Rent3D:
Floor-Plan Priors for Monocular Layout Estimation

Chenxi Liu\textsuperscript{1,\,*} Alexander Schwing\textsuperscript{2,\,*} Kaustav Kundu\textsuperscript{2}
Raquel Urtasun\textsuperscript{2} Sanja Fidler\textsuperscript{2}

\textsuperscript{1}Tsinghua University, \textsuperscript{2}University of Toronto
How Many Times Have You Looked for Apartments?
How Many Times Have You Looked for Apartments?

- **United States:**
  - 11.7% per year

- **Craigslist:**
  - 90,000 rental ads per day only in New York
  - 10 million people visit the website per day
How Many Times Have You Looked for Apartments?

- Chenxi: 2 times
- Alex: 3 times
- Kaustav: 2 times
- Raquel: 4 times
- Sanja: 5 times
Finding an Apartment/House is a Pain...

- Particularly during a winter in Toronto
Renting Apartments

5 bedroom apartment for sale
One Hyde Park, Knightsbridge, SW1X

£64,999,950

This property is marketed by:

Aylesford International, Chelsea
440 Kings Road, London, SW10 0LH

View properties from this agent

Request Details

or call: 020 8012 4022

Save property

Add notes

Print

Send to Friend

Share this property

Don’t miss out

75% of home-movers in

Liu, Schwing, Kundu, Urtasun, Fidler

Rent3D
Example Rental Data

- Plus some meta information e.g. wall height
Rent3D: View Rental Ads in 3D
Camera localization within apartment
Related Work

- Room layout estimation
  - Hedau et al., 2009, 2012
  - Lee et al., 2010
  - Schwing et al., 2012, 2013
  - Del Pero et al., 2011, 2012
  - Choi et al., 2013

- Virtual tours
  - Xiao & Furukawa, 2012

- 3D indoor reconstruction from large photo collections or video
  - Cabral & Furukawa, 2014
  - Brualla et al., 2014

- Indoor localization (video, depth sensors)
  - Project Tango
  - SLAM work
Related Work

- Room layout estimation
  - Hedau et al., 2009, 2012
  - Lee et al., 2010
  - Schwing et al., 2012, 2013
  - Del Pero et al., 2011, 2012
  - Choi et al., 2013

Our work:
3D indoor reconstruction and localization using monocular imagery

- Cabral & Furukawa, 2014
- Brualla et al., 2014

Indoor localization (video, depth sensors)
- Project Tango
- SLAM work
Overview
Accurate camera localization:

- Scene cues
Accurate camera localization:

- Scene cues
- Semantic cues
Overview

Accurate **camera localization**:

- **Scene cues**
- **Semantic cues**
- **Geometric cues** by exploiting the dimension information
Formulation

- $r \in \{1, \ldots, R\}$ ... discrete random variable representing the room
Formulation

- $r \in \{1, \ldots, R\}$ ... discrete random variable representing the room

Front wall is the plane defined by $v_{P_0}$ and $v_{P_1}$
Formulation

- \( r \in \{1, \ldots, R\} \) ... discrete random variable representing the room
- \( c_r \in \{1, \ldots, |C_r|\} \) ... a discrete variable representing within room \( r \) which wall the picture is facing (\(|C_r|\) the number of walls in a room)
Formulation

- $r \in \{1, \ldots, R\}$ ... discrete random variable representing the room
- $c_r \in \{1, \ldots, |C_r|\}$ ... a discrete variable representing within room $r$ which wall the picture is facing ($|C_r|$ the number of walls in a room)
Formulation

- \( r \in \{1, \ldots, R\} \) \ldots discrete random variable representing the room
- \( c_r \in \{1, \ldots, |C_r|\} \) \ldots a discrete variable representing within room \( r \) which wall the picture is facing (\(|C_r|\) the number of walls in a room)
Formulation

- $r \in \{1, \ldots, R\}$ ... a discrete random variable representing the room
- $c_r \in \{1, \ldots, |C_r|\}$ ... a discrete variable representing within room $r$ which wall the picture is facing ($|C_r|$ the number of walls in a room)
Formulation

