SELF-ORGANIZAITON FOR IMAGES FROM A MOVING CAMERA

Yanpeng Cao and John McDonald

Department of Computer Science
National University of Ireland, Maynooth, Ireland
{ycao, johnmcd} @cs.nuim.ie

ABSTRACT

Given a set of unsorted views captured in a wide area, an effective solution is proposed for image self-organization. The method starts with an initialization step where a small number of key frame pairs are selected to set up a global reference. Given a query image we automatically relate it to the existing key frames based on their pair-wise similarity evaluation. Four major enhancements are made in this step to achieve better performance. Firstly, a recently developed technique, SURF, is applied for robust feature detection. Secondly, an efficient coarse-to-fine matching strategy is implemented. Thirdly, an improved global representation is defined over each image for accurate and fast similarity evaluation. Finally, the method is constantly updated by adding more query images. Experiments were carried out to evaluate the performances of image self-organization by using a large number of images captured from our university's campus.

1. INTRODUCTION

Structure from Motion (SFM) is an active research topic in computer vision and has been extensively studied in last two decades [3], [5], [6]. For SFM estimation to be robust and accurate, at least two spatially related images are needed. In a time-indexed video, enough spatial continuity can be guaranteed due to the short time interval between frames (e.g. 40 ms). Unfortunately, this is not the case for a sparse set of images taken from arbitrary view points in a wide area.

In this paper, we provide a robust technique for the self-organization of a large number of images without spatial ordering. Without extra information from other sensors or human interaction, this is not an easy problem to solve. An effective solution to this problem is proposed through the use of Speeded-Up Robust Feature (SURF) [1]. Unlike other detectors (e.g. the Harris corner detector), SURF not only defines the position of a feature point but also provides a 64-dimensional descriptor which allows reliable matching and comparison on local basis. Given a set of SURF descriptors for each image, their spatial relationship is robustly determined based on similarity evaluation. An improved global representation is defined over each SURF descriptor set for accurate and efficient similarity evaluation in this step. Also, a good strategy is proposed to use fewer SURF features detected in a coarser level for quick image matching. Finally, constant query images will be utilized to improve the results of image self-organization.

Our method consists of two major steps: initialization and image self-organization. Details of each step will be described in Section 2 and 3, respectively. Experimental evaluations are reported in Section 4. Finally, the conclusion and future research direction are given in Section 5.

2. INIALISATION

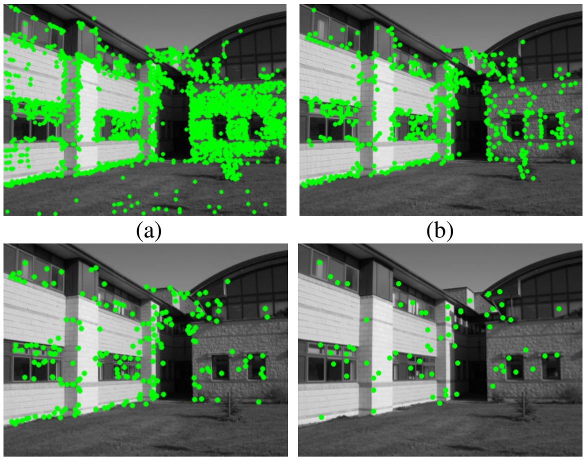
To achieve robust Structure from Motion (SFM) results, a sparse set of well-conditioned key frame pairs are firstly selected to set up a global reference. In [5], a good guideline was provided for key frame selection as follows: (1) baseline between a pair of key frames is large enough to recover their epipolar geometry; (2) sufficient feature correspondences are obtained; (3) Image-based distance $b = median(d(H\mathrm{x}, \mathrm{x}'))$ is large so that the selected frames are not near degeneration, where H is the best fitting homography that transfers the first image to the second one.

Then, Speeded-Up Robust Features (SURF) are used for robust feature detection in key frames. For each detected feature point, SURF gives a 64-dimensional descriptor which is invariant to scale, illumination, and rotation changes. It allows us to perform robust image matching and registration. Another impressive advantage of SURF is its low computational cost due to the use of integral images. In our evaluations, more than 2000 features and descriptors can be computed in a 2272×1704 image within 500ms. For similar amount of work, another popular feature detector, SIFT [4], takes more than 2 seconds. For more details about SURF and its evaluations, please refer to [1].

