Multi-modal Registration of Visual Data

Massimiliano Corsini

Visual Computing Lab, ISTI - CNR - Italy

Overview

- Introduction and Background
- Features Detection and Description (2D case)
- Features Detection and Description (3D case)
- Image-geometry registration
- Recent Advances and Applications

Overview

- Introduction and Background
- Features Detection and Description (2D case)
- Features Detection and Description (3D case)
- Image-geometry registration
- **Recent Advances and Applications**
  - **2D-to-3D for image-based localization and video registration**
  - **Video navigation/exploration**
  - **Paintings-to-3D models**
  - **Advancements in object detection/recognition and understanding of joint properties of 3D models and images**

Recent Advances and Applications

Image-based Localization

- *Where my photo is ? (location, position and orientation)*
- Modern solution: image reconstruction (SFM) + search for 2D-to-3D correspondences
  - High accuracy (more than GPS + camera orientation).
  - Methods can be distinguished between *indirect* and *direct*.

Image-based Localization

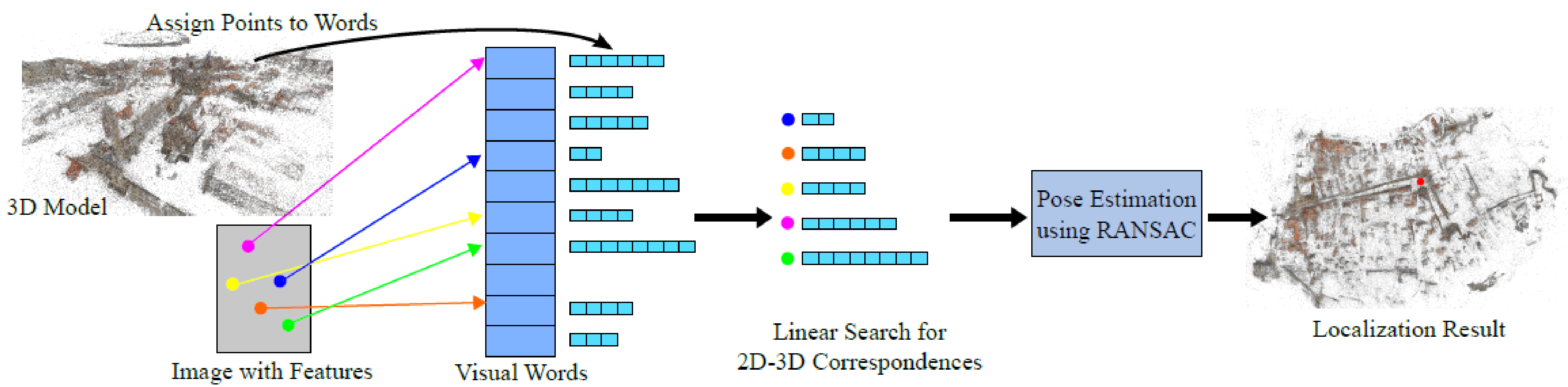
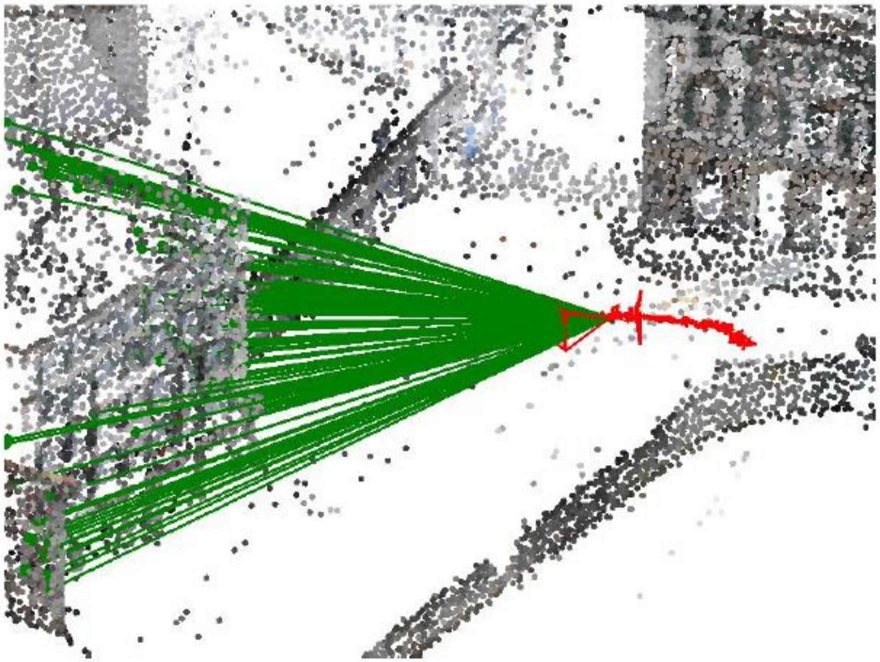
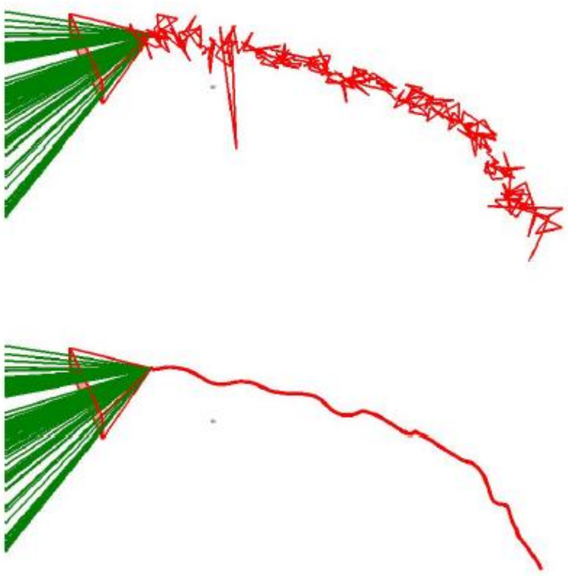


Image from Torsten Sattler, Bastian Leibe, Leif Kobbelt, "Fast Image-Based Localization using Direct 2D-to-3D Matching", *Proc. of ICCV2011*, 2011.

Video-to-SFM Registration

- 2D-to-3D solutions are not feasible (e.g. image-based localization) → ad hoc solutions are necessary (positional error dominates the estimation).





Average position error is 50x the true average position steps.

Image from Till Kroeger and Luc Van Gool, "Video Registration to SFM Models", *Proc. of ECCV2014*, 2014.

