Pacific Graphics 2017
J. Barbic and O. Sorkine-Hornung
(Guest Editors)

Volume 36 (2017), Number 7

Interactive Point Set Joint Registration and Co-segmentation for Indoor Scenes

Siyu Hu [1]

[1] USTC

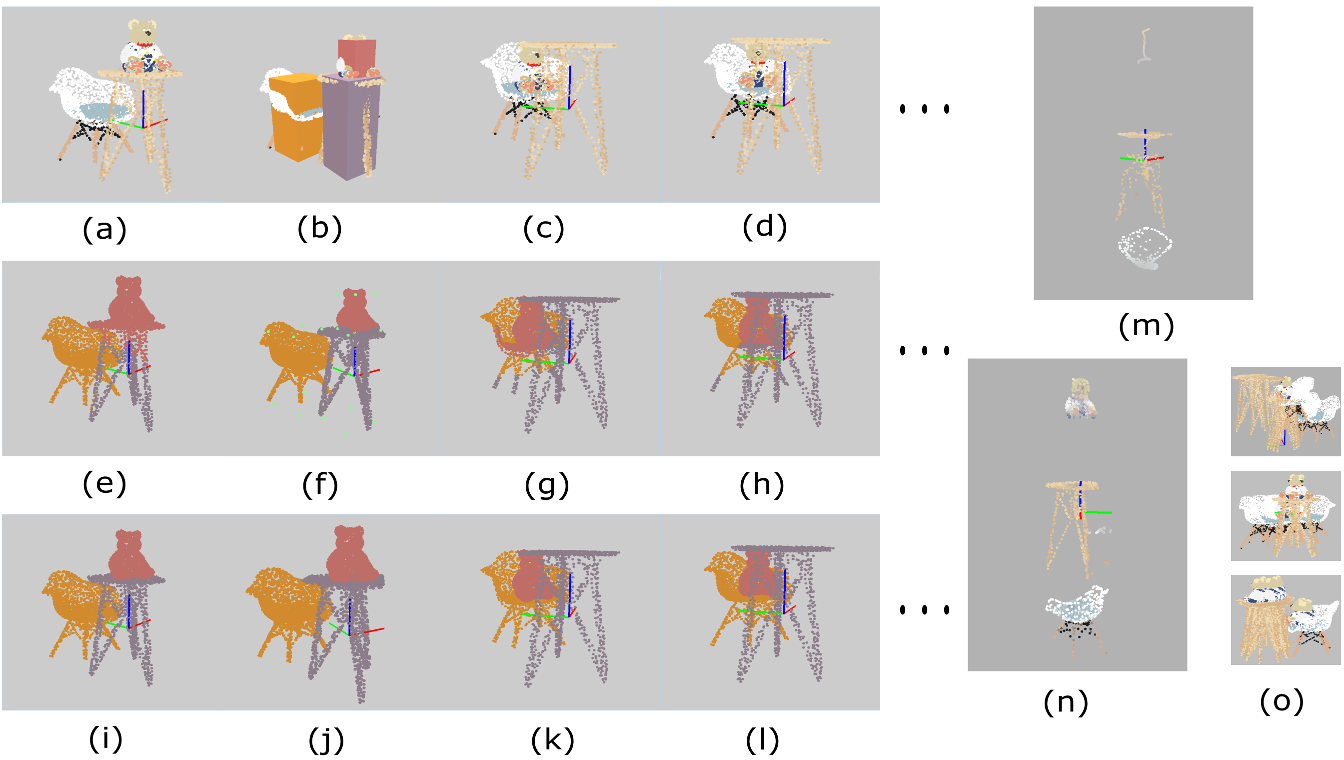

Figure 1: *(a)(b)(c)(d) are input point sets and user have initialized layout for (b) by interactively placing boxes in it. (m) shows the centroids of latent model. (e)(f)(g)(h) are segmentation result based on the instant parameters of latent model. (i)(j)(k)(l) are final segmentation result. (n) shows the final centroids of latent model. (o) verifys the accuracy of final transformation result by aligning input sets to each object.*

