Coupled Joint Registration and Co-segmentation for Indoor Rigid Object Sets

Siyu Hu*

Abstract

Keywords: Co-segmantion, Joint Registration

Concepts: •**Computing methodologies** → **Image manipulation;** *Computational photography;*

1 Introduction

In many researches and applications of indoor scenes the data of segmented and even annotated 3D indoor scenes are required as either data base or training data (e.g.[Nan et al. 2012][Dema and Sari-Sarraf 2012][Fisher et al. 2012][Chen et al. 2014][Fisher et al. 2015]).

One way to build such data base is to interactively compose scenes from 3D shape models resulting in scenes with object segmentation and annotation naturally available, or to mannually segment and annotate existing scenes. This procedure can be tedious and time consuming, despite the efforts to improve the interaction experience(e.g.[Merrell et al. 2011][Xu et al. 2013]).

Another way is to automatically generate scenes from 3D shape models according to the input RGB or RGB-D images(e.g.[Liu et al. 2015][Chen et al. 2014]). In such methods, a retrievial procedure is usually needed and inevitablely limit the result to a certain set of 3D models despite the actual 3D model in the input images.

We prefer a approach that helps us build such data set directly from the captured data. One of the major gap between the required data set and available scene capturing framework(e.g.[Izadi et al. 2011]) is the general object level segmentation. We want to stress that a general object level segmentation problem should not be treated as an equivalence of multilabel classification problem since it is not limited to a certain set of objects. For 3D data, [Jia et al. 2015] used some simplified physical prior knowledge (i.e. the block based stability) to help acheiving the general object segmentation, while the work of [Xu et al. 2015] proposes a practical and rather complete framework to close the gap between the required data set and available scene capturing method. One of the observation in [Xu et al. 2015] is that the motion consistency of rigid object can serve as a strong evidence of general objectness. To exploit this fact, they employ a robot to do proactive push and use the movement tracking to verify and iteratively improve their object level segmentation result. Our work presented in this paper is trying to exploit the same observation from a different approach.

We intend to use the motion consistency that is naturally revealed by human activities along the time. Down to this approach, we are facing the choice of scanning scheme. One way is to record the change of the scene along with the human activities, another is to arrange a daily or even a once every half day sweep to only record the result of human activities but avoid the instant of human motion. The main challenge brought in by the second scheme is that we may not be able to solve the object correspondence by a local search due to the sparse sampling over time, but the very same challenge exists in the first scheme due to the exclusion caused by human bodies not to mention other additional process(e.g. tracking with severe oclussion) needed for human bodies. With the second scanning scheme, our original intention of building 3D scene data set from capturing naturally leads us to the problem of coupled joint registration and co-segmentation.

In this problem, registration and segmentaion are entangled in each other. On one hand the segmentation depends on the registration to connect the point clouds into series of rigid movement so that the object level segmentation can be done based on the motion consistency, on the other hand, the registration depends on the segmentation to break the problem into a series of rigid joint registration instead of a joint registration with non-coherent point drift(A pair of points is close to each other in one point set but their correspondent pair of points in another point set is far from each other, in other words, the point drift of this pair is non-coherent. This happens when this pair of points actually belong to different objects.)

To model the problem, we employ a group of gaussian mixture models and each of these gaussian mixture models represents a potential objects. This modeling handles the entanglement of registration and segmentation in the way that

*e-mail:sy891228@mail.ustc.edu.cn

SIGGRAPH 2016 Posters, July 24-28, 2016, Anaheim, CA
ISBN: 978-1-4503-ABCD-E/16/07
DOI: http://doi.acm.org/10.1145/9999997.9999999

2 Related Work

2.1 Point Set Registration with GMM Representation

[Chui and Rangarajan 2000]
[Myronenko and Song 2010]
[Jian and Vemuri 2011]
Our work is most related to [Evangelidis et al. 2014]. We actually extend the formulation of [Evangelidis et al. 2014] to simultaneously handle joint registration and co-segmentation.

