6, dan FC-7) diujikan terhadap dataset York Univ dengan 17 kategori citra lokasi indoor. *Confusion matrix* yang dihasilkan dari metode TL-AlexNet pada Conv-5, FC-6, dan FC-7 ditampilkan pada Gambar 6. Secara visual, ketiga metode TL-AlexNet tersebut juga menunjukkan performa yang cukup baik dalam mengkategorikan citra lokasi indoor. Hal ini terlihat dari rendahnya nilai citra yang terklasifikasi secara salah serta cukup tingginya nilai akurasi, yaitu pada bagian diagonal confusion matrix. Satu-satunya kategori citra yang memiliki nilai akurasi rendah adalah pada kategori Plant-Room.

Perbandingan nilai akurasi pada masing-masing metode TL-AlexNet untuk proses kategorisasi citra lokasi dataset York Univ disajikan pada Gambar 7. Hasil eksperimen tersebut memperlihatkan bahwa secara umum metode TL-AlexNet pada FC-6 mengungguli metode dengan layer Conv-5 dan FC-7. Performa secara keseluruhan metode TL-AlexNet pada Conv-5, FC-6, dan FC-7 ditunjukkan pada Tabel 2. Metode TL-AlexNet pada FC-6 unggul dalam hal nilai



Gambar 6. Visualisasi *confusion matrix* dari pengujian terhadap dataset York Univ (17 kategori) dengan metode *transfer learning* AlexNet pada layer Conv-5, FC-6, dan FC-7, dimana sumbu Y merepresentasikan kategori citra sebenarnya (*actual category*) dan sumbu X merepresentasikan kategori citra hasil prediksi (*predicted category*)

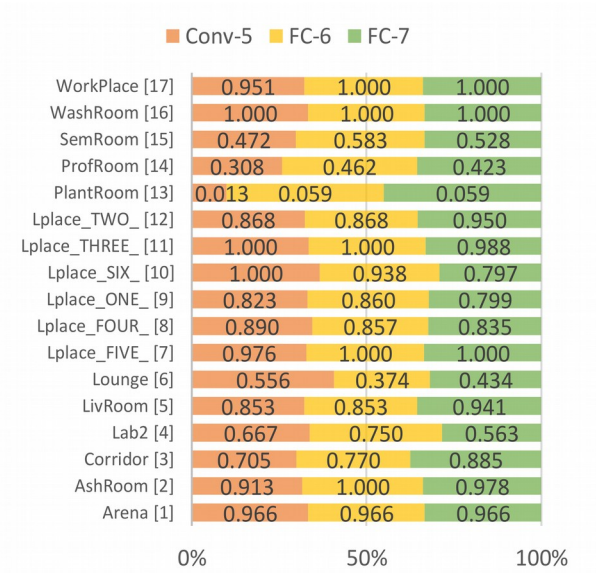
accuracy (78,5%), precision (78,4%), recall (83,3%), dan F1 Score (80,8%).

Secara lebih detail, nilai sebaran akurasi pada masing-masing kategori citra ditampilkan pada Gambar 5. Dari pengujian yang dilakukan, terlihat bahwa secara umum nilai akurasi pada masing-masing kategori untuk metode TL-AlexNet FC-6 lebih dominan mengungguli Conv-5 dan FC-7. Nilai akurasi Conv-5 dan FC-7 hanya unggul pada beberapa kategori saja, sedangkan FC-6 unggul pada lebih dari 50% jumlah kategori yang ada.

Guna membandingkan performa secara keseluruhan masing-masing metode TL-AlexNet, nilai *accuracy*, *precision*, *recall*, dan *F1 score* disajikan pada Tabel 1, sesuai dengan perhitungan pada Persamaan 1-4. Dari Tabel 1 tersebut, terlihat bahwa performa keseluruhan dari metode TL-AlexNet pada FC-6 mengungguli dua layer lain (Conv-5 dan FC-7), yaitu dengan *accuracy* 54,2%, *precision* 54,2%, *recall* 61,2%, dan *F1 score* 57,5%. Dengan keunggulan tersebut, maka dapat disimpulkan bahwa metode TL-AlexNet paling optimal berada pada layer ke-6, yaitu *Fully-Connected Layer* 6, untuk dataset ISR-CVPR dengan 67 kategori citra.

