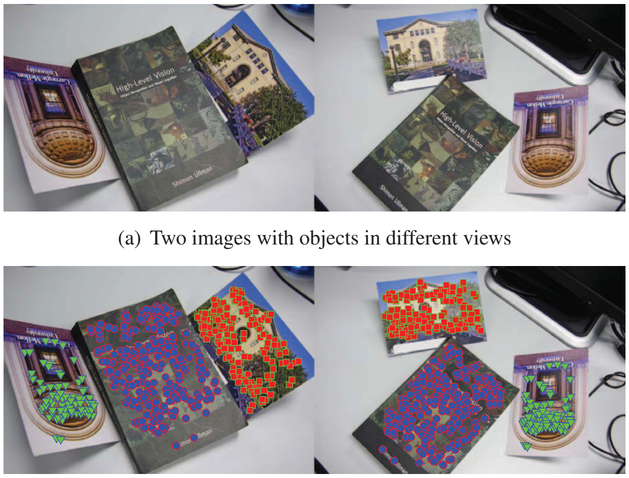
2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance

CLUMOC: Multiple Motion Estimation by Cluster Motion Consensus

Yinan Yu, Weiqiang Ren, Yongzhen Huang, Kaiqi Huang, Tieniu Tan
NLPR, Institute of Automation, CAS
95 Zhongguancun East Road, 100190, Beijing, China
{ynyu, wqren, yzhuang, kqhuang, tnt}@nlpr.ia.ac.cn

Abstract

In this paper, we present techniques for robust multiple motions estimation based on dual consensus via clustering in both the image spatial space and the motion parameter space. Starting from traditional Random Samples Consensus algorithm, we novelly propose the CLUster MOtion Consensus (CLUMOC) to extract robust motions. The proposed algorithm has two advantages: 1), instead of random samples, the CLUMOC employs clustering in initial sample selection, which can remove outliers from correct pairs of motion; 2), CLUMOC automatically decides the number of motions, by employing competition among motion and samples, that each motion needs to compete for matching pairs and each pair of matching competes for motions. The experimental results show that the proposed method is effective and efficient under various situations.



(a) Two images with objects in different views

(b) Correspondences

Figure 1. A view of the task of multiple motion estimation and the performance of the proposed method.

1. Introduction

The task of multiple motion estimation is to find the correspondences between two images with same objects in different view, as shown in Figure 1. It is a fundamental problem in computer vision. Given a set of initial matched points, which always full of noises (or ”outliers” in litera-tures), we are trying to find the correct correspondences, estimate individual motions and segment key points into different objects simultaneously. The matched points are usually extracted by detectors, e.g. Harris [4], SIFT [6], SURF [1] or HLSIFD [15].

The traditional image matching algorithm [14] or photo panorama system [2] mainly focuses on single object recognition. A popular algorithm for single motion estimation is the RANdom SAmple Consensus (RANSAC) [3] algorithm. The RANSAC algorithm iteratively estimates the parameters of the motion, which is usually a homography matrix or a fundamental matrix. There are two assumptions behind the RANSAC algorithm. The first one is that the data contains one subset of inliers at least, which are consecutive in the parameter space of the fitting model. The second assumption is that the inliers are not few in number comparing to the whole dataset. The RANSAC algorithm converges to the model which can be fit by most of the data due to the random sample strategy.

A drawback of RANSAC algorithm is that it can only tackle single motion estimation. When there are more than one motions, RANSAC may fail to extract either one. Considering the real situation that the whole matched set in object motion estimation may contain several motions, the traditional RANSAC algorithm is extended by iteratively estimating the most salient inlier from the rest of the data [10]. However, iteratively using RANSAC (or its extension) to estimate multiple motions has several drawbacks [9]. Reconsidering the problem of single and multiple motion estimation, the RANSAC and its various extensions, e.g. [11, 7, 12], are trying to estimate the maximum consensus subset or multiple different consensus motions in the parameter space.