Abstract

This paper presents a method of joint registration and co-segmentation for point sets of indoor scenes. We view the joint registration and co-segmentation as two problems heavily entangled with each other. To model such entangled problems, we treat the input point sets as samples from a latent generated model and bring up with a novel formulation based on Gaussian mixture model. By maximizing the posterior probability of the samples, we gradually recover the latent object model and object level segmentation and align the objects to the latent model(solve the registration). Alone with the formulation, we design a procedure of interaction that can help users to intuitively initialize the optimization. Our evaluation shows that our novel method is helpful and effective to do the joint registration and co-segmentation on point sets of indoor scenes.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [COMPUTER GRAPHICS]: Applications—I.4.8 [IMAGE PROCESSING AND COMPUTER VISION]: Scene Analysis—Range data

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. [NXS12] [DSS12] [FRS*12] [CLW*14] [FSL*15]).
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. [MSL*11] [XCF*13]).
Another way is to automatically generate scenes from 3D shape models according to the input RGB or RGB-D images(e.g. [LZW*15] [CLW*14]). 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. [IKH*11]) 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, [JGSC15] used some simplified physical prior knowledge (i.e. the block based stability) to help acheiving the general objectness segmentation, while the work of [XHS*15] 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 [XHS*15] 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 schedule 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 objectness 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 object. This model unentangle the registration and segmentation in the way that the segmentation can be done by evaluate the probability of points belongs to the Gaussian mixture models and the registration can be done by evaluate rigid registration against each gaussain mixture models.
In summary our work makes following contributions:
Firstly, as far as we know we are the first work that bring up with the problem of point set joint registration and co-segmentation for indoor scenes.
Secondly, we come up with a Gaussian mixture model based formulation to simultaneously model both the registration and co-segmentation problem.
Thirdly, targeting the disadvantages of our formulation, we design a procedure of interaction and provide a practical tool for point set joint registration and co-segmentation based on it. We release the tool at `https://github.com/samhu1989/DevBundle`

2. Related Work

In this section we explain how our work is related to the previous works and how we draw experience from these previous works.

2.1. Point Set Registration with GMM Representation

There are a serious of works that uses gaussion mixture model as representation for point set to formulate the registration problem. [MS10] consider the registration of two point sets as a probability density estimation problem. They force the Gaussian mixture model centroids to move coherently as a group to preserve the topological structure of the point sets. Their method is appliable to both rigid registration and non-rigid registration. As we highlighted in section 1, our problem is different from the non-rigid registration considered in [MS10], the point drift could be non-coherent in our probelm. [JV11] summarized the works for point set registration using Gaussian mixture models and present a unified framework for the rigid and nonrigid point set registration problem. These works select one of the point set as the “model”. Unlike these works, [EKBHP14] treats all the point sets as data: they are all realizations of a Gaussian mixture and the registration is cast into a clustering problem. Comparing to these works, our work is most related to [EKBHP14]. Our formulation can be seen as an extention of the formulation of [EKBHP14] to simultaneously handle joint registration and co-segmentation.

2.2. Image segmentation and co-segmentation

[RKB04] is an influential work for interactive image segmentation. It uses two Gaussian mixture model, one for foreground and one for background. To initialize these two Gaussian mixture models, [RKB04] let user place a rectangle that contain the foreground. Our design of interaction draw on the experientce from [RKB04]. The difference is that our interaction is designed for 3D space and can handle multiple objects segmentation rather than foreground-background segmentation. [TSS16] jointly recover cosegmentation and dense per-pixel correspondence in two images. Our work solve a similar problem for multiple 3D point sets.

2.3. Segmentation from Motion

The idea that motion can be strong hint for segmentation is used in many works. [XHS*15] employs a robot to do proactive push and track the motion to learn object segmentation. [LPR*16] use the motion in video and use the motion edge as training data to learn an edge detector. These methods lean on the motion that is continuous in time and can be tracked. Our method can handle motion that is not continuous in time.