Selanjutnya, ketiga buah metode TL-AlexNet (Conv-5, FC-6, dan FC-7) akan diujikan terhadap dataset York Univ dengan 17 kategori citra lokasi *indoor*. Confusion matrix yang dihasilkan dari metode TL-AlexNet pada Conv-5, FC-6, dan FC-7 ditampilkan pada Gambar 6. Secara visual, ketiga metode TL-AlexNet tersebut juga menunjukkan performa yang cukup baik dalam mengkategorikan citra lokasi *indoor*. Hal ini terlihat dari rendahnya nilai citra yang terklasifikasi secara salah serta cukup tingginya nilai akurasi pada bagian diagonal *confusion matrix*. Satu-satunya kategori citra yang memiliki nilai akurasi rendah adalah pada kategori Plant-Room.

Perbandingan nilai akurasi pada masing-masing metode TL-AlexNet untuk proses kategorisasi citra lokasi dataset York Univ disajikan pada Gambar 7. Hasil eksperimen tersebut memperlihatkan bahwa secara umum metode TL-AlexNet pada FC-6 mengungguli metode dengan layer Conv-5 dan FC-7. Selanjutnya, performa secara keseluruhan metode TL-AlexNet pada Conv-5, FC-6, dan FC-7 ditampilkan pada Tabel 2. Pada Tabel 2 tersebut terlihat bahwa

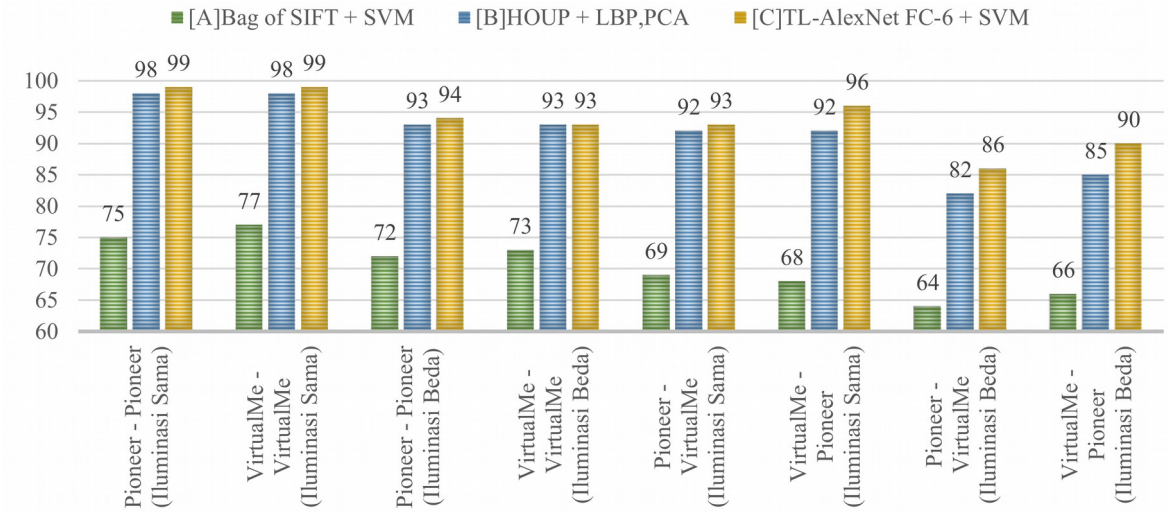


Gambar 7. Perbandingan nilai akurasi dalam rentang 0 – 1 (0 – 100%) pada masing-masing kelas citra dataset York University (*17 places*), menggunakan metode pengujian TL-AlexNet pada FC-7, FC-6, dan Conv-5

metode TL-AlexNet pada FC-6 unggul dalam hal nilai *accuracy* (78,5%), *precision* (78,4%), *recall* (83,3%), dan *F1 Score* (80,8%).