As known in Functional Analysis, not only pair of matching can been seen as a point in spatial space, but also

978-0-7695-4797-8/12 $26.00 © 2012 IEEE
DOI 10.1109/AVSS.2012.19

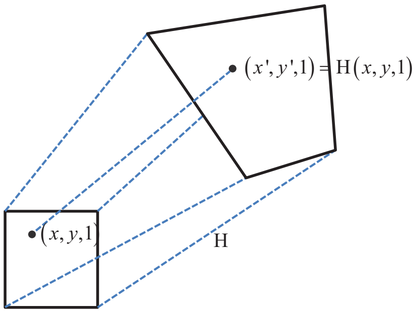


Figure 2. The homography transformation.

motions can be seen as points in motion parameter space. Take homography for example, the motion is usually formulated as a 3 by 3 matrix:

$$H_{ab} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{1}$$

and $h_{33} = 1$ usually. The homography describes the relationship between a plane in different view, as shown in Figure 2

In this paper, we extend the traditional RANdom SAmple Consensus (RANSAC) algorithm, considering the consensus in motion parameter space, and propose a CLUster MOtion Consensus (CLUMOC) algorithm to automatically extract multiple motions and remove the outliers. Different to the traditional RANSAC-like algorithms, our algorithm firstly cluster the matched points into several clusters (subsets). For each cluster, we estimate the model parameter. Then the models of each cluster are agglomerated adaptively together with two criteria: the similarity between models which inherits from the traditional RANSAC and the similarity between the data subsets. Every subset competes others to attract data into its cluster under an fixed toleration error threshold. The data in unsanitary clusters will be rejected or absorbed by other clusters. After several iterations, the clusters will converge and the multiple parameters are extracted from each of them. Our method have some advantages: 1) it simultaneously estimates the multiple motions and reject outliers; 2) with the data space clustering, the proposed method is robust to high number of outliers and errors caused by rough matching; 3) the proposed method can distinguish false transformations and similar transformations which usually caused by iterative RANSAC.

This paper is organized as follows: Section 2 will briefly introduce the traditional RANSAC algorithm. In Section 3, we will discuss the some issues of the data and the motion. Section 4 proposes our method. Section 5 shows the experimental results and Section 6 concludes this paper.

2. RANSAC Review

Let $P = \{p_i = (m_i = (x_i, y_i, 1), m'_i = (x'_i, y'_i, 1)|i = 1, 2, ..., n\}$ be a set of matched points between two images I and I' obtained by [6]. The initial matching method does not consider any spatial or transformation information, and always leads to a lot of outliers. Given a fixed model and its parameters, the RANSAC algorithm assumes that there is a subset (”inliers”) of the observed data, which fit a parameter of the model. The algorithm randomly samples a subset of the data, denoted as S . The given model then fitted by this subset. All data in P are tested by the fitted model and the well fitted data will reconstruct the subset S. This procedure iterates several times until the parameter of the model converges.

For a data point x in P and a model parameter H. The fitting error from the model H to the data point is defined as $L(H, x)$, which can be:

$$L_i = \|Hm_i - m'_i\| \tag{2}$$

for homography model and $| \cdot |$ denotes the L_2 norm here. And given a set of points P, the error between a model H and the set C is:

$$L = \sum_i L_i = \sum_i \|Hm_i - m'_i\| \tag{3}$$

3. Correspondence between points and motions

A matching pair can be seen as a new point $p = (x, y, x', y')$. And all pairs of matching consist a set $P = \{p_i|i = 1, ..., n\}$. Take a sub-set of P, the set S for example, we can extract a motion from S by minimize the following loss function:

$$\tilde{H} = f(S) = \arg\min_H \left(\sum_{p_i \in S} \|Hm_i - m'_i\| \right) \tag{4}$$

where $m = (x, y, 1)$ and $m' = (x', y', 1)$. It can be found that there is a mapping f from a set S to a motion H.

Given a motion H and a error criterion ε, we can find the matching pairs which belong to the motion:

$$S' = g_\varepsilon(H) = \{p| \|Hm - m'\| < \varepsilon, p \in P\} \tag{5}$$

where $p = (x, y, x', y')$, $m = (x, y, 1)$ and $m' = (x', y', 1)$. Therefore, the set S' is a subset of P and there is a mapping g_ε from a motion to a subset G.

The two mappings f and g_ε consist the correspondence between subset of matching pairs and motion parameter, as shown below:

$$S \overset{f}{\Rightarrow} H \overset{g}{\Rightarrow} S' \overset{f}{\Rightarrow} H' \cdots \tag{6}$$

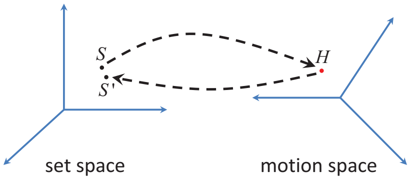


Figure 3. The correspondence between set of matches and motions.