- \( r \in \{1, \ldots, R\} \ldots \) discrete random variable representing the room
- \( c_r \in \{1, \ldots, |C_r|\} \ldots \) a discrete variable representing within room \( r \) which wall the picture is facing (\( |C_r| \) the number of walls in a room)
- \( y \ldots \) rays representing a room layout

Typical parametrization for room layout [Hedau et al., 2009]:

- Room is a 3D cuboid
- \( y = (y_1, y_2, y_3, y_4) \)
- 4 rays needed to define it
Formulation

- $r \in \{1, \ldots, R\}$ ... discrete random variable representing the room
- $c_r \in \{1, \ldots, |C_r|\}$ ... a discrete variable representing within room $r$ which wall the picture is facing ($|C_r|$ the number of walls in a room)
- $y$ ... rays representing a room layout

We formulate the problem as inference in a Conditional Random Field with the following energy:

$$E(r, c_r, y) = E_{\text{scene type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y)$$
Energy Terms: Scene Type

\[ E(r, c_r, y) = E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]

- **Potential**: Score of a scene classifier predicting scene type (e.g., bedroom, kitchen, reception)
Energy Terms: Scene Type

\[ E(r, c_r, y) = E_{\text{scene-type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]

**Potential**: Score of a scene classifier predicting scene type (e.g., bedroom, kitchen, reception)
Energy Terms: Layout

\[ E(r, c_r, y) = E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]

Orientation Map [Lee et al., 2009]  
Geometric Context [Hedau et al., 2009]
Energy Terms: Layout

\[ E(r, c_r, y) = E_{scene\_type}(r) + E_{layout}(r, c_r, y) + E_{win}(r, c_r, y) \]

Orientation Map [Lee et al., 2009]

- **Potential**: Counts of blue, red, etc, pixels inside and outside of each wall
- Fast computation using *integral geometry* [Schwing et al., 2012]
Energy Terms: Layout

\[ E(r, c_r, y) = E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]
Energy Terms: Layout

\[ E(r, c_r, y) = E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]

- \( y = (y_1, y_2, y_3, y_4) \)
- \( y_4 = f(r, c_r, y_1, y_2, y_3) \)
\[ E(r, c_r, y) = E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]

- \( y = (y_1, y_2, y_3, y_4) \), \( y_4 = f(r, c_r, y_1, y_2, y_3) \)
- Additional constraint on \( y \): Camera is inside the room
Energy Terms: Windows

\[ E(r, c_r, y) = E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]

- Window-background segmentation
Energy Terms: Windows

\[ E(r, c_r, y) = E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \]

- Window-background segmentation
- **Potential**: count window pixels inside and outside the window area
Learn and Inference

- We are minimizing the energy:

\[
(r^*, c_r^*, y^*) = \arg\min_{r, c_r, y} \left( E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \right)
\]
Learning and Inference

- We are minimizing the energy:

\[
(r^*, c_r^*, y^*) = \arg\min_{r, c_r, y} \left( E_{\text{scene-type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \right)
\]

- Inference:
  - Exhaustive enumeration of \( r \) and \( c_r \)
  - Exact branch and bound inference for \( y \) [Schwing & Urtasun, 2012]
Learning and Inference

- We are minimizing the energy:

\[
(r^*, c_r^*, y^*) = \arg\min_{r, c_r, y} \left( E_{\text{scene\_type}}(r) + E_{\text{layout}}(r, c_r, y) + E_{\text{win}}(r, c_r, y) \right)
\]

- Inference:
  - Exhaustive enumeration of \( r \) and \( c_r \)
  - Exact branch and bound inference for \( y \) \cite{Schwing2012}

- We use S-SVM for training
Dataset

- We crawled a London apartment rental site

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># apartments</td>
<td>215</td>
</tr>
<tr>
<td># of images</td>
<td>1570</td>
</tr>
<tr>
<td># of indoor images</td>
<td>1259</td>
</tr>
<tr>
<td># images without GT alignment</td>
<td>82</td>
</tr>
<tr>
<td>avg. # rooms per apt</td>
<td>6</td>
</tr>
<tr>
<td>avg. # walls per apt</td>
<td>31</td>
</tr>
<tr>
<td>avg. # windows per apt</td>
<td>6</td>
</tr>
<tr>
<td>avg. # doors per apt</td>
<td>9</td>
</tr>
</tbody>
</table>
Apartments in Central London Are Not Small