Dari pengujian tersebut, dapat diperoleh kesimpulan bahwa metode *Transfer Learning* menggunakan *pre-trained* CNN AlexNet memiliki performa yang maksimal jika dilakukan pada layer ke-6, yaitu *Fully-Connected Layer* 6 (FC-6). Untuk selanjutnya, metode TL-AlexNet FC-6 tersebut akan digunakan sebagai pembanding terhadap 2 jenis metode konvensional yang menggunakan *handcrafted-features*, yaitu BoW dengan fitur SIFT pengklasifikasi SVM dan HOUP fitur LBP.

Pada pengujian, factor perubahan iluminasi dan format citra yang berbeda juga akan menjadi parameter pengujian, sehingga pada pengujian akan digunakan dataset York Univ dengan 11 kategori. Hal ini disebabkan dataset dengan 11 kategori memiliki parameter yang lebih variatif, yaitu selain diambil pada siang dan malam hari, dataset tersebut juga diambil



Gambar 8. Perbandingan nilai akurasi 3 jenis algoritma pengenalan tempat (dalam %), yakni yang menggunakan *handcrafted features*: [A] Metode *Bag of Words* menggunakan fitur SIFT (*Scale Invariant Feature Transform*) dengan pengklasifikasi menggunakan *Support Vector Machine* (SVM) dan [B] *Histogram of Oriented Uniform Pattern* (HOUP) menggunakan *Local Binary Patterns* (LBP) dan *Principal Component Analysis* (PCA) [ ], serta yang menggunakan *learned features* yakni: [C] Metode *Transfer Learning* menggunakan *Convolutional Neural Network* berarsitektur Alex-Net pada *Fully Connected Layer* ke-6, dengan kondisi uji iluminasi bervariasi.

menggunakan 2 jenis robot yang berbeda, yaitu robot Pioneer dan Virtual Me.

Hasil pengujian ditampilkan pada Gambar 8 dan Tabel 3. Pada Metode BoW dengan fitur SIFT memiliki performa paling rendah, sedangkan metode HOUP dan TL-AlexNet FC-6 bersaing ketat (Gambar 8). Pada citra training dan uji dengan iluminasi sama pada robot yang sama, HOUP cenderung memiliki performa yang hampir sama dengan TL-AlexNet FC-6, namun ketika kondisi iluminasi berbeda dan dengan robot berbeda untuk training dan tes, performa HOUP menurun secara signifikan. Sedangkan metode TL-AlexNet FC-6 relatif lebih stabil dengan nilai akurasi yang cukup tinggi. Hal ini dapat dilihat pada grafik komparasi paling kanan, yaitu ketika pengujian robot berbeda dan iluminasi berbeda, TL-AlexNet FC-6 masih mencapai 90% akurasi atau 5% di atas akurasi HOUP.

Nilai performa tertinggi dicapai ketika kondisi pencahayaan sama dan dengan robot yang berbeda, pada kondisi tersebut metode TL-AlexNet FC-6 mampu mencapai nilai *accuracy* 99%, *precision* 98%, *recall* 98%, dan *F1 score* 98%. Perubahan iluminasi berpengaruh terhadap menurunnya nilai performansi metode pengenalan tempat berbasis citra pada semua jenis metode. Namun, pada metode TL-AlexNet FC-6 penurunan nilai performansi tidak terlalu signifikan, bahkan dengan robot yang berbeda (perbedaan sudut pandang pengambilan citra) dan dengan kondisi iluminasi yang berbeda. Metode TL-AlexNet mampu mempertahankan nilai *accuracy*, *precision*, *recall*, dan *F1 score* di atas 87% (Tabel 3). Dengan hasil pengujian tersebut, metode yang dikembangkan ini memiliki performa yang lebih baik dibandingkan terhadap metode konvensional dalam [17] dan dengan metode BoW dalam [10].

