Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2010, Article ID 469563, 11 pages
doi:10.1155/2010/469563

Research Article

Feature-Based Image Comparison for Semantic Neighbor Selection in Resource-Constrained Visual Sensor Networks

Yang Bai and Hairong Qi

Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, TN 37996, USA

Correspondence should be addressed to Yang Bai, ybai2@utk.edu

Received 28 December 2009; Revised 22 May 2010; Accepted 20 September 2010

Academic Editor: Li-Qun Xu

Visual Sensor Networks (VSNs), formed by large number of low-cost, small-size visual sensor nodes, represent a new trend in surveillance and monitoring practices. Sensor collaboration is essential to VSNs and normally performed among sensors having similar measurements. The directional sensing characteristics of imagers and the presence of visual occlusion present unique challenges to neighborhood formation, as geographically-close neighbors might not monitor similar scenes. In this paper, we propose the concept of forming semantic neighbors, where collaboration is only performed among geographically-close nodes that capture similar images, thus requiring image comparison as a necessary step. To avoid large amount of data transfer, we propose feature-based image comparison as features provide more compact representation of the image. The paper studies several representative feature detectors and descriptors, in order to identify a suitable feature-based image comparison system for the resource-constrained VSN. We consider two sets of metrics from both the resource consumption and accuracy perspectives to evaluate various combinations of feature detectors and descriptors. Based on experimental results obtained from the Oxford dataset and the MSP dataset, we conclude that the combination of Harris detector and moment invariants presents the best balance between resource consumption and accuracy for semantic neighbor formation in VSNs.

1. Introduction

At the convergence of advances in computer vision, imaging, embedded computing, and sensor networks, Visual Sensor Networks (VSNs) emerge as a cross-disciplinary research field and are attracting more and more attention. A VSN is a group of networked visual sensors with image/video capturing, computing, and wireless communication capabilities powered by on-board batteries [1, 2]. From the perspective of a single node, it is built at low cost, resulting in limited sensing, processing, and communication capabilities. However, the low cost of single sensor nodes makes it possible to randomly deploy a large number of them, and through *collaboration*, some challenging high-level vision tasks can be achieved. The application of VSNs has spanned a wide spectrum of domains, including environmental surveillance, human behavior monitoring, and object tracking and recognition [3, 4].

Generally speaking, collaboration among sensors is achieved through sensing the same area from similar or different perspectives as well as exchanging/fusing their measurements, based on which a more panoramic understanding toward the sensing area can be built. In order to collaborate, sensors should firstly form clusters such that within each cluster, they would have similar measurements for the same scene. This cluster formation process is relatively easy for wireless sensor networks using scalar sensors, whose sensing range is normally omnidirectional, and the geographical closeness of scalar sensors implies the similarities among their measurements. However, for visual sensors, this process is a lot different. On one hand, the sensing range of a visual sensor, that is, the Field of View (FOV), is directional. On the other hand, visual occlusions ubiquitously exist. Hence, geographical closeness can no longer guarantee measurement similarity.

Figure 1 illustrates how directional sensing and visual occlusion can affect sensor clustering in a VSN. Four visual sensors, A, B, C, and D, are geographically close but have different orientations. The FOV of each visual sensor is indicated by the area spanned by the two radials originated

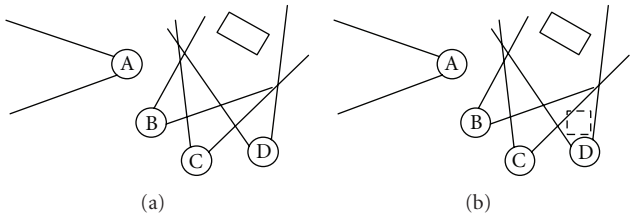


Figure 1: Illustration of how (a) directional sensing and (b) visual occlusion can affect sensor collaboration in a VSN and the formation of semantic neighborhood.

from the sensor. The rectangles indicate different objects with the object of dash rectangle occluding the view of the object of solid rectangle for sensor D. From Figure 1(a), we observe that although sensor A is geographically close to sensors B, C, and D, because it points to a different direction, the scene captured by A is completely different from those of B, C, and D. Therefore, for collaboration purpose, only sensors B, C, and D should form a neighborhood or cluster. From Figure 1(b), we further observe that although sensor D points to the same direction as sensors B and C, because of visual occlusion, the "content" of the image from sensor D would be completely different from that of B and C. Therefore, in this scenario, only B and C would have similar measurements to be able to form a cluster. We refer to the nodes with only geographical closeness as "g-neighbors" and "g-neighbors" who also possess *similar* measurements as "s-neighbors".

In this paper, we propose to form the semantic neighborhood for collaboration purpose based on image comparison among g-neighbors. Direct image comparison using raw image data is almost prohibited in VSNs due to their resource-constrained nature. The communication bandwidth in VSNs is typically low and is not equipped for frequent large amount of data transfer. In addition, due to the fact that the radio transceiver accounts for most of the energy consumption in all activities (e.g., sensing, processing, etc.) on a typical sensor node [5], the limited on-board battery could not support large amount of data transfer. Therefore, we propose feature-based image comparison for s-neighbor formation as image features can be more compact than the raw image in terms of data volume but still provide good representation of the image content. And we will examine 10 combinations of feature detectors and descriptors, including the detectors used in SIFT and SURF, wavelet-based detector, Harris detector and the descriptors used in SIFT and SURF, and moment invariants, to identify a suitable feature-based image comparison system for the resource-constrained VSN.

The organization of this paper is as follows. Section 2 reviews the related work in feature-based image comparison algorithms and neighbor selection for Wireless Sensor Networks; Section 3 explains the proposed feature-based image comparison method; Section 4 provides performance comparison based on two datasets and examines how image overlap affects the performance of the feature-based image comparison method; Section 5 concludes this paper.