Based on the correspondence, as shown in Figure 3, we can find that the subset of matching pairs and the point in motion parameter space are dual mathematically, therefore, a good motion estimation should consider both the consensuses in spatial space and the motion parameter space, that the sets S, S', *etc*. and the motions H, H', *etc*. should both consensus.

Before proposing our algorithm, we define several distance below:

- **Set to Set Distance**: Let S and S' be two subsets of P, the distance between them is defined as below:

$$d_s(S, S') = 1 - \frac{|S \cap S'|}{\min\left(|S|, |S'|\right)} \tag{7}$$

- **Set to Set Distance**: Let H and H' be two motions (homography matrixes) which estimated by the RANSAC algorithm from subset S and S' under threshold ε, the similarity of H and H' is defined as follow:

$$\begin{aligned} d_h(H, H') &= d_s(g_\varepsilon(H), g_\varepsilon(H')) \\ &= 1 - \frac{|g(H) \cap g(H')|}{\min(|g(H)|, |g(H')|)} \end{aligned} \tag{8}$$

Because the motion parameter space is not an Euclid space, therefore the motion distance is defined based on the set distance. The definitions benefit us to measure the distance (or similarity) between two subsets of matches or the motions.

4. The Proposed Method

Our algorithm can start with any local feature detector, descriptor and initial matching algorithm. In this paper, we use SIFT [6] and the key points are matched by the NN-DT (Nearest Neighbor Distance Threshold) algorithm [6, 13, 9]. Let the matched points pairs be $P = \{p_i = (x_i, y_i, x'_i, y'_i) | i = 1, 2, ..., N\}$. The task of our algorithm is to automatically estimate the motion(s) and reject outliers.

The original RANSAC algorithm assumes that the data points are uniformly distribute in the data space, and ignores the spatial relationship between the data points. Considering that the points matched in the 2D image plane, the points locating nearly should probably have same transformation. Thus, instead of random sample points from the whole set, our algorithm starts with clustering the data by location. We employ K-means clustering [16] to randomly generate M' subsets for T times, and get M = $M' * T$ subsets $A = \{S_1, S_2, ..., S_{M' * T}\}$. Using clustering instead of random sample has some advantages: 1) the samples in one cluster are more probably belonging to one consensus. 2) it can reject most of the outliers at the beginning with RANSAC. 3) it increases the convergency speed. Besides, the performance is robust to M' and T.

We calculate the homography matrix H for each subset and get a group of joint data $J = \{D_i\} = \{(S_1, H_1), (S_2, H_2), ..., (S_{M_1}, H_{M_1})\}$ where $M_1 \leq M$. We take away the subsets which fail to estimate a motion by RANSAC. Each of the joint point in J represents one homography transformation. Some of them may represent the same one. We call this step as estimation step (Step I).

The next step is similar to the RANSAC algorithm. We test every points x_i in C by each of the transformation calcuated in the Step I. The points x_j which fitting error by H_i is less than ε will gathered together to reconstruct the support data of H_i, denoted as $A' = S'_1, S'_2, ..., S'_{M1}$. The joint data set J is updated by the fitting test: $J' = \left\{{D_i}'\right\} = \left\{\left({S_1}', H_1\right), \left({S_1}', H_1\right), ..., \left({S_{M1}}', H_{M1}\right)\right\}$. We call this step as reconstruction step (Step II).

The major difference between our method and the traditional RANSAC algorithm is the following step, which we call the clustering step (Step III). The similarity measurement in the joint space depends on the two elements: S and H. We separate them to measure the similarities of the joint point. As we defined, for each point D'_i and D'_j in J', the set to set similarity of them is $d_s(S, S')$. The similarity is symmetric and we calculate every pair points in J' to generate the set to set similarity matrix $G_s = (d_s(S'_i, S'_j))$. Similar, we calculate the parameter similarity of every pair points in J' to generate the parameter similarity matrix $G_p = (d_h(H_i, H_j))$. Then two strategies are used to cluster the joint points in J':

1. If the similarity of H_i and H_j is less than a threshold ε_p, we will merge S'_i and S'_j together to generate a new data set Si''. By testing every two transformations in the joint data set, we get a new subset group $A'' = \{S''_1, S''_2, ..., S''_{M_2}\}$ and $M_2 \leq M_1$.

2. The second strategy is used on the subset group A''. If the similarity of S''_i and S''_j is lower than a threshold ε_s, S''_i and S''_j will be merged together and we get A'''. After the two strategies, the initial subset group is updated to the new subset group $A = A'''$.