In our work, detected SURF descriptors are used for pair-wise similarity evaluation and image self-organization. To achieve better efficiency, a coarse-to-fine strategy is applied for image matching. It's noted that fewer SURF features are detected in lower resolution images while some distinctive features constantly appear in each level (see Fig. 1). Therefore, we propose to use the fewer features detected in a lower resolution for fast similarity evaluation. Image matching is initially performed at the lowest resolution and

978-1-4244-3610-1/09/$25.00 ©2009 IEEE.

proceeds through to the higher resolution until sufficient similarities are found. In this way, if enough evidence is gathered at a lower resolution, the self-organization will stop there to save computational cost. The effectiveness of this coarse-to-fine matching scheme is evaluated in Section 4.



(a) (b)

(c) (d)

Fig. 1. Detected SURF features in images of different resolutions (a) 2482 points are found in the original image, (b) 886 points in 1/2 resolution image, (c) 216 points in 1/4 resolution, and (d) 71 points in 1/8 resolution.

3. IMAGE SELF-ORGANISATION

Starting from a randomly captured image, the challenge is how to relate it to the selected key frames. More specifically, we need to identify the key frames it overlaps with. We solve this problem using a two-step approach. Firstly a global representation is defined over each image, and then image correspondences are determined based on the similarity evaluation of their global representations. The 64-dimensional SURF descriptor set is a good choice to represent an image [1]. However similarity evaluation based on SURF descriptor sets has two major drawbacks. First of all, it's very time consuming to search possible matches in two large sets of descriptors (more than 1000 descriptors are usually computed in an 800×600 image). Also, it's difficult to find a proper distance threshold to determine correct matches.

In this work an improved global representation is defined over an image as suggested by Grauman and Darrell [2]. We consider two sets of SURF descriptors $D_1 = \{d_1, \dots, d_{n1}\}$ and $D_2 = \{d_1, \dots, d_{n2}\}$ derived from two images, where d_i is a 64-dimensional descriptor. The global representation is defined over each descriptor set as:

$$\Psi(D_i) = [H_0(D_i), H_1(D_i) \dots, H_L(D_i)] \qquad (1)$$

where H_i is a histogram vector that records the number of descriptors fall into 64-dimensional bins (corresponding to the 64-dimensional SURF descriptor) of side length 2^i. The bins in the finest level 0 are small enough that each feature descriptor falls into its own bin, while all descriptors fall into one single bin at the coarsest level L. Then the similarity between two images is measured by comparing their corresponding histograms in different levels as:

$$S(D_1, D_2) = \sum_{i=1}^{L} \omega^i (\Upsilon_{i+1}(D_1, D_2) - \Upsilon_i(D_1, D_2)) \qquad (2)$$

where Υ_i is the overlap of histograms at level i, and ω^i is the weight coefficient which gives more credits to the overlap found in a smaller size bin.

Compared to the descriptor set D_i, $\Psi(D_i)$ is a better global representation. Firstly, it's derived from SURF descriptors thus it keeps all the good features of SURF such as invariance to rotation, scale, and illumination. Secondly, it offers better computational efficiency. Two sets of descriptors can be inserted in parallel into some pre-structured (not pre-created) multi-size bins, and their similarity is immediately obtained by counting the number of descriptors falling into the same bins. In this way, computational complexity is largely reduced from a set-to-set matching (polynomial dependence on the descriptor number) to two set-to-bin matching (linear dependence on the descriptor number). Finally, its pyramid matching procedure provides a good understanding of the similarity evaluation results between two sets of descriptors (e.g. number of matches found under different selection criteria). We use this information to select certain query images for updating the self-organization scheme.

A coarse-to-fine searching scheme is also implemented in our work for better efficiency as follows:

(1) Image is downsized to the lowest resolution (1/8 of the original size) and SURF features are computed;

(2) In this level, similarity evaluations are performed between the query image and each key frame. The query image will be related with the frame which has the largest similarity. It's noted that an image should always share high similarity with two adjacent key frames due to the spatial continuity. We imposed this spatial constraint to eliminate false results;

(3) If the similarity variation found is below a pre-defined threshold, we increase the image resolution and repeat the above steps until we reach the highest resolution. If the similarity variation is still too low, the query image is classified as an outlier (image covers sky, trees, road, or other buildings).

Over time a number of query images are selected for updating the self-organization scheme. If a query image cannot be correctly organized in its lowest resolution, its global representation is recorded. If lots of overlaps start to appear in bigger size bins (it means enough similarity can be only found when we loose the selection criteria), we also take that image in account. When a large number of images are recorded in the scheme, we will increase the sample rate based on image distribution for quick image matching and self-organization.

