The 3D jigsaw puzzle: mapping large indoor spaces

Ricardo Martin-Brualla[1], Yanling He[1], Bryan C. Russell[2], and Steven M. Seitz[1]

[1]University of Washington [2]Intel

Abstract. We introduce an approach for analyzing annotated maps of a site, together with Internet photos, to reconstruct large indoor spaces of famous tourist sites. While current 3D reconstruction algorithms often produce a set of disconnected components (3D pieces) for indoor scenes due to scene coverage or matching failures, we make use of a provided map to lay out the 3D pieces in a global coordinate system. Our approach leverages position, orientation, and shape cues extracted from the map and 3D pieces and optimizes a global objective to recover the global layout of the pieces. We introduce a novel crowd flow cue that measures how people move across the site to recover 3D geometry orientation. We show compelling results on major tourist sites.

Keywords: Indoor scene reconstruction, maps, 3D jigsaw puzzle.

1 Introduction

Recent breakthroughs in computer vision now allow us to model our world in 3D with extraordinary accuracy and visual fidelity from just about any set of overlapping photos [1–3]. However, a limitation of state-of-the-art 3D reconstruction techniques from Internet photos is that large scenes tend to break up into a collection of disconnected pieces due to gaps in the depicted scene coverage or matching failures. Rather than a single, fully-connected Vatican model, for instance, we get a collection of smaller 3D pieces for different rooms, such as the Sistine Chapel, the Raphael Rooms, and the Hall of Maps, each having their own 3D coordinate system. A major challenge is to automatically put these 3D pieces together correctly into a global coordinate frame. This is akin to solving a *3D jigsaw puzzle*, where the scale, rotation, and translation of the 3D pieces must be recovered with respect to the global coordinate frame.

Solving the 3D jigsaw puzzle is extremely difficult using image information alone due to the aforementioned coverage and matching failures. Instead, we seek to leverage readily available map data to solve the 3D jigsaw puzzle. Such data provides additional information that helps constrain the spatial layout of the 3D pieces. For example, a map of the Vatican shows an annotated floorplan of the different rooms, with a legend providing the names of the rooms and any objects located inside the rooms. Such maps are plentiful and widely available, for example in tourist guidebooks (e.g. Rick Steves, Lonely Planet, Baedeker) and online (e.g. planetware.com).

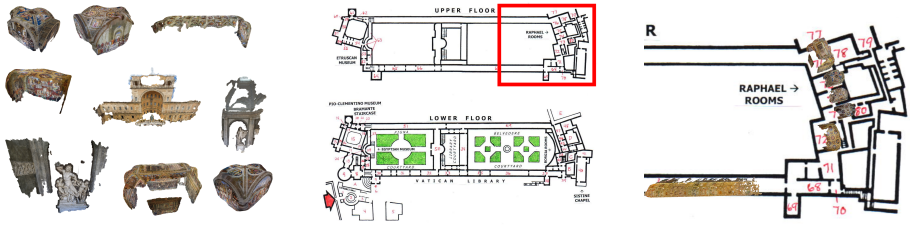


Fig. 1. Given a set of disconnected reconstructed 3D models of a large indoor scene, for example the Vatican (left), we jointly reason about a map of the site (middle) and the 3D pieces to produce a globally consistent reconstruction of the entire space (blow up at right).

Automatically leveraging map data for the 3D jigsaw puzzle is challenging as the pieces are unlabeled and lack absolute position, orientation, and scale. The 3D Wikipedia system provided one approach to automatically link objects described in text to their spatial location in a 3D model [4]. While [4] can be used to link the 3D pieces to text on an annotated map, it does not provide information on how to *place* the pieces in a global coordinate system. Moreover, most maps provide only 2D cues (e.g., via a floorplan), with objects and dimensions placed only approximately. Finally, we must cope with rooms having orientation ambiguities (e.g. square or rectangular rooms), which the map and 3D piece geometry alone cannot disambiguate.

The key to our approach is to extract and integrate position, orientation, and scale cues from the 3D pieces and the map. These include the room shape, map annotations, cardinal direction (available as compass measurements provided in the image EXIF data used to reconstruct the 3D pieces), and crowd flow through the different rooms of the site. The latter crowd flow cue, which measures the dominant direction of travel through the 3D pieces, provides information on the orientation of the pieces. For example, in the Vatican Museum tourists tend to go from the entrance toward the Sistine Chapel, passing through the Gallery of Candelabra, Gallery of Tapestries, and Hall of Maps along the way. We formulate the 3D jigsaw puzzle problem as an integer quadratic program with linear constraints to globally solve for the 3D layout of the pieces. Ours is the first system to reconstruct large indoor spaces of famous tourist sites from Internet photos via joint reasoning with map data and disconnected 3D pieces returned by structure-from-motion.

We show compelling results on four major sites. Our system reliably assigns and orients many of the 3D pieces relative to the input maps (we provide a detailed analysis of assignment precision/recall and orientation accuracy). Moreover, we show an integrated visualization of the map and reconstructed 3D geometry, where a user can interactively browse and fly to different rooms of the site (please see the video [5]).

