3D Ordinal Constraint in Spatial Configuration for Robust Scene Recognition

Ching Lik Teo
DSO National Laboratories
tchingli@dso.org.sg

Shimiao Li, Loong-Fah Cheong, Ju Sun
Department of Electrical and Computer Engineering
National University of Singapore
{*shimiao, eleclf, idmsj*}*@nus.edu.sg*

Abstract

This paper proposes a scene recognition strategy that integrates the appearance based local SURF features and the geometry based 3D ordinal constraint. Firstly, we show that spatial ordinal ranks of 3D landmarks are well correlated across large camera viewpoint and view direction changes and thus serve as a powerful tool for scene recognition. Secondly, ordinal depth information is acquired in a simple and robust manner when the camera undergoes a bio-mimic 'Turn-back-and-Look'(TBL) motion. Thirdly, a scene recognition strategy is proposed by combining local SURF feature matches and global 3D rank correlation coefficient into the scene recognition decision process. The performance is validated and evaluated over four indoor and outdoor databases.

1 Introduction

The problem of scene recognition is defined as the ability to recognize matches between an input query (test) image and reference images in an image database. It has important applications in robot visual navigation [9] and image retrieval [10].

A successful scene recognition strategy should possess two crucial characteristics. Firstly, the strategy should be robust to various types of image distortions due to changes of viewpoint, view direction, illumination and even moderate change of the scene environment itself. Secondly, it should be able to reliably discriminate dissimilar scenes. Unfortunately, these two characteristics usually conflict each other – to achieve better robustness, one has to sacrifice the algorithm's discrimination power and vice versa. Another important issue is the generality of the recognition algorithm. A general scene recognition strategy should be able to deal with both indoor and outdoor environments.

To deal with the dilemma, many researchers have been combining the appearance based local feature information [8] [1] together with the geometry based semi-local or global spatial information [2] [7]. Most of the works are limited to 2D spatial configurations, which are unreliable when the perspective effect is obvious. On the other hand, epipolar constraint and homography seem to be the few 3D spatial constraints used in feature point based scene recognition [4]. However, these 3D constraints are computationally expensive and not robust, especially when large number of mismatches exist.

In this paper, we propose a scene recognition strategy that integrates the appearance based local SURF features [1] and the geometry based global *3D ordinal constraint*. In **Sec.2**, the 3D ordinal constraint based on qualitative spatial information of the scene is proposed. In **Sec.3**, ordinal depth is acquired in a simple and robust manner when the camera undergoes a bio-mimic 'Turn-back-and-Look' (TBL) motion. **Sec.4** presents the details of the scene recognition strategy. In **Sec.5**, the performance is validated on four indoor and outdoor databases. Results show that 3D ordinal constraint improves the scene recognition performance compared to a simple matching based method and a matching plus epipolar constraint method.

2 3D Ordinal Constraint in Spatial Configuration

In this section, we propose the *3D ordinal constraint* which is based on 3D qualitative spatial information and show that it can serve as a powerful tool in scene recognition tasks.

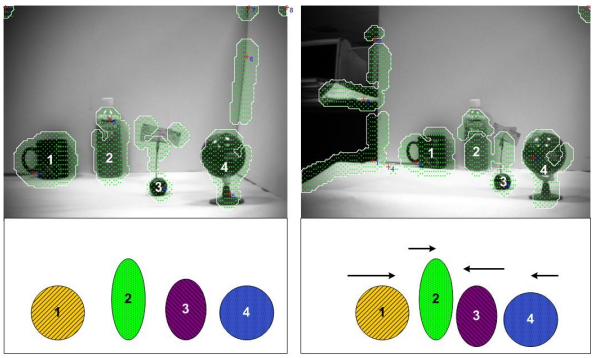

Figure 1. Landmark ranks (in x dimension) under slight viewpoint change.

