3D Wikipedia: Using online text to automatically label and navigate reconstructed geometry

Bryan C. Russell et al.
SIGGRAPH Asia 2013

Presented by YoungBin Kim

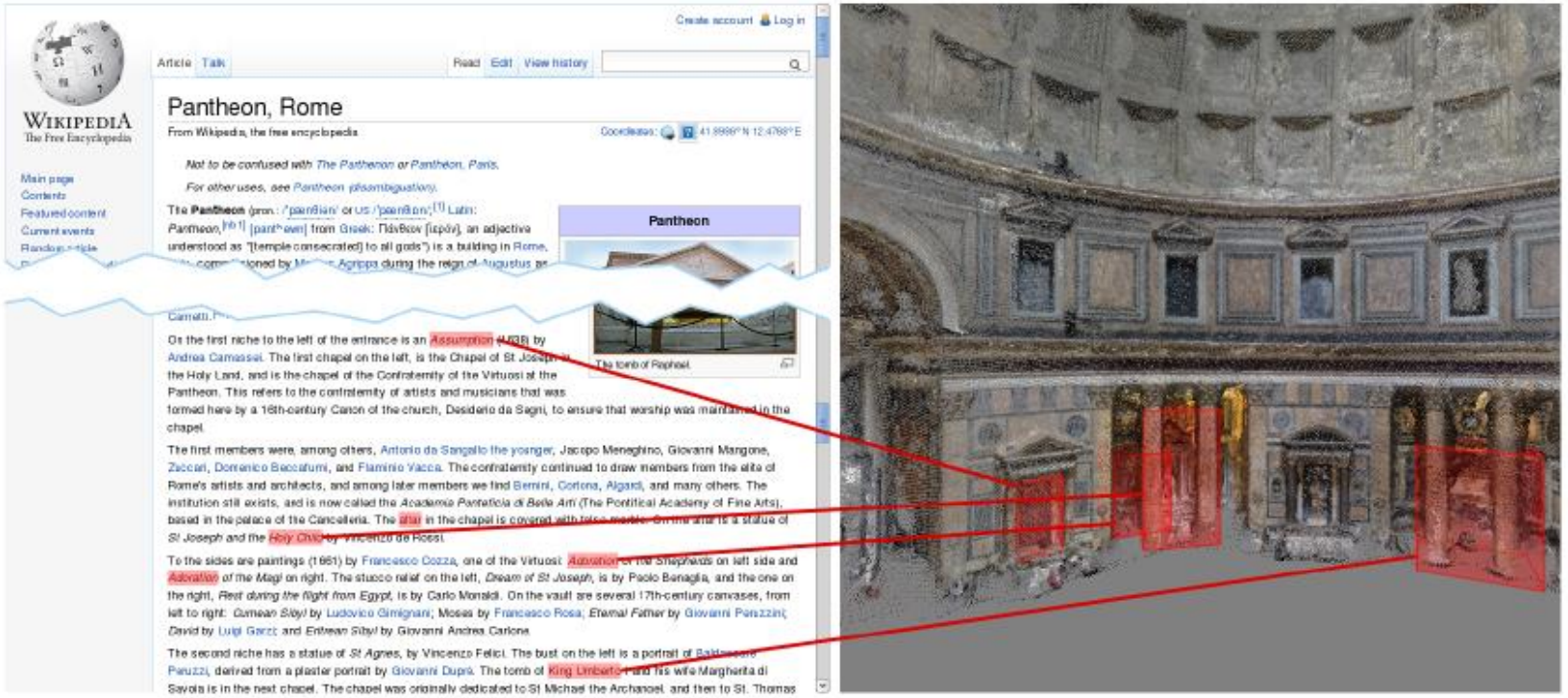
2014. 01. 10

Computer Graphics @ Korea University

KOREA UNIVERSITY

KUCG

Abstract



- **Produce annotated 3D models of famous tourist sites**
  - Analyzing Wikipedia and other text
  - Together with online photos
  - Automatically discover and link objects in text and 3D geometry

Introduction

Introduction

- Guidebooks
  - Packed with interesting historical facts
  - Descriptions of site specific objects and spaces
    - Difficult to fully visualize the scenes they present
    - Difficult to understand the spatial relationships between each image viewpoint
- Online sites
  - Do not have space restrictions
    - But similarly sparse and disconnected visual coverage

Introduction

- Our goal
  - Interactive, photorealistic visualization
    - Wikipedia page is shown next to a detailed 3D model of the described site
  - Create such a visualization completely automatically
    - Analyzing the Wikipedia page itself
    - Together with many photos of the site available online

Introduction

- Automatically creating such a visualization
  - Formidable challenge
    - Text and photos
      - Provide only very indirect cues about the structure of the scene
    - Automatically extracting the names of objects is not trivial
      - Name the artist that created the object, or other unrelated concept
    - Determining the precise 3D location of each described object
      - Even more challenging

Introduction

- Key to our approach
  - Mine text and photo co-occurrences across all of the Internet
    - ex) photo anywhere on the Internet with the caption "Annunciation, Pantheon"
  - This simple strategy does not completely solve the problem
    - Hence, we treat the image results as a noisy signal

Annunciation, Pantheon

Coronation of the Virgin, Pantheon

Tuscan School, Pantheon

Introduction

- Our reconstruction and visualization approach
  - Inspired by Photo Tourism
    - [Snavely et al., SIGGRAPH 2006]
  - Employ similar techniques to generate 3D models
  - Render transitions to photos within those models
    - VisualSFM : A Visual Structure from Motion System
    - http://homes.cs.washington.edu/~ccwu/vsfm/

- We show compelling results for several major tourist sites
  - Able to reliably extract many of the objects in each scene
    - With relatively few errors

Related work

Related work

- Natural language processing and 3D computer vision
  - Very fertile area with little prior research

