HAL
open science

Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models

Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan C. Russell, Josef Sivic

To cite this version:

Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan C. Russell, Josef Sivic. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models. CVPR 2014 - Compter Vision and Pattern Recognition, Jun 2014, Colombus OH, United States. hal-01057240

HAL Id: hal-01057240

https://inria.hal.science/hal-01057240v1

Submitted on 21 Aug 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models

Mathieu Aubry[1,*] Daniel Maturana[2] Alexei A. Efros[3,*] Bryan C. Russell[4] Josef Sivic[1,*]

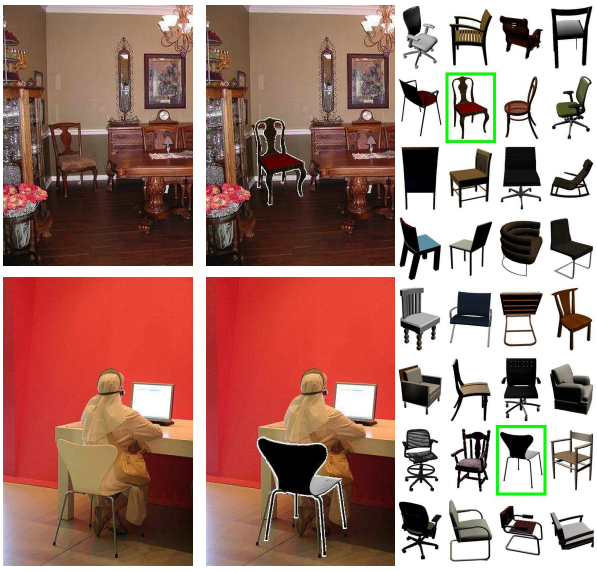
[1]INRIA [2]Carnegie Mellon University [3]UC Berkeley [4]Intel Labs

Abstract

This paper poses object category detection in images as a type of 2D-to-3D alignment problem, utilizing the large quantities of 3D CAD models that have been made publicly available online. Using the "chair" class as a running example, we propose an exemplar-based 3D category representation, which can explicitly model chairs of different styles as well as the large variation in viewpoint. We develop an approach to establish part-based correspondences between 3D CAD models and real photographs. This is achieved by (i) representing each 3D model using a set of view-dependent mid-level visual elements learned from synthesized views in a discriminative fashion, (ii) carefully calibrating the individual element detectors on a common dataset of negative images, and (iii) matching visual elements to the test image allowing for small mutual deformations but preserving the viewpoint and style constraints. We demonstrate the ability of our system to align 3D models with 2D objects in the challenging PASCAL VOC images, which depict a wide variety of chairs in complex scenes.

1. Introduction

From its very beginnings [32] and up until the early nineties [29], object recognition research has been heavily geometry-centric. The central tenet of the time was *alignment*[1], and the act of recognition was posed as correctly aligning a 3D model of an object with its 2D depiction in the test image [21, 26]. The parameters recovered during alignment (object pose, object scale, etc.) served as the output of the recognition process, to be used, for instance, in the perception-manipulation loop in robotics applications. Unfortunately, the success of these 3D model-based methods was largely limited to instance recognition tasks for objects with well-pronounced rectilinear structures (e.g. staplers were a favorite example). As the field moved toward category recognition and objects with more complex appearance, 3D model-based object recognition has been replaced by the new 2D appearance-based methods (e.g. [9, 14, 37]). These methods forgo 3D and operate directly on the 2D image plane. Thus, instead of a 3D model of an object, they use a large dataset of 2D views of the object class from different viewpoints, as the model. These methods have shown steadily improving performance on a number of challenging tasks, such as the PASCAL VOC dataset [13]. However, their main drawback is that the result of a successful recognition is typically just the name of the object that was found (e.g. "chair") and a bounding box to indicate its rough location



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

Figure 1: Given an input image (left), our algorithm searches a database of 1,393 3D chair models to detect any depicted chairs in the image. The algorithm returns a 3D model matching the style of the chair and recovers its viewpoint relative to the camera (outlined in green, right). We overlay a projection of the returned 3D model onto the input image (middle). Notice the agreement of the returned model with the depicted chair style and pose.

*WILLOW project-team, Département d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548, Paris, France.