2. Related Works

This paper studies the feature-based image comparison method for s-neighbor selection in visual sensor networks. Hence, this section reviews literatures from two perspectives, image comparison using features and neighbor selection in sensor networks in general. We also describe the difference between this paper and our previously published work. Moreover, we compare the feature detection/description techniques used for image comparison in VSNs and those developed for general Content-Based Image Retrieval (CBIR) systems.

2.1. Feature-Based Image Comparison. In computer vision terminologies, image feature is a mathematical description of the raw image. Generally speaking, comparing the image features is more effective and accurate than comparing raw images. Three steps are involved for feature-based image comparison: feature detection, feature description, and feature matching.

Feature detection is to detect feature points such as corner points, from the image. Harris and Stephens defined image corner points as pixels with large autocorrelation. They formulated the Harris matrix [6], H, to calculate the autocorrelation, which is in turn used to derive the corner strength measure, R,

$$H = \begin{bmatrix} \widehat{I_x^2} & \widehat{I_x I_y} \\ \widehat{I_x I_y} & \widehat{I_y^2} \end{bmatrix}, \qquad R = \det(H) - \kappa \mathrm{Trace}(H), \tag{1}$$

where $\hat{\cdot}$ denotes the average intensity within a window surrounding the pixel. Harris and Stephens used a circular Gaussian averaging window, so that the corner strength measure can suppress the noise and the calculation is isotropic.

Fauqueur et al. proposed a keypoint detection method that applies Dual Tree Complex Wavelet Transform (DTCWT) [7]. They built a "Keypoint Energy Map" for localizing keypoints from decimated DTCWT coefficients, and the keypoint scale parameter is determined by the gradient minima of the keypoint energy map in its vicinity. However, the DTCWT analysis is redundant, that is, the data volume of the decomposition structure is more than that of the original image and thus incurs additional computational cost and data storage.

In Scale Invariant Feature Transform (SIFT), Lowe proposed to use the Difference of Gaussian (DoG) operator to detect keypoints (blob-like features) in multiscale space [8]. DoG is an approximation of Laplacian of Gaussian (LoG) intending to speed up the computation, and the latter was used to detect edges in images. To filter out the edge responses from DoG, Lowe added a postprocessing step that relies on analyzing eigenvalues of the Hessian matrix in a similar way as Harris did in his detector. To speed up image feature detection, Bay et al. proposed the Speeded-Up Robust Feature (SURF) [9]. SURF relies on the Hessian matrix for interest point detection and approximates it by a series

of predefined templates suitable for fast convolution using integral images [10].

The second step in image comparison is to describe the detected feature points. The descriptor should be distinctive and invariant to noise, geometric, and photometric deformations. In SIFT, the feature descriptor is a histogram of the oriented local gradients around the keypoint. The bins of the histogram are stored in a vector typically with 128 entries. In SURF, the descriptor is composed of the Haar wavelet responses with typically 64 entries. Compared to the 128-entry or 64-entry feature descriptors, there is also the moment invariants feature descriptor which is of much lower dimension [11]. The "moment invariants" is a vector containing 7 entries and is invariant to translation, scale, and rotation changes.

The third step in image comparison is feature matching, both SIFT and SURF use Euclidean distance as a similarity measure because all entries in the feature descriptor have comparable magnitudes and thus can be properly normalized. For moment invariants, Mahalanobis distance is used.

2.2. Neighbor Selection in Sensor Networks. In sensor networks, a neighborhood is defined as a group of nearby sensors taking the same or similar measurements. Depending on the characteristics of sensing devices, the neighbor selection algorithms can be very different. In traditional sensor networks where scalar sensing is used, the sensors are generally omnidirectional, and their measurements are mostly related to the distance between the event point and the sensor, which implies that the closer two sensors stand, the more similar their measurements are. That is, the neighborhood is determined based on distances between sensors.

The concept of neighbor selection originated from sensor networks under the name "cooperative strategies" and was used to prolong the network lifetime by arranging an efficient sleep-wakeup selection among neighbor nodes [12]. This method puts some nodes within the neighborhood into the sleep mode for energy conservation purpose when its neighbors are capable of taking measurements. Chen et al. introduced a power-saving technique named SPAN which assumes that every node is aware of its active neighbors within a 2-hop distance and exchanges this active-neighbor information with its peers [13]. All nodes can thus use this information to decide which node to turn off. Ye et al. proposed a Probing Environment and Adaptive Sleeping (PEAS) protocol that puts the node into the sleep mode as long as possible if there is a nearby active node within its neighborhood [14]. Younis and Fahmy presented a Hybrid Energy-Efficient Distributed (HEED) clustering protocol to organize the ad hoc network in clusters and select cluster head based on node's power level and proximity to its neighbors [15]. Among all these studies, the neighborhood information is regarded as trivial since it is only determined by the geographical distance between node pairs.

On the other hand, visual sensors are usually directional sensing devices. Therefore, their measurement similarity depends not only on their distances, but also on the content of the images captured. For example, two surveillance cameras on the two sides of a wall pointing at opposite directions are geographically close, but they are not capturing the same scene and hence should not be defined as neighbors. In this paper, we use feature-based image comparison method for s-neighbor selection where the s-neighbors should satisfy *two* conditions, geographically close and having similar measurements. The second condition essentially imposes the orientation and occlusion-free constraints.

Medeiros et al. distinguished the differences between scalar sensor networks and camera networks and proposed a light-weight event-driven sensor clustering protocol for a wireless camera network [16]. But the focus of their work is in clustering protocol design, and a simple color histogram feature is used to find semantic neighbors.

2.3. Comparison to Our Previous Work. The feature-based image comparison framework in VSNs was first proposed as a conference publication by the authors [17]. In that work, we focused on the design of a feature description algorithm, Extended-SURF, for better representation of the feature points, which uses a 128-element vector to describe a feature point.