2 Related work

Our work is related to prior research leveraging auxiliary information, such as geographic data, human path of travel, and text, to augment 3D reconstruction and image localization. Geographic data, such as Google Street View and Google Earth 3D models, has been used to georegister point clouds generated from Internet images [6]. Maps have been used in conjunction with visual odometry for self localization [7, 8]. Human path of travel has been used for image geolocalization [9] and predicting tourists' path of travel [10]. Note that in this work we use human path of travel to recover 3D piece orientation. Finally, text has been used to automatically label objects in reconstructed geometry [4].

Most related is prior work that matched free space from a 3D reconstruction to white space in a map [11]. However, [11] addressed a particularly simple case where the 3D jigsaw puzzle has only one large piece (the 3D model), and the floor plan is accurate. While aligning 3D geometry shards has been explored by other authors (e.g., Stanford's Forma Urbis project [12]), our problem is more challenging as the scale of each piece is unknown and we do not have access to complete scans. Also related are approaches for solving 2D jigsaw puzzles [13–15], which operate entirely on 2D rectangular puzzle pieces and determine puzzle piece locations through absolute position cues (e.g. corners, edges, color) and adjacency cues (e.g. shape). The analogy in our case is that label correspondences provide absolute cues and tourist flow provides adjacency.

Reconstructing large indoor spaces is a challenging problem due to lack of texture on many surfaces and the difficulty of systematically scanning every surface of a site. Major efforts to scan and reconstruct large indoor scenes include the Google art project [16], museum reconstruction via constructive solid geometry [17], and human-operated systems to scan a large site [18, 19].

3 System overview

In this paper we present a system to solve the 3D jigsaw puzzle via joint reasoning over 3D geometry and annotated map data. Our system takes as inputs: (i) one or more reconstructed 3D pieces for a site and (ii) an annotated map corresponding to a set of 2D map points of interest (associated with rooms and named objects), with corresponding 2D map regions (the extent of rooms and objects in the map) and text annotations (the legend).

Our system begins by generating a discrete set of candidate placements of the 3D pieces to the map points of interest (Section 4.2). 3D pieces are assigned to the map by querying Google Image Search using the extracted text annotations from the map and linking the returned images to the 3D pieces via camera resectioning. This provides links between a given map point of interest to candidate 3D locations on the 3D pieces. Note that the links are noisy as the returned images may depict incorrect locations of the site. Given the links, a discrete set of candidate 3D transformations to the global coordinate system are generated for each 3D piece.

Given the candidate placements, we optimize an objective function that seeks a globally consistent layout of the 3D pieces by integrating cues extracted over the points of interest, their 2D map regions, and the 3D pieces, described in Section 4. The objective integrates cues about the shape of the rooms, cardinal direction, crowd flow through the site, and mutual exclusion of the 3D pieces. We show results of our system in Section 5.

4 Model for the 3D jigsaw puzzle

Given a discrete set of candidate placements of 3D pieces to map points of interest, we seek a globally consistent layout of the 3D pieces. Let $p \in \{1, \dots, P\}$ index the map points of interest, $m \in \{1, \dots, M\}$ the 3D models, $q_m \in \{1, \dots, Q_m\}$ 3D locations on 3D model m, and $t_m \in \{1, \dots, T_m\}$ candidate 3D transformations of 3D model m to the global coordinate system. A candidate placement is the tuple (p, m, q, t), where we omit the subindices for brevity.

A solution to the 3D jigsaw puzzle is a selection of 3D piece placements from the candidate set. We define binary variables $x_{p,m,q,t} \in \{0, 1\}$ to indicate whether the candidate placement appears in the solution set and auxiliary binary variables $y_{m,t} \in \{0, 1\}$ to indicate that 3D model m is placed in the global coordinate system under 3D transformation t. We formulate the 3D jigsaw puzzle as an integer quadratic program with linear constraints where vector b and matrix A encode unary and pairwise cues over the position, scale, and orientation of the candidate placements (described in Section 4.1):

$$\max_{x,y} \quad x^T A x + b^T x \tag{1}$$

$$\text{s.t.} \quad \forall p \quad \sum_{m,q,t} x_{p,m,q,t} \leq 1 \qquad \forall q \quad \sum_{p,m,t} x_{p,m,q,t} \leq 1 \tag{2}$$

$$\forall m \quad \sum_{t} y_{m,t} \leq 1 \qquad \forall p, m, q, t \quad x_{p,m,q,t} \leq y_{m,t} \tag{3}$$

Constraints (2) enforce mutual exclusion of the 3D puzzle pieces. We require each point of interest p to be assigned to at most one 3D location q on a model, and vice versa. We find that enforcing mutual exclusion is critical for our problem since we are reconstructing unique object instances of a site. Constraints (3) enforce each model m to be placed in the global coordinate system under a single 3D transformation t.

Given pairwise and unary coefficients A and b, we optimize Objective (1) using mixed-integer quadratic programming [20]. Note that while it has been shown that solving jigsaw puzzles with uncertainty in the piece compatibility is NP-hard [21], the small size of our datasets, of up to a few dozen pieces, enables us to express the mutual exclusion constraints exactly. This is in contrast to recent work in modeling 2D jigsaw puzzles that have formulated the problem as a Markov Random Field with mutual exclusion constraints approximated by a set of local pairwise terms due to large problem size [13, 14].