- Scene segmentation using the wisdom of crowds
  - [Simon and Seitz, ECCV '08]
  - Segmenting and labeling 3D point clouds
    - Analyzing SIFT feature co-occurence in tagged Flickr photos
  - Flickr tags are notoriously noisy
  - Far less informative
    - Compared to Wikipedia and other authoritative guides

Related work

- In the 2D domain
  - Correlating regions in images/video to captioned text or keywords
    - [Barnard et al. 2003; Laptev et al. 2008; Cour et al. 2011]
  - Generating sentences or captions for specific images
    - [Farhadi et al. 2010; Berg et al. 2012; Mitchell et al. 2012].
  - Relatively small set of object classes (e.g. car, boat)
  - Require captioned photographs during the training of their model

Related work

- Reconstructing 3D models of tourist sites
  - From Internet photo collections
  - Structure-from-motion
    - [Snavely et al. 2008; Agarwal et al. 2011; Raguram et al. 2011]
  - Multi-view stereo
    - [Furukawa and Ponce 2010; Furukawa et al. 2010; Goesele et al. 2007]
- Recognition in RGB-D and range-scan data
  - Advent of commodity depth sensors
    - Like Kinect
    - [Ren et al. 2012; Silberman et al. 2012; Ladick´y et al. 2012]

- We focus on labeling instances

Related work

- Recognizing images of specific objects or places (instances)
  - Large-scale image retrieval
    - [Sivic and Zisserman 2003; Chum et al. 2007; Philbin et al. 2008]
  - Matching local features computed at interest points
    - Between an input image and a database of labeled images
    - [Lowe 2004]
  - GPS-tagged images
    - [Crandall et al. 2009; Hays and Efros 2008]
  - Require a database of labeled objects as reference

- Our focus is to create such a database
  - From joint analysis of text and images

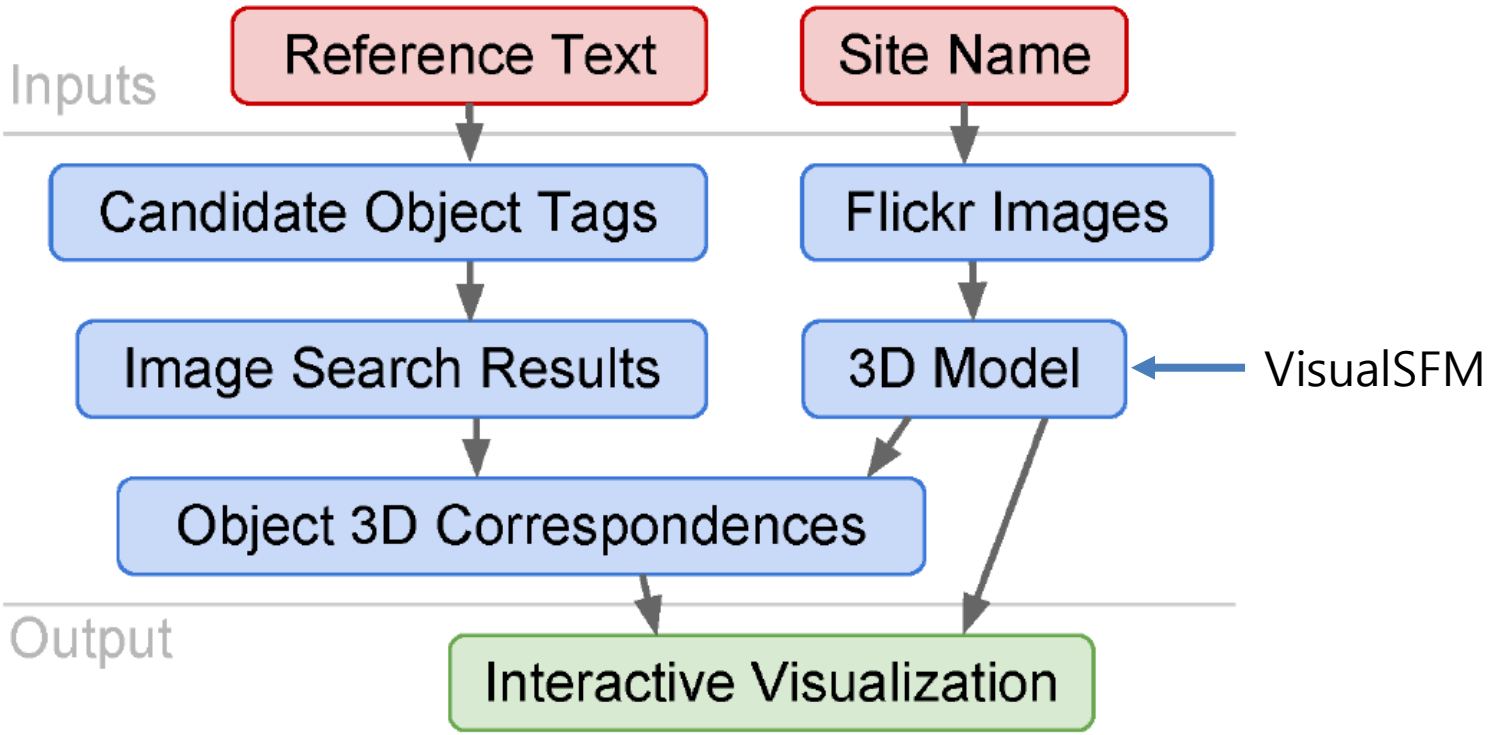
System Overview

System Overview

- We present a fully automatic system
  - Generates interactive visualizations
    - Link authoritative text sources with photorealistic 3D models
  - System requires two types of inputs
    - One or more reference text sources
      - Such as Wikipedia
    - Unique name for the site to reconstruct
      - Such as the Pantheon in Rome

System Overview

- Overview of the complete approach



Automatic labeling of 3D models from text

Automatic labeling of 3D models from text

- Algorithm consists of three steps
  1. Generate an over complete list
     - Candidate object hypothesis from the text
  2. Obtain their likely location on the 3D model
  3. Filter the large number of false positive detections
     - By training a classifier over features