In this paper, we focus more on the comprehensive performance evaluation regarding various combinations of feature detectors and descriptors, in order to arrive at an appropriate design option suitable for the resource-constrained VSN.

2.4. Comparison to Features Used in Content-Based Image Retrieval. Content-Based Image Retrieval (CBIR) is a technique that uses the visual content of an image as the query to retrieve similar images from an image database. The visual content of each image is represented by *image signatures* built from various image features [18]. A group of local features, such as color-based features, texture features, shape features, and local invariant-based features are broadly adopted in CBIR systems, among which the local invariant-based features, including SIFT and SURF, are receiving much attention because of their robustness to scale and affine variances.

Although both a CBIR system and a VSN adopt image feature detection/description algorithms, their selection criteria are different. For a CBIR system, the feature detection and description can be performed offline, leaving less concern on the computational cost of the algorithms. For a VSN, all feature detection and description processes need to be carried out in real time within the network, demanding feature detectors/descriptors with low computational cost and the resulting image features to be compact and effective.

3. Feature-Based Image Comparison

In this section, we detail the three steps, that is, feature detection, feature description, and feature matching, in feature-based image comparison.

3.1. Feature Detection. We evaluate the performance of four feature detectors, including the feature detectors used in SIFT

and SURF, Harris corner detector, and an improved DWT-based detector.

The feature detection method in SIFT involves two steps [8]. The first step is to calculate the Difference of Gaussian of the image and find local extremes in scale space. This step is to detect all possible keypoints that bear blob-like patterns. The second step is to calculate the Hessian matrix at each keypoint location from step one and reject those keypoints if their Hessian matrices have two significantly different eigenvalues. Essentially this operation excludes the keypoints detected on edges.

The feature detection method in SURF detects the interest points by finding the local extremes of the determinant of Hessian matrix at each pixel. SURF utilizes integral images for fast convolution, and the Hessian matrix at each pixel is approximated by a series of predefined templates. Compared to SIFT, the computational efficiency of feature detection in SURF has been improved to a great extent. However, they still cannot compete with the Harris detector or the improved DWT-based detector.

The Harris matrix H contains only the first-order derivatives of each pixel, and the corner strength measure requires only the calculations of the determinant and the trace of a 2×2 Harris matrix.

Here, we propose a computationally light-weight corner detector based on Discrete Wavelet Transform (DWT) that is robust to both scale change and noise. The algorithm includes two steps, with the first step localizing the corner points and the second step determining the scale parameter for each corner detected. In the first step, the DWT coefficients computed through Fast Wavelet Transform are used to build a corner strength map at each decomposition level and these maps are upsampled by Gaussian kernel interpolation to the original image size. These interpolated corner strength maps are then summed to build the corner strength measure. The local maxima of the corner strength measure are detected as corners. In the second step, the scale of the corner point is calculated to eliminate small-scale corners on discrete edges. The scale of a corner point is defined as a function of the corner distribution in its vicinity. We provide a fast calculation of the scale parameters utilizing a Gaussian kernel convolution.

3.2. Feature Description. As discussed in Section 2.1, SIFT uses a 128-element vector to describe each feature calculated from the image, and SURF uses a 64-element vector. When there are many feature points detected, transmission of the feature vectors is nontrivial. At the extreme case, the amount of data transferred in feature-based image comparison can be larger than that in raw data-based comparison. To further improve the transmission efficiency, we resort to the moment invariants, that is, a 7-element vector, to represent each corner point detected.

The 7-element feature descriptor is invariant to rotation, translation, and scale, making it a good candidate to overcome the two shortcomings of Harris detector, that is, corner point displacement and variant to rotation. It is calculated based on the texture-based statistical descriptor moment. For a detailed definition and calculation of the invariant moments, please refer to [19].

3.3. Feature Matching. The feature-based image comparison is based on the assumption that if two feature points in different images are corresponding to the same real-world point, then the two feature descriptors should be "close" enough and they form a pair of matching features. If two images have enough number of matching features, we claim the two images are similar, implying they capture a similar scene.

There have been two popular distance metrics used, Euclidean distance and Mahalanobis distance. Mahalanobis distance is generally more accurate since it takes into account the covariance among the variables and is scale invariant. However, the calculation of the covariance matrix does incur additional cost. The closeness between the SIFT and SURF descriptors is evaluated using the Euclidean distance since all entries in the feature descriptor have similar scale and thus can be properly normalized. However, the Mahalanobis distance has to be adopted when evaluating the closeness among moment invariant descriptors as the scale of the 7 entries in the descriptor tends to vary a lot.

4. Performance Evaluation

In this section, we perform thorough comparative study on feature-based image comparison with various combinations of the four feature detectors (i.e., SIFT detector, SURF detector, Harris detector, and the improved DWT-based detector) and the three feature descriptors (i.e., SIFT descriptor, SURF descriptor, and the moment invariants).

4.1. Datasets. The performance is evaluated on two image datasets. The Oxford dataset contains images from different scenes and is for general image feature evaluation purpose [20]. The same dataset was used in [21] to provide a performance evaluation of local descriptors. The second dataset is gathered from a small-size VSN testbed constructed in the Advanced Imaging and Collaborative Information Processing (AICIP) lab at the University of Tennessee. The images from the second dataset are taken in an office setup. Unlike the Oxford dataset, where images from different groups bear distinctive differences, all images taken from the small-size VSN have similarity to some degree, making feature-based image comparison more challenging.

To further investigate into how the degree of overlap between images affect the performance of feature-based image comparison, we adopt another dataset, the Columbia Object Image Library (COIL-100) set. The COIL-100 is a dataset containing images of 100 objects. Each object was placed on a motorized turntable against a black background, and the turntable rotated 360° to vary the object's pose with respect to a fixed camera. An image was taken every time the turntable rotated 5° (separation angle), generating 72 images for each object with different poses. The image overlap can be evaluated by the separation angle between two images, the smaller the separation angle, the more overlap they have.