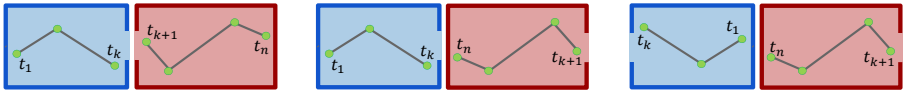


Fig. 2. Illustration of the crowd flow cue for two adjacent rooms on the map. For a sequence of photos captured by a particular user, green points show the location where the images where taken and $t_1, \ldots, t_n$ their ordered time stamps. Here, the user moved from the blue to the red room from left to right. Our goal is to orient the rooms to be consistent with the direction of travel. Left: room orientations are consistent with the user path through both rooms. Middle: the red room is inconsistent with the user path. Right: both rooms are inconsistent with the user path.

4.1 Cues for position, scale, and orientation

In this section we describe the cues that are used to pose the 3D pieces relative to the map. These cues encode the crowd flow through the space, number of registered image search results to 3D pieces, cardinal direction of the pieces, and room shape.

Crowd flow potential As previously noted [10], for many popular places people tend to visit different parts of the scene in a consistent order. For example, in the Vatican Museum, most tourists walk from the entrance toward the Sistine Chapel, passing through the Gallery of Candelabra, Gallery of Tapestries, and Hall of Maps along the way. We seek to harness the "flow of the crowd" to help disambiguate the orientation of the 3D pieces.

We wish to characterize the crowd flow within each 3D piece m and between 3D pieces m and m'. We start by considering the sets of photos taken by individual Flickr users that were aligned to the 3D pieces and sort the photos based on their timestamps. These aligned images indicate the users' direction of travel within the 3D pieces (e.g., tourists move from right to left of the main painting inside the Hall of the Immaculate Conception) and across the 3D pieces (e.g., tourists visit the Galeria of Candelabra before the Gallery of Maps). We say that the candidate placements of two 3D pieces agree with the crowd flow if the dominant direction of travel *across* the two pieces is oriented in the same direction as *within* the pieces after placing them onto the global coordinate system. We illustrate the crowd flow cue in Figure 2.

More concretely, given the camera locations for the images for a particular user i in model m, let $d^m_{i,k}$ be a unit vector in the direction of travel between consecutive images k and $k+1$ in the sequence, which corresponds to how the user moved between shots. For candidate placement $\alpha = (p, m, q, t)$, we define the dominant direction of travel within 3D piece m as $\delta_\alpha = H_t(norm(\sum_{i,k} d^m_{i,k}))$ where $H_t(\cdot)$ is the 3D transformation for t and $norm(\cdot)$ normalizes the input vector to unit length.

To estimate the dominant direction of travel across 3D pieces m and m', we count the number of users $u_{m,m'}$ that took a picture first in m and later in

m'. For candidate placements α and α' with $m \neq m'$, we denote the dominant direction of travel across the two pieces in the global coordinate system as the unit vector $\delta_{\alpha,\alpha'} = sign(u_{m,m'} - u_{m',m}) \cdot norm(H_{t'}(c_{m'}) - H_t(c_m))$ where c_m is the 3D centroid of 3D piece m. Note that if most users travel from m to m', $\delta_{\alpha,\alpha'}$ will point in the direction from 3D piece m to m' in the global coordinate system. We define the crowd flow cue for candidate placements α and α' as the sum of inner products:

$$A_{\alpha,\alpha'} = < \delta_{\alpha,\alpha'}, \delta_\alpha > + < \delta_{\alpha,\alpha'}, \delta_{\alpha'} > \tag{4}$$

Unary potentials For each candidate placement we extract unary potentials for assignment $\phi^{assign}(\alpha)$, cardinal direction $\phi^{card}(\alpha)$, and room shape $\phi^{shape}(\alpha)$. We concatenate these potentials into vector $\Phi(\alpha)$ and, given weights w, define the unary coefficients b as:

$$b_\alpha = w^T \Phi(\alpha) \tag{5}$$

We wish to leverage the vast amounts of labeled imagery online to connect the map points of interest to their locations in the 3D pieces. Using the text annotation for each point of interest in the map, we issue a query to Google Image Search concatenating the annotation text with the site name, followed by registering the returned images to the 3D pieces. We define $\phi^{assign}(\alpha) = count(p, m, q)$ as the number of images retrieved by querying for the text associated with map point of interest p that are registered to the 3D location q in model m.

A small fraction of Flickr images contain heading information in EXIF tags (e.g., via compass). Although we have found such data to be sparse and not always reliable, we can exploit it when available. The cardinal direction potential $\phi^{card}(\alpha)$ measures the compatibility of compass measurements corresponding to images used to reconstruct a 3D piece to a cardinal direction given on the map (e.g. "north"). Let $C_m > 0$ be the number of images used to reconstruct 3D piece m having a heading and $C_{m,t}$ be the number of such images that agree on the orientation of the provided cardinal direction within τ degrees after applying 3D transformation t into the global coordinate system. We define the potential to be $\phi^{card}(\alpha) = C_{m,t}/C_m$.