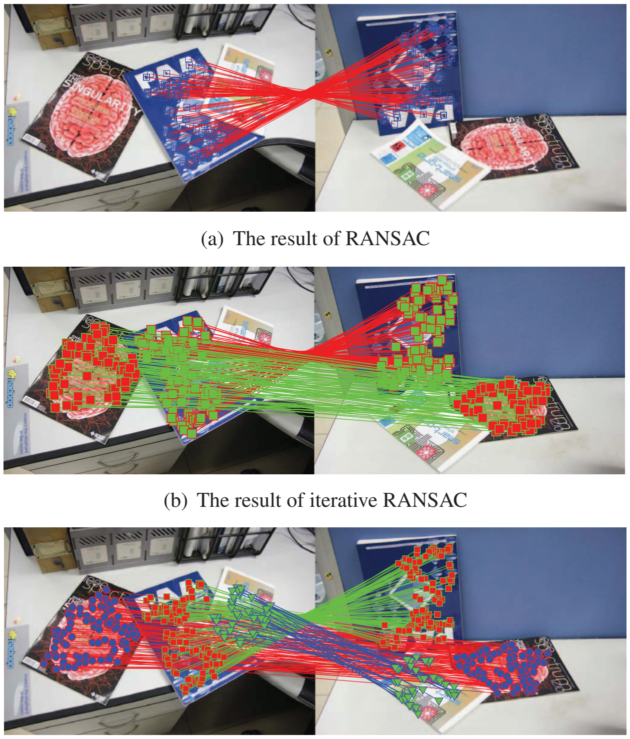
We iterate these three steps several times until the subset group A converges. The number of the element in the subset group A is the number of the motion. All the data points will be retested by the estimated transformations. The error between every data point and transformation is calculated using equation (2), noted as $E = (E(i, j))$, and $E(i, j) = L(H_j, x_i)$. If $E(i, j)$ multiplied by a threshold t is no greater than the error between x_i and any other transformations and $E(i, j)$ is less than the threshold ε, we say that the data point x_i distinctively belong to transformation H_j. The threshold is set to 1.2 empirically in our experiments. Finally, we reconstruct the motions by the data which distinctively belong to them.

5. Experimental results

In this section, we illustrate the stage results of the proposed method. The key points are extracted and matched following the algorithm in [6], implemented by Andrea Vedaldi and Brian Fulkerson [13] with the default parameters.

Firstly, We compare the proposed algorithm with the traditional RANSAC and the iterative RANSAC, as shown in Figure 4. The traditional RANSAC only extract the most salient motion from all pairs of matching. The other two motions are ignored. The iterative RANSAC can extract one more motion, while it still fails in another one. We analyze that two reason makes it failing: 1), the procedure of motion estimation is one by one, therefore there is no competition between motions to fight for matches; 2), because of no consideration of motion consensus, the RANSAC or iterative RANSAC algorithm is not robust to the matches including much noise. While, the proposed algorithm works well in this situation.

We show the important steps of the proposed algorithm. We use two images with three motions to illustrate the proposed method, as shown in Fig 5-(a). The initial matching results are full of errors and we can not distinguish good matches from the outliers, as shown in Fig 5-(b). For multiple motions problem, the traditional RANSAC algorithm can only extract one of them, and is always not stable because of the random sampling strategy. We apply RANSAC to the initial matching and get the result shown in Figure 5-(c). Only one motion is found and others are considered as ”outliers” and rejected. Using k-means cluster the data into several groups in the spatial space instead of random sampling is an important step in our algorithm. We cluster the data by the location of the points in the first image, as shown in Figure 5-(d). The points near each other always belong to same motion. We calculate the homography matrix H_i for each group by RANSAC algorithm. The clusters which contain less than 10 pairs of matches will be ignored. Then, all the matching pairs are tested by each of the homography matrix to find which points belong to a specific motion, as shown in Figure 5-(e) in different colors and markers. Here, a pair of matching points could belong two or more motions, we does not distinguish which one it should be in, but leave this problem in the end, as mentioned in last section. After several iterations, we get the final resutls, as shown in Figure 5-(f). All the three motions are estimated correctly and the outliers are all rejected by our algorithm. The similar motions are clustered together by the proposed method and the whole process is less than 40s including drawing markers and lines.