Obtaining object hypotheses from text

- Seek to automatically obtain a list of candidate descriptive phrases
  - For each site
  - Texts come from two sources
    - Freely available online
    1. Articles from Wikipedia
    2. Text from other, site specific, third-party web pages

Obtaining object hypotheses from text

- Use the syntactic structure of the language
  - To define the set of possible descriptive phrases
    - Primarily leveraging the fact that noun phrases
      - Can name physical objects in English

- Extract noun phrases
  - Use the Stanford parser
    - [Klein and Manning 2003]
    - Achieves near state-of-the-art performance
    - Available as public-domain software
      - Ran the parser with the default parameter settings

Obtaining object hypotheses from text

- Boost recall
  - Extract prepositional phrases that are immediately followed by a noun phrase
    - e.g. a fresco of the Annunciation
  - Merge adjacent noun phrases
    - e.g. a canvas by Clement Maioli of St. Lawrence and St. Agnes
  - These additional phrases allow us to overcome parsing errors

- Reduce false positives
  - Remove phrases containing only a single stop word
    - As defined by a commonly used stop word list
  - Remove phrases containing only numerals

Obtaining object hypotheses from text

"The first chapel on the right, the Chapel of the Annunciation, has a fresco of the Annunciation attributed to Melozzo da Forli."

- annunciation
- annunciation attributed to melozzo da forli
- chapel
- chapel of the annunciation
- da forli
- first chapel
- first chapel on the right
- first chapel on the right the chapel of the annunciation
- forli
- fresco
- fresco of the annunciation
- fresco of the annunciation attributed to melozzo da forli
- melozzo
- melozzo da
- melozzo da forli
- right

Link objects to 3D geometry

- Generate proposal regions for their 3D location within the site
  - Given the automatically obtained list of candidate named objects

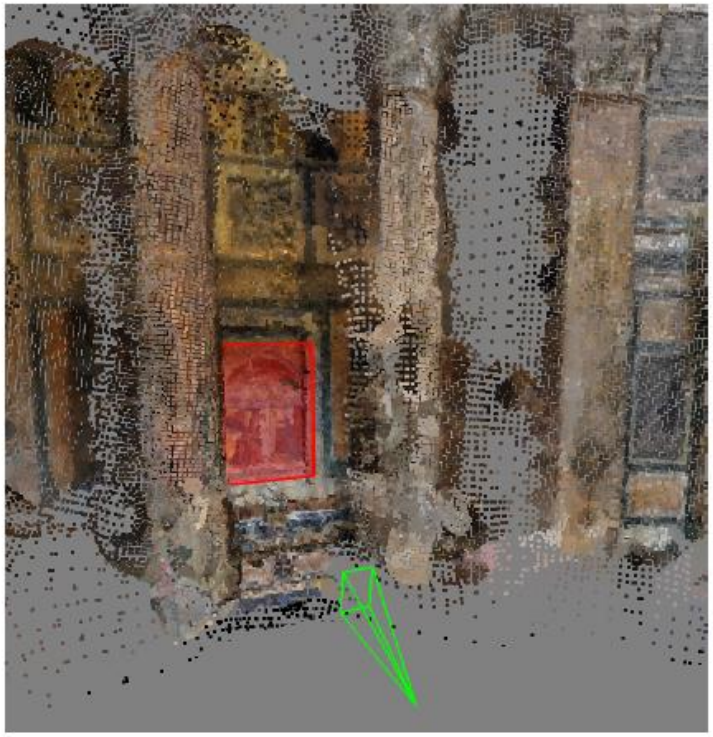
- Search for and download images using Google image search
  - For each candidate named object
  - Construct the query terms
    - By concatenating the extracted noun phrase with the place name
      - e.g. central figure Trevi Fountain

Link objects to 3D geometry

- To find candidate regions within the 3D model for the site
  - Build upon the success of feature matching and geometric verification
    - "Modeling the world from Internet photo collections"
      - [Snavely et al., International Journal of Computer Vision, 2008]
  - Match SIFT key points extracted from the downloaded images
    - To the inlier SIFT key points corresponding to 3D points within the 3D model

Link objects to 3D geometry

- Recover the camera parameters
  - Using the putative 2D-3D point correspondences
  - Camera resectioning
    - [Hartley and Zisserman 2004]



Link objects to 3D geometry

- Perform camera resectioning for the top 6 images
  - For each search query

- Keep the alignment
  - If camera resectioning finds at least 9 inlier correspondences
  - At least 9 inlier features
    - Almost always yields a correct alignment to the 3D model
    - Using fewer yields incorrect alignments
  - This requirement
    - Discards many images that do not depict the site at all
    - Maintains a high recall for valid images that do depict the site

Model for filtering hypotheses

- Internet image search
  - Returns many valid images for the candidate object tags
  - Remains a high number of false positives
- Over-generated list of candidate objects resulting
  - Output of the natural language processing parser

- Our goal
  - Extract good object detections
    - From the hypothesis set of object-region pairs

Model for filtering hypotheses

- Start by merging highly-overlapping camera frustra
  - Corresponding to the aligned images for a given object tag
  - First project each frustrum onto a reference image
    - i.e. panorama or perspective photograph
    - Depicting the site that has been registered to the 3D model
  - Form a bounding box by taking the maximum x, y
    - Extent of the projected frustrum
  - Merge two frustra
    - If their relative overlap exceeds 0.5
      - Ratio of intersection area to their union
    - Mean of their bounding boxes returned
  - Results in a set of object tag and detection frustrum pairs
    - Dubbed the *candidate pool*

Model for filtering hypotheses

- Extract features
  - From the candidate pool and the original text
  - Visual features
    - Number of merged frustra for the candidate
    - Rank number for the top-ranked image search result
      - That aligned to the 3D model
    - Total number of frustra across all object tags
      - That overlap the candidate frustrum
        - High number indicates a generic viewpoint of the site