4.2. Metrics. We use five metrics for performance evaluation purpose, that is, number of feature points detected, number of bytes transferred, computation time, recall, and precision. The first three metrics are straightforward and used to evaluate resource consumption of the algorithm. The latter two reflect algorithm performance compared to the ground truth.

Let True Positive (TP) represent the detected correct matching image pairs, False Positive (FP) the unmatching image pairs but have been detected as matching, and the False Negative (FN) the matching image pairs but have been detected as unmatching. Then *Recall* is defined as the ratio between the number of TPs and the total number of TPs and FNs. *Precision* is defined as the ratio between the number of TPs and the total number of TPs and FPs as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

The higher the recall, the less "misses" compared to the true match; and the higher the precision, the less "false alarms".

We also propose a so-called *rank-based score* function, where the performance of each detector/descriptor combination is ranked in terms of data volume, recall, and precision. Since there are altogether 10 combinations of detector/descriptor, the rank for each of the three metrics is a number within the (1, 10) range with 1 being the best and 10 the worst. By taking the summation of the three ranks (i.e., data volume, recall, and precision) for each detector/descriptor combination, a single score (in the (3, 30) range) can be obtained, making it more convenient to evaluate detector/descriptor combinations across different performance metrics.

4.3. Window Size. A free parameter in the feature-based image comparison system is the size of local area (or window size) used to calculate local statistics around a detected feature point. We use the default window size from the SIFT and SURF descriptors, which is 16×16 and 20×20, respectively. The window size for calculating the invariant moments is determined through empirical study. We use 6 "graffiti" images from the Oxford dataset, formed into 15 image pairs whose perspective change can be represented by a homography matrix. The homography matrices are provided as the ground truth, so that we can calculate the correspondence between feature points. The recalls and precisions are averaged over resutls from the 15 image pairs.

Figure 2 shows the *recall* and *precision* as functions of the half window size for calculating the invariant moments in feature description. We observe the consistent trend between "recall versus window size" and "precision versus window size" and that different window size does affect both the *recall* and *precision*. We choose to use 23 as the half window size because it provides a good enough recall/precision rate. While increasing the window size to, for example, 38, it would improve the performance to some extent, it would also incur more computational cost.

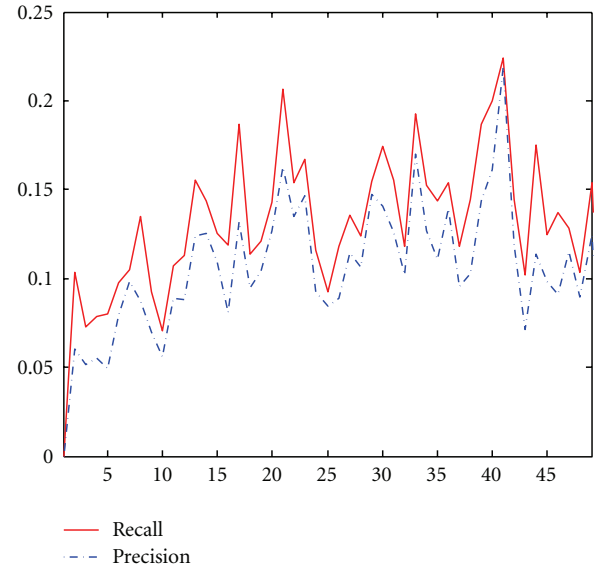


Figure 2: Precision and recall as functions of the half window size for calculating the invariant moments in feature description.

4.4. Finding Correspondence between Feature Points. When comparing the similarity between two images, the first step is to find the correspondence of detected feature points (or features) across the two images. The second step is to count the number of corresponding feature pairs. If this number is large enough, the two images are claimed to be "similar".

Given two sets of feature descriptors generated by analyzing two images, for each feature descriptor in the first image, the distances between this descriptor and all feature descriptors in the second image are calculated. If the ratio between the smallest distance and the second smallest distance is below a threshold, the feature pair with the smallest distance is claimed to be a matching feature pair or that the two feature points that these two descriptors are describing correspond to each other. The threshold has been experimentally determined to be 0.3. Both SIFT and SURF use this threshold in their implementations.

4.5. Feature-Based Image Comparison on the Oxford Dataset. The Oxford image dataset contains 8 groups of images from 8 totally different surroundings. Each group has 6 images taken from the same scene but under some variations. Figure 3 shows examples of the Oxford dataset, where six image transformations are presented, including blur (Figures 3(a) and 3(b)), viewpoint change (Figures 3(c) and 3(d)), scale and rotation changes (Figures 3(e) and 3(f)), illumination change (Figure 3(g)), and JPEG compression (Figure 3(h)).

We apply various combinations of feature detector and descriptor on the Oxford dataset. Firstly, image features are detected and descriptors are calculated from the 48 images, out of which one set is selected as a query. Secondly, the query feature set is compared to other 47 record feature sets and records that have large enough number of matching features with the query set will be returned as retrievals. Assume the maximum number of matching features between