Next we wish to encode how well the 3D piece matches the shape of a given 2D region on the map. We encode the shape by projecting the structure-from-motion points of model m onto the map via transformation t and rasterize the points into a grid of cells. The shape potential $\phi^{shape}(\alpha)$ is a weighted sum of three terms: (i) the ratio of intersection area over union between the 2D region and occupied grid cells, (ii) average truncated distance of each grid cell to the 2D map region edge, and (iii) fraction of grid cells that lie outside of the region.

4.2 Generating candidate placements

In this section we describe how to generate the set of candidate placements of 3D pieces to map points of interest. First, we parse the map into a set of regions

Fig. 3. Left: A 3D piece of our system corresponding to the Hall of the Immaculate Conception. Middle: Colored 2D regions extracted from the floorplan. Number 72 in purple corresponds to the ground truth location of the 3D piece. Right: Candidate placements of the 3D piece to the 2D region.

and points of interest with accompanying text, described in Appendix A. Then we describe how we assign and align the 3D pieces to the map regions and points of interest.

Given the extracted text annotations from the map, we align images downloaded from Google image search to the 3D pieces. We cluster the set of inlier 3D points across all queries and set the 3D locations q as the centers of mass of the clusters. We orient the vertical direction of each 3D piece by aligning its z-axis with its up vector and setting the ground plane ($z = 0$) at the bottom of the piece. The up vector is the normal direction of a plane fitted to the inlier camera centers of the piece, oriented towards the cameras' up vectors.

A map may provide labels for only the room and/or for multiple objects in a room. For example, the Vatican Museums have only the rooms labeled, whereas the Pantheon has objects labeled within the main room. We wish to account for both cases when generating candidate placements. When only the room is labeled, we generate multiple candidate placements by finding local maxima of the unary shape potential $\phi^{shape}(\alpha)$. When multiple objects are labeled, we use the candidate assignments between the 3D locations on the models and the 2D points of interest on the map as putative matches. We then estimate a similarity transformation given the matches to yield the candidate placements. Example candidate placements are shown in Figure 3.

5 Results

We evaluated our system on four major tourist sites: the Vatican Museums, St. Peter's Basilica in Rome, Pantheon in Rome, and the Hearst Castle. We collected maps for each site and reconstructed 3D models for the sites by downloading images from Flickr by querying for the site name and running VisualSFM [22, 23]. In addition, for each reconstructed Flickr photo, we downloaded all photos taken by the same user within 2 hours and match them to the reconstructed pieces, yielding a much larger image set (factor of 5-10). For visualization purposes we use PMVS for multi-view stereo [24] and Poisson Surface Reconstruction [25] to generate colored meshes. Note that all these packages are freely available online.

Table 1. Site statistics: # POIs – number of points of interest in the map, # GT POIs – number of points of interest in the map with ground truth 3D model assignments, # GT Orientations – number of points of interest in the map with ground truth 3D model orientation assignments, # Images – number of images used in the 3D reconstruction, # 3D Pieces – number of reconstructed 3D pieces.

Site	# POIs	# GT POIs	# GT Orientations	# Images	# 3D Pieces
Vatican Museums	75	30	11	11K	68
Hearst Castle	22	5	5	3K	30
Pantheon	9	8	8	705	11
St. Peter's	34	13	11	3K	55

We collected ground truth assignments between the pieces and the map legends by finding information in authoritative sites, such as Wikipedia articles and specialized sites about the landmarks, like the official website of the Vatican Museums or saintpetersbasilica.org. Collecting ground truth orientations of the 3D pieces is challenging given that images alone do not disambiguate between orientations. Fortunately some authoritative sites contain more detailed maps for a small section of a landmark that place different objects inside the rooms or enumerate the views with their cardinal orientations. We can also infer the orientation of some rooms from official museum itineraries by correlating the direction of travel of the 3D pieces with the observed direction of travel from the Flickr users. We summarize the ground truth dataset statistics in Table 1.

The Vatican Museums and the Hearst Castle datasets are examples of very large multiroom scenes where most pieces correspond to complete rooms in the site, like the Sistine Chapel or the Raphael Rooms in the Vatican Museums. Figures 4 and 5 show the recovered layout of the different 3D pieces using the annotated maps for the Vatican Museums and Hearst Castle, respectively. Notice that we are able to correctly position and scale/orient many of the 3D pieces. While our 3D model coverage appears sparse in some regions, particularly the lower floor of the Vatican and 2nd floor of Hearst Castle, we correctly place most of the most visited and well-photographed rooms, such as the Raphael Rooms and the 2nd floor galleries of the Vatican Museums. Indeed, the correctly aligned pieces account for 75% and 73% of all reconstructed images for the Vatican Museums and Hearst Castle respectively. Note that some pieces are incorrectly scaled, like the Pigna Courtyard, due to the lack of a complete model of the room, as well as errors in the map parsing.