Model for filtering hypotheses

- Extract features
  - Text features
    - Whether a non-spatial preposition
      - (ago, as, because of, before, despite, during, for, like, of, since, until)
      - Resides in the same sentence as the extracted noun phrase
        - Which often corresponds to historical descriptions
    - Whether the tag corresponds to an author
    - Whether an author appears in the same sentence as the tag

- Presence of an author
  - As a feature
    - Authorship of an object is often described together
  - Detect the presence of an author
    - Analyzing prepositional *by* dependencies

Model for filtering hypotheses

- Train a linear classifier
  - Using logistic regression across a set of training sites
- Test on the remaining sites

- Construct the training set
  - Project each frustra in the candidate pool for the site
    - Onto the reference image
  - Intersect the projected frustra with objects
    - That have been manually labeled via LabelMe
      - [Russell et al. International Journal of Computer Vision 2008]

Model for filtering hypotheses

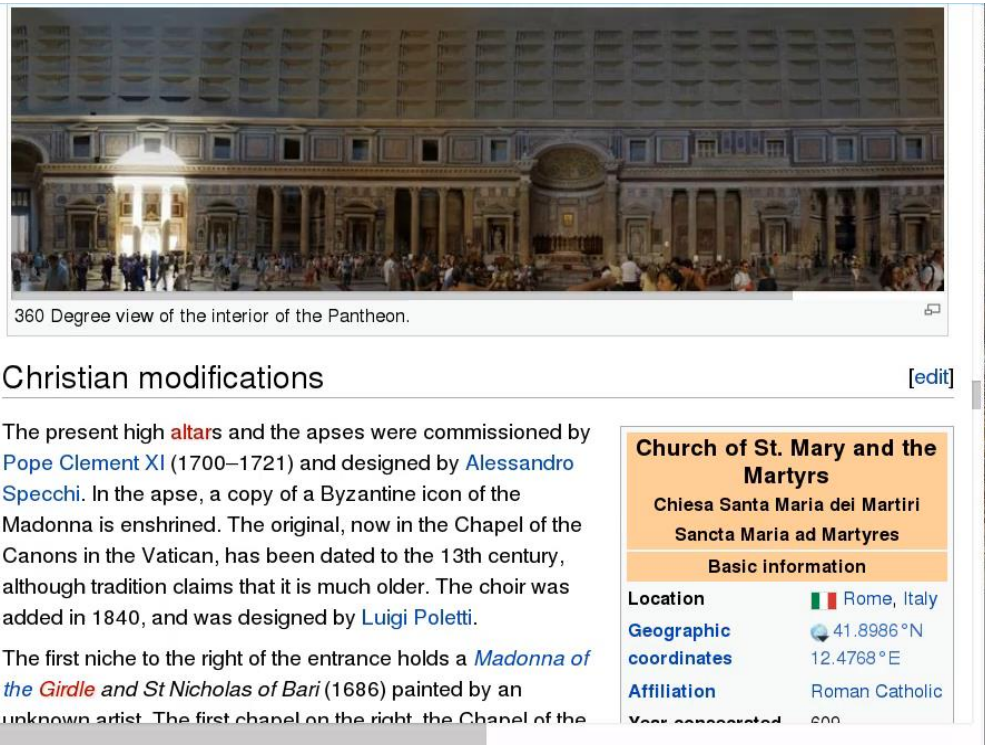
- For each labeled object
  - Keep the object tag/detection frustrum pair
    - Has highest word F-score
      - When comparing the object and labeled tags
    - Having the center of their bounding boxes
      - Residing in the other's bounding box
- Form the set of positive examples
  - Tag/frustrum pairs
    - That match to a ground truth label
- Form the set of negative examples
  - Tag/frustrum pairs
    - That do not have tag or frustrum overlap with any of the positive training examples

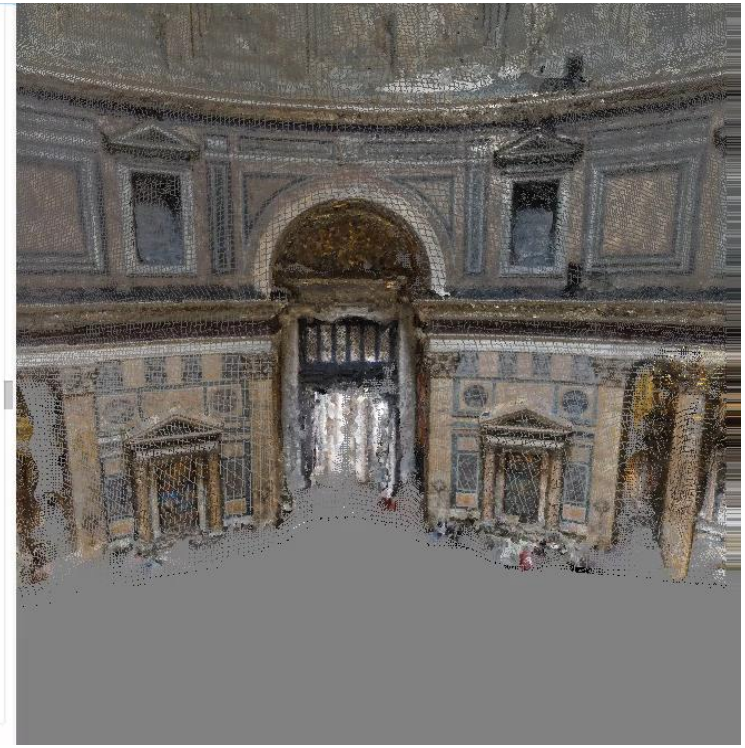
Model for filtering hypotheses

- During testing, perform non-maximum suppression
  - Suppress detections
    - If a higher confidence frustrum overlaps a lower confidence one
      - Their relative overlap exceeds 0.3 and their centers reside in the other's bounding box
    - If any of the tag words overlap in the same sentence