Video-to-SFM Registration

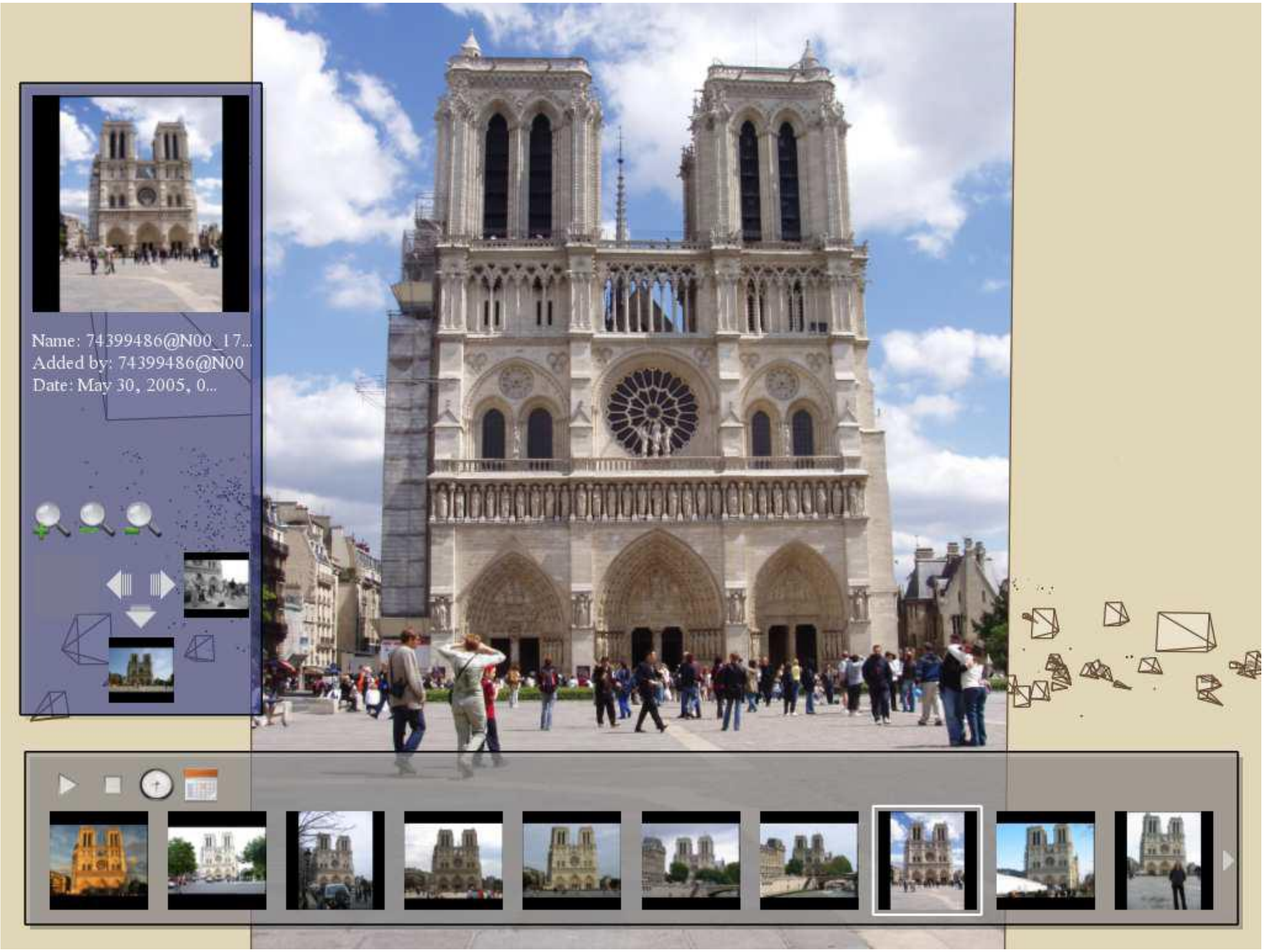
- SLAM is a related problem but..
  - Typically small environments (e.g. not an entire city)
  - 3D scene is not known a priori
  - Feature tracking is performed jointly to the reconstruction/camera position estimation

Video Navigation/Exploration

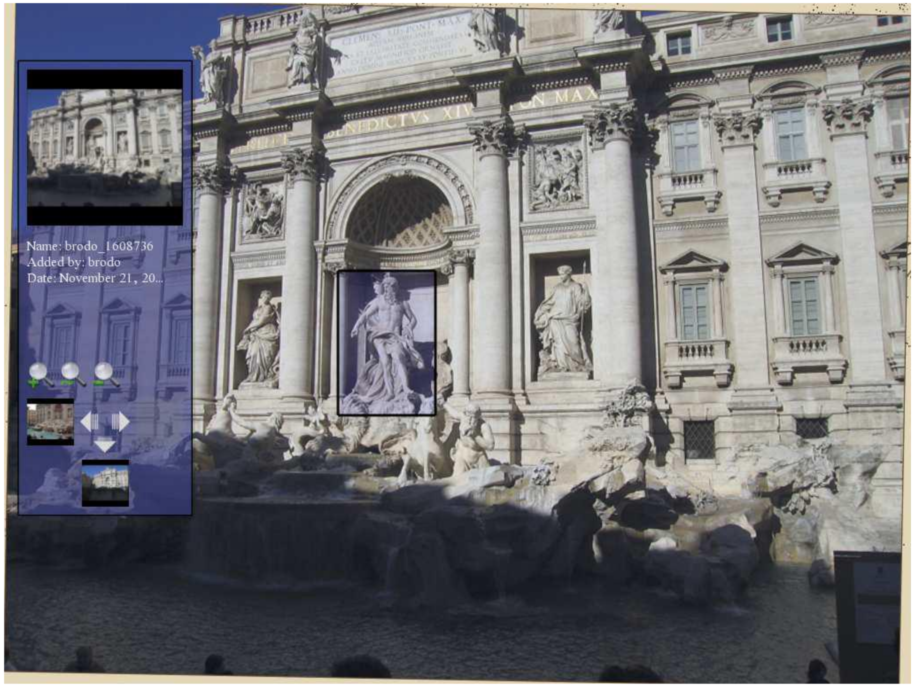
- Solutions to navigate between digital photographs have been developed.
- Snavely et al.[1] proposed one of the first system of this type called *PhotoTourism* (now *PhotoSynth*).

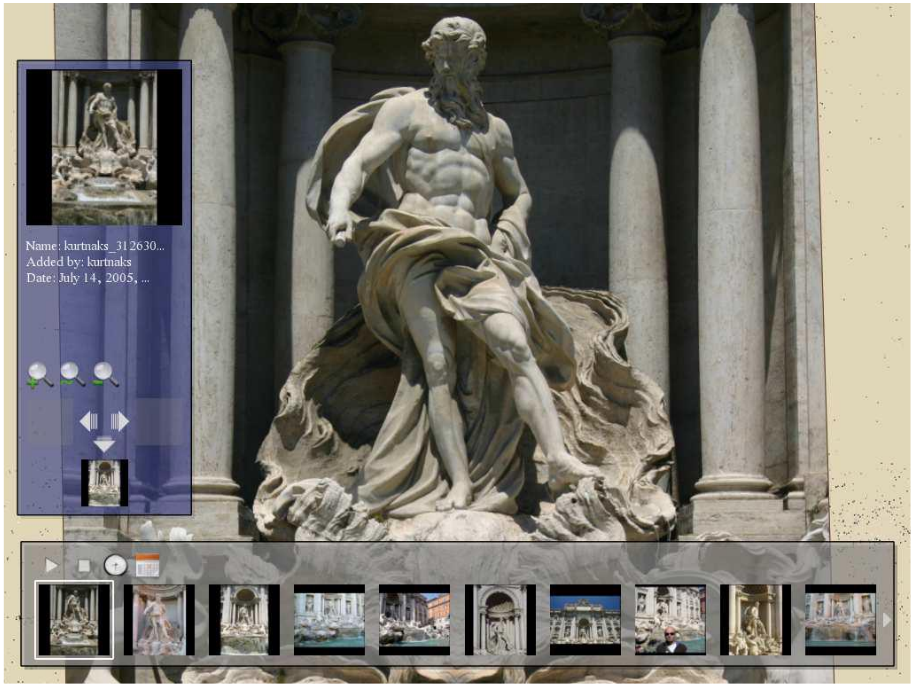
[1] Noah Snavely, Steven M. Seitz, and Richard Szeliski, "Photo Tourism: Exploring photo collections in 3D", ***ACM Transactions on Graphics*****, Vol. 25(3), August 2006.**

PhotoTourism



PhotoTourism





PhotoCloud[2],[3]

- Joint navigation of 3D model/point cloud.
- Navigation bar:
  - suitable for large image set
  - Permit joint 3D navigation/2D browsing

[2] P. Brivio, M. Tarini, F. Ponchio, P. Cignoni, R. Scopigno, "PileBars: Scalable Dynamic Thumbnail Bars", ***VAST 2012 Symp. Proc.*****, pp. 49-56, 2012.**

[3] P. Brivio, L. Benedetti, M. Tarini, F. Ponchio, P. Cignoni, R. Scopigno, "PhotoCloud: Interactive Remote Exploration of Joint 2D and 3D Datasets", ***IEEE Computer Graphics and Applications*****, Vol. 33(2), pp. 86-96, 2013.**

PhotoCloud[2],[3]