The Pantheon and St. Peter's Basilica are examples of single large rooms, where the annotated maps detail the specific objects names present in the site. Both sites contain large open spaces that enable the 3D reconstruction process to create a mostly complete 3D model of the entire site. Figures 6 and 7 show the recovered layout for both sites. The Pantheon model was aligned to the map by the assignment of 7 of its objects to points of interest in the map. In the St. Peter's case, three objects contained in the large 3D model were assigned to points of interest as well as other smaller models, such as Michelangelo's Pieta and the Chapel of Presentation.

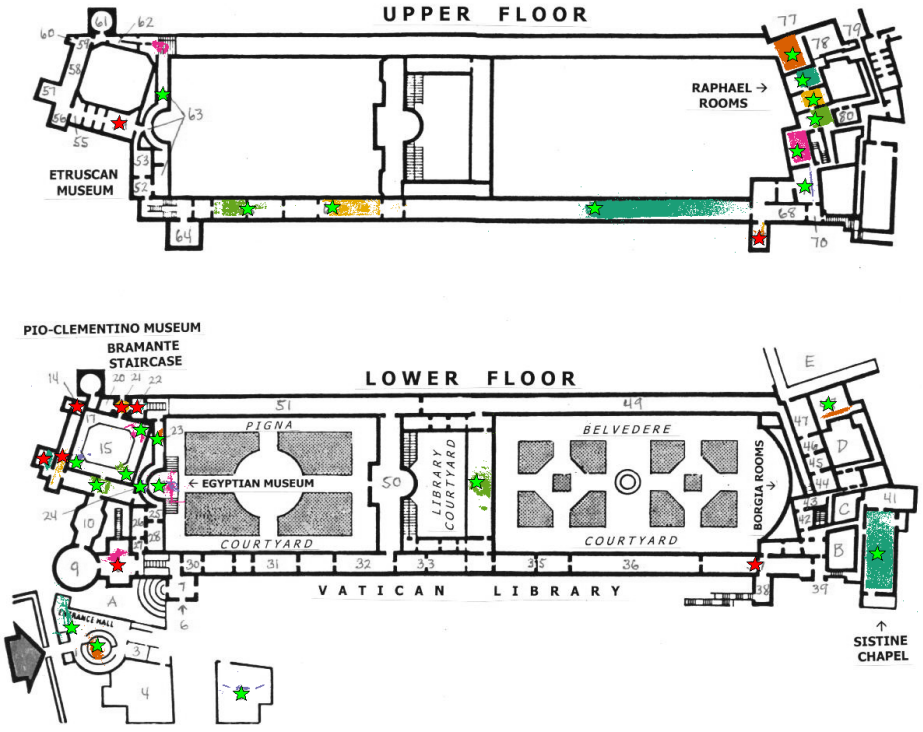


Fig. 4. Results for the Vatican Museums. 3D pieces are shown as the projection of the SfM points on the map, with different colors for each model. Green stars represent correct assignments, red stars incorrect ones. Please zoom in on the electronic version to see the details.

We quantitatively evaluate the assignments of 3D pieces to the points of interest in the map and the orientation of those assignments in Table 2. As a baseline we use only the assignment potential score $\phi^{assign}(\alpha)$ described in Section 4.1, which ignores the mutual exclusion constraint. Our system consistently improves the precision of the assignment over the baseline. The orientations proposed by our system for the correctly assigned points of interest are correct in 25 out of 33 cases across all sites.

We perform an ablative study over the orientation cues for the sites with multiple rooms (Vatican Museums and Hearst Castle). Note that the Pantheon is a single large room and St. Peter's has stand-alone objects (e.g. the Pieta, the Altar of St. Jerome), plus one central room. In Table 3 we show statistics of the data collected for the cues and orientation accuracy values using the crowd flow cue, the cardinal direction cue and the joint model. The crowd flow cue disambiguates cases such as the galleries in the second floor of the Vatican, but fails on 3D pieces representing objects, such as statues or paintings, since users don't move in a predetermined path of travel when photographing them. The compass cue is powerful when enough data is available, but is ineffective for

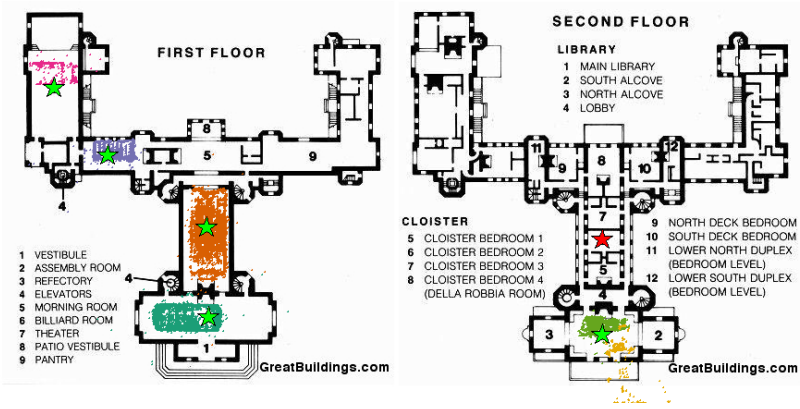


Fig. 5. Results for the Hearst Castle. 3D pieces are shown as the projection of the SfM points on the map, with different colors for each model. Green stars represent correct assignments, red stars incorrect ones.