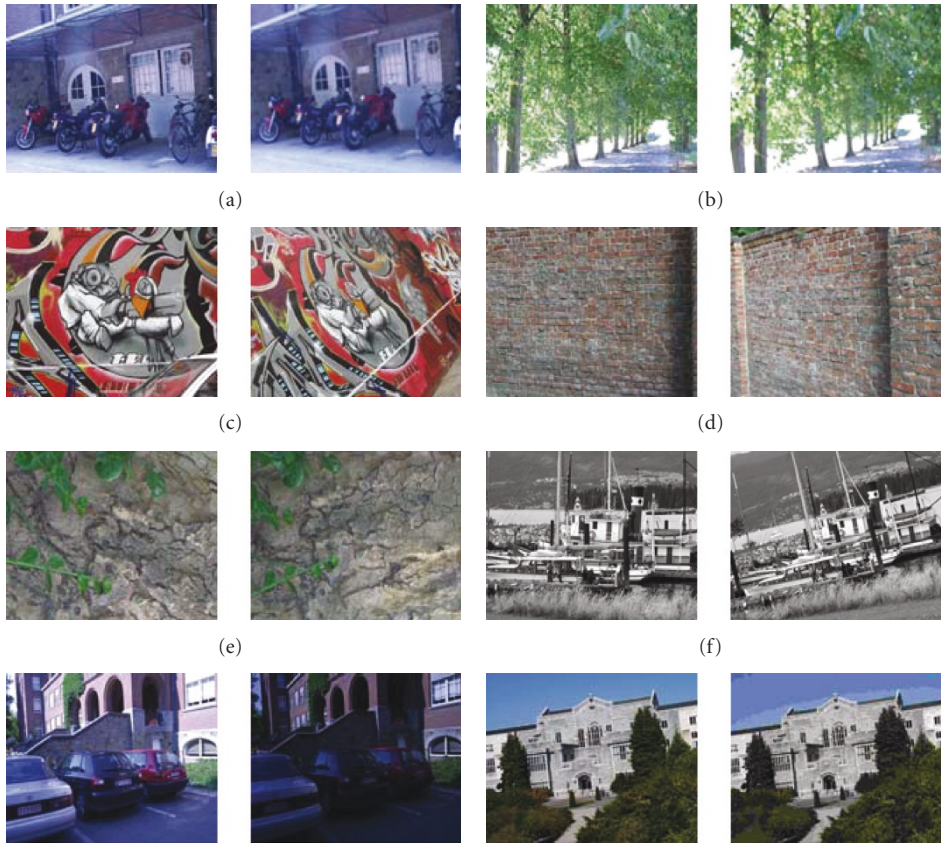


Figure 3: The Oxford dataset. Examples of images from each group: (a) and (b) blur; (c) and (d) viewpoint change; (e) and (f) zoom plus rotation; (g) illumination change and (h) JPEG compression.

the query and the record is N, then we set up the threshold as $N \times 0.85$, such that any record having matching features more than this threshold will be "claimed" as a retrieval. Thirdly, we compare the retrieved feature set(s), or image(s), to the ground truth and compute the TP, FP, FN, TN, recall, and precision. The ground truth is based on the fact that all images from the same group are similar but not otherwise. We apply a 48-fold cross validation such that every image from the dataset will be used as the query once, and we average out all the results, which are presented in Table 1.

From the aspect of computational speed, Harris detector takes 0.2346 seconds to detect 442 corners on each image; the improved DWT-based detector takes 0.3727 seconds to detect 1145 corners on each image; SIFT takes 3.5855 seconds to detect 5938 keypoints on each image; SURF takes 0.8211 seconds to detect 1478 interest points on each image. Note that the Harris detector and the improved DWT-based detector are implemented in Matlab while the SIFT [22] and SURF [23] implementations are in C code. Considering the difference in implementation efficiency between Matlab and C, the speed between the Harris detector and the others would differ even more.

From Table 1, we observe that in terms of resource consumption, SIFT is the worst, considering it actually creates even larger volume of data than the raw image, with the slowest computational speed. Except for the combination of the DWT-based detector and the SIFT descriptor, which generates comparable data volume as the raw image, all other combinations of feature detector/descriptor provide savings in data volume and therefore data transmission, especially those using moment invariants.

In terms of performance accuracy, SIFT and SURF provide highest precision although the recalls are low. Harris detector performs well when used together with SIFT or SURF descriptor, generating better balance between recall and precision. The DWT-based corner detector does not perform as well as the Harris detector when used with SIFT or SURF descriptor. This is because the DWT-based detector is more sensitive to discretization noise caused by image rotations which happen to be very common in the Oxford dataset (4 out of the 8 groups contain strong rotational variations, another 3 contain weak rotational variations).

In addition, by observing the fourth to seventh rows, we find out that the SIFT descriptor is superior to the SURF descriptor in terms of the precision. And from the eighth to eleventh rows, we observe that the moment invariants-based feature descriptors are all inferior to SIFT or SURF descriptors in terms of precision. Considering the

Table 1: Recall and precision of each feature detector/descriptor combination. The test run on the Oxford dataset and results in the 4th through 6th columns are for each image on average. For reference, the average raw image volume is 574.2 KB. Assume that a float type number takes 4 Bytes in storage. (The number in the parentheses in columns 4, 5, and 6 indicates the "rank" of the detector/descriptor combination in terms of the corresponding performance metric in that column. The "rank-based score" in the last column is the summation of the three ranks along the row.)

	Descr. length	No. of feature points	Feature data vol. (KB)	Recall	Precision	Rank-based score
SIFT	128	5938	3040.3(10)	0.3542(10)	1.0000(1)	21
SURF	64	1478	378.4(8)	0.3958(8)	1.0000(1)	17
Harris + SIFT descr.	128	442	226.3(6)	0.4583(4)	0.9375(3)	13
Harris + SURF descr.	64	442	113.2(4)	0.7083(3)	0.6970(5)	12
DWT + SIFT descr.	128	1145	586.2(9)	0.4375(5)	0.9063(4)	18
DWT + SURF descr.	64	1145	293.1(7)	0.4375(5)	0.5758(6)	18
SIFT detec.+ M. I.	7	5938	166.3(5)	0.4167(7)	0.5084(7)	19
SURF detec.+ M. I.	7	1478	41.4(3)	0.3958(8)	0.4029(8)	19
Harris + M. I.	7	442	12.4(1)	0.7292(2)	0.2014(9)	12
DWT + M.I.	7	1145	32.1(2)	0.8958(1)	0.1566(10)	13

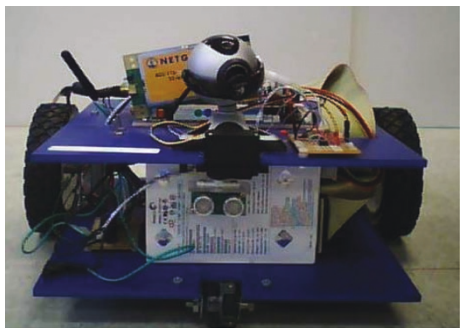


Figure 4: The mobile sensor platform (MSP).

definitions of recall and precision, a comparable recall value but extremely low precision value implies that there are too many False Positives (or false alarms).