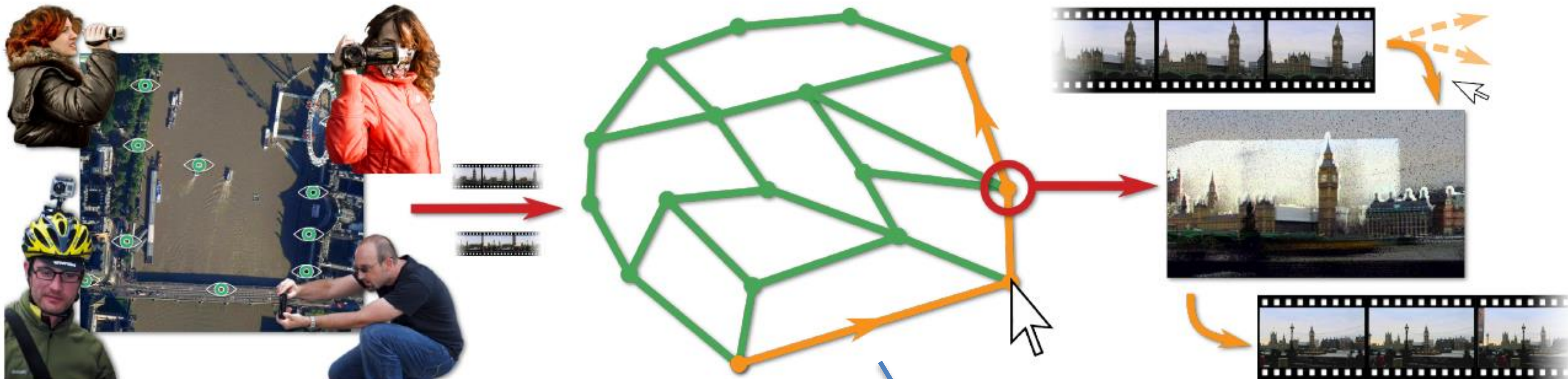
Video Navigation

- Performance capture by an audience[4]
  - Separate the background from the foreground
  - Foreground subjects are modeled with billboards
  - View interpolation
- Casually captured videos in a large area (e.g. London center)[5]

[4] Luca Ballan, Gabriel J. Brostow, Jens Puwein, Marc Pollefeyes, "Unstructured Video-Based Rendering: Interactive Exploration of Casually Captured Videos", *Siggraph 2010*.

[5] James Tompkin, Kwang In Kim, Jan Kautz, and Christian Theobalt "Videoscapes: exploring sparse, unstructured video collections", *Siggraph 2012*.

Video Navigation – Videoscapes

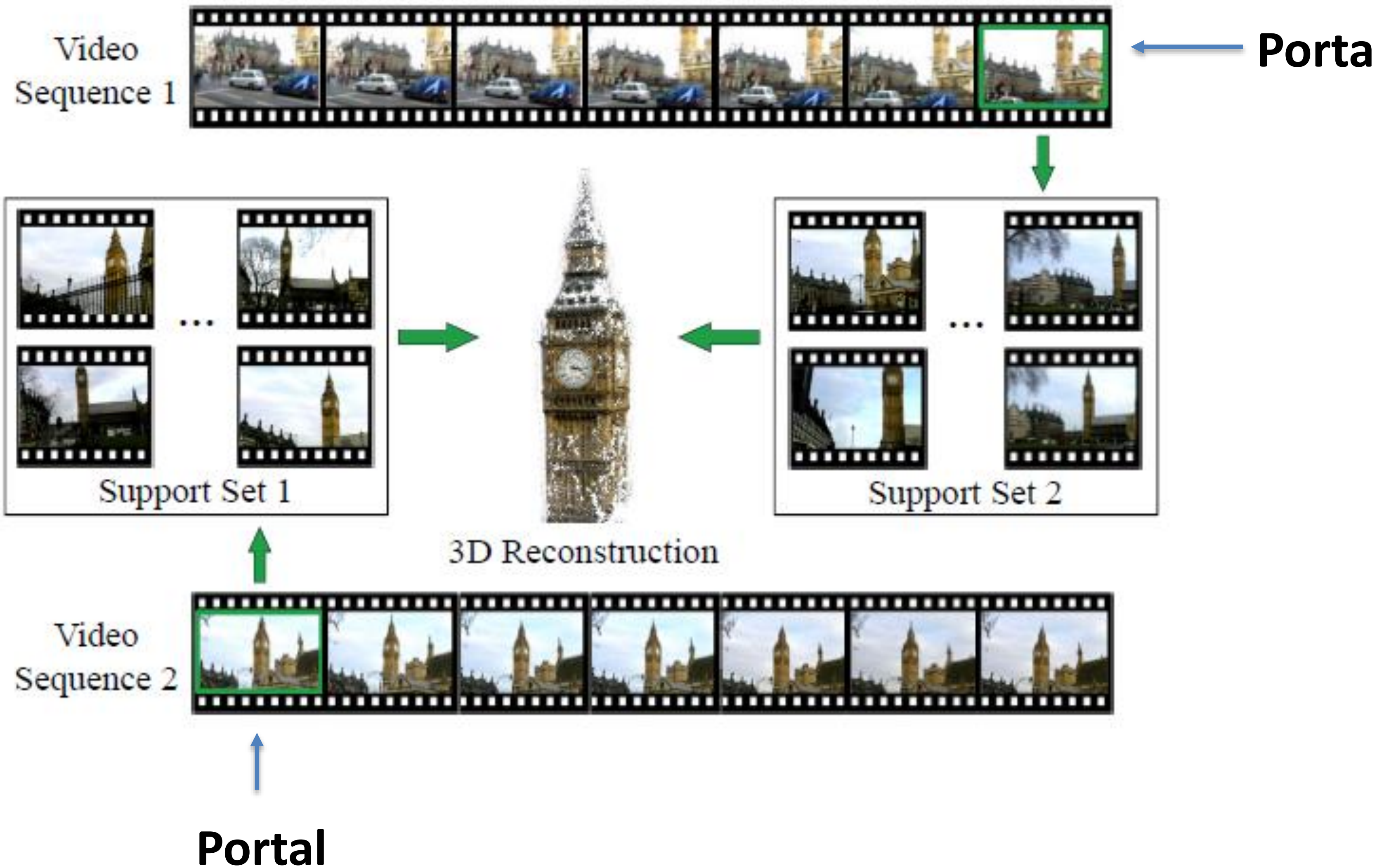


***Edge*: a part of a video sequence**

***Node*: a possible video transition point (a *portal*)**

***Path*: is a newly generated video coming from different captured video sequences**

Videoscapes



Identification of Portals

- GPS information + frames without significative movement are discarded (25% of accumulated *optical flow* are get).
- Holistic Matching (global similarity) + Feature Matching (SIFT + RANSAC).
- Context refinement → a graph representing pairwise matches is build and analyzed to evaluate match's quality.

Paintings-to-3D Models

- A very challenging problem → significative geometric (drawing errors, missing elements) and appearance differences (different textures, no physical lighting, different seasons).
- Russell et al.[6]: automatic alignment of non-photographic depictions of a scene.
- Aubry et al.[7]: paintings, drawings and architectural photographs registered on a 3D model.

[6] Bryan C. Russell, Josef Sivic, Jean Ponce, Helene Dessales, "Automatic alignment of paintings and photographs depicting a 3D scene" *ICCV 2011*.

[7] Mathieu Aubry, Bryan Russell Josef Sivic, "Painting-to-3D Model Alignment Via Discriminative Visual Elements", *Siggraph 2014* .

Russell et al.[6] – Goal

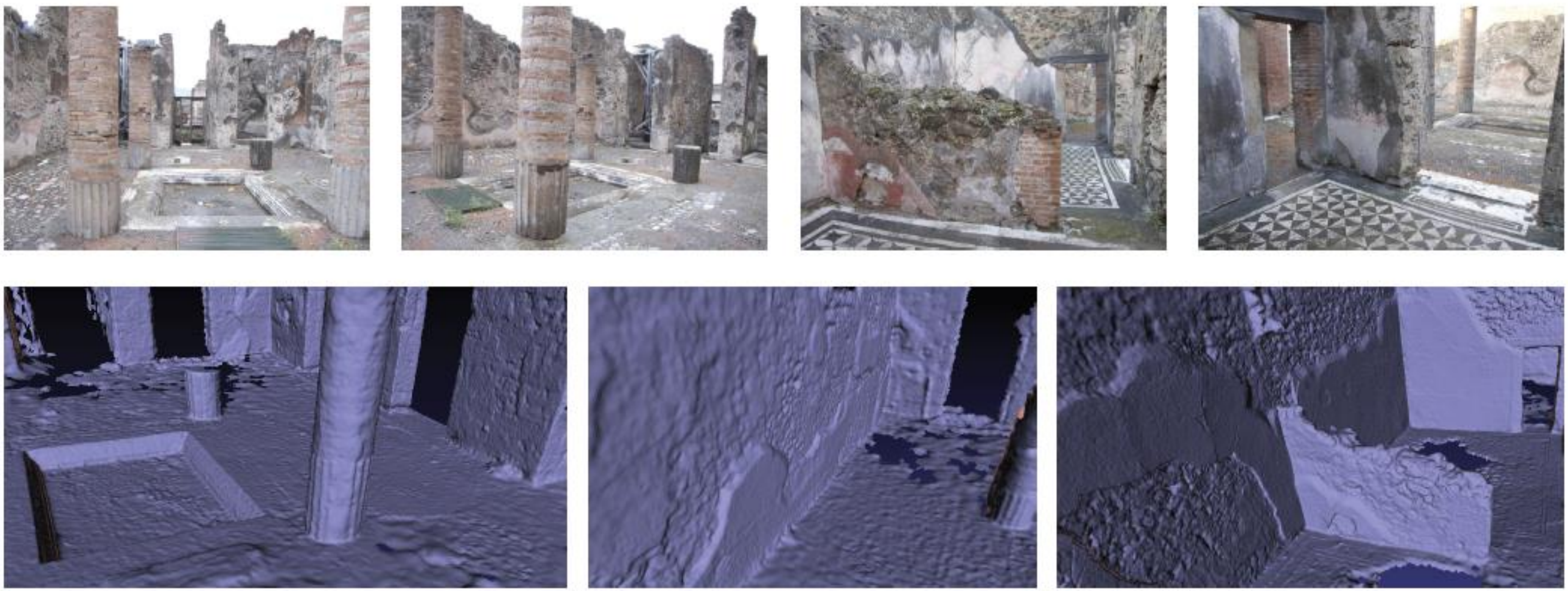
- ***Goal***: automatic alignment of non-photographic depictions of a scene
- *Case study*: alignment of the XIXth Century architectural watercolors of the Casa di Championnet in Pompei with modern photographs.