Table 2. For each site, we report assignment precision/recall values with respect to all annotated points of interest in the map for our model and a baseline (see text), and orientation accuracy of our model.

Site	Assignment				Orientation
	Baseline		Model		Model
	Precision	Recall	Precision	Recall	Accuracy
Vatican Museums	53%	57%	73%	43%	91%
Hearst Castle	83%	27%	83%	27%	60%
Pantheon	67%	89%	100%	78%	100%
St. Peter's	45%	59%	70%	29%	50%

Table 3. For each site, we report orientation accuracy using the crowd flow cue, the cardinal direction cue and the joint model.

Site	Crowd flow	Cardinal Direction	Joint Model
Vatican Museums	27%	72%	91%
Hearst Castle	40%	40%	60%

datasets with fewer photos, like the Hearst Castle, where we only match 3 out of the 16 photos with compass heading to the assigned 3D pieces. Augmenting the image dataset by downloading more photos for the set of users is critical for the crowd flow and cardinal direction cues, as it vastly increases the number of reconstructed photos and also the number of reconstructed photos per user. In Table 4 we report statistics of the dataset expansion.

For each dataset, the integer quadratic program contained up to a thousand variables and was solved within 5 seconds on a single workstation with 12 cores.

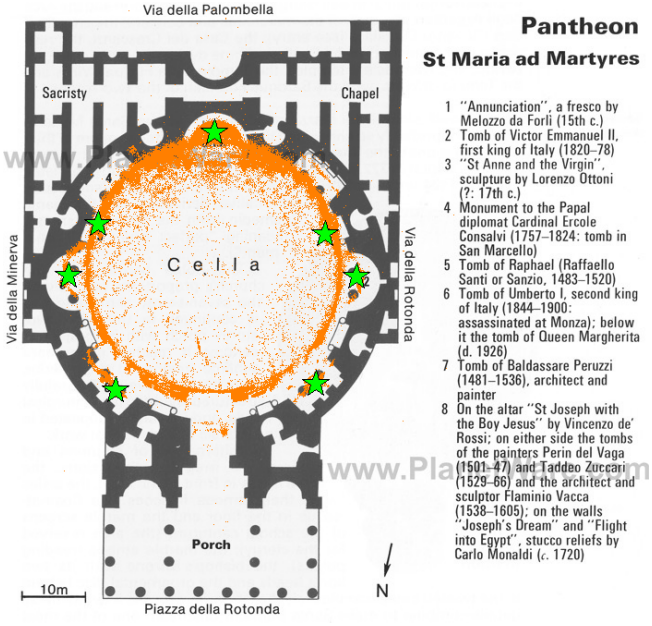


Fig. 6. Results for the Pantheon. 3D pieces are shown as the projection of the SfM points on the map, with different colors for each model. Green stars represent correct assignments, red stars incorrect ones.

Table 4. For each site, we report the number of Flickr users, number of photos before and after the dataset expansion and number of reconstructed photos before and after dataset expansion.

Site	Users	Photos		Recons. Photos	
		Before	After	Before	After
Vatican Museums	2112	11K	99K	4K	11K
Hearst Castle	367	3K	16K	828	3K

5.1 Failure cases

We have observed different failure cases of our system, showcasing the challenges of reconstructing indoor spaces from 3D pieces.

In some cases, the annotated text in the map may yield noisy image search results, leading to incorrect assignments. For example, in Figure 8(a), we show the model recovered for the point of interest labeled as "Round Vestibule" in the Vatican Museums that is actually the "Circular Hall", which is located in the same Pio Clementino Museum.

Another interesting case are the recovered 3D pieces corresponding to individual objects, such as the painting in the "Sobieski Room", shown in Figure 8(b). The room that contains the painting is rectangular and provides no cues for precise alignment of the object, even when the orientation is recovered from