Tabel 2. Nilai *accuracy*, *precision*, *recall*, dan *F1 score* untuk masing-masing proses klasifikasi dataset York Univ (17 kategori) menggunakan metode TL dengan mengambil fitur CNN AlexNet pada layer 5 (Conv-5), 6 (FC-6), dan 7 (FC-7)

CNN Layer	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Conv-5	0,762	0,761	0,778	0,770
FC-6	0,785	0,784	0,833	0,808
FC-7	0,774	0,773	0,797	0,785

Tabel 3. Nilai *precision*, *recall*, dan *F1 score* untuk masing-masing proses klasifikasi dataset York University (*11 places*) menggunakan metode TL-AlexNet FC-6 dengan skema variasi iluminasi.

Robot *(Iluminasi)*	Precision	Recall	F1 Score
Pi - Pi *(Sama)*	0,984	0,985	0,984
VMe – VMe *(Sama)*	0,985	0,986	0,986
Pi – Pi *(Beda)*	0,966	0,966	0,966
VMe – VMe *(Beda)*	0,927	0,934	0,930
Pi – VMe *(Sama)*	0,902	0,913	0,908
VMe – Pi *(Sama)*	0,957	0,959	0,958
Pi – VMe *(Beda)*	0,826	0,857	0,841
VMe – Pi *(Beda)*	0,878	0,894	0,886

Pada tahap selanjutnya, pengujian pengenalan tempat dilakukan dengan menggunakan metode TL-AlexNet FC-6 dengan *video sequences* secara *offline*, sebagaimana *screenshoot* hasil pengujian yang ditampilkan pada Gambar 9. Beberapa kategori citra memiliki kemiripan yang cukup tinggi, seperti beberapa sudut ruang *WorkPlace* yang sekilas mirip dengan

Gambar 9. Beberapa hasil proses pengenalan dan klasifikasi citra tempat *indoor* menggunakan dataset York Univ dengan kategori 11 jenis ruangan. Bagian atas merupakan *screenshoot* ketika algoritma berhasil mengenali tempat dengan benar, sedangkan bagian bawah saat algoritma salah dalam mengenali jenis ruangan.

SemRoom, atau sudut ruang berpintu pada *LivRoom* yang sekilas mirip *Corridor*. Namun, pada pengujian, secara umum kesalahan pengenalan tempat (mis-klasifikasi) cukup rendah, yakni rata-rata bernilai tidak lebih dari 7%. Waktu komputasi untuk algoritma TL-AlexNet FC-6 pada Matlab adalah antara 0,07-0,10 detik untuk memproses dan mengklasifikasikan citra lokasi indoor. Hasil ini cukup baik karena pada prakteknya algoritma bisa memproses *video sequence* dalam ~10 FPS. Hal ini disebabkan arsitektur AlexNet yang cukup dalam, yaitu terdiri dari 8 *layer*.

IV. KESIMPULAN

Metode *Transfer Learning* (TL) dari *pre-trained* CNN AlexNet pada FC-6, sebagai salah satu metode pengenalan tempat berbasiskan *learned-features*, cukup handal untuk pengenalan tempat pada robot-bergerak. Dalam hal akurasi dan kepresisian, metode ini lebih unggul dibandingkan dengan metode-metode konvensional lain berbasiskan *handcrafted-features* seperti BoW dan HOUP. Selain itu, metode ini memiliki keuntungan, antara lain lebih *robust* terhadap perubahan iluminasi cahaya, *adaptable*, dan dapat digunakan pada sudut pandang pengambilan citra yang berbeda. Namun, kekurangannya adalah dalam hal waktu komputasi. Metode ini belum cukup cepat untuk implementasi *real time* pada aplikasi robot-bergerak karena hanya mampu memproses video kurang lebih 10 FPS saja, padahal robot-bergerak bernavigasi lincah membutuhkan minimal 25-30 FPS. Ke depannya, untuk aplikasi *real-time* pada robot-bergerak, metode *transfer learning* dapat menggunakan arsitektur CNN lain yang memiliki jumlah *layer* lebih sedikit.