[1]Indeed, one of the oft-told stories is that when a student asked Takeo Kanade what are the three most important problems in computer vision, his reply was: "Alignment, alignment, alignment!".

within the image. While this type of result is reasonable for tasks such as retrieval (e.g. "find all chairs in this dataset"), it is rather unsatisfying for doing any deeper reasoning about the scene (e.g. "what's the pose of the chair?", "can I sit on it?", "what is this chair occluding?", "how can I fold this chair?", etc). All these questions could be trivially answered, if only we had a 3D model of the chair aligned with the image!

The work presented in this paper aims to combine some of the benefits of the 3D model-based instance alignment methods with the modern, appearance-based object category tools towards getting a best-of-both-worlds object recognition engine. The idea is to use a large library of textured 3D object models that have become publicly available on the Internet to implicitly represent both the 3D shape of the object class, as well as its view-dependent 2D appearance. Our approach can be considered as a marriage between part-based discriminative models [14] and exemplar-based matching [28]. Like part-based models, we represent objects using a set of connected appearance parts. But, like exemplar-based methods, we avoid explicitly training an object model by relying on a large dataset of object instances that serve as their own model, both in 2D as well as 3D.

We picked the "chair" category as the running example in this paper because: 1) it is very hard even for the state-of-the-art methods [14], achieving only 0.13–0.20 average precision (AP) on PASCAL VOC [1]; 2) it is a category well-represented in publicly-available 3D model collections (e.g. Google/Trimble 3D Warehouse), 3) chairs have huge intra-class variation – whereas there are perhaps only hundreds of types of cars ever made, there are thousands of different types of chairs!

1.1. Related Work

Our paper is part of an emerging trend towards reclaiming some of the early successes in 3D recognition, and combining them with modern visual recognition tools.

3D geometry with multi-view constraints has served as a strong alignment oracle for *instance-level* recognition [24, 33] and retrieval [8]. Images are typically represented using local invariant features such as SIFT [27], which work best for textured objects such as building facades [24, 31, 35]. More recent work has seen the re-emergence of contour-based representations for matching skylines [5] and smooth objects such as furniture pieces [25] or statues [3], however, these efforts have also largely focused on instance recognition.

In *category-level* recognition, recent work has explored recognition, alignment and fine pose estimation for cars and bicycles in outdoor scenes using low-dimensional parametric deformable 3D models combined with a relatively small number of learnt part detectors [20, 30, 41]. Others have explored simplified 3D cuboid models for reasoning about outdoor scenes [18] or detecting and estimating pose for box-like objects such as printers, beds or sofas [15, 39], often using simplifying 3D box-layout constraints for indoor scenes [7, 11]. Indoor objects have been also represented by simplified models consisting of a small set of planar parts [38]. Finally, a collection of registered images of 22 cars has been used in a Hough-like voting procedure to estimate the poses of cars [17]. In a separate line of research, a discriminative, exemplar-based object detection framework demonstrated (anecdotal) results for *transferring* 3D object geometry onto a test image once 2D-to-2D alignment with an exemplar was established [28]. However, [28] models an object with a single global template, thus requiring a large number of exemplars to represent categories with high intra-class variation. Non-parametric representations have also shown promising results for 2D-to-3D matching of indoor scenes [34], but again, due to the use of a global scene descriptor, this method works only for highly structured scenes, such as (tidy) bedrooms.

Addressing the need for descriptors that capture less than a full object/scene, but are more descriptive than low-level features such as SIFT [27], are a new breed of *mid-level* visual elements. Recently, discriminatively trained mid-level representations have shown great promise for various visual recognition tasks, both in a supervised regime [6], as well as unsupervised [22, 23, 36], including 2D-to-3D alignment of object instances across large changes in appearance [4]. While here we build on the latter work, which aligned a single image to a *single* 3D model, we address the much more difficult problem of category-level recognition, which must align a single image to over *1300* 3D chair models.