Approx. Gross Internal Area  2696 Sq Ft - 250.46 Sq M

Biggest apartment in dataset: 16 rooms, 5 bedrooms, 88 walls
Apartments in Central London Are Not Small

Approx. Gross Internal Area 2696 Sq Ft - 250.46 Sq M

Rent: £25,000 per month

Biggest apartment in dataset: 16 rooms, 5 bedrooms, 88 walls.
Results: Layout Estimation

- We assume we know which wall the camera is facing
- **Metrics**: Pixel accuracy for predicting 5 walls

<table>
<thead>
<tr>
<th></th>
<th>Layout error</th>
<th>Evaluations</th>
<th>Test time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schwing’12</td>
<td>13.88</td>
<td>16012.4</td>
<td>0.0208</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>11.81</strong></td>
<td><strong>1269.5</strong></td>
<td><strong>0.0019</strong></td>
</tr>
</tbody>
</table>

Liu, Schwing, Kundu, Urtasun, Fidler
Results: Layout Estimation

- We assume we know which wall the camera is facing
- **Metrics**: Pixel accuracy for predicting 5 walls

<table>
<thead>
<tr>
<th></th>
<th>Layout error</th>
<th>Evaluations</th>
<th>Test time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schwing'12</td>
<td>13.88</td>
<td>16012.4</td>
<td>0.0208</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>11.81</strong></td>
<td><strong>1269.5</strong></td>
<td><strong>0.0019</strong></td>
</tr>
</tbody>
</table>

- 2% reduction in layout error
Results: Layout Estimation

- We assume we know which wall the camera is facing
- **Metrics:** Pixel accuracy for predicting 5 walls

<table>
<thead>
<tr>
<th></th>
<th>Layout error</th>
<th>Evaluations</th>
<th>Test time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schwing'12</td>
<td>13.88</td>
<td>16012.4</td>
<td>0.0208</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>11.81</strong></td>
<td><strong>1269.5</strong></td>
<td><strong>0.0019</strong></td>
</tr>
</tbody>
</table>

- 2% reduction in layout error
- 10 times less branching operations
Results: Layout Estimation

- We assume we know which wall the camera is facing

**Metrics**: Pixel accuracy for predicting 5 walls

<table>
<thead>
<tr>
<th></th>
<th>Layout error</th>
<th>Evaluations</th>
<th>Test time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schwing’12</td>
<td>13.88</td>
<td>16012.4</td>
<td>0.0208</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>11.81</strong></td>
<td><strong>1269.5</strong></td>
<td><strong>0.0019</strong></td>
</tr>
</tbody>
</table>

- 2% reduction in layout error
- 10 times less branching operations
- 10x speedup
Results: Camera Localization

- **Metrics**: % of correct assignments of front wall to the apartment wall

<table>
<thead>
<tr>
<th></th>
<th>Aspect</th>
<th>+Scene</th>
<th>+Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.0328</td>
<td>0.1138</td>
<td>0.1954</td>
</tr>
<tr>
<td>Ours (no windows)</td>
<td>0.0686</td>
<td>0.1945</td>
<td>0.2654</td>
</tr>
<tr>
<td>Ours (windowGT)</td>
<td>0.2128</td>
<td>0.4737</td>
<td>0.5995</td>
</tr>
<tr>
<td>Ours (window)</td>
<td>0.1670</td>
<td>0.3982</td>
<td>0.5080</td>
</tr>
</tbody>
</table>
Results: Camera Localization

- **Metrics**: % of correct assignments of front wall to the apartment wall

<table>
<thead>
<tr>
<th></th>
<th>Aspect</th>
<th>+Scene</th>
<th>+Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.0328</td>
<td>0.1138</td>
<td>0.1954</td>
</tr>
<tr>
<td>Ours (no windows)</td>
<td>0.0686</td>
<td>0.1945</td>
<td>0.2654</td>
</tr>
<tr>
<td>Ours (windowGT)</td>
<td>0.2128</td>
<td>0.4737</td>
<td>0.5995</td>
</tr>
<tr>
<td>Ours (window)</td>
<td>0.1670</td>
<td>0.3982</td>
<td>0.5080</td>
</tr>
</tbody>
</table>