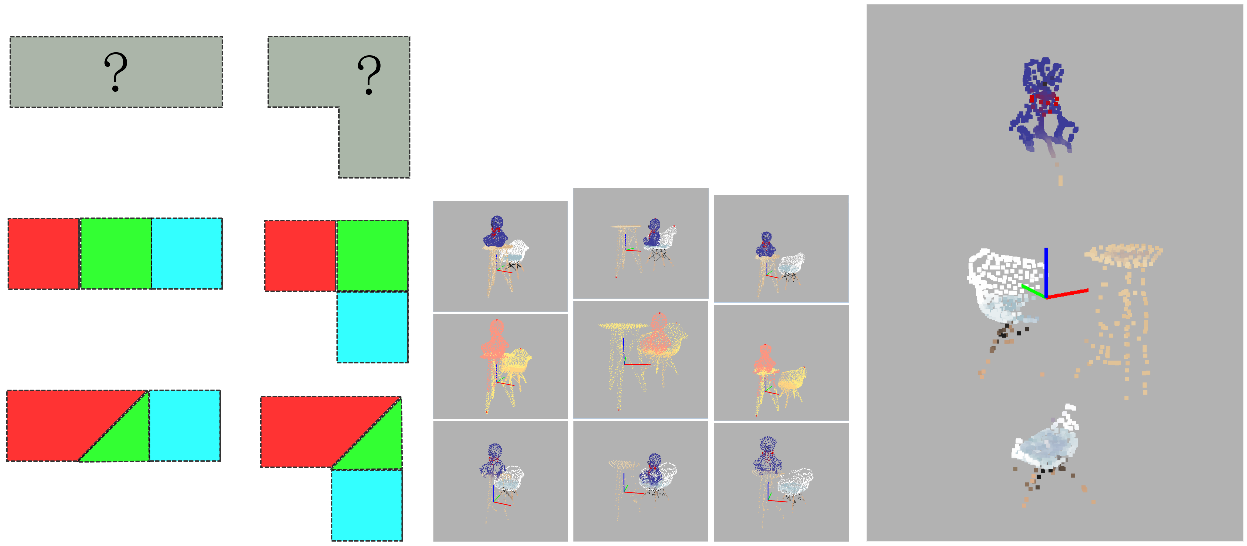
2.2 Functional Mapping

The coupled joint registration and co-segmentation problem comes with a latent problem of point-to-point correspondence problem. A series of work based on the functional maps representation advocated in [Ovsjanikov et al. 2012] have be done. In one of the most recent work [Maron et al. 2016], a convex relaxation technique was used to better approximate the global minimal for both rigid and non-rigid registration problem.

3 Method Overview

3.1 Problem Statement

Given a set of point clouds which record the same group of rigid indoor objects with different layout. We intend to samutaneously partition the point clouds into objects and align the points of same object to recover layouts for corresponding object. Figure **??** shows an example of input point clouds set.



(a) Multiple Explanation for the Same Set of Observation (b) Three Sample Frame:from top to down are input, segmentation and composed with latent model (c) latent model

3.2 Baseline Formulation

To formulate the relation between the unknown object set and the input point clouds. We come up with a generation model as follows:

$$P(v_{mi}) = \sum_{k=1}^{K_n} p_k N(v_{mi}|\phi_{mn}(x_k), \Sigma_k) \tag{1}$$

which means, The observed point clouds are generated by N object model. Each object model is represented by a gaussian mixture model with K_n centroids. Our goal is to maximize the probability of the expected compelete-data log-likelihood. The object function can be written as:

$$\Theta = \operatorname{argmax} \sum_{Z} P(Z|V, \Theta) \ln P(V, Z; \Theta) \tag{2}$$

in which:

$$\Theta = \{\{p_k, x_k, \Sigma_k\}_{k=1}^{\sum K_n}, \{\phi_{mn}\}_{m=1,n=1}^{MN}\}$$

is the parameters of the generation model.
p_n is the prior probability that the point is generated by the n-th object.
p_k is the weight of the k-th Gaussian.
x_k is the center of the k-th Gaussian.
Σ_k is the standard deviation of the k-th Gaussian.
There are $\sum K_n$ Gaussian model in total and among them, K_n Gaussian models belongs to object n.
V is the M input point clouds.
v_{mi} is the i-th point of the m-th point cloud.
Z is a latent variable set defined as:

$$Z = \{z_{ij}|j = 1...M, i = 1...N_j\}$$

among which if $z_{ij} = k(k = 1...\sum K_n)$ assign the observation of $\phi_{mn}(v_{mi})$ to the k-th component of Gaussian mixture model.Such formulation can be seen as an extention of joint registration formulation in [Evangelidis et al. 2014], upon which we add several gaussian mixture model together to express a group of objects. By solving this new problem we simutaneously solve the object co-segmentation of given observation.

3.3 Bilateral Formulation

When considering features, we can develop it into a bilateral GMM formulation.

$$P(v_{mi}, f_{mi}) = \sum_{k=1}^{K_n} p_k N(v_{mi}|\phi_{mn}(xv_k), \sigma v_k) N(f_{mi}|xf_k, \sigma v_f)) \tag{3}$$

we measure the feature difference by a gaussian with diagnal Σ, to make this measurement valid we rotate and scale the feature space with PCA.