1.2. Approach Overview

Our representation consists of a large set of 3D CAD models, which captures both the large variety of chair styles and their different possible viewpoints. Chair detection in new images is accomplished by finding an alignment between the 2D chair and the most similar 3D chair model rendered at the most appropriate camera viewpoint, as shown in Figure 1. Aligning photographed 2D objects with the most similar (but not identical) computer-generated 3D model is a very hard problem. Here we address it by representing the collection of all 3D models by more than 800,000 calibrated view-dependent mid-level visual elements learned in a discriminative fashion from rendered views. At test time, all the learned visual elements are applied to the test image in parallel. The most spatially and appearance-wise consistent alignment is found, while preserving the style and viewpoint-consistency constraints. The details of the algorithm are described in Section 3.

Contributions. Posing object *category* detection in images as a 3D *instance* alignment problem, we: 1) develop an exemplar-based 3D category representation capturing vari-

ations across object style and viewpoint; 2) establish the correspondence between computer-generated 3D models and 2D photographs using a large collection of mid-level visual elements; 3) collect a dataset of more than 1,300 3D chairs from the Internet, publicly available at [2].

2. Large dataset of 3D chairs

We explicitly represent the shape variation of an object category by a large collection of 3D CAD models. The 3D models used in this work have been downloaded from the Google/Trimble 3D Warehouse, an on-line repository of publicly available, user-contributed 3D graphics content created using Google SketchUp. We initially searched the repository for "chair" and downloaded over two thousand 3D models. However, many of them were not of good visual quality; some weren't even chairs. After manually culling the data, we were left with $1,393$ high-quality 3D chair models, representing a variety of chair styles.

Each 3D chair model was rendered on a white background from 62 different viewpoints sampled over the upper half of the viewing sphere centered on the chair. The resulting chair dataset contains a total of $1,393 * 62 = 86,366$ synthesized images, each annotated with the chair ID indicating the different style, as well as the viewing orientation (θ, ϕ). This dataset is our non-parametric representation of the object category explicitly representing the large variety of chair styles and the set of possible viewing positions.

3. Building Discriminative Visual Element Detectors

Explicitly representing and synthesizing the fine-grained style and viewpoint of the 3D object category significantly simplifies the difficult task of 2D-to-3D alignment. Nevertheless, reliably matching a synthesized view of an object to a real photograph is still challenging due to differences in, e.g., texture, materials, color, illumination or geometry. Furthermore, it is well known that computer generated images have different statistical properties than real photographs. To address these issues we cast the matching problem as a classification task, and represent the collection of 3D models using a large set of *mid-level visual elements* – linear classifiers over HOG features learnt from the rendered views in a discriminative fashion (section 3.1). As each of the hundreds of thousands of visual elements is learnt individually, calibrating their matching scores becomes a critical issue. We address this by learning a linear calibrating function for each element on a common dataset of negative images (section 3.2). Finally, we wish to be tolerant to small geometric deformations (such as a chair with longer legs or shorter armrests). We develop a matching procedure (section 3.3) that allows for small deformations in the spatial configurations of the matched visual elements while preserving consistent viewpoint and style.

3.1. Large-scale learning of visual element detectors

The goal is to match a square image patch q, represented as a HOG descriptor [9], in a rendered 3D view to its corresponding patch in the input image. Following [4], we treat matching as a classification task where we wish to find the patch x^* in the input image that maximizes the following linear classifier score that is dependent on q:

$$S_q(x) = w_q^T x. \tag{1}$$

Here w_q is a vector of learned weights for the classifier dependent on q. In contrast to measuring similarity via Euclidean distance between q and x, the weights w_q emphasize feature dimensions that are discriminative for the patch's appearance.

As we are given only a single rendered view for each 3D chair viewpoint, following [28] we learn the weights w_q by training an exemplar classifier using the patch q as a single positive example (with label $y_q = +1$) and a large number of negative patches x_i for $i = 1$ to N (with labels $y_i = -1$). As we have a large number of patches q (in our case close to 1 million patches), we compute w_q analytically via Linear Discriminant Analysis (LDA). Recent works [16, 19] have shown that linear HOG-based object detectors computed with LDA can reach similar object detection accuracy as detectors learnt by expensive iterative SVM training [28]. The resulting w_q has the following form [4, 16, 19]