UCAPAN TERIMA KASIH

Penulis berterima kasih kepada Dr. Bogdan Kwolek dari AGH University of Science and Technology Krakow, Polandia, yang telah menjadi pembimbing sekaligus rekan diskusi dalam penelitian yang telah dilakukan. Penulis juga mengucapkan terima kasih kepada UNESCO dan AGH University of Science and Technology Krakow yang telah mendanai program penelitian pada UNESCO/Poland *Co-Sponsored Fellowship Programme in Engineering* 2017 yang dilakukan oleh penulis. Hasil penelitian ini merupakan salah satu bagian dari luaran program *fellowship* yang telah dilakukan.

DAFTAR PUSTAKA

[1] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual Simultaneous Localization and Mapping: A Survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55-81, 2015.

[2] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual Place Recognition: A Survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1-19, 2016.

[3] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, Greece, Sept. 1999, pp. 1150-1157.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up Robust Features," in *European Conference on Computer Vision*, Berlin, Heidelberg, 2006, pp. 404-417.

[5] S. Leutenegger, M. Chli, and R. Y. Siegwart, ”BRISK: Binary Robust Invariant Scalable Keypoints,“ in *2011 International Conference on Computer Vision, Barcelona*, Spain, Nov. 2011, pp. 2548-2555.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 2564-2571.

[7] Z. Wang, F. Wu, and Z. Hu, "MSLD: A Robust Descriptor for Line Matching," *Pattern Recognition*, vol. 42, no. 5, pp. 941-953, 2009.
[8] C. G. Harris and M. Stephens, "A Combined Corner and Edge Detector.," in *Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147-151.
[9] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The Devil is in the Details: an Evaluation of Recent Feature Encoding Methods," in *The 22nd British Machine Vision Conference*, England, Sept. 2011, pp. 76.1-76.12.
[10] D. Gálvez-López and J. D. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188-1197, 2012.
[11] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual Place Recognition with Repetitive Structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2346-2359, 2013.
[12] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place Recognition with Convnet Landmarks: Viewpoint-robust, Condition-robust, Training-free," in *Proceedings of Robotics: Science and Systems XI*, Rome, Italy, July 2015.
[13] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative Evaluation of Hand-crafted and Learned Local Features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[14] Q. Li, K. Li, X. You, S. Bu, and Z. Liu, "Place Recognition Based on Deep Feature and Adaptive Weighting of Similarity Matrix," *Neurocomputing*, vol. 199, pp. 114-127, 2016.
[15] L. Tai and M. Liu, "Deep-Learning in Mobile Robotics-From Perception to Control Systems: A Survey on Why and Why Not," *arXiv preprint arXiv:1612.07139 [cs]*, Dec. 2016.
[16] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.
[17] R. Sahdev and J. K. Tsotsos, "Indoor Place Recognition System for Localization of Mobile Robots," in *13th Conference on Computer and Robot Vision (CRV)*, Victoria, Canada, Jun. 2016, pp. 53-60.
[18] A. Quattoni and A. Torralba, "Recognizing Indoor Scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, Jun. 2009, pp. 413-420.
[19] S. Krig, "Interest Point Detector and Feature Descriptor Survey," in *Computer Vision Metrics*, Berkeley: Apress, 2014, pp. 187-246.
[20] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," in *Proc. of the 18th ACM International Conference on Multimedia*, New York, USA, Oct. 2010, pp. 1469-1472.
[21] E. Fazl-Ersi and J. K. Tsotsos, "Histogram Of Oriented Uniform Patterns for Robust Place Recognition and Categorization," *International Journal of Robotics Research*, vol. 31, no. 2, pp. 468-483, 2012.
[22] O. Russakovsky et al., "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 1, pp. 211-252, 2015.
[23] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional Neural Networks for Matlab," in *Proceedings of the 23rd ACM International Conference on Multimedia*, Brisbane, Australia, Oct. 2015, pp. 689-692.