Visualization tool for browsing objects in online text

Text-to-3D navigation





3D-to-text navigation

Plan of the architectural elements, real and illusionary

Real [edit]

The Sistine Chapel is 40.5 metres long and 14 metres wide. The ceiling rises to 20 metres above the main floor of the chapel. The vault is of quite a complex design and it is unlikely that it was originally intended to have such elaborate decoration. Pier Matteo d'Amelia provided a plan for its decoration with the architectural elements picked out and the ceiling painted blue and dotted with gold stars, similar to that of the Arena Chapel decorated by Giotto at Padua.[25]

The chapel walls have three horizontal tiers with six windows in the upper tier down each side. There were also two windows at each end, but these have been closed up above the altar when Michelangelo's Last Judgement was painted, obliterating two lunettes. Between the windows are large *pendentives* which support the vault. Between the pendentives are triangularly shaped arches or *spandrels* cut into the vault above each window. Above the height of the pendentives, the ceiling slopes gently without much deviation from the horizontal.[nb 5] This is the *real* architecture. Michelangelo has elaborated it with illusionary or *fictive* architecture.

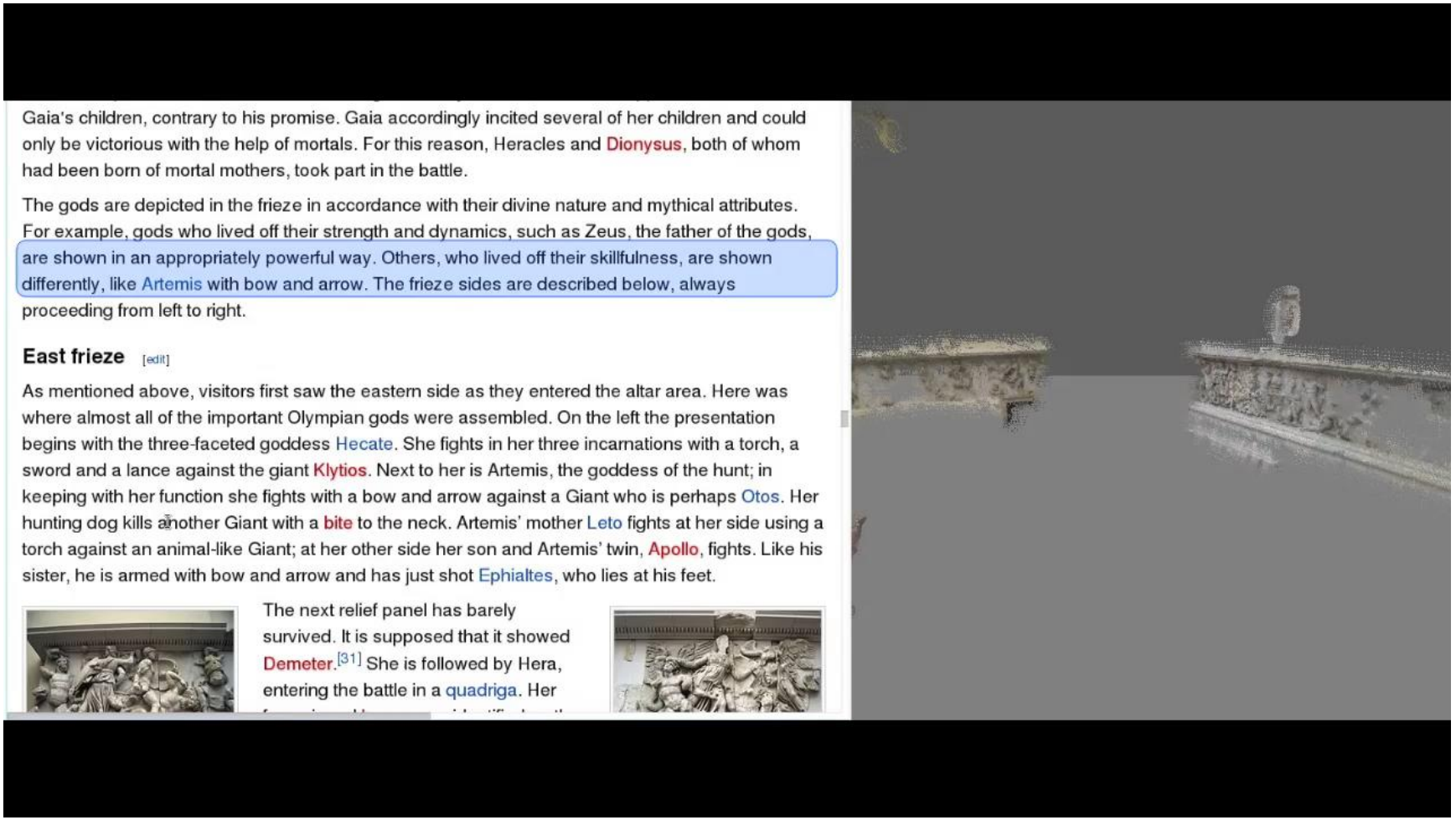
Illusionary [edit]

The first element in the scheme of painted architecture is a definition of the *real* architectural elements by accentuating the lines where spandrels and pendentives intersect with the curving vault. Michelangelo painted these as decorative courses that look like sculpted stone moldings.[nb 6] These have two repeating motifs, a formula common in Classical architecture.[nb 7] Here, one motif is the acorn, the symbol of the family of both Pope Sixtus IV who built the chapel and Pope Julius II who commissioned Michelangelo's work.[nb 8][26] The other motif is the scallop shell, one of the symbols of the Madonna, to whose assumption the chapel was dedicated in 1483.[nb 9][27] The crown of the wall then rises above the spandrels, to a strongly projecting painted cornice that runs right around the ceiling, separating the pictorial areas of the biblical scenes from the figures of Prophets, Sibyls and Ancestors, who literally and figuratively support the narratives. Ten broad painted crossribs of travertine cross the ceiling and divide it into alternately wide and narrow pictorial spaces, a grid that gives all the figures their defined place.[28]