Using the rank-based overall score, four combinations perform noticeably better than the others, including Harris + SIFT, Harris + SURF, Harris + Moment Invariants, and DWT-based + Moment Invariants. We see that 3 out of the 4 best combinations involve Harris detector, and moment invariants, as a descriptor, appear twice in the top 4.

4.6. Feature-Based Image Comparison on the MSP Image Dataset. This section reports the performance evaluation on the image dataset from a small size Visual Sensor Network. This VSN is composed of 12 Mobile Sensor Platforms (MSPs) built at the AICIP lab [24]. The structure of MSP is based on two 12×12 inch PVC sheets. The parts mounted on the platform include a Mini-ITX motherboard with 1.5 GHz processor, 1 GB RAM, 40 GB hard drive, 802.11 g wireless card, two 12VDC gearhead motors, and H-Bridge circuit to control the motors. The on-board sensing devices include a Logitech Quickcam 4000 Pro webcam and an ultrasonic range sensor, connected via USB and the parallel port, respectively. A complete assembled MSP is shown in Figure 4.

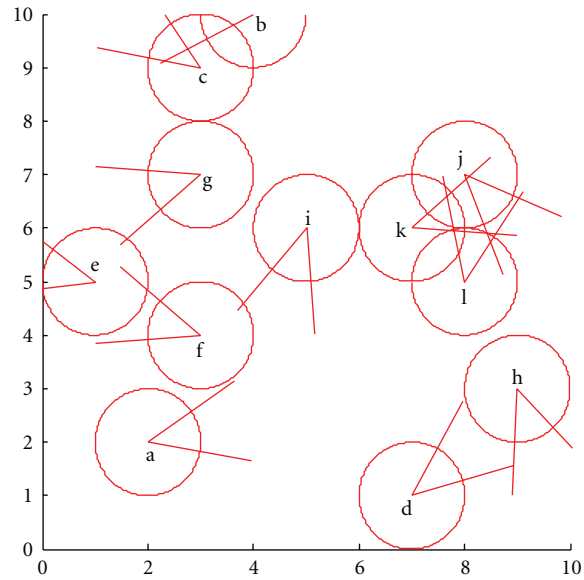
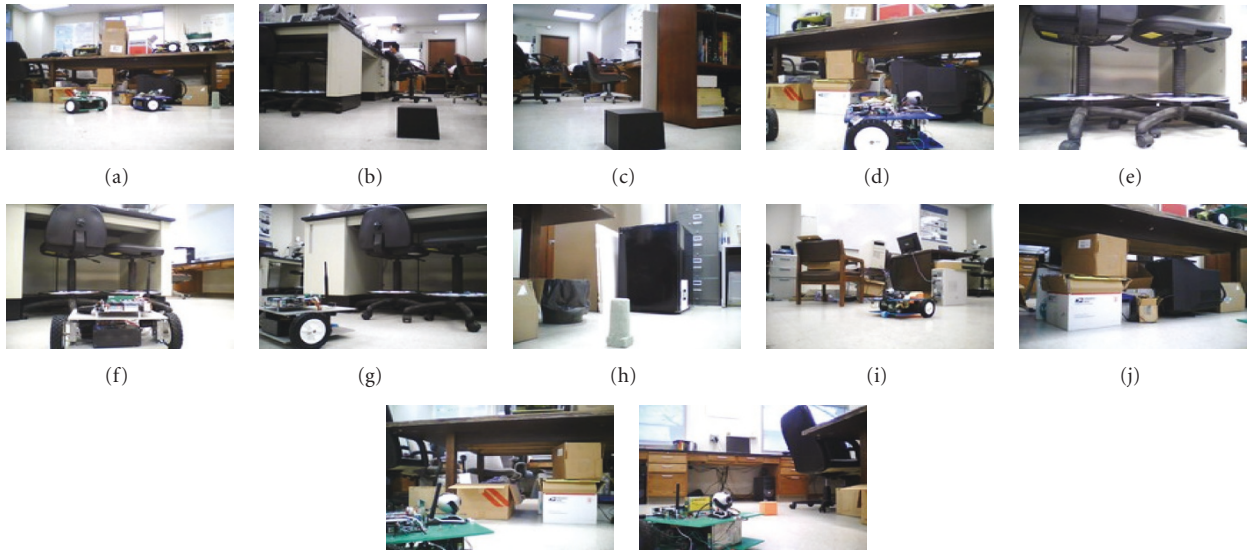


Figure 5: 12 MSPs in a 10-by-10 grid map. The letter within each circle indicates the image label in Figure 6.

In this experiment, we deploy 12 MSPs within a 10-by-10 grid area in an office setup. The deployment map is shown in Figure 5. The position and orientation of every MSP is randomly assigned.

Figure 6 shows the 12 images in this dataset. Different from the Oxford image dataset which contains images from totally different surroundings, the images in this dataset are all taken from the same office but from different viewpoints. Therefore, there are many common subregions (like the floor, the wall, and the ceiling) in images even when the cameras are not shooting at the same direction. This would result in a higher False Positive rate compared to the Oxford dataset. Another noteworthy problem is that for some MSPs, even they are shooting the similar scene, the overlapped



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l)

Figure 6: 12 images taken from 12 MSPs.