$$w_q = \Sigma^{-1}(q - \mu_n), \tag{2}$$

where $\mu_n = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_n)(x_i - \mu_n)^T$ are estimated from a large set of HOG descriptors $\{x_i\}$ extracted from patches that are sampled from a set of ("negative") photographs independent of all images considered in this work. As noted in [4, 16] the weights obtained from LDA can also be derived analytically from a least-squares cost function, with the resulting weights differing from the LDA weights by a constant scale factor. Using weights (2), the similarity score (1) can be re-written [4] as

$$S_q(x) = \Phi(q)^T\Phi(x) - \Phi(q)^T\Phi(0), \tag{3}$$

where $\Phi(x)$ is the "whitening" transformation given by

$$\Phi(x) = \Sigma^{-\frac{1}{2}}(x - \mu_n). \tag{4}$$

Note that the first term in (3) is a dot product between whitened q and x and the second term is an additive scalar factor, which is independent of x. As a result, another way to see our matching method is that, for a given q, it ranks patches x using the dot product between whitened q and x as the similarity measure. Other similarity measures such as

normalized cross-correlation between whitened q and x also could be potentially used [12].

To cope with the millions of patches in the set of rendered views of the 3D chairs, we aim to select a few of the most discriminative ones in each view (dubbed *discriminative visual elements*). Similar to [4] we assume that patches, which have the smallest probability under the Gaussian approximation of the negative patch distribution, are the most discriminative. Since the "whitening" transformation given by eq. (4) takes this Gaussian to a unit covariance Gaussian centered at the origin, patches in the whitened space furthest from the origin are considered most discriminative. For a patch q from a rendered view, this can be measured by taking the square of the norm of the whitened patch $\|\Phi(q)\|^2$. This whitened norm can also be viewed as the Bhattacharyya distance between the Gaussian approximations of the positive and negative data. Finally, it can be shown [4] that patches with larger whitened norm are better separated from the negative training data in terms of the least squares cost.

To select the discriminative visual elements, we densely compute the squared whitened norm response at multiple spatial scales. We select 10 visual elements per rendered 3D view having highest response after non-maximum suppression. For non-maximum suppression, we used an intersection area to union ratio of 0.25. We require the visual elements to be square patches and filter out patches less than 100×100 pixels. This filtering removes small spurious noisy patches while ensuring good coverage of the object. We used the publicly available implementation of HOG from [14] and extracted contrast insensitive features from 10×10 HOG cells, resulting in a 900 dimensional feature vector. We also set to zero components of w_q corresponding to spatial bins with sum of the absolute value across the HOG channels less than 0.01. This effectively downweights the background (white) pixels in the rendered views and focuses the detector weights on the foreground chair. We found without this normalization the visual element detector tends to have high score for constant intensity regions in the input image.

Discussion. As demonstrated in section 4, this procedure outputs reliable matches between the rendered views and cluttered photographs. The success can be attributed to three main factors. First, both images and synthesized views are represented by mid-size HOG descriptors. This captures object boundaries in a "soft" manner without attempting to explicitly detect object edges, which is difficult to do reliably, especially in cluttered photographs. Second, the discriminative matching procedure further enhances the salient boundaries in the specific image region by re-weighting the different dimensions of the HOG descriptor. Finally, the visual elements representing each view are chosen to be distinctive and thus more reliably detectable in the photographs.

3.2. Calibrating visual element detectors

As noted in prior work [4, 28], calibration of matching scores across different visual element detectors is important for the quality of the final detection outputs. As we learn more than 800K element detectors independently, we found careful calibration of their scores to be critical. We address this issue by running all detectors on a common large dataset of 200K negative patches that do not contain the object of interest. In detail, we use an affine calibration of the linear classifier of Equation (1):

$$S'_q(x) = a_q \, S_q(x) + b_q, \tag{5}$$

where for each visual element detector we seek to find the scalars a_q and b_q.

We recover the calibration parameters with respect to two operating points. To select the first operating point, we run the visual element detector on 200K patches that are randomly sampled from the negative image set, which is known to not contain any chairs. We select the negative patch x_n that yields a false positive rate of 0.01%, i.e., x_n is the patch having the 99.99 percentile detection score. We choose as the second operating point μ_n the mean HOG feature vector. Given these two points, we set $S'_q(x_n) = 0$ and $S'_q(\mu_n) = -1$. This calibration leads to the expected false positive rate of 0.01% when $S'_q(x) = 0$. We found this to be a good compromise between representing the tail of the score distribution and the amount of time to scan the negative data.