3. Method Overview

3.1. Problem Statement

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

3.2. Basic 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:

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

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 [EKBHP14], 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.

3.4. Interaction Design

Unfortunately, there are serveral parameters that can not be easily initialized in our formulation . In this subsection we first introduce our design of interaction, which is intuitive for users to input the semantic prior this way. We then explain how we can easily initlialize those parameters for our optimization based on the manual input.
As demonstrated in Figure 2, we let user choose one of the point sets and placing and editing boxes in it to indicate the layout for this point set. From this, we can easily initialize the total number of objects N and determine $\{K_n\}$ which is the numbers of Gaussian mixture models used to represent each object. These two paremeters are difficult to be initialized without semantic prior, but with the input of the users we can naturally initialize the N as the number of different color label and the K_n as

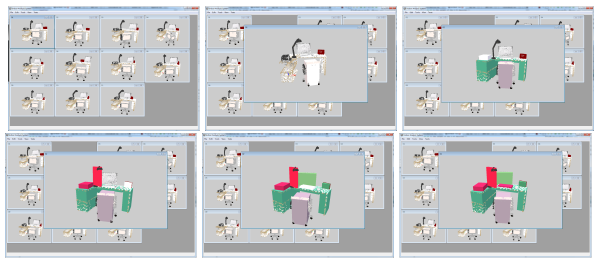

Figure 2: *From the first to the nineth, the nine images show the procedure of interaction: the user pick one point set and place boxes in it to indicate the layout for this point set. The box in white is the box currently under editing. The boxes in other colors are boxes placed to represent object layouts. One color represent one object. The interaction allows multiple boxes to represent same object.(e.g. the desk is represented by three boxes in same color)*

$$K_n = \frac{V_n}{\sum V_n} K_{all} \tag{4}$$

in which the V_n represent the total volume of the boxes in the n-th color and the K_{all} is initialized as $K_{all} = 0.5 * median(I_m)$ and $\{I_m\}$ are point numbers of M input point set. This is an emperical choice borrowed from [EKBHP14].

4. Algorithms and Implementation Details

4.1. Expectation Conditional Maximization

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

$$\varepsilon(\Theta|V,Z) = \sum_{m,i,k} \alpha_{mik}(\log p_k + \log P(\phi_{nm}(v_{mi})|z_{ji} = k;\Theta)) \quad (5)$$

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 flexibly 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. Experiment and Discussion

5.1. Evaluation for Segmentation

5.2. Evaluation for Registration

5.3. Stress Tests on Noisy Data

5.4. User Study for Interaction

References

[CLW*14] CHEN K., LAI Y.-K., WU Y.-X., MARTIN R., HU S.-M.: Automatic semantic modeling of indoor scenes from low-quality rgb-d data using contextual information. *ACM Trans. Graph. 33*, 6 (Nov. 2014), 208:1–208:12. URL: http://doi.acm.org/10.1145/2661229.2661239, doi:10.1145/2661229.2661239. 2

[DSS12] DEMA M. A., SARI-SARRAF H.: 3d scene generation by learning from examples. In *Multimedia (ISM), 2012 IEEE International Symposium on* (Dec 2012), pp. 58–64. doi:10.1109/ISM.2012.19. 2

[EKBHP14] EVANGELIDIS G. D., KOUNADES-BASTIAN D., HORAUD R., PSARAKIS E. Z.: *A Generative Model for the Joint Registration of Multiple Point Sets*. Springer International Publishing, Cham, 2014, pp. 109–122. URL: http://dx.doi.org/10.1007/978-3-319-10584-0_8, doi:10.1007/978-3-319-10584-0_8. 2, 3