Russell et al.[6] – Algorithm

- Stage 1: Recovering a 3D model of the scene
  - Bundler+PMVS+Poisson surface reconstruction
- Stage 2: Coarse alignment by view-sensitive retrieval
  - Viewpoint generation
  - Matches using GIST (minimum L2 distance)
- Stage 3: Fine alignment by matching view-dependent contours
  - Ridges, valleys and occlusion contours are extracted from the matching viewpoint ; edges from the painting (gPB detector)
  - ICP-like refinement

Russell et al.[6] – First Stage

3D Model reconstruction (using 563 photographs)

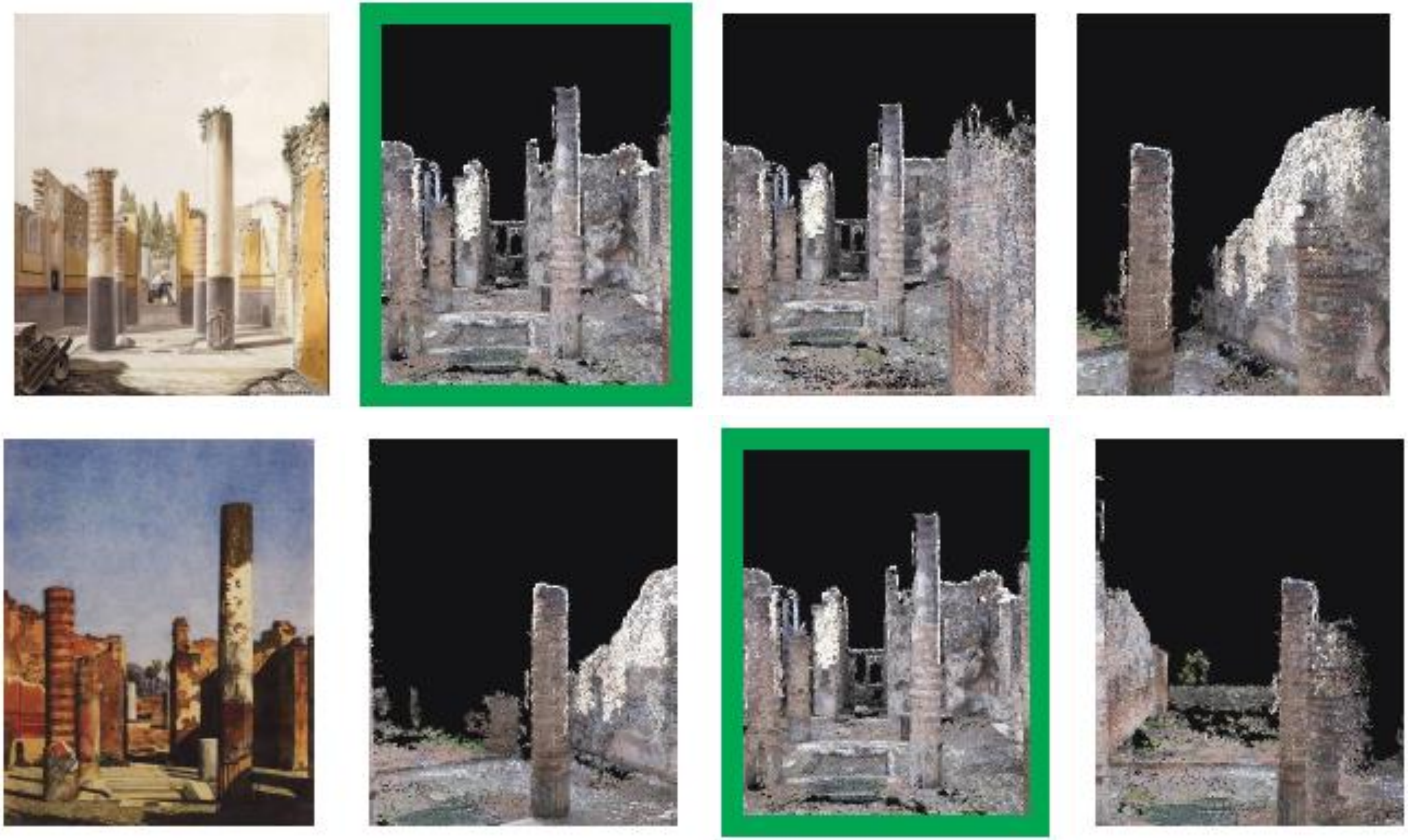


Russell et al.[6] – 2nd Stage

- Viewpoint generation:
  - Height is set at eye-level
  - Upright
  - 12 orientations
  - Rendering: PMVS points on an uniform background
- *GIST*[8] is a global descriptor.

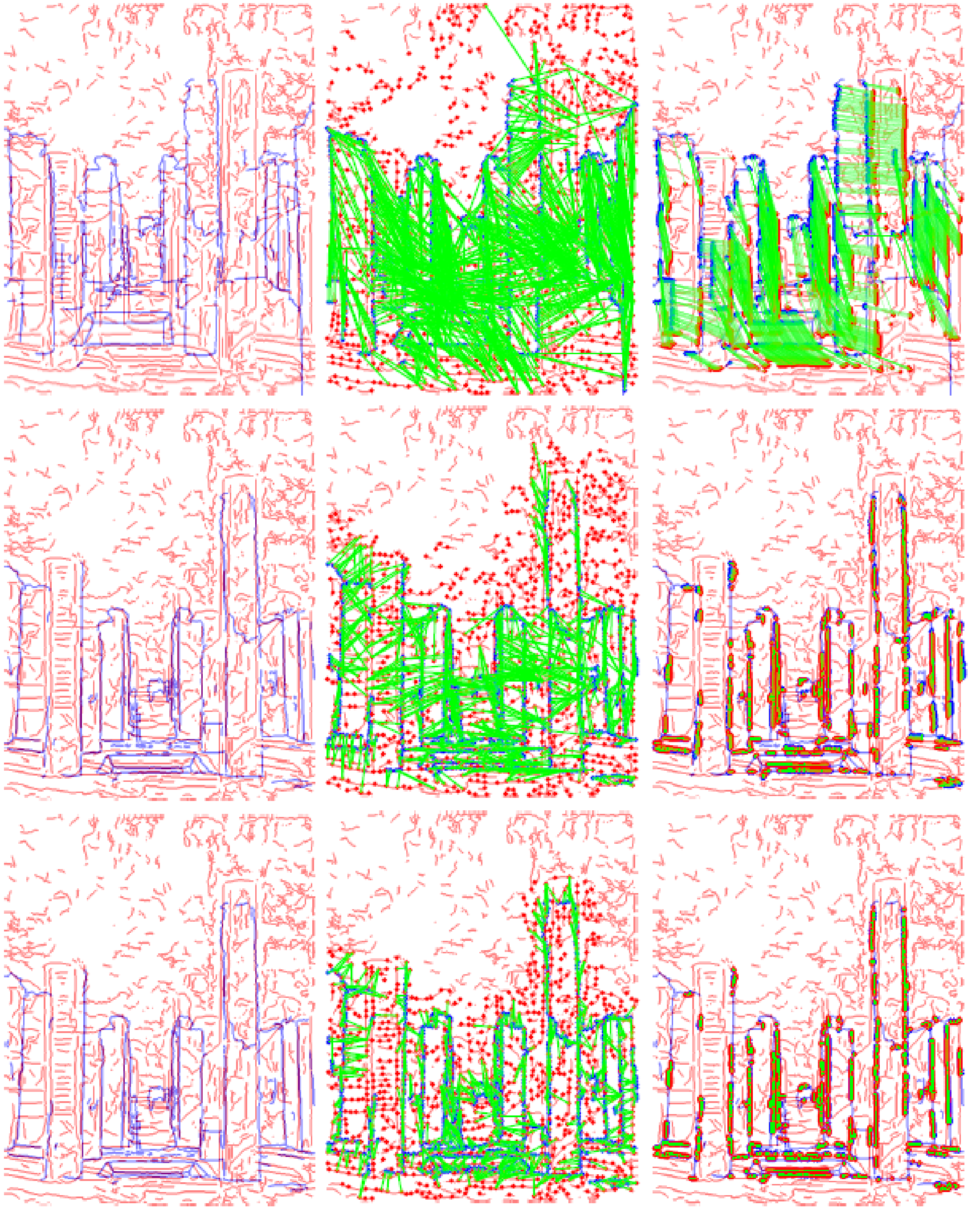
[8] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope", IJCV 42(3), pp. 145-175, 2001.