2.1 Landmark Ranks and Rank Correlation Coefficient (RCC)

For simplicity, we use an intuitive way to illustrate the idea of *landmark ranks*. In fig. 1, two images of a scene taken from different viewpoints are shown. The image on the right is taken when the camera has undergone a slight viewpoint change to the right. The numbers show the ranks of the landmarks according to the x coordinate of the landmarks' position in the image. The metric measurement of the landmarks' position is converted to an ordinal scale [11]. Hence only the ranks of the landmarks' positions are of concern. The arrows in Fig.1 show clearly how the positions of the four landmarks have changed as the camera shifts to the right. Although the landmarks' absolute positions change in the x direction, their ranks remain invariant.

In practice, however, due to viewpoint changes and mismatches, ranks of landmark matches between different views of a scene may suffer from some perturbations. Nonetheless, the landmark ranks calculated from the reference image remains correlated to some degree to the landmark ranks calculated from a positive test image. This correlation imposes an *ordinal constraint* on the spatial configuration of the landmarks. In statistics, correlation between different ranks is measured by the *rank correlation coefficient (RCC)* [6], which is a variable within $[-1, 1]$. 1 (or -1) indicates perfect agreement(or disagreement) between the two ranks. 0 indicates complete independence. Many methods have been proposed to calculate RCC, here we use the two most widely used - Spearman's ρ and Kendall's τ [6].

2.2 3D Rank Correlation Coefficient (3D RCC)

In this paper, we measure the landmark ranks and RCC in three dimensions: two image dimensions x and y, and the depth dimension Z. We denote X, Y, Z as axis in the camera coordinate system. It is intuitive and can be mathematically proven that perturbations in the landmark ranks on different dimensions (x, y, Z) are likely to occur under different types of viewpoint change and in different types of scene configurations. We summarize the characteristics as follows:

* Ranks in x (or y) dimension are likely to be perturbed by camera translation. The sensitivity depends on the *in-depthness of the scene*: $D_X = \frac{\triangle Z}{\triangle X}$ ($D_Y = \frac{\triangle Z}{\triangle Y}$)(see Fig. 2). The bigger the D_X (D_Y), the higher the sensitivity. The ranks are however invariant to pure camera rotation around the camera X or Y axis.

* Ranks in Z dimension are likely to be perturbed by camera rotation around the X or Y axis. The bigger the D_X and D_Y, the lower the sensitivity. In this case, the ranks are invariant to pure camera translation.

* Camera cyclotorsion (rotation around Z axis) may perturb ranks in x and y dimension but does not perturb ranks in Z dimension.

Therefore, RCC in different dimensions compensates one another. One way to obtain a robust tool for recognition for scenes of various in-depthness under all kinds of viewpoint changes is to combine the RCC in three dimensions. In this way, a *3D Ordinal Constraint* is imposed on the landmarks. Take Spearman's ρ for example (similar for Kendall's τ), we define the *3D Rank Correlation Coefficeint (3D RCC)* as:

$$\rho_{3D} = w_Z\rho_Z + w_x\rho_x + w_y\rho_y \qquad (w_Z + w_x + w_y = 1) \tag{1}$$

The weights w_Z, w_x,w_y can be chosen particularly if some prior knowledge of the reference database and the camera view change are known. If not, we can simply average the three ($\overline{\rho_{3D}} = \frac{1}{3}(\rho_Z + \rho_x + \rho_y)$). If the ordinal depth information is not available, w_Z can be set to zero, and the 3D RCC then degenerates to 2D.