God dividing the waters, showing the illusionary architecture, and the positions of the *Ignudi* and shields

A great number of small figures are integrated with the painted architecture, their purpose apparently purely decorative. These include two faux marble *putti* below the cornice on each rib, each one a male and female pair; stone rams-heads are placed at the apex of each spandrel; copper-skinned nude figures in varying poses, hiding in the shadows, propped between the spandrels and the ribs like animated bookends; and more putti, both clothed and unclothed strike a variety of poses as they support the nameplates of the Prophets and Sibyls.[29] Above the cornice and to either side of the smaller scenes are an array of round shields, or medaillons. They are framed by a total of twenty more figures, the so-called *Ignudi*, which are not part of the architecture but sit on

Tour navigation



Implementation details

- Automatically create the 3D models from Internet photo collections
  - VisualSFM [Wu et al., CVPR 2011]
    - For generating a sparse point cloud
  - PMVS [Furukawa and Ponce, PAMI 2010]
    - For generating a dense point cloud
  - Poisson Surface Reconstruction [Kazhdan et al. SGP 2006]
    - To generate a mesh

Implementation details

- Then delete
  - Small connected components of the mesh
  - Vertices that lie far away from the PMVS points
- Then color the mesh vertices
  - According to the closest PMVS points
  - Keep the vertices of the mesh as our final point cloud

- We only use the vertices from the mesh for visualizations
  - Although we generate colored meshes
  - Point cloud is visually more forgiving of artifacts

Implementation details

- To highlight the objects in the 3D model
  - Generate 3D bounding boxes for each object
    - That are rendered semi-transparently in our visualizations
  - First compute the mode m
    - Over the distribution of the normals of the points
      - That lie in the frustra of the images
  - Then choose a coordinate unit-vector x
    - In the world-coordinate frame of the reconstruction
      - That is approximately orthogonal to the mode
  - Finally, calculate the other axis vector
    - $y = \frac{\bar{y}}{\|\bar{y}\|}$ with $\bar{y} = m - (m \cdot x)x$
    - $z = x \times y$

Evaluation

Evaluation

- Manually evaluate the accuracy of the correspondences

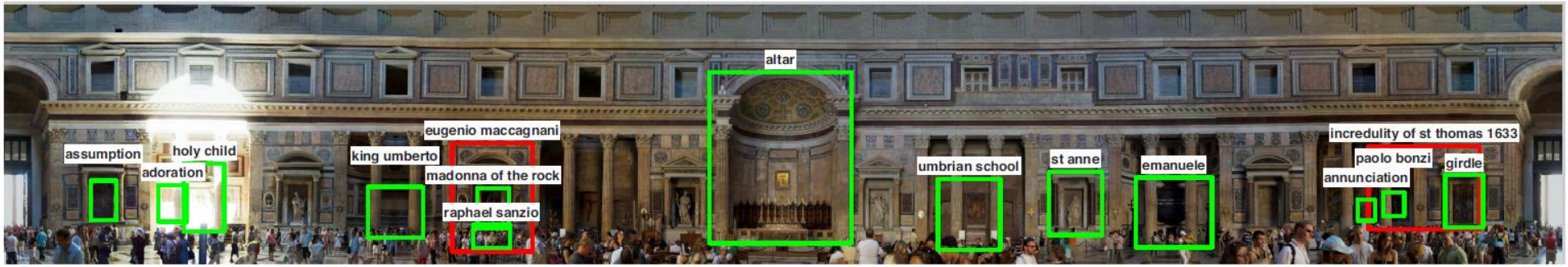
Site	Pantheon, Rome	Trevi Fountain	Sistine Chapel	US Capitol Rotunda	Pergamon Altar
# 3D points	146K	208K	121K	84K	55K
# ground truth	31	16	31	38	49
# noun phrases	1796	821	3288	2179	2949
# image matches	510	348	2282	884	1600

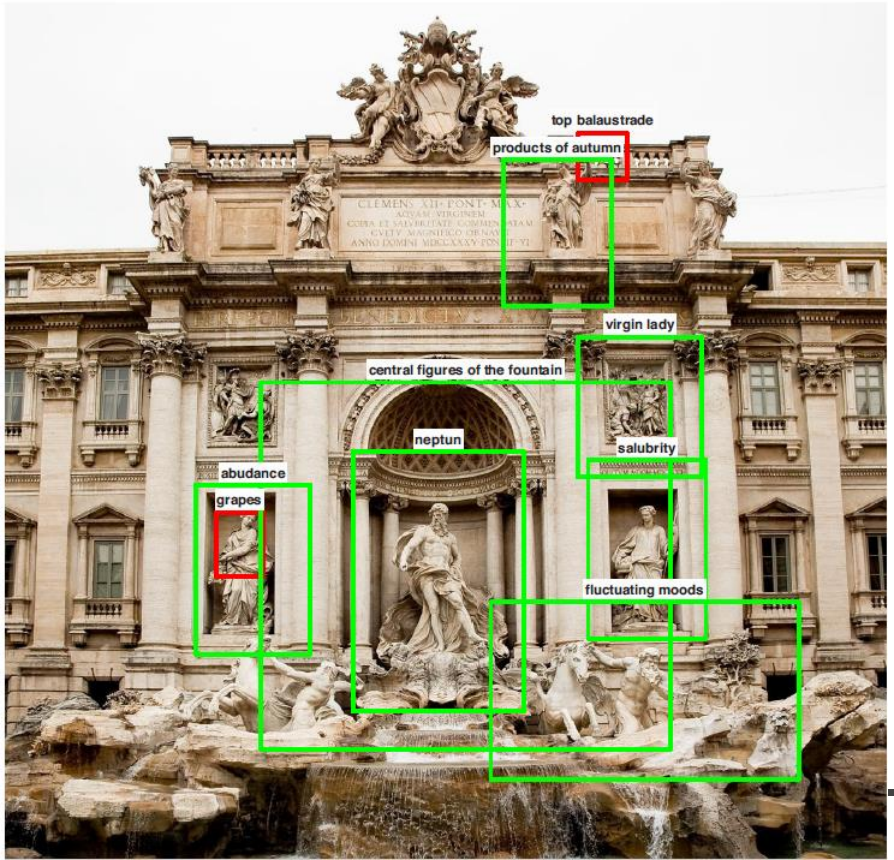
  - Evaluate performance relative to a set of canonical views
    - Set of registered panoramas
    - Set of Perspective photographs depicting most or all of a site
      - Manually labeled the name and a bounding box