3.4 Hough Based Formulation

In order to constrain the shape of objects and allow optimization of the shape, we assume that the surface of any indoor object is composed of five axist aligned rectangle plates.

$$P(v_{mi}) = \sum_{k=1}^{K_n} p_k N(v_{mi}|\phi_{mn}(Y_k), \Sigma_k) \tag{4}$$

where Y_k is a rectangle plate and the $N(v_{mi}|\phi_{mn}(Y_k), \Sigma_k)$ are expanded as

$$(\sqrt{2\pi})^{-3}|\Sigma_k|^{-1} \exp(-0.5 \inf_{x \in Y_k} \{||v_{mi} - \phi_{mn}(x)||^2_{\Sigma_k}\})$$

in which the

$$\inf_{x \in Y_k} \{||v_{mi} - \phi_{mn}(x)||^2_{\Sigma_k}\}$$

means that the distance from a point to a plate is defined as the infimum of distances from this point to points inside the plate. For plate model, with other parameter fixed the update of ϕ_{mn} can be solved as a problem

$$\begin{cases} \min\limits_{R_{mn}, t_{mn}} & ||(R_{mn}W_{mn} + t_{mn}e^T - C(\{Y\}_n))\Lambda_{mn}||_F^2 \\ & s.t.\ R_{mn}R_{mn}^T = I,\ |R_{mn}| = 1 \end{cases} \tag{5}$$

in which $C(\{Y_k\}_n)$ is a 3×5 matrix with each column being a center coordinates of a plate in nth object and $W_{mn} =$

$[w_{mk_1}, ..., w_{mk_5}]_{k_j \in O_n}$ is matrix with size of 3×5 and each column being a virtual 3D point given by

$$w_{mk_j} = \frac{\sum_{i=1}^{N_m} \alpha_{mik} v_{mi}}{\sum_{i=1}^{N_m} \alpha_{mik}}$$

3.5 Sparsity in Height of Supporting Plane and Gravity Axis Translation

The height of supporting plane and the translation of the object along the gravity axis is sparse in a physical world.
For translation:

$$\vec{t_z}(\phi_{mn}) =< \vec{w_t}, \vec{h} > \quad |\vec{w_t}| = 1 \tag{6}$$

$$\vec{z}_{plane}(X) =< \vec{w_x}, \vec{h} > \quad |\vec{w_x}| = 1 \tag{7}$$

where $\vec{h}$ is the space of supporting plane and always have a zero element as the floor is always a supporting plane. $\vec{t_z}(\phi_{mn})$ is the translation component of the transformation ϕ_{mn}. $\vec{z}_{plane}$ is a height of supporting plane in latent object model. $\vec{h}$ can be constructed by detecting horizontal planes in the observations.

4 Algorithms and Implementation

4.1 Expectation Conditional Maximization

Assuming the observed point clouds $\{V_m\}$ are independent and identically distributed, we can then write the (2) as:

$$\varepsilon(\Theta|V,Z) = \sum_{m,i,k} \alpha_{mik}(\log p_k + \log P(\phi_{nm}(v_{mi})|z_{ji} = k; \Theta)) \tag{8}$$

In which the $\alpha_{mik} = P(z_{mi} = k|v_{mi}; \Theta)$,

Algorithm 1 Joint Registration and Co-segmentation (JRCS)

Input:
$\{V_m\}$:Observed point clouds
$\{\alpha^0_{mik}\}$:Initial posterior probabilities
Output:
Θ^q:Final parameter set

1. $q \leftarrow 0$
2. **repeat**
3. CM-step-a: Use α^q_{mik}, x^{q-1}_k to estimate $\{R^q_{mn}\}$ and $\{t^q_{mn}\}$
4. CM-step-b: Use α^q_{mik}, $\{R^q_{mn}\}$ and $\{t^q_{mn}\}$ to estimate the Gaussian centers x^q_k
5. CM-step-c: Use α^q_{mik}, $\{R^q_{mn}\}$ and $\{t^q_{mn}\}$ to estimate the covariances Σ^q_k
6. CM-step-d: Use α^q_{mik} to estimate the priors p^q_k
7. E-step: Use Θ^{q-1} to estimate posterior probabilities. $\alpha^q_{mik} = P(z_{mi}|v_{mi}; \Theta^{q-1})$
8. $q \leftarrow q + 1$
9. **until** Convergence
10. **return** Θ^q

4.2 Initialization Techniques

A key advantage motivates our formulation is that the soft correspondence can be initialized more flexiblely comparing to the typical initialization techniques such as landmark point pairs in registration.
The result of Clustering:

$$P(B_{mj} \in C_n)$$

Soft Correspondence Initialization
Then the α is initialized as:

$$\alpha_{ijk} = P(B_{mj} \in C_n)$$

on the condition that:

$$v_{ij} \in B_{mj} \wedge x_k \in O_n$$

5 Experiments and Discussion

Acknowledgements

To Robert, for all the bagels.