Figure 2 shows 100 3D random points observed under typical viewpoint changes in the scene recognition tasks. Figure 3 shows ($\overline{\rho_{3D}}$) between the reference view C_0 and the test views $C_i, i = 1, 2, \ldots, 9$ in scenes of various in-depthness. The black line shows the 3D RCC between two different scenes under random viewpoints, which is zero. This shows that 3D RCC can discriminate between dissimilar scenes. Furthermore, the

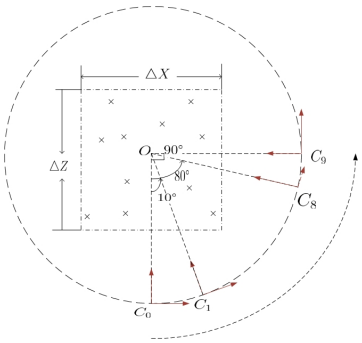

Figure 2. **Simulation of 100 3D random points scene under several camera view changes.**

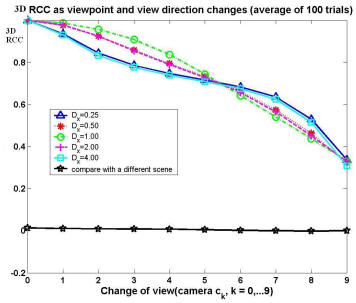

Figure 3. **3D RCC scores across the camera views for different in-depthness of the scene (average of 100 random trials).**

invariance of 3D RCC to different scene in-depthness highlights the robustness of the measurement to different scene configurations, making it a powerful tool for scene recognition.

3 Robust Ordinal Depth Acquisition - TBL Motion

Since ranks in the depth dimension are used to compute ρ_Z in 3D RCC, acquiring ordinal depth in a simple and robust way is a critical issue. In this paper, we recover ordinal depth from the bio-mimic *Turn-Back-and-Look (TBL) motion* using standard structure from motion(SFM) technique.

In the face of errors in SFM, Cheong and Xiang [3] showed that there exists generic motion types that render depth recovery more robust and reliable. For lateral movement, although Euclidean reconstruction is difficult, the distortion caused by errors in the motion estimates preserves the depth relief, which means that ordinal depth is still preserved. The ecological relevance of lateral motion is underlied by the prevalence of lateral motion used by different animals in nature to appreciate distances [12]. In the case of bees and wasps, this type of motion is known as *Turn-Back-and-Look (TBL) motion*. The TBL motion consists of a series of arcs centered about an object of interest, with the direction of translation perpendicular to the line of sight of the insect. TBL is believed to be important for the bees to recognize the scenes on their return trip [10].

We exploit camera TBL motion to recover the ordinal depth information required for scene recognition purpose. This strategy can also be readily extended to stereo camera systems where the two cameras or eyes are frontally placed. To get reliable ordinal depth information in this work, all reference and test images in the database consist of two frames taken under camera TBL motion. Depth (Z_{TBL}) of each feature point is then recovered from the optical flow computed from the two frames. The recovered depth may be inaccurate metrically, however, the ordinal scale depth information is trustworthy and can be used to compute ρ_Z in 3D RCC.

4 Scene Recognition Strategy

Our scene recognition strategy(SRS) is a SURF based method combining 3D geometry verification using 3D RCC.

4.1 Constructing the Scene Matrix

The first step of our SRS is to extract and encode useful regions in the scene for reliable recognition in the later stages. We use a modified version of a computational model of human visual saliency [5]. Various image morphological operations are then applied on the saliency map to extract the salient regions which are then encoded using SURF [1] to form 'salient-SURF' features (see fig. 4). For improved robustness against illumination changes, the SURF features are detected over the HSV color space. The final N salient-SURF features are then saved into a compact matrix known as the *Scene Matrix*, $s_m(c)$ for the c^{th} color space

$$s_m(c) = [locs_{N\times 5}\ des_{N\times 65}\ Z_{TBL}]_{N\times 71}\,, \quad (2)$$

where $c \in \{h, s, g\}$ represents the H, S, V color components. $locs$ is a 5D location vector of one SURF feature while des encodes its corresponding 65D SURF descriptor [1]. Z_{TBL} is the depth associated with this feature, obtained through camera TBL motion.