Site	Pantheon, Rome	Trevi Fountain	Sistine Chapel	US Capitol Rotunda	Pergamon Altar
Recall	0.39	0.31	0.71	0.21	0.18
Raw Precision	0.80	0.31	0.46	0.35	0.56
Full Precision	**0.87**	**0.78**	**0.79**	**0.65**	**0.94**

  - Raw Precision : labeled ground truth as a guide
  - Full Precision : manually verified detections

Evaluation


altar
assumption
holy child
adoration
king umberto
eugenio maccagnani
madonna of the rock
raphael sanzio
umbrian school
st anne
emanuele
incredulity of st thomas 1633
paolo bonzi
annunciation
girdle

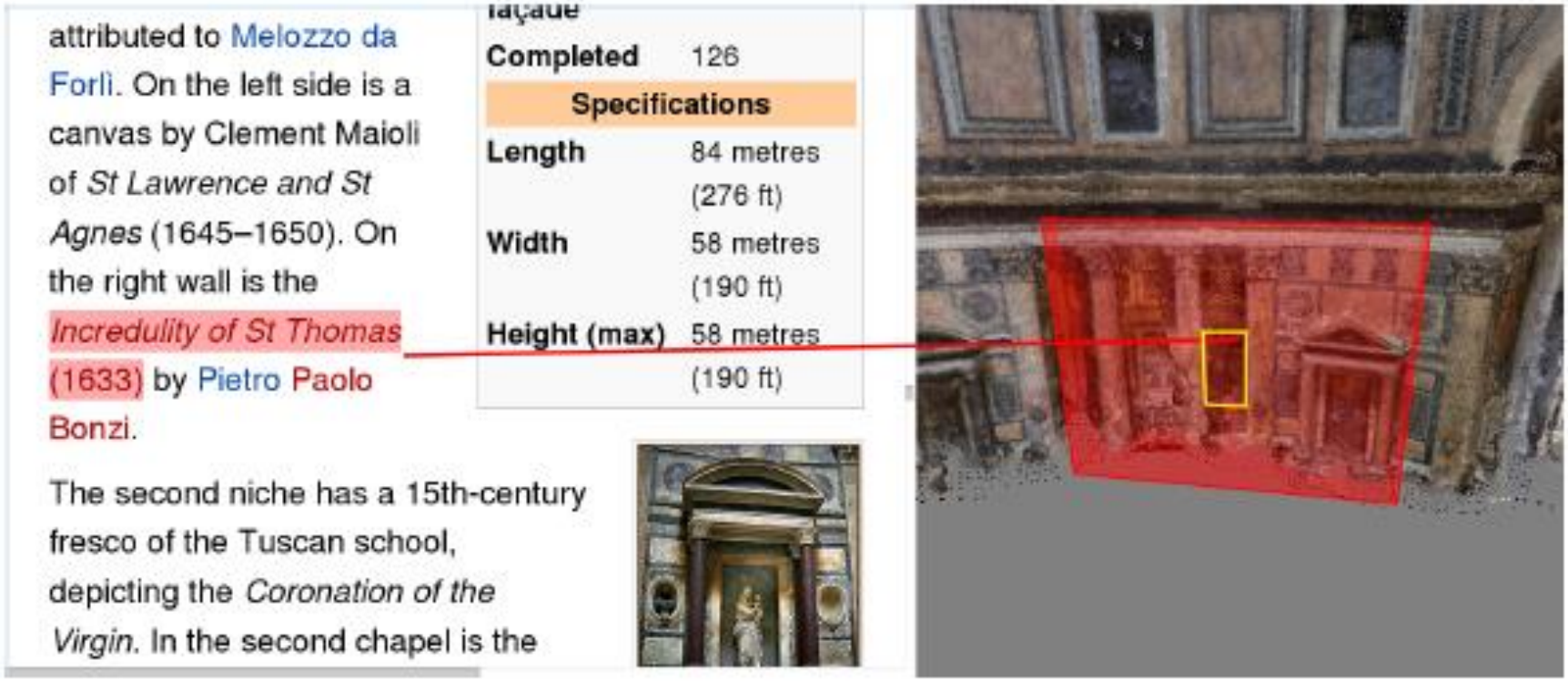

top balaustrade
products of autumn
virgin lady
central figures of the fountain
neptun
salubrity
abudance
grapes
fluctuating moods

Evaluation

- Object recall for the sites in which we retained the Flickr tags
  - Pantheon – 0.06
  - Trevi Fountain – 0
  - US Capitol Rotunda – 0.21
  - Flickr baseline performs significantly worse