References

BOUAZIZ, S., TAGLIASACCHI, A., AND PAULY, M. 2013. Sparse iterative closest point. In *Proceedings of the Eleventh Eurographics/ACMSIGGRAPH Symposium on Geometry Processing*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SGP '13, 113–123.

BRONSTEIN, M. M., AND KOKKINOS, I. 2010. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 1704–1711.

CHEN, K., LAI, Y.-K., WU, Y.-X., MARTIN, R., AND HU, S.-M. 2014. Automatic semantic modeling of indoor scenes from low-quality rgb-d data using contextual information. *ACM Trans. Graph. 33*, 6 (Nov.), 208:1–208:12.

CHUI, H., AND RANGARAJAN, A. 2000. A new algorithm for non-rigid point matching. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2, 44–51 vol.2.

DEMA, M. A., AND SARI-SARRAF, H. 2012. 3d scene generation by learning from examples. In *Multimedia (ISM), 2012 IEEE International Symposium on*, 58–64.

EVANGELIDIS, G. D., KOUNADES-BASTIAN, D., HORAUD, R., AND PSARAKIS, E. Z. 2014. *A Generative Model for the Joint Registration of Multiple Point Sets*. Springer International Publishing, Cham, 109–122.

FISHER, M., RITCHIE, D., SAVVA, M., FUNKHOUSER, T., AND HANRAHAN, P. 2012. Example-based synthesis of 3d object arrangements. *ACM Trans. Graph. 31*, 6 (Nov.), 135:1–135:11.

FISHER, M., SAVVA, M., LI, Y., HANRAHAN, P., AND NIESSNER, M. 2015. Activity-centric scene synthesis for functional 3d scene modeling. *ACM Trans. Graph. 34*, 6 (Oct.), 179:1–179:13.

IZADI, S., KIM, D., HILLIGES, O., MOLYNEAUX, D., NEWCOMBE, R., KOHLI, P., SHOTTON, J., HODGES, S., FREEMAN, D., DAVISON, A., AND FITZGIBBON, A. 2011. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, ACM, New York, NY, USA, UIST '11, 559–568.

JIA, Z., GALLAGHER, A. C., SAXENA, A., AND CHEN, T. 2015. 3d reasoning from blocks to stability. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*, 5 (May), 905–918.

JIAN, B., AND VEMURI, B. C. 2011. Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 8 (Aug), 1633–1645.

KRÄHENBÜHL, P., AND KOLTUN, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 109–117.

LEVY, B. 2006. Laplace-beltrami eigenfunctions towards an algorithm that "understands" geometry. In *IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*, 13–13.

LIU, Z., ZHANG, Y., WU, W., LIU, K., AND SUN, Z. 2015. Model-driven indoor scenes modeling from a single image. In *Graphics Interface Conference*.

MARON, H., DYM, N., KEZURER, I., KOVALSKY, S., AND LIPMAN, Y. 2016. Point registration via efficient convex relaxation. *ACM Trans. Graph. 35*, 4 (July), 73:1–73:12.

MERRELL, P., SCHKUFZA, E., LI, Z., AGRAWALA, M., AND KOLTUN, V. 2011. Interactive furniture layout using interior design guidelines. *ACM Trans. Graph. 30*, 4 (July), 87:1–87:10.

MYRONENKO, A., AND SONG, X. 2010. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 12 (Dec), 2262–2275.

NAN, L., XIE, K., AND SHARF, A. 2012. A search-classify approach for cluttered indoor scene understanding. *ACM Trans. Graph. 31*, 6 (Nov.), 137:1–137:10.

NEAL, R. M., AND HINTON, G. E. 1998. *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*. Springer Netherlands, Dordrecht, 355–368.

OVSJANIKOV, M., BEN-CHEN, M., SOLOMON, J., BUTSCHER, A., AND GUIBAS, L. 2012. Functional maps: A flexible representation of maps between shapes. *ACM Trans. Graph. 31*, 4 (July), 30:1–30:11.

PAPON, J., ABRAMOV, A., SCHOELER, M., AND WRGTTER, F. 2013. Voxel cloud connectivity segmentation - supervoxels for point clouds. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2027–2034.

RODOL, E., BUL, S. R., AND CREMERS, D. 2014. Robust region detection via consensus segmentation of deformable shapes. *Computer Graphics Forum 33*, 5, 97–106.

WANG, F., HUANG, Q., AND GUIBAS, L. 2013. Image co-segmentation via consistent functional maps. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 849–856.

XU, K., CHEN, K., FU, H., SUN, W.-L., AND HU, S.-M. 2013. Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models. *ACM Trans. Graph. 32*, 4 (July), 123:1–123:15.

XU, K., HUANG, H., SHI, Y., LI, H., LONG, P., CAICHEN, J., SUN, W., AND CHEN, B. 2015. Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM Trans. Graph. 34*, 6 (Oct.), 177:1–177:14.