**Aspect**: Only aspect ratio information (and not scene) used
Results: Camera Localization

- **Metrics**: % of correct assignments of front wall to the apartment wall

<table>
<thead>
<tr>
<th></th>
<th>Aspect</th>
<th>+Scene</th>
<th>+Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.0328</td>
<td>0.1138</td>
<td>0.1954</td>
</tr>
<tr>
<td>Ours (no windows)</td>
<td>0.0686</td>
<td>0.1945</td>
<td>0.2654</td>
</tr>
<tr>
<td>Ours (windowGT)</td>
<td>0.2128</td>
<td>0.4737</td>
<td>0.5995</td>
</tr>
<tr>
<td>Ours (window)</td>
<td>0.1670</td>
<td>0.3982</td>
<td>0.5080</td>
</tr>
</tbody>
</table>

+Scene: Aspect information and scene classifier are used
Results: Camera Localization

- **Metrics**: % of correct assignments of front wall to the apartment wall

<table>
<thead>
<tr>
<th></th>
<th>Aspect</th>
<th>+Scene</th>
<th>+Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.0328</td>
<td>0.1138</td>
<td>0.1954</td>
</tr>
<tr>
<td>Ours (no windows)</td>
<td>0.0686</td>
<td>0.1945</td>
<td>0.2654</td>
</tr>
<tr>
<td>Ours (windowGT)</td>
<td>0.2128</td>
<td>0.4737</td>
<td>0.5995</td>
</tr>
<tr>
<td>Ours (window)</td>
<td>0.1670</td>
<td>0.3982</td>
<td>0.5080</td>
</tr>
</tbody>
</table>

+**Room**: We know which room the picture was taken in.
### Results: Camera Localization

**Metrics**: \% of correct assignments of front wall to the apartment wall

<table>
<thead>
<tr>
<th></th>
<th>Aspect</th>
<th>+Scene</th>
<th>+Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.0328</td>
<td>0.1138</td>
<td>0.1954</td>
</tr>
<tr>
<td>Ours (no windows)</td>
<td>0.0686</td>
<td>0.1945</td>
<td>0.2654</td>
</tr>
<tr>
<td>Ours (windowGT)</td>
<td>0.2128</td>
<td>0.4737</td>
<td>0.5995</td>
</tr>
<tr>
<td>Ours (window)</td>
<td>0.1670</td>
<td>0.3982</td>
<td>0.5080</td>
</tr>
</tbody>
</table>
Results: Camera Localization

- **Metrics**: % of correct assignments of front wall to the apartment wall

<table>
<thead>
<tr>
<th></th>
<th>Aspect</th>
<th>+Scene</th>
<th>+Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.0328</td>
<td>0.1138</td>
<td>0.1954</td>
</tr>
<tr>
<td>Ours (no windows)</td>
<td>0.0686</td>
<td>0.1945</td>
<td>0.2654</td>
</tr>
<tr>
<td>Ours (windowGT)</td>
<td>0.2128</td>
<td>0.4737</td>
<td>0.5995</td>
</tr>
<tr>
<td>Ours (window)</td>
<td>0.1670</td>
<td>0.3982</td>
<td>0.5080</td>
</tr>
</tbody>
</table>
Results: Joint Layout and Localization

Red arrow: Groundtruth camera    Green arrow: Predicted camera
Results: Joint Layout and Localization

Red arrow: Groundtruth camera  
Green arrow: Predicted camera
Results: Reconstruction

Window + Aspect
- 1 images out of 4
- 2 walls out of 8

+ Scene
- 4 images out of 4
- 8 walls out of 8

+ Room
- 4 images out of 4
- 8 walls out of 8

Ground-truth
- -
Summary

- Problem of apartment 3D reconstruction from monocular imagery
- Model that jointly solves for localization and room layout estimation by exploiting floor-plans
- Real-time inference
- Results:
  - We improve layout prediction over past work
  - Achieve good localization performance
- Dataset with 215 apartments and all annotations available:

  http://www.cs.toronto.edu/~fidler/projects/rent3D.html
Alex on the Market Next Year
Thank You
Welcome to our poster at #9!