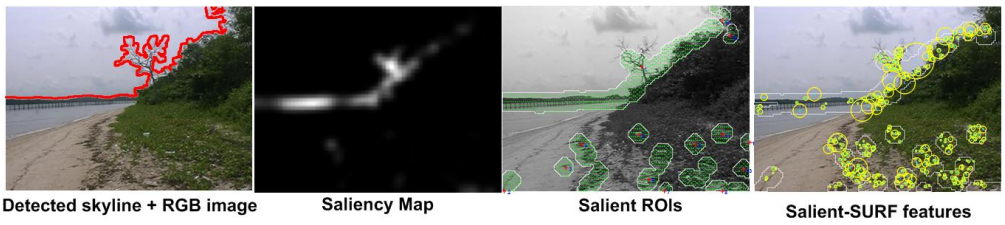

Figure 4. From the original RGB image, we extract the skyline and compute the saliency map. The salient ROIs are then extracted and encoded using SURF features augmented with ordinal depth to form the salient-SURF features.

4.2 Measuring Scene Similarity

Given the reference and the test scenes encoded by their respective scene matrices (s_m^r, s_m^t), where s_m represents the scene matrices over the three color spaces, we use the correspondence matching algorithm for SIFT descriptors [8] for efficient matching of the SURF features. To measure the similarity of the reference and the test, we define a similarity measure G_c, known as the *Global Scene Correlation Coefficient*:

$$G_c(s_m^r, s_m^t) = (N_{match}/N_{tot}) \times R. \quad (3)$$

where N_{match} and N_{tot} are the number of matches and the total number of features in the test image respectively. $R = \frac{\overline{\rho_{3D}} + \overline{\tau_{3D}}}{2}$ is the spatial 3D rank correlation of the matches between the reference and the test images. Equation (3) effectively weighs the 3D RCC by $\frac{N_{match}}{N_{tot}}$, because a large 3D RCC may occur between 2 dissimilar scenes when N_{match} is small. For ideal cases, G_c is close to 1 for positive scenes and close to 0 for negative scenes. In practice, however, the effects of viewpoint changes, mismatches and occlusions often degrade G_c.

4.3 Recognition Decision Module

A database of N_{ref} reference scenes will require N_{ref} pairwise comparisons with the input test scene each of which will use (3) to compute a measure of scene similarity. The results are summarized into a *match statistic* matrix m_s:

$$m_s = [N_{match}/N_{tot}\ R\ G_c]_{N_{ref} \times 3} \quad (4)$$

The *candidate match* G_{max}, is the reference scene that yields the largest G_c in m_s. The decision process is as follows:
if $G_{max} < t_d$, *reject the test image;*
elseif $G_{max}/G_{2ndmax} < t_a$, *reject the test image;*
else *accept the test image.*
where t_d and t_a are preset thresholds; G_{2ndmax} is the 2nd largest G_c in m_s.

Table 1. The four databases. $(N_{ref}, N_{pos}, N_{neg})$ refers to the number of reference scenes, positive test scenes, negative test scenes respectively.

Database	$(N_{ref}, N_{pos}, N_{neg})$	Type
IND	$(18, 25, 21)$	Indoor
UBIN	$(20, 63, 69)$	Outdoor coastal
SBWR	$(15, 15, 16)$	Outdoor enclosed
NS	$(20, 41, 52)$	Outdoor varied

Table 2. Recognition results of the Proposed SRS(SURF + 3D Ordinal Constraint) (%)

Database	(t_a, t_d)	P_{acc}	P_{rej}	$P_{overall}$
IND	(1.1, 0.03)	84.00	90.48	87.24
UBIN	(1.3, 0.02)	84.13	91.30	87.72
SBWR	(1.2, 0.02)	93.30	100.00	96.65
NS	(1.2, 0.01)	92.68	92.31	92.49

5 Experimental Results

In order to validate the performance of our SRS, four challenging indoor and outdoor databases, with significant image distortions between the test and reference scenes are created (see table 1 and fig. 5).