4. EXPERIMENTAL RESULTS

In this section we test the performance of the proposed methods for image self-organization. Two sets of images are used for evaluation. The first one is the standard Leuven Castle image sequence [7], which contains 27 consecutive frames (image resolution is 768×576 pixels) captured by a hand-held camera. The second data set was captured along the Engineering Building within the campus at National University of Ireland, Maynooth. Over 1000 frames (resolution is 2272×1704 pixels) are recorded in both indoor and outdoor environments. The raw inputs from camera are color images, while SURF implemented in our method only takes in the gray-level images. Some representative images are demonstrated in Fig. 2.

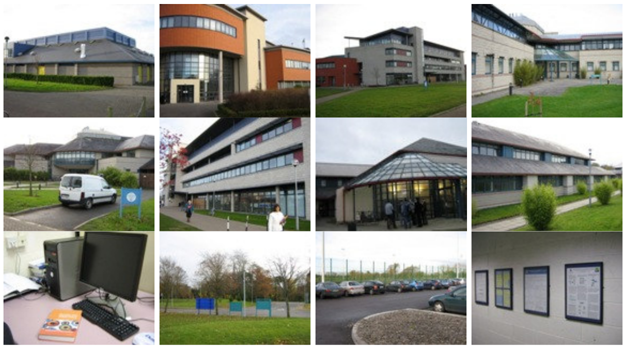


Fig. 2. Some representative frames. Many challenging images were captured for testing our method (e.g. large illumination changes, occlusions , irrelevant images).

Firstly, the method is tested in the Leuven Castle image sequence [7]. Frame 1, 9, 18 and 27 are selected as the query frames, while the rest are treated as key frames. Full-size images are downsized twice (1/2 and 1/4) and SURF features are detected at each resolution. The number of detected SURF features for each image in different resolutions is shown in Fig. 3.

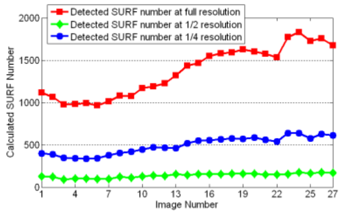
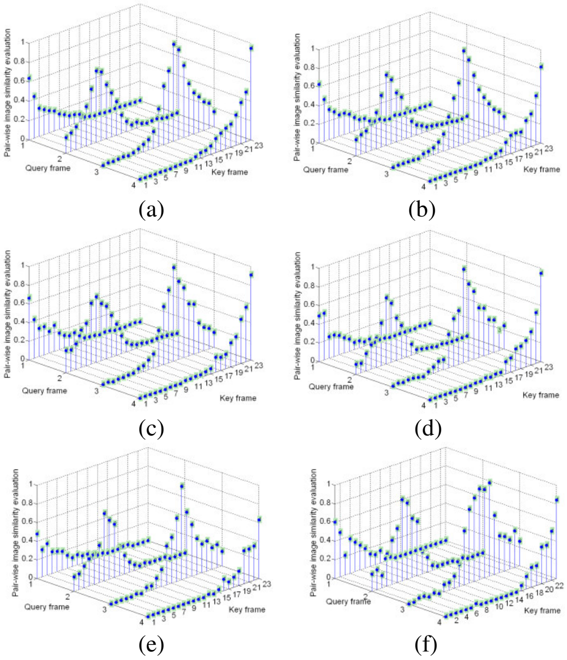


Fig. 3. The number of detected SURF features in different resolutions.

Given SURF descriptor set, image self-organization was performed using the method described in Section 3. The results were compared with the ones based on SURF descriptor set and Nearest Neighbour (NN) matching [4]. As shown in Fig. 4 , all the query frames can be correctly organized into the image sequence based on similarity evaluation even in the lowest resolution (e.g. query frame 2 (frame 9 in the sequence) has the highest similarities with key frames 7 (the 8th frame in the sequence) and 8 (the 10th frame in the sequence)). This is a quite good result, especially taking into account that displacements between consecutive frames are very insignificant. The accuracy of our method is better than the one based on SURF and Nearest Neighbour matching (see Fig. 4 (e) and (f) for comparison). Tab. I shows the processing efficiency. Since the implementations were run in Matlab, we set the processing time of our method for the lowest resolution image as the reference *T*. Significant computational costs were saved due to the coarse-to-fine matching scheme and a better global representation used in the method.