Russell et al.[6] – 2nd Stage



Russell et al.[6] 3rd Stage

- Ridges/valleys/occlusion contours are extracted using the algorithm by Ohtake et al.
- Image edges are extracted using the global probability of boundary (gPB) detector.
- ICP-like refinement



Russell et al.[6] – Results

Aubry et al.[7]

- The aim is to align paintings, historical drawings, and architectural photographs on the input 3D model.
- *Assumption:* the painting is at least an approximation of a perspective rendering.
- *Main idea*: automatically discovering discriminative visual elements of the 3D scene and use them for the registration.

Aubry et al.[7]

- Discriminative visual elements is a mid-level patch such that:
  - Visually discriminant w.r.t the “visual world”
  - Distinctive
  - It can be reliably matched

Aubry et al.[7] – Overview

- Rendering representative views
- Finding discriminative visual elements
- Filtering unstable visual elements
- Recovering viewpoint

Aubry et al.[7] – Rendering Viewpoints

- Identify ground plane, then sample on a regular grid 24 orientations (12 horizontal and 2 elevation)

Aubry et al.[7] – Finding Discriminative Visual Elements

- Matching as classification: given a patch *q* (of the rendered view), we learn to discriminate it between negative examples with an SVM classifier (we learn weights *w*).
- The patch *x* in the input image with the highest score $s(x) = w^T x + b$ is the matching one.
- Similar to *per-exemplar SVM classification*..

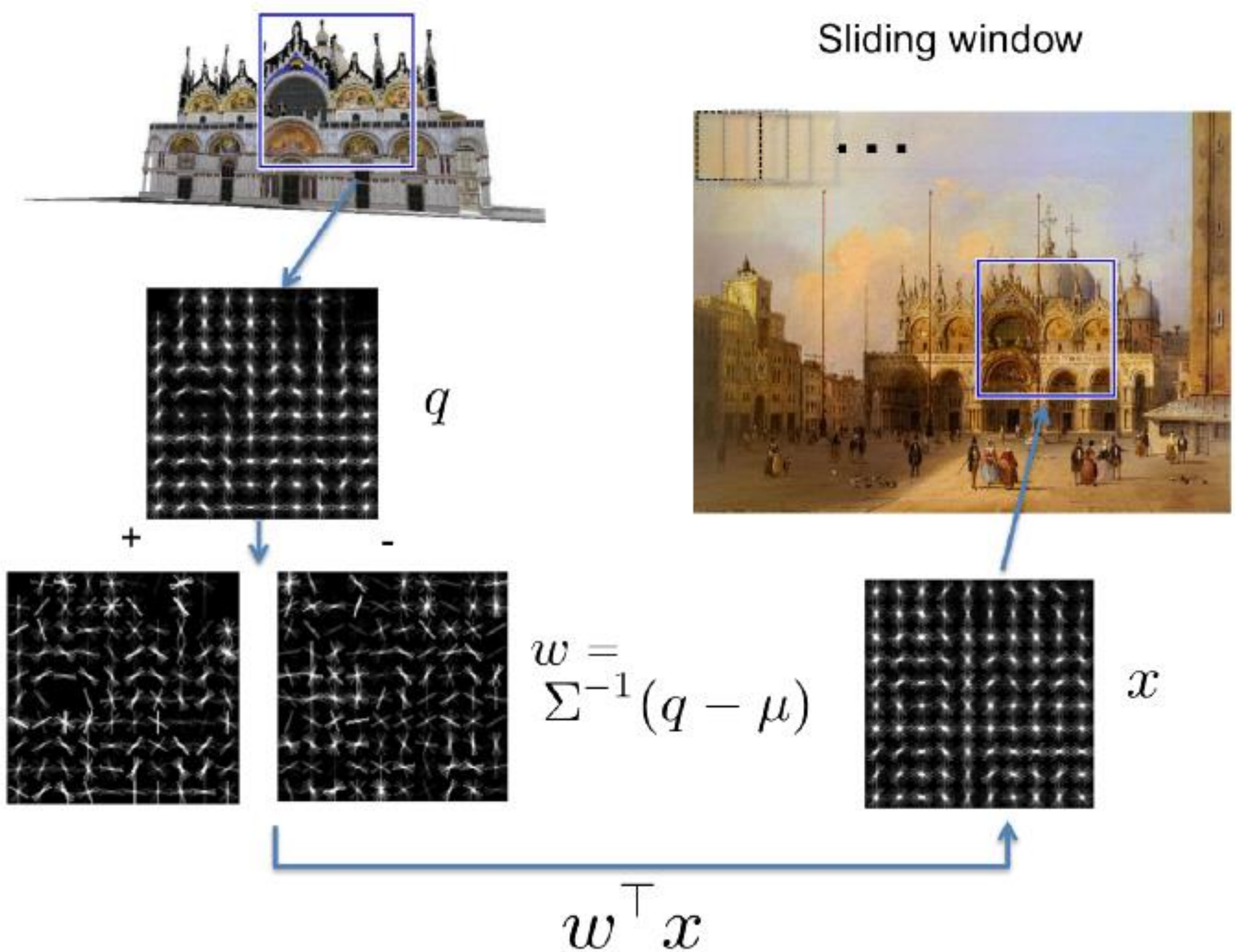
Aubry et al.[7] – Finding Discriminative Visual Elements

- *Matching as classification*: given a patch *q* (of the rendered view) with associated HOG descriptor, we learn to discriminate it between negative examples with an SVM classifier (we learn weights *w*).

$$E\left(w,b\right)=L\left(1,w^{T}q+b\right)+\frac{1}{N}\sum_{i=1}^{N}L\left(-1,w^{T}x_{i}+b\right)$$

- The patch *x* in the input image with the highest score $s(x) = w^T x + b$ is the matching one.
- Similar to *per-exemplar SVM classification*..