(a) The result of RANSAC

(b) The result of iterative RANSAC

(c) The proposed algorithm

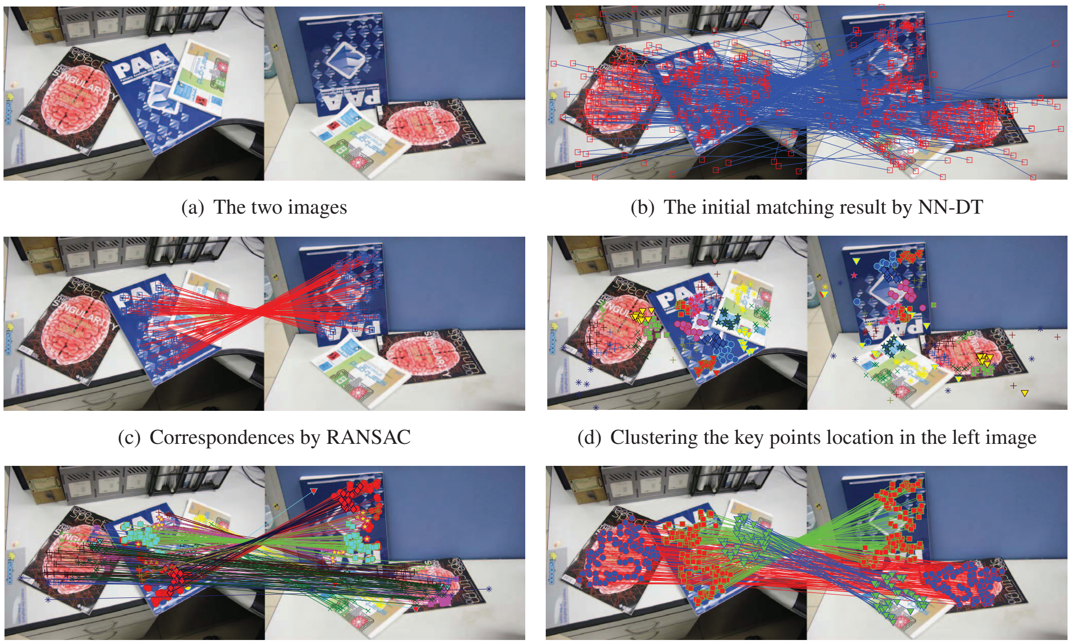
Figure 4. These figures are comparison among RANSAC, iterative RANSAC and the proposed algorithm.

We also test our algorithm in some images, as shown in Figure 6. The images in Figure 6-(a),(b) and (c) come from [8] and [9] respectively. Figure 6-(d) shows two images with two same objects in different view with two different objects. Our algorithm find the correct motions. Figure 6-(f) shows two images with three same objects in similar views. Our algorithm can distinguish them with effectively. The proposed method is more effective for objects with rich textures, as shown in Figure 6-(e).

The parameters in our method is set as follows: $M = 40$, $T = 2$, $\varepsilon = 0.001$, $\varepsilon_p = 1$, $\varepsilon_s = 0.2$, $t = 1.2$. We employ vlfeat [13] and the computer vision libs of Kovesi [5] to implemente our algorithm.

6. Conclusion

In this paper, we proposed a novel algorithm termed as CLUMOC (CLUster Motion Consensus). The motion we