(a) (b)

(c) (d)

(e) (f)

Fig. 4. Similarity evaluation results comparison. Fig. 4 (a), (c), (e) - results of our method at resolution 768×576, 384×288 and 192×144, Fig. 4 (b), (d), (f) - results of the method based on SURF and Nearest Neighbour matching (NN) at resolution 768×576, 384×288 and 192×144.

Res. Method	768×576	384×288	192×144
Our method	~25*T*	~10*T*	*T*
SURF + NN	~200*T*	~50*T*	~1.4*T*

Tab. I. Processing time of similarity evaluation based on our method and SURF+NN at different image resolutions.

Next, we tested the method for realistic images captured in the campus. Full-size images (2272×1704 pixels) are downsized three times (1/2, 1/4, and 1/8) and SURF features are detected at each level. Three different experiments were organized as follows:

(1) Accuracy test without update. 500 inlier frames (images covering the Engineering Building) were captured at

different times during the day. Many images contained occlusions such as trees, vehicles, and pedestrians. Ground truth about where the images were taken is obtained through human observation and prior knowledge. Based on the criteria described in section 2, another 30 key frames were manually selected to set up a global reference. Then image self-organization was performed using the method described in Section 3. In this experiment the test was performed without any updating (no additional images were recorded during the test). Results are reported in Tab. II.

Res. Result	284×213	568×426	1136×852	2272×1704
Number of Positive	56	278	87	44
Correct Detection	50	253	75	40
Detection Percentage	11.2%	55.6%	17.4%	8.8%
Note: Accuracy rate =83.6%, False positive = 35 (7%)				

Tab. II. Image organization results without updating

(2) Accuracy test with update. This time we used same image dataset in the Test 1 for image organization, while allowing the method automatically store new images for updating. We divided 500 query frames into 2 subsets (250 frames each) and processed them sequentially to demonstrate the effect of updating. The result is shown in Tab. III. Obvious improvements were noticed after more images were saved for updating.

First 250 Images

Res. Result	284×213	568×426	1136×852	2272×1704
Number of Positive	78	97	44	21
Correct Detection	72	91	40	20
Detection Percentage	31.2%	38.8%	17.6%	8.4%
Note: 53 query images were recorded Accuracy rate =89.2%, False positive = 10 (4%)				

Second 250 Images

Res. Result	284×213	568×426	1136×852	2272×1704
Number of Positive	103	90	37	17
Correct Detection	99	87	35	17
Detection Percentage	41.2%	36%	14.8%	3.8%
Note: 26 query images were recorded Accurate rate =95.2%, False positive = 3 (1.2%)				

Tab. III. Image organization results with updating

(3) Robustness test in presence of outliers (images covering trees, sky, other buildings, and indoor scenes). 100 outlier frames were recorded, which contains images of trees (25 frames), road and vehicle (25 frames), indoor office and people (25 frames), other buildings (25 frames). The updated image self-organization method from the Test 2 was used for this evaluation. The result is reported in Tab. IV.

Type Result	Tree	Road/ Vehicle	Indoor/ People	Other buildings
Correct Detection	25	25	23	20

Tab. IV. The result of robustness test in presence of outliers

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose an effective self-organization method for a large set of unsorted images taken over a wide area. Several useful improvements were undertaken in the method and encouraging results were reported in terms of accuracy and efficiency. Next step we plan to further consider the constraint imposed by the epiploar geometry [3] to improve the robustness. Then we will evaluate the method in some more complex and larger scale environments. Finally the method will be applied as an important component in applications such as SFM recovery, intelligent navigation, and augmented reality.

ACKOWNLEDGE

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, “Speeded-Up Robust Features (SURF),” *International Journal of Computer Vision*, 110(3): 346-359, 2008.

[2] K. Grauman and T. Darrell, “The pyramid match kernels: Discriminative classification with sets of image features,” *IEEE Conf. on Computer Vision and Pattern Recognition*, 1458–1465, 2005.

[3] R. Hartley and A. Zisserman, “*Multiple View Geometry in Computer Vision*,” Cambridge University Press, 2004.

[4] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, 60(2): 91–110, 2004.

[5] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, R. Koch, “Visual modeling with a hand-held camera,” *International Journal of Computer Vision*, 59(3): 207-232, 2004.

[6] N. Snavely, S. M. Seitz, R. Szeliski, “Modeling the World from Internet Photo Collections,” *International Journal of Computer Vision*, 80(2):189-210, 2008.

[7] Leuven Castle Sequence, *http://www.cs.unc.edu/~marc/.*