Aubry et al.[7] – Finding Discriminative Visual Elements



Aubry et al.[7] – Finding Discriminative Visual Elements

- Computationally very expensive..
- A closed-form solution to estimate the weights exists (change hinge loss with a square loss):

$$w_{LS} = \frac{2}{2 + \|\Phi(q)\|^2} \Sigma^{-1}(q - \mu) \quad E^*_{LS} = \frac{4}{2 + \|\Phi(q)\|^2}$$

- This solution depends on the “whitening” transformation:

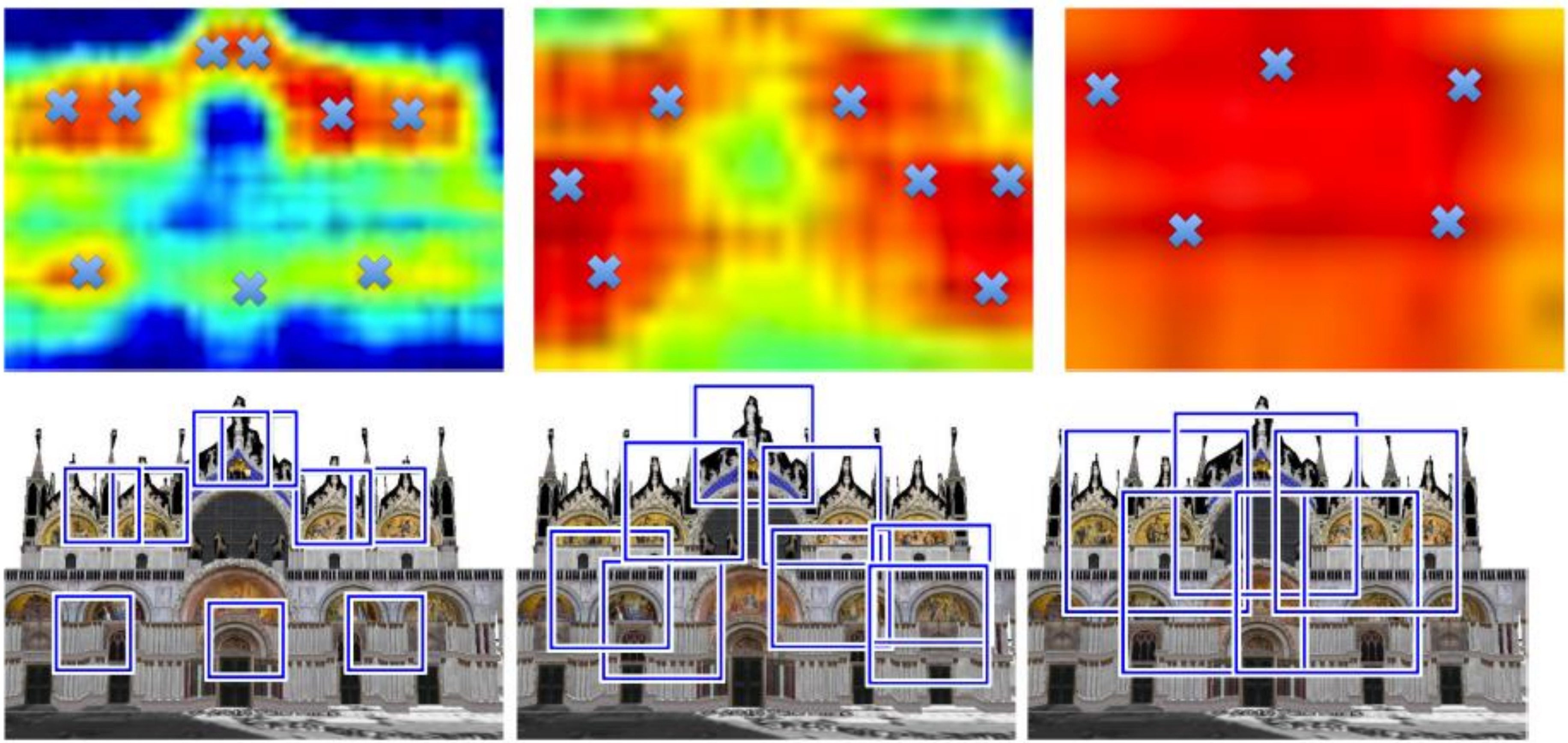
$$\Phi(q) = \Sigma^{-\frac{1}{2}}(q - \mu)$$

Aubry et al.[7] – Finding Discriminative Visual Elements

- So, at the end, the whitened norm $\|\Phi(q)\|^2$ for each patch is evaluated and used to selected the discriminative visual element.
- *Whitened norm high → training cost low → high discriminability.*

Aubry et al.[7] – Finding Discriminative Visual Elements

Selection at different scales



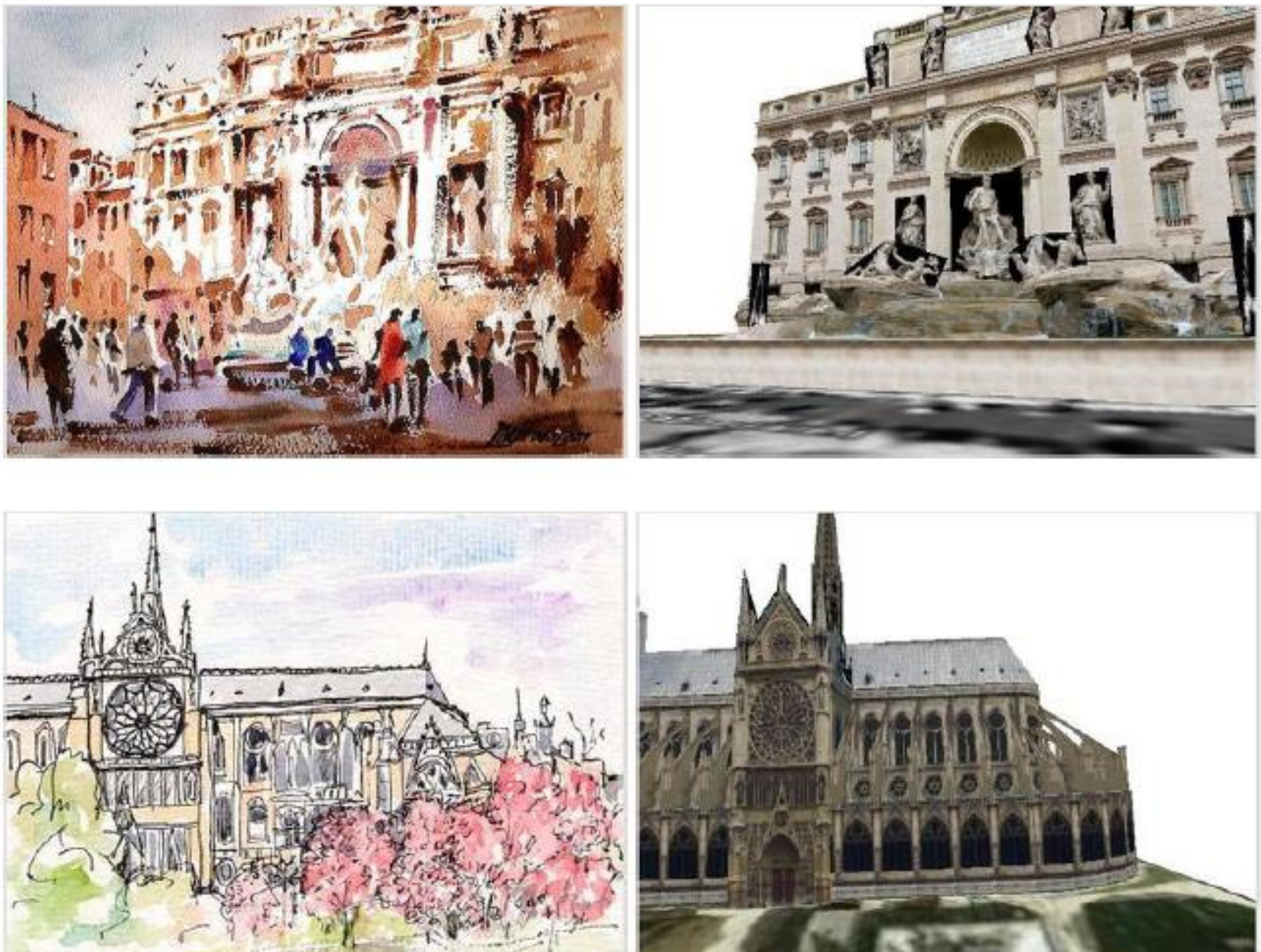
Aubry et al.[7] – Filtering Unstable Visual Elements

- Nearby viewpoints are identified (using a measure of *visual overlap*).
- Candidate visual elements are searched inside the nearby viewpoints → elements that cannot be matched reliably are discarded.