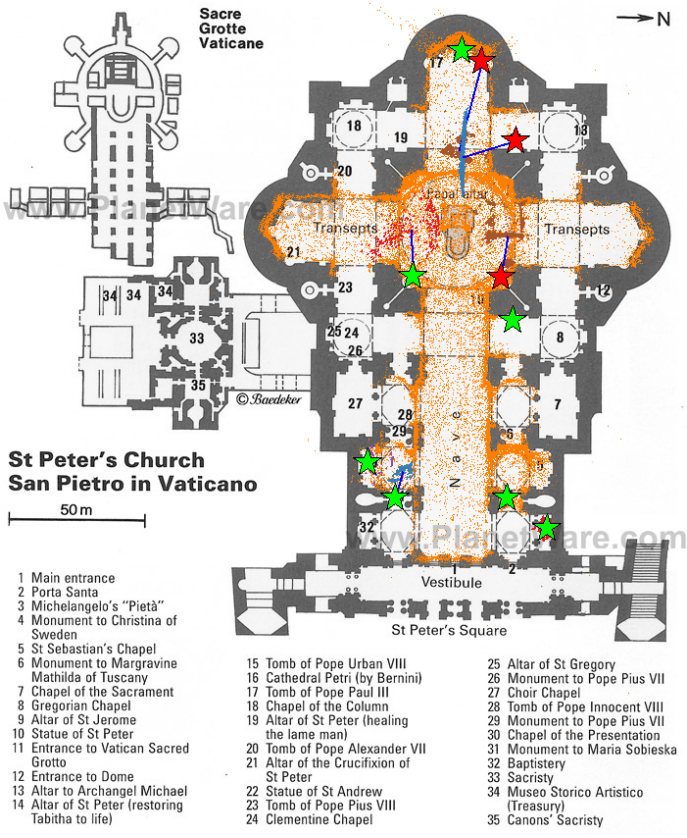


Fig. 7. Results for the St. Peter's Basilica. 3D pieces are shown as the projection of the SfM points on the map, with different colors for each model. Green stars represent correct assignments, red stars incorrect ones.

heading measurements. Our system can still provide a plausible alignment of the object along one of the walls, but the object might be scaled incorrectly.

Our system also fails to produce precise alignments to the walls of the rooms, such as the "Raphael Rooms" shown in Figure 8(c), due to inacurracies of the map. In the annotated map of the Vatican Museum dataset, the first three Raphael Rooms appear to have a 2:1 aspect ratio, although our 3D models indicate an aspect ratio closer to 1:1. By consulting other maps from different sources, we are able to determine that the aspect ratio of our models is actually correct, i.e., the map is wrong. Being able to register multiple maps together and detect these map inaccuracies is a promising direction for future work.

5.2 Navigation

We showcase the results of our indoor reconstructions via an interactive web visualization tool. We illustrate the interactions of the visualization tool in Fig-

(a) (b) (c)

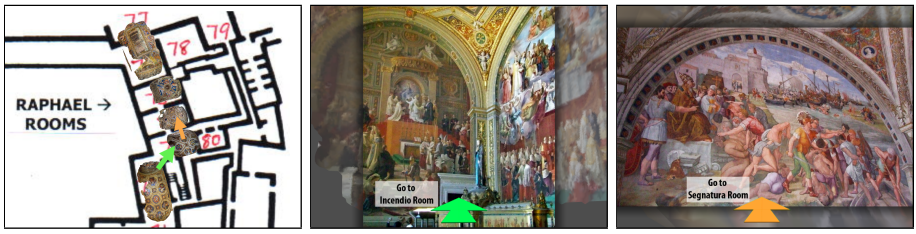
Fig. 8. Failure modes: (a) Incorrectly placed 3D model of the "Round Hall"; assigned point of interest marked in red, correct one in green, (b) ambiguous placement of object due to lack of scale and orientation information, (c) inaccurate map with incorrect aspect ratio for the rooms.

ure 9, but we refer the reader to the video available at the project website [5]. We feature two navigation modes to explore the map and reconstructed geometry. In map navigation mode, we allow common panning and zooming capabilities of the map. When you click on a room that has been assigned a 3D piece, the visualization automatically flies into the aligned 3D piece. You can navigate through the piece via an image-based rendering visualization, similar to the one in PhotoTourism [26]. When you look towards a neighbouring room, an arrow appears on the bottom of the screen pointing towards it. When you click on the arrow, the visualization transitions between the two rooms, recreating the experience of moving from one room to another.

6 Conclusion

This paper introduced the first system to reconstruct large indoor spaces of famous tourist sites from Internet photos via joint reasoning with map data and disconnected 3D pieces returned by structure-from-motion. We framed the problem as a 3D jigsaw puzzle and formulated an integer quadratic program with linear constraints that integrate cues over the pieces' position, scale, and orientation. We also introduced a novel crowd flow cue that measures how people travel through a site. Experiments on multiple sites showed consistently high precision for 3D model assignment and orientation relative to the input map, which allows for high quality interactions in the visualization tool. Our system works on popular tourist sites as it requires lots of images, text, and image metadata.

Acknowledgements: The research was supported in part by the National Science Foundation (IIS-1250793), the Intel Science and Technology Center for Visual Computing (ISTC-VC), the Animation Research Labs, and Google.



(a) (b) (c)

Fig. 9. Screenshots of our interactive visualization: (a) The annotated map is shown with the aligned 3D models rendered on top. When the user clicks on the model of the Hall of Immaculate Conception, the visualization flies into the room showing a photo taken in it (b). An arrow points to the location of the next room, the Incencio Room, and when clicked, the visualization flies the user to that room (c).

A Map parsing

Given an annotated map of a site, we seek to extract the spatial layout of the different rooms and objects depicted on the map. Automatically parsing a map is an interesting problem, but not strictly necessary for our task, as it would be straightforward to have manual workers parse maps for all leading tourist sites, or have future map-makers generate maps with the requisite annotations. For completeness, we describe a semi-automatic method of extracting the spatial layout and the object labels. We have restricted ourselves to annotated maps depicting the floorplan of a space, with referenced rooms and objects in the map appearing as text in a legend, as illustrated in Figure 6.