Error Analysis

- Bounding box is too large

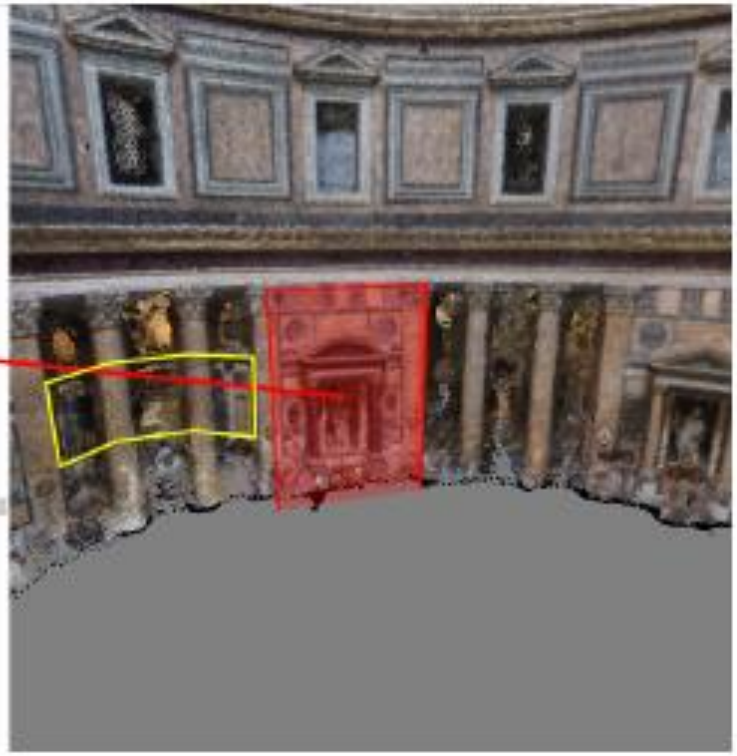


Error Analysis

- Incorrect object tag
  - Typically come from noisy co-occurrences
    - Between images and text in the online sources

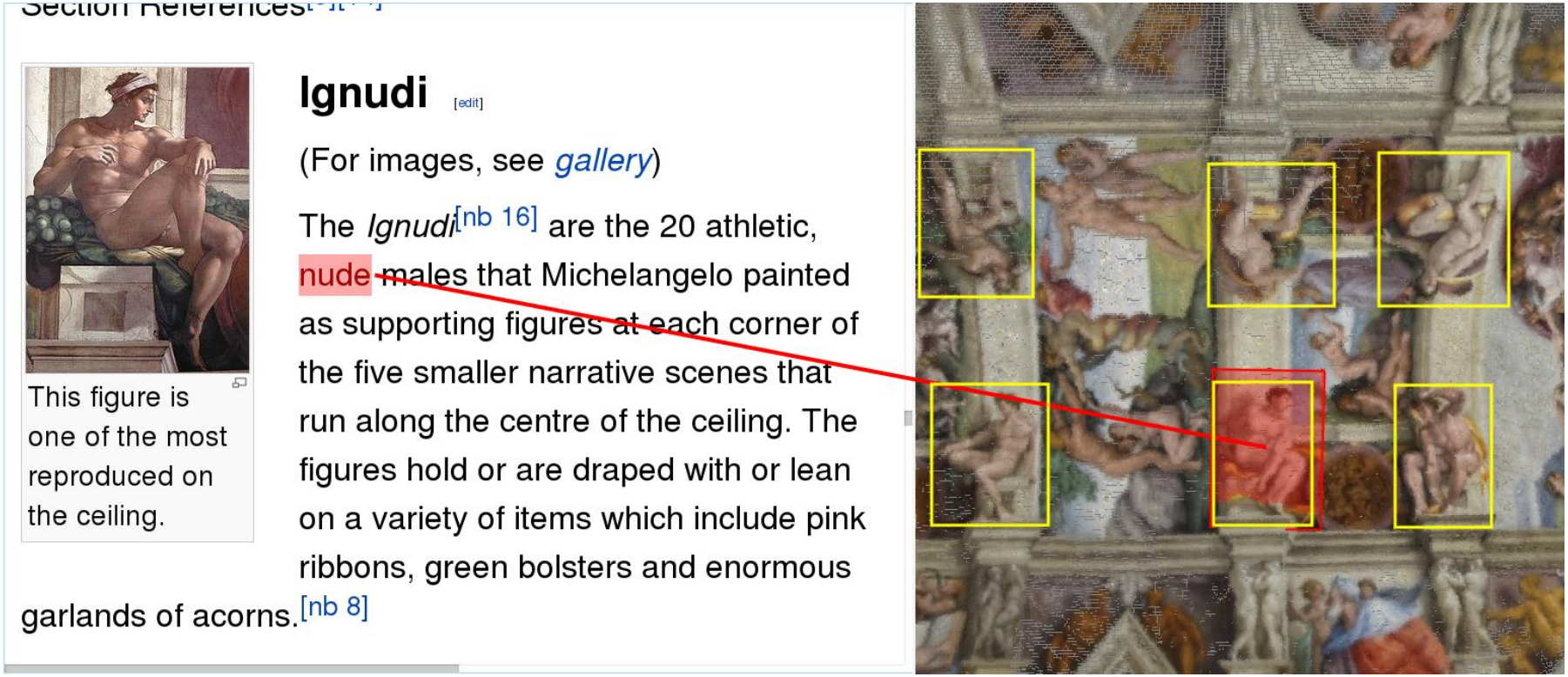
Archangel, and then to St. Thomas the Apostle. The present design is by Giuseppe Sacconi, completed after his death by his pupil Guido Cirilli. The tomb consists of a slab of alabaster mounted in gilded bronze. The frieze has allegorical representations of *Generosity*, by Eugenio Maccagnani, and *Munificence*, by Arnaldo Zocchi. The royal tombs are maintained by the National Institute of Honour Guards to the Royal Tombs, founded in 1878. They also organize picket guards at the tombs. The altar with the royal arms is by Cirilli.

The third niche holds the mortal remains – his Ossa et cineres, "Bones and ashes", as the inscription on the sarcophagus says – of the great artist Raphael. His



Error Analysis

- Multiple object class instances



Conclusion

Conclusion

- First system
  - Automatically build immersive 3D visualizations of popular sites
    - Using online text and photo collections
  - Built using off-the-shelf ingredients
    - Ideas and the system are new
      - Based on crowd-sourced data on the Internet
      - Insight of using text parsing + Google image search
        - Connect web text to 3D shape data
      - Viable approach
        - Incorporating a series of sophisticated steps

- Room for improvement
  - To leverage *spatial* terms
  - People to assist in the labeling task