Table 2: Recall and precision of each feature detector/descriptor combination. The test run on the MSP dataset and results in the 4th through 6th columns are for each image on average. For reference, the average raw image volume is 19.2 KB. Assume a float type number takes 4 Bytes in storage. (The numbers in the parentheses in columns 4, 5, and 6 indicate the "rank" of the detector/descriptor combination in terms of the corresponding performance metric in that column. The "rank-based score" in the last column is the summation of the three ranks along the row.)

	Descr. length	No. of feature points	Feature data vol. (KB)	Recall	Precision	Rank-based score
SIFT	128	139	71.2(10)	0.4000(3)	0.3077(4)	17
SURF	64	72	18.4(6)	0.4500(1)	0.5294(1)	8
Harris + SIFT descr.	128	39	20.0(7)	0.3500(6)	0.3043(5)	18
Harris + SURF descr.	64	39	10.0(5)	0.4500(1)	0.1915(9)	15
DWT + SIFT descr.	128	87	44.5(9)	0.4000(3)	0.3636(3)	15
DWT + SURF descr.	64	87	22.2(8)	0.2500(8)	0.2000(7)	23
SIFT detec.+ M. I.	7	139	3.9(4)	0.1500(9)	0.2000(7)	20
SURF detec.+ M. I.	7	72	2.0(2)	0.1500(9)	0.1071(10)	21
Harris + M. I.	7	39	1.1(1)	0.4000(3)	0.4211(2)	6
DWT + M. I.	7	87	2.4(3)	0.3000(7)	0.2609(6)	16

scene region among their images is quite small, due to the differences in viewpoints. This would result in a higher False Negative rate compared to the Oxford dataset.

We apply the 10 combinations of feature detector/descriptor on the 12 images, and every feature set will be selected as the query to be compared with the others. The ground truth is found from the knowledge of the deployment map, through which we would know which MSP is shooting at the same scene as other MSPs and therefore build the image matching relations.

Table 2 lists the results of resource consumption and performance accuracy. In terms of resource consumption, SIFT and DWT-based detector plus SIFT descriptor give us the worst result by generating larger data volume than the raw image. On the other hand, the moment invariants-based descriptor generates the most compact feature sets.

In terms of performance accuracy, as we expected, all combinations of feature detector/descriptor give us worse results compared to those generated from the Oxford dataset. As a reference, the most mature image features, SIFT and SURF, could only provide a very low precision of 0.3077 and 0.5294, respectively, with SURF providing the highest recall. By observing rows 4 through 7, we find out that SIFT descriptors are superior to SURF descriptors in terms of precision, which is consistent with the Oxford dataset. According to the eighth to eleventh rows, the moment invariants-based descriptors are still inferior to SIFT or SURF descriptors, but for this dataset, the margin is much smaller.

Figure 7: Images of object 5 from the COIL-100 dataset, at view angles 0°, 10°, ..., and 170°, respectively.

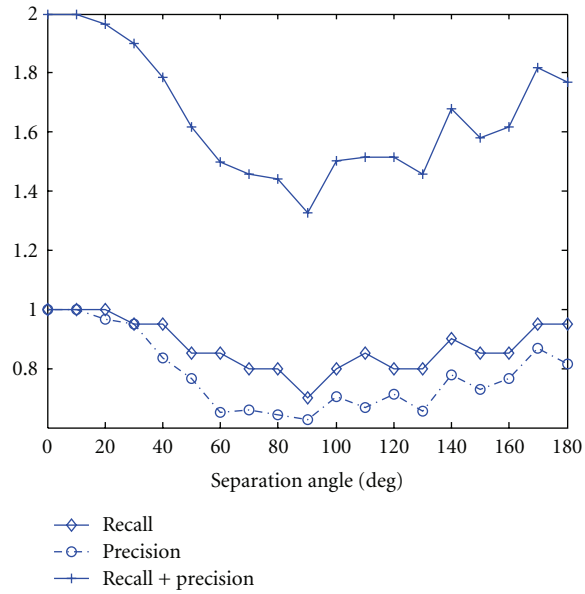


Figure 8: Recall and precision as a function of image overlap (measured by separation angle).

For detectors, Harris detector and DWT-based detector are comparable in terms of recall and precision. This conclusion is different from that of the Oxford dataset, because in the MSP dataset, all images are taken with MSPs laying down on the ground, and there is little rotational variation in images.

Measured from the rank-based overall score, Harris detector plus Moment Invariants prevails and SURF ranks the second.

4.7. Feature-Based Image Comparison on the COIL-100 Dataset. By comparing Tables 1 and 2, the recall and precision rates for the MSP dataset are severely worse than those for the Oxford dataset. Although both datasets contain image pairs having similar scenes, the degree of overlap in the similar image is different. The similar images in Oxford dataset have much more overlap than those in the MSP dataset. The third experiment is designed to reveal the relationship between the amount of image overlap and the algorithm performance in terms of recall and precision rates.

We select the first 10 objects from the COIL-100 dataset, one of which at different view angles is shown in Figure 7. For each object, we use two images: the first image is the one taken at angle 0° and the second is the one taken at angle θ. For the first object, we use its image taken at angle 0° as a reference and compare it to the rest 19 images using feature-based image comparison, recording the recall and precision rates. The Harris detector plus the SIFT descriptor combination is used in this experiment because it demonstrated good performance in the Oxford dataset. This process is repeated for all 10 objects and the average is taken. Then we vary the θ value from 10° to 180° with a step size of 10° and plot the results in Figure 8.

The recall and precision, as well as their summations, are plotted separately in Figure 8. Recall and precision are two contradictory indices in that when other conditions are the same, lower recall rate would result in higher precision rate and vice versa. Therefore, their summation is used to illustrate the general trend of the performance degradation. We use the summation instead of the average of recall and precision to provide an offset, so that the curve can be better visualized. The summation curve stays stable when there is a 10° separation angle between the two images of the same object; when the separation angle increases to 40°, the curve drops to 11%. The curve drops to its minimum when the separation angle is 90°, with a performance degradation of 34%, after which the performance starts to improve with the curve climbing back up.