Aubry et al.[7] – Recovering Viewpoint

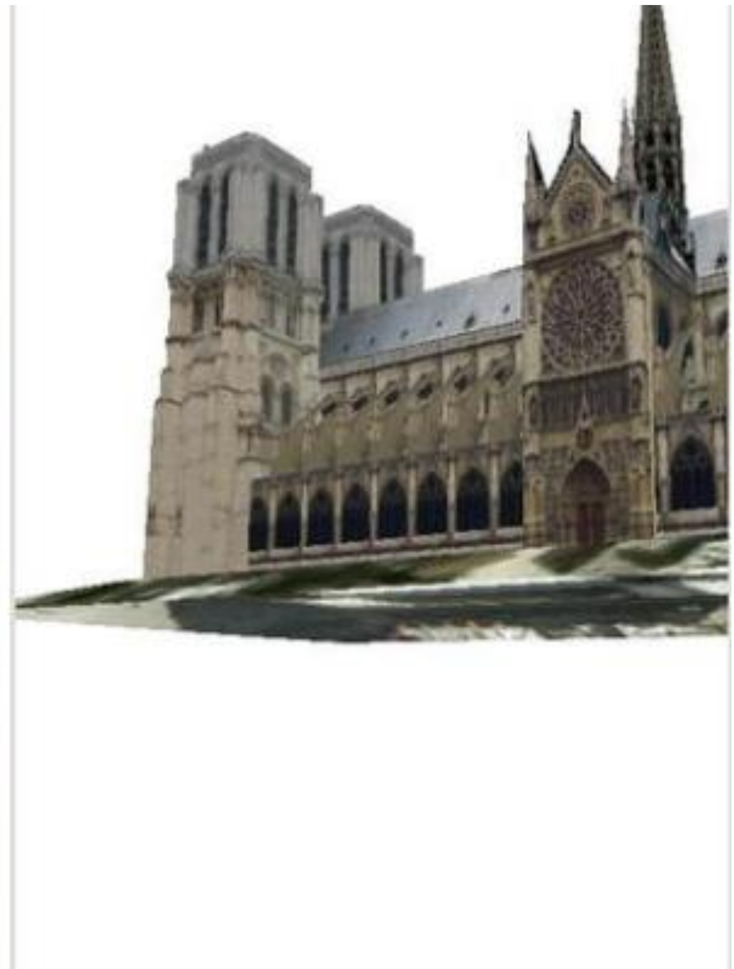
- “Rough + fine” approach.
- *Coarse registration*: use direct matching between the discriminative visual elements (5 putative correspondences for each element).
- *Fine registration*: HOG-based ICP-like refinement.

Aubry et al.[7] – Result



Aubry et al.[7] – Result

Aubry et al.[7] – Result



Advancements in object detection/recognition and understanding of joint properties of 3d models and images

- “Seeing 3D chairs”[9] → Object category detection as a part-based 2D/3D alignment problem.
- CROSSLINK[10] → Joint understanding/processing of image collections and 3D models collections
- RenderCNN[11] → viewpoint estimation in images.

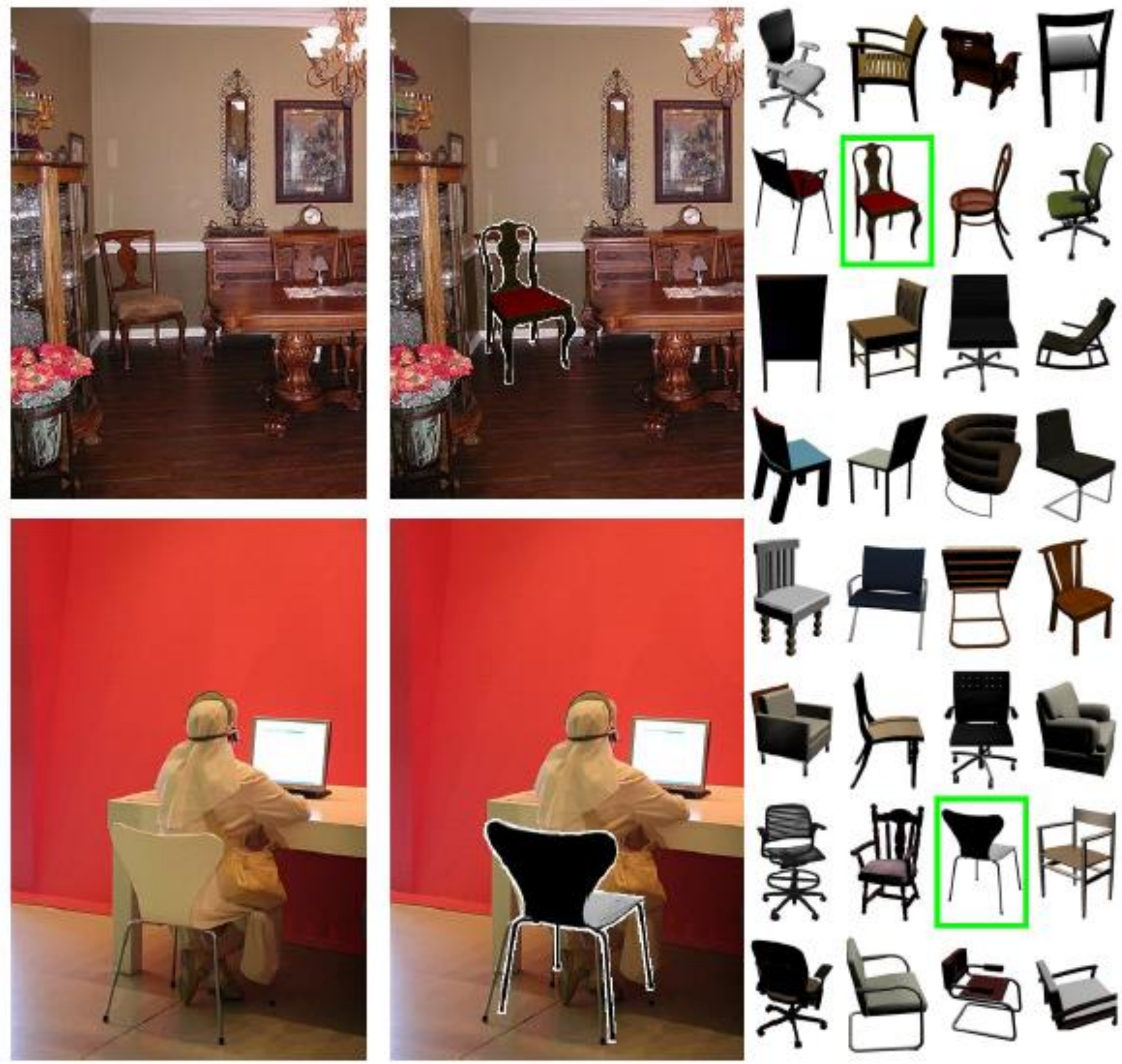
[9] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, J. Sivic, “Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models”, *Proc. Of CVPR2014, 2014*.

[10] M. Hueting, M. Ovsjanikov, N. J. Mitra, “CROSSLINK: Joint Understanding of Image and 3D Model Collections through Shape and Camera Pose Variations”, *Siggraph Asia 2015*.

[11] H. Su, C. R. Qi, Y. Li, L. J. Guibas, “Render for CNN: Viewpoint Estimaton in Images Using CNNs Trained with Rendered 3D Model Views”, *Proc. of ICCV15,* 2015.

"Seeing 3D chairs"[9]

- Object category detection as a type of 2D-to-3D alignment.



(a) Input images (b) Aligned output (c) 3D chair models

“Seeing 3D chairs”[9]

- Why chairs ?
  - Hard
  - Huge intra-class variations
- 1300 chairs collected from Internet (Google/Trimble 3D Warehouse).
- 800,000 view dependent distinctive visual elements are computed from renderings (as in [7]).
- Visual elements detector must be calibrated (!)
- Part-based matching (spatial configuration is taken into account).

CROSSLINK[10]

- 3D model collections and image collections provide *complementary* information.
- NO manual intervention ; NO assumption about the dataset (e.g. clean dataset).
- The idea is to retrieve, using Bing Image Search and Trimble 3D Warehouse two collection with the same keyword and perform a *joint analysis*.