[FRS*12] FISHER M., RITCHIE D., SAVVA M., FUNKHOUSER T., HANRAHAN P.: Example-based synthesis of 3d object arrangements. *ACM Trans. Graph. 31*, 6 (Nov. 2012), 135:1–135:11. URL: http://doi.acm.org/10.1145/2366145.2366154, doi:10.1145/2366145.2366154. 2

[FSL*15] FISHER M., SAVVA M., LI Y., HANRAHAN P., NIESSNER M.: Activity-centric scene synthesis for functional 3d scene modeling. *ACM Trans. Graph. 34*, 6 (Oct. 2015), 179:1–179:13. URL: http://doi.acm.org/10.1145/2816795.2818057, doi:10.1145/2816795.2818057. 2

[IKH*11] IZADI S., KIM D., HILLIGES O., MOLYNEAUX D., NEWCOMBE R., KOHLI P., SHOTTON J., HODGES S., FREEMAN D., DAVISON A., FITZGIBBON A.: 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* (New York, NY, USA, 2011), UIST '11, ACM, pp. 559–568. URL: http://doi.acm.org/10.1145/2047196.2047270, doi:10.1145/2047196.2047270. 2

[JGSC15] JIA Z., GALLAGHER A. C., SAXENA A., CHEN T.: 3d reasoning from blocks to stability. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*, 5 (May 2015), 905–918. doi:10.1109/TPAMI.2014.2359435. 2

[JV11] JIAN B., VEMURI B. C.: Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 8 (Aug 2011), 1633–1645. doi:10.1109/TPAMI.2010.223. 2

[LPR*16] LI Y., PALURI M., REHG J. M., DOLLAR P., UNDEFINED, UNDEFINED, UNDEFINED, UNDEFINED: Unsupervised learning of edges. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 00* (2016), 1619–1627. doi:doi.ieeecomputersociety.org/10.1109/CVPR.2016.179. 3

[LZW*15] LIU Z., ZHANG Y., WU W., LIU K., SUN Z.: Model-driven indoor scenes modeling from a single image. In *Graphics Interface Conference* (2015). 2

[MS10] MYRONENKO A., SONG X.: Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 12 (Dec 2010), 2262–2275. doi:10.1109/TPAMI.2010.46. 2

[MSL*11] MERRELL P., SCHKUFZA E., LI Z., AGRAWALA M., KOLTUN V.: Interactive furniture layout using interior design guidelines. *ACM Trans. Graph. 30*, 4 (July 2011), 87:1–87:10. URL: http://doi.acm.org/10.1145/2010324.1964982, doi:10.1145/2010324.1964982. 2

[NXS12] NAN L., XIE K., SHARF A.: A search-classify approach for cluttered indoor scene understanding. *ACM Trans. Graph. 31*, 6 (Nov. 2012), 137:1–137:10. URL: http://doi.acm.org/10.1145/2366145.2366156, doi:10.1145/2366145.2366156. 2

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A.: "grabcut": Interactive foreground extraction using iterated graph cuts. In *ACM*

SIGGRAPH 2004 Papers (New York, NY, USA, 2004), SIGGRAPH '04, ACM, pp. 309–314. URL: http://doi.acm.org/10.1145/1186562.1015720, doi:10.1145/1186562.1015720. 2

[TSS16] TANIAI T., SINHA S. N., SATO Y.: Joint recovery of dense correspondence and cosegmentation in two images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016). 3

[XCF*13] XU K., CHEN K., FU H., SUN W.-L., HU S.-M.: Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models. *ACM Trans. Graph. 32*, 4 (July 2013), 123:1–123:15. URL: http://doi.acm.org/10.1145/2461912.2461968, doi:10.1145/2461912.2461968. 2

[XHS*15] XU K., HUANG H., SHI Y., LI H., LONG P., CAICHEN J., SUN W., CHEN B.: Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM Trans. Graph. 34*, 6 (Oct. 2015), 177:1–177:14. URL: http://doi.acm.org/10.1145/2816795.2818075, doi:10.1145/2816795.2818075. 2, 3