We conclude from Figure 8 that the feature-based image comparison method (specifically, using Harris detector plus SIFT descriptor) can perform well (within 11% degradation) as long as the similar images have a certain degree of overlap that corresponds to a 40° separation angle in viewpoints. But as the amount of overlap decreases, the performance drops to 34%. Note that these measurements have not considered the factor of distance variations between the camera and the object, which would further reduce the performance. The performance improvements beyond the separation angle of 90° are due to the fact that the objects used in our experiment all show circular symmetry.

5. Discussions

A Visual Sensor Network environment is a more challenging setup for feature-based image comparison algorithms. First of all, it poses difficulties for image comparison itself by having to differentiate images of much similarity. Second, it requires the algorithms to be computationally of light weight, so that the low-end processor can afford running it. Finally, it requires the feature set to be compact enough, so that the transmission overhead is low.

From our experiments, although mature image features, SIFT or SURF, could perform well on general image comparison problems, such as the Oxford dataset, their performances degrade severely in the VSN environment, where images have high similarity. Moreover, their computational

burden and large volume of feature sets are strictly prohibitive for the VSN. On the other hand, simple feature detectors, like the Harris detector or the proposed improved DWT-based detector, combined with the compact moment invariants-based feature descriptor, show their advantages in terms of low resource consumption and comparable performance over the mature image features.

The low performance accuracy for feature-based image comparison in VSNs is due to the fact that the images have high similarities as the common background exists in all images. Therefore, one way to increase the performance accuracy is to perform background removal techniques prior to image feature detection.

6. Conclusion

The paper studied the feature-based image comparison methods for Visual Sensor Networks where the resources are stringent. The purpose of this study is to cluster visual sensors into different semantic neighborhoods to facilitate collaboration. Four image feature detectors, Harris detector, the improved DWT-based corner detector, SIFT detector, and SURF detector, as well as three image feature descriptors, SIFT descriptor, SURF feature descriptor, and moment invariants-based feature descriptor were examined, and different detector/descriptor combinations were tested on two image datasets. Although for general image comparison problems SIFT and SURF provide better performance in terms of accuracy, our experiment showed that for VSNs, where all images have high similarity, simple feature detectors, like Harris detector, DWT-based detector, and compact feature descriptors, like the moment invariants-based descriptor, can provide comparable performance in terms of accuracy but much superiority in terms of resource consumption. Overall, the Harris detector plus the moment invariant provides the best balance between resource consumption and performance accuracy.

Acknowledgment

This work was supported in part by National Science Foundation under Grant NSF ECS-0449309 and NSF CNS-1017156.

References

[1] D. Kundur, C.-Y. Lin, and C.-S. Lu, "Visual sensor networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 21515, 2007.

[2] S. Soro and W. Heinzelman, "A survey of visual sensor networks," *Advances in Multimedia*, vol. 2009, Article ID 640386, 21 pages, 2009.

[3] K. Obraczka, R. Manduchi, and J. J. Garcia-Luna-Aveces, "Managing the information flow in visual sensor networks," in *Proceedings of the 5th International Symposium on Wireless Personal Multimedia Communications*, vol. 3, pp. 1177–1181, 2002.

[4] A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa, "Object detection, tracking and recognition for multiple smart cameras," *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1606–1624, 2008.

[5] D. Estrin, "Tutorial on wireless sensor networks: sensor network protocols," in *Proceedings of the 8th Annual International Conference on Mobile Computing and Networking (MobiCom '02)*, 2002.

[6] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, pp. 77–116, 1988.

[7] J. Fauqueur, N. Kingsbury, and R. Anderson, "Multiscale keypoint detection using the dual-tree complex wavelet transform," in *Proceedings of IEEE International Conference on Image Processing*, pp. 1625–1628, 2006.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] H. Bay, T. Tuytelaars, and L. van Gool, "SURF: speeded up robust features," in *Proceedings of the 9th European Conference on Computer Vision (ECCV '06)*, vol. 3951 of *Lecture Notes in Computer Science*, pp. 404–417, 2006.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, pp. I511–I518, December 2001.

[11] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, pp. 179–187, 1962.

[12] D. M. Blough and P. Santi, "Investigating upper bounds on network lifetime extension for cell-based energy conservation techniques in stationary ad hoc networks," in *Proceedings of the Annual International Conference on Mobile Computing and Networking (MOBICOM '02)*, pp. 183–192, Atlanta, Ga, USA, September 2002.

[13] B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, "Span: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks," *Wireless Networks*, vol. 8, no. 5, pp. 481–494, 2002.

[14] F. Ye, G. Zhong, J. Cheng, S. Lu, and L. Zhang, "PEAS: a robust energy conserving protocol for long-lived sensor networks," in *Proceedings of the 23th IEEE International Conference on Distributed Computing Systems*, pp. 28–37, May 2003.

[15] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[16] H. Medeiros, J. Park, and A. Kak, "A light-weight event-driven protocol for sensor clustering in wireless camera networks," in *Proceedings of the 1st ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '07)*, pp. 203–210, Vienna, Austria, September 2007.

[17] Y. Bai and H. Qi, "Redundancy removal through semantic neighbor selection in visual sensor networks," in *Proceedings of the 3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '09)*, pp. 1–8, Como, Italy, August-September 2009.

[18] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, article 5, 2008.

[19] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 2007.

[20] http://www.robots.ox.ac.uk/~vgg/research/affine/.

[21] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[22] http://www.cs.ubc.ca/~lowe/keypoints/.
[23] http://www.vision.ee.ethz.ch/~surf/.
[24] C. Beall and H. Qi, "Distributed self-deployment in visual sensor networks," in *Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision (ICARCV '06)*, pp. 1–6, December 2006.