Our map parsing procedure begins by recovering a set of 2D regions from the floorplan corresponding to rooms, hallways, courtyards and other features of the site. We extract the floor plan of the map by clustering the pixel values found in the map image by K-means. We generate 2-4 clusters and manually select the cluster corresponding to the floorplan to form a binary image. To extract regions corresponding to the rooms we must close small gaps in the floor plan corresponding to doors and passages, which we achieve by simple morphological operations. We recover a segment for the room region by flood filling seeded by the room annotation marker on the map.

While OCR systems (e.g. Tesseract [27]) have shown much success in reading text in images, automatically recognizing text labels and markers in maps is still very difficult since the text is not generally structured into lines and may appear in different orientations, thus violating critical assumptions made by these systems. Moreover, markers and other visual elements appearing on the floorplan confuse the text line detection algorithms. The application of recently developed scene text recognition systems [28–30] to annotated maps remains outside the scope of this work and an interesting topic for future work. For our purposes we have manually annotated the map using LabelMe [31] by marking each text label or marker with the appropriate text label.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building rome in a day. In: International Conference on Computer Vision. (2009)
2. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Commun. ACM **54**(10) (October 2011) 105–112
3. Shan, Q., Adams, R., Curless, B., Furukawa, Y., Seitz, S.M.: The visual Turing test for scene reconstruction. In: Joint 3DIM/3DPVT Conference (3DV). (2013)
4. Russell, B.C., Martin-Brualla, R., Butler, D.J., Seitz, S.M., Zettlemoyer, L.: 3D Wikipedia: Using online text to automatically label and navigate reconstructed geometry. ACM Transactions on Graphics (SIGGRAPH Asia 2013) **32**(6) (November 2013)
5. http://grail.cs.washington.edu/projects/jigsaw3d
6. Wang, C.P., Wilson, K., Snavely, N.: Accurate georegistration of point clouds using geographic data. In: 3DV. (2013)
7. Levin, A., Szeliski, R.: Visual odometry and map correlation. In: IEEE Computer Vision and Pattern Recognition or CVPR. (2004)
8. Brubaker, M., Geiger, A., Urtasun, R.: Lost! Leveraging the crowd for probabilistic visual self-localization. In: IEEE Computer Vision and Pattern Recognition or CVPR. (2013)
9. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A.A., Hertzmann, A.: Image sequence geolocation with human travel priors. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV '09). (2009)
10. Simon, I.: Scene Understanding Using Internet Photo Collections. PhD thesis, University of Washington (2010)
11. Kaminsky, R., Snavely, N., Seitz, S.M., Szeliski, R.: Alignment of 3D point clouds to overhead images. In: Workshop on Internet Vision. (2009)
12. Koller, D., Trimble, J., Najbjerg, T., Gelfand, N., Levoy, M.: Fragments of the city: Stanford's digital forma urbis romae project. J. Roman Archaeol. Suppl. **61** (2006) 237–252
13. Cho, T.S., Avidan, S., Freeman, W.T.: The patch transform. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(8) (2010) 1489–1501
14. Cho, T.S., Avidan, S., Freeman, W.T.: A probabilistic image jigsaw puzzle solver. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)
15. Gallagher, A.: Jigsaw puzzles with pieces of unknown orientation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012)
16. : Google art project. http://www.google.com/culturalinstitute/project/art-project
17. Xiao, J., Furukawa, Y.: Reconstructing the world's museums. In: Proceedings of the 12th European Conference on Computer Vision. (2012)
18. Liu, T., Carlberg, M., Chen, G., Chen, J., Kua, J., Zakhor, A.: Indoor localization and visualization using a human-operated backpack system. In: Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on. (2010)
19. Xiao, J., Owens, A., Torralba, A.: SUN3D: A database of big spaces reconstructed using SfM and object labels. In: International Conference on Computer Vision. (2013)
20. Bemporad, A.: Hybrid Toolbox - User's Guide (2004) http://cse.lab.imtlucca.it/ bemporad/hybrid/toolbox.
21. Demaine, E., Demaine, M.: Jigsaw puzzles, edge matching, and polyomino packing: Connections and complexity. Graphs and Combinatorics **23** (2007)

22. Wu, C., Agarwal, S., Curless, B., Seitz, S.M.: Multicore bundle adjustment. In: In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2011) 3057–3064
23. Wu, C.: VisualSFM - a visual structure from motion system. http://ccwu.me/vsfm/ (2011)
24. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. IEEE Trans. Pattern Analysis and Machine Intelligence **32**(8) (2010) 1362–1376
25. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the 4th Eurographics Symposium on Geometry Processing (SGP). (2006) 61–70
26. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. ACM Trans. Graph. **25**(3) (July 2006) 835–846
27. : tesseract-ocr. https://code.google.com/p/tesseract-ocr/
28. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. 2013 IEEE Conference on Computer Vision and Pattern Recognition **0** (2010) 2963–2970
29. Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. CoRR **abs/1312.6082** (2013)
30. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: Photoocr: Reading text in uncontrolled conditions. In: The IEEE International Conference on Computer Vision (ICCV). (December 2013)
31. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. Int. J. Comput. Vision **77**(1-3) (May 2008) 157–173