CROSSLINK[10]

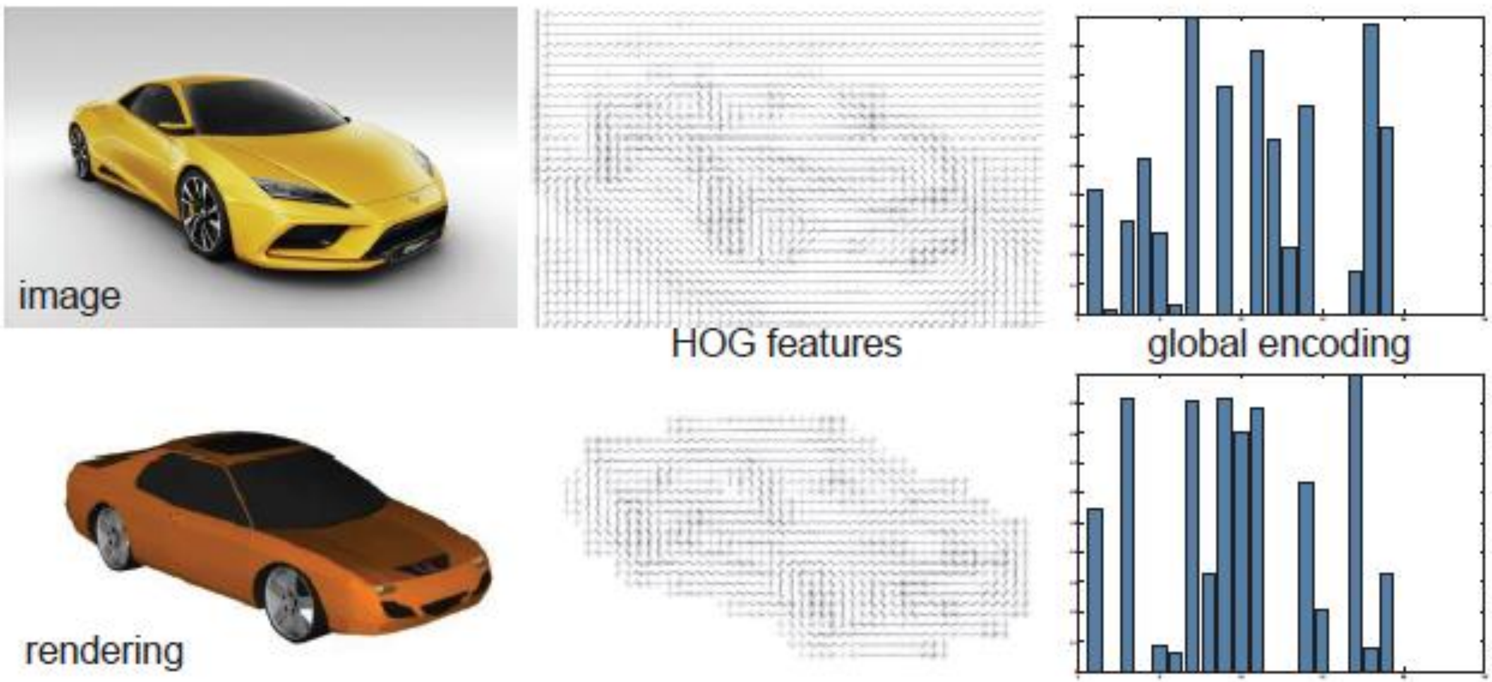
- This *joint analysis* permits to:
  - Improve 3D search (through re-ordering)
  - Improve the organization of the images according to shape attributes (in particular, *viewpoint* and *width/height ratio*).
- A method to *co-align* the re-ordered 3D collection is also provided.
- A tool for the joint exploration of a collection of 3D models and a collection of images.

CROSSLINK[10]

- Views generation through rendering
  - 3D models are rendered with step of 10 degrees (36 orientations at a fixed elevation)
- Features are extracted from the images and from these views:
  - KC-encoded HOG
  - CNN features (last layer of a STAR CNN)

CROSSLINK[10]

KC-encoding



CROSSLINK[10] – Improve 3D Search

Original search

Improved search exploiting the image-views matching

CROSSLINK[10] – Co-Alignment

3D Models Filtered

3D Models Filtered and Co-Aligned

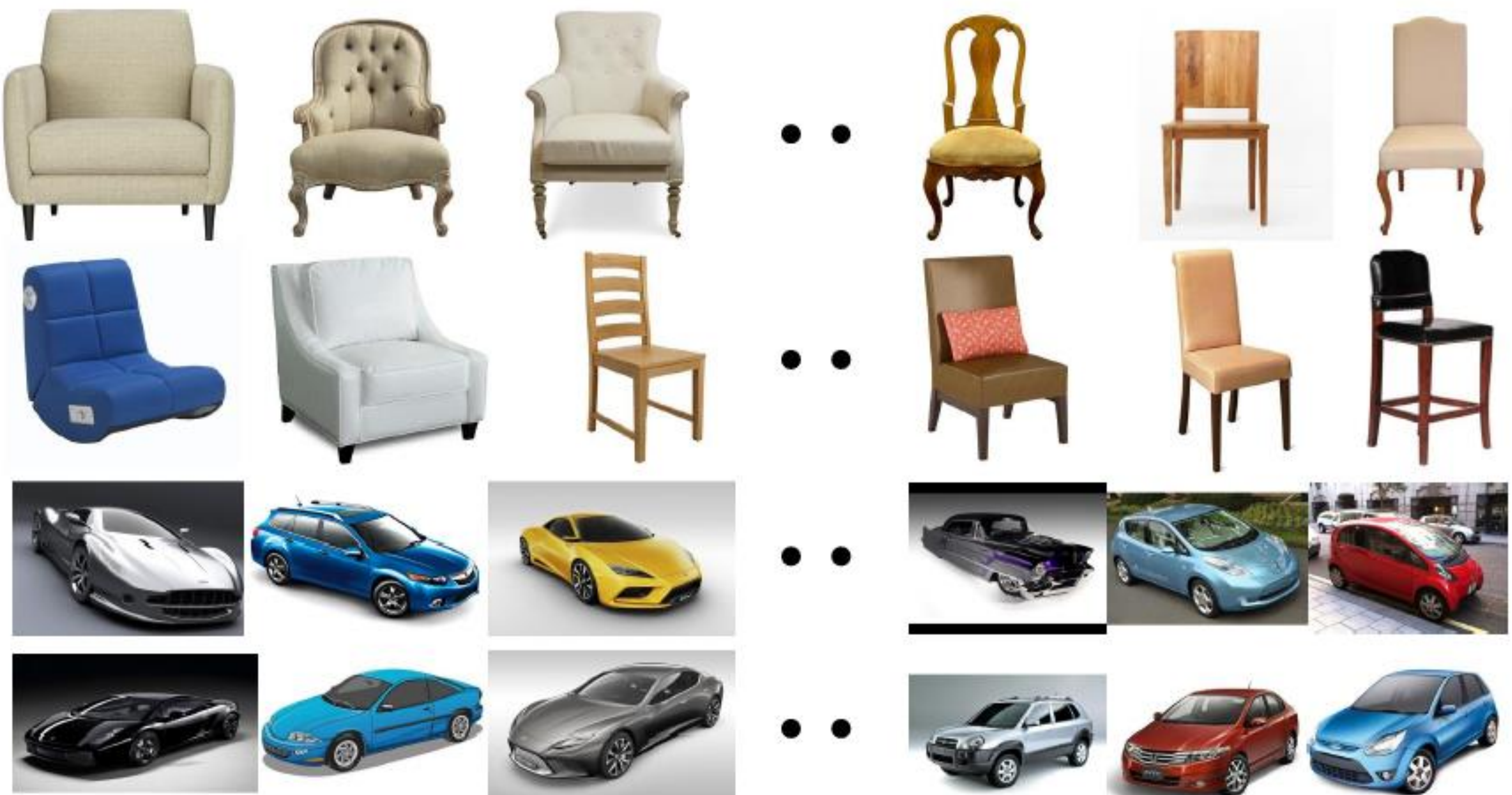
CROSSLINK[10] – Camera Pose Estimation

$$P_{\text{view}} := \mathrm{F}(\boldsymbol{V}_c^{\theta}) \qquad N_{\text{view}} := \mathrm{F}(\boldsymbol{V}_c \setminus \boldsymbol{V}_c^{\theta})$$

Positive examples **Negative examples**

A weighted sum of probabilities is evaluated
For each image (output SVM → probability)

CROSSLINK[10] – Image sorting according to shape attributes



Aspect (h/w ratio)

RenderCNN[11]

- Many annotated image dataset for image detection/recognition task exists.
- Viewpoint annotation is poor in large image dataset (largest is *PASCAL3D* – 22K images).
- **Main idea:** rendering 3D models, train a CNN, and learn to estimate the viewpoint of a real image.
- To exploit the information provides by the 3D model (through rendering) to annotate the real images automatically.
- *Rendering.. in which way ?*

Conclusions

- *Registration* is a fundamental task in Computer Vision and Computer Graphics.
- Image-image registration (even image with very different appearance) is a mature field.
- Geometry registration depends heavily on the specific task and on the type of data.
- Image-geometry registration → many solutions but only few *general* and *robust*.
- Results in object recognition/detection are very interesting also in the field of 2D/3D registration.

Questions ?