(a) The two images

(b) The initial matching result by NN-DT

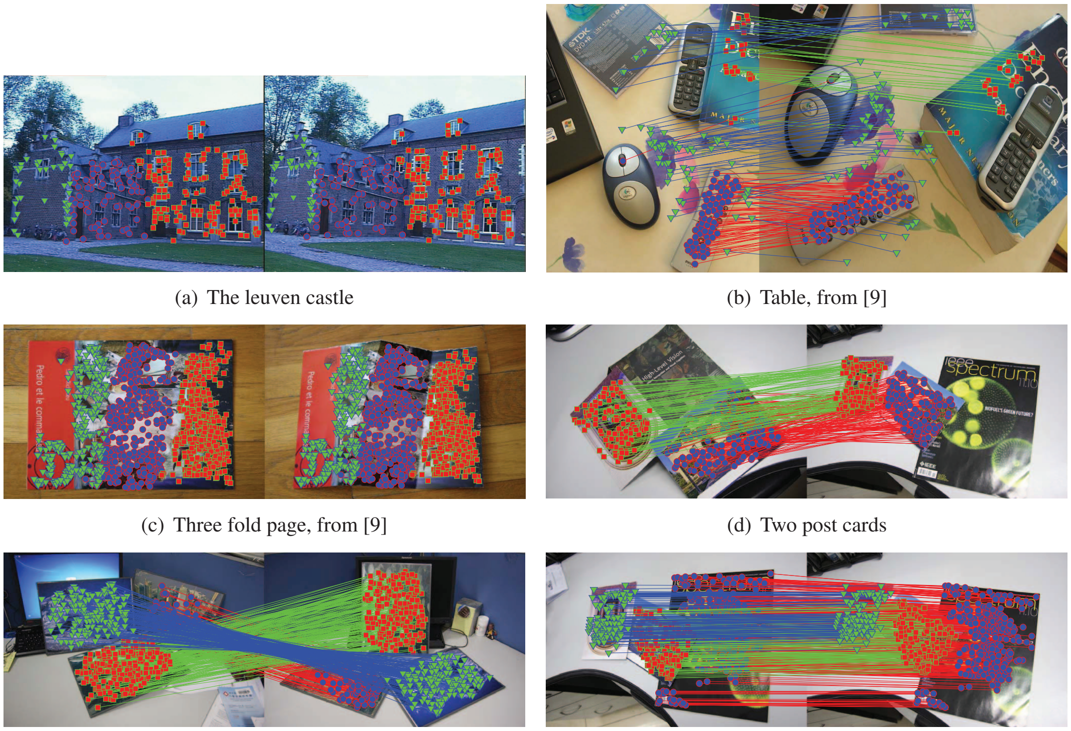
(c) Correspondences by RANSAC

(d) Clustering the key points location in the left image

(e) The motions estimated from the clusters

(f) The final result of the proposed method

Figure 5. These figures show the stage results of the proposed method. (This figure is best view in color and 400% magnification.)



(a) The leuven castle

(b) Table, from [9]

(c) Three fold page, from [9]

(d) Two post cards

(e) Three boards

(f) Two post cards and a book

Figure 6. These figures show the results of the proposed method. (This figure is best view in color and 400% magnification.)

extract is not only consensus in data space, but also in motion parameter space. We novelly consider the multiple motions estimation as bottom up clustering problem and employ two distance measurement to cluster the sub samples together: the similarity of motion in parameter space and the similarity of the set in data set space. We use K-means to cluster the samples instead the traditional ”Random Sample” strategy to make the algorithm more efficient and effective. The empirical study shows that our method is very effective in various situations. Our method is implemented for the homography model, and we will apply the proposed framework for the epipolar geometry model in future work.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

[2] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 1(74):59–73, 2007.

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981.

[4] C. Harris and M. Stephens. A combined corner and edge detection. *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[5] P. D. Kovesi. MATLAB and Octave functions for computer vision and image processing. Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia. Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[7] L. Moisan and B. Stival. A Probabilistic Criterion to Detect Rigid Point Matches Between Two Images and Estimate the Fundamental Matrix. *International Journal of Computer Vision*, 57(3):201–218, 2004.

[8] M. Pollefeys. Leuven castle image sequence. `www.cs.unc.edu/~marc/data/castlejpg.zip`.

[9] J. Rabin, J. Delon, Y. Gousseau, and L. Moisan. MAC-RANSAC: a robust algorithm for the recognition of multiple objects. *3D’PVT, 2010*, 2010.

[10] C. V. Stewart. Bias in robust estimation caused by discontinuities and multiple structures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:818–833, 1997.

[11] P. H. S. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78:2000, 2000.

[12] R. Tron and R. Vidal. A benchmark for the comparison of 3D motion segmentation algorithms. In *In CVPR*, volume 1, 2007.

[13] A. Vedaldi and B. Fulkerson. VLFEAT: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[14] Y. Yu, K. Huang, W. Chen, and T. Tan. A novel algorithm for view and illumination invariant image matching. *Image Processing, IEEE Transactions on*, 21(1):229 –240, jan. 2012.

[15] Y. Yu, K. Huang, and T. Tan. A harris-like scale invariant feature detector. *The ninth Asian Conference on Computer Vision*, 2009.

[16] D. Zeimpekis and E. Gallopoulos. TMG: A matlab toolbox for generating term document matrices from text collections. `http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/`, 2006.

7. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No.61135002,61175007), the National Basic Research Program of China (2012CB316302), the Strategic Priority Research Program of Chinese Academy of Sciences (XDA06030300). Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation (Y2U1011MC1).