We evaluate the performance of our **proposed SRS** by computing the recognition accuracy in terms of positive acceptance P_{acc} (%), and negative rejection P_{rej} (%). Averaging out the results yields $P_{overall}$ (%), the overall recognition accuracy. The results are summarized in table 2 (the thresholds shown yield the best overall performance over a range of threshold values).

We compare the results with that of **simple**

Figure 5. **Various challenging test and reference scenes for the four databases. IND: ambiguous scenes(t) and viewpoint changes(b). UBIN: clear vs hazy overcast skies(t) and shadows vs leaves swept up(b). SBWR: numerous occlusions due to dense vegetation. NS: non-uniform illumination(t) and changes in scene content due to rain and tree fall(b). All the test scenes shown here are successfully recognized.((t)op,(b)ottom).**

Table 3. **Recognition results of the Simple SRS(SURF Only) (%)**

Database	(t_a, t_d)	P_{acc}	P_{rej}	$P_{overall}$
IND	(1.2, 0.03)	76.00	71.43	73.71
UBIN	(1.2, 0.01)	69.84	68.12	68.98
SBWR	(1.2, 0.02)	80.00	81.25	80.63
NS	(1.5, 0.02)	70.73	98.08	84.40

Table 4. **Recognition results of the Epipolar SRS(SURF + RANSAC based Epipolar Constraint) (%)**

Database	(t_a, t_d)	P_{acc}	P_{rej}	$P_{overall}$
IND	(1.2, 0.01)	75.20	71.43	73.32
UBIN	(1.2, 0.01)	57.78	68.12	62.95
SBWR	(1.2, 0.01)	80.00	81.25	80.63
NS	(1.2, 0.01)	80.00	80.77	80.38

SRS(table 3). Simple SRS uses only the percentage matches $\frac{N_{match}}{N_{tot}}$ as a similarity measure. One can see that the recognition accuracy of the proposed SRS is higher and more consistent over the four databases than that of the simple SRS. 3D ordinal constraint thus contributes positively in the recognition accuracy.

For further evaluation, we compare the results with that of **epipolar SRS**(table 4), which uses SURF matching plus RANSAC epipolar constraint to eliminate mismatches. It can be seen that the proposed SRS outperforms the epipolar SRS significantly. In addition, the proposed SRS is computationally more efficient than the iterative RANSAC based epipolar SRS. Therefore, 3D ordinal constraint is more effective and efficient than the commonly used epipolar constraint in the scene recognition task.

References

[1] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *ECCV*, pages 404–471, May 2006.

[2] G. Carneiro and A. D. Jepson. Flexible spatial configuration of local image features. *PAMI*, 29(12):2089–2104, 2007.

[3] L. F. Cheong and T. Xiang. Characterizing depth distortion under different generic motions. *IJCV*, 44(3):199–217, 2001.

[4] R. Hewitt, L. Goncalves, and M. E. Muhich. Enforcing 3-d constraints to improve object and scene recognition. *BMVC*, pages 192–201, Sep. 2007.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998.

[6] M. Kendall and J. Gibbons. *Rank Correlation Methods, 5th edition*. Edward Arnold, 1990.

[7] S. Lazebnik. *Local, Semi-local and Global Models for Texture Object and Scene Recognition*. PhD thesis, UIUC, 2006.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–100, 2004.

[9] S. Se, D. Lowe, and J. Littler. Vision-based mobile robot localization and mapping using scale-invariant features. *ICRA*, pages 2051–2058, May 2001.

[10] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrival at the end of the early years. *PAMI*, 22(12):1349–1380, 2000.

[11] S. S. Steven. On the theory of scales of measurements. *Science*, 103:677–680, 1946.

[12] R. Voss and J. Zeil. Active vision in insects: an analysis of object-directed zig-zag flights in wasps (odynerus spinipes, eumenidae). *J. of Comp. Physio. A*, 182(3):377–387, 1998.