3.3. Matching visual element spatial configurations

For the final object detection, we wish to enforce a consistent spatial layout of the visual element detections corresponding to a rendered 3D view. We assume a star model for the spatial layout, similar in spirit to [14]. The star model provides spatial constraints, while allowing for small spatial deformations. Moreover, the star model can be run in a sliding-window fashion, enabling detection of multiple object instances in the scene.

More concretely, for all of the visual elements for a rendered 3D view we compute a dense response map across different spatial scales of the 2D input image. For each visual element, we consider the 2D patch locations whose response exceeds a threshold $\tau = 0.2$. We anchor these patches as the root for the star model and compute the detection scores of the remaining visual elements centered around their expected 2D locations relative to the root patch. We perform a local search in a 3×3 HOG cell grid around the expected locations and return the maximum response. The final score is the sum of the visual element detection scores given by Equation (1). Note that we require all visual element detections to come from the same synthesized view, which provides a strong constraint on the viewpoint and style consistency of the detected chair. We found this matching procedure to work well,

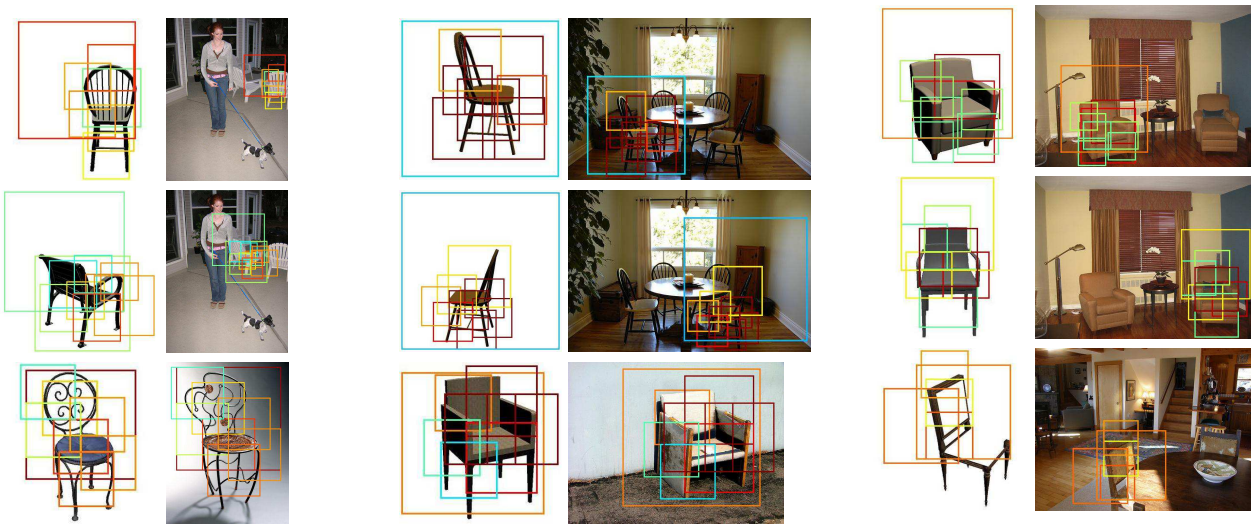


Figure 2: Our output alignments. Parts are colored based on their detection confidence (warmer colors are more confident). **Please see additional results on the project webpage [2].**

though the view and style consistency constraints can be potentially relaxed to accumulate matches across multiple close-by views or models with a similar style.

We perform non-max suppression on the resulting bounding boxes in the following manner. Starting from the most confident detection in the image we (i) remove all other detections that overlap this detection with more than 0.5 area ratio overlap but only (ii) downweight (not remove) the remaining detections with non-zero overlap. This procedure is then iterated starting from the next remaining most confident detection. We found this strategy removes well overlapping false positives while preserving highly confident (and correct) close-by detections (e.g. chairs around a table). This motivates a system that would reason about the entire scene jointly.

4. Results

In this section we show qualitative output alignment results of our system and quantitatively evaluate our approach on images from the challenging PASCAL VOC dataset.

Qualitative results. In Figure 2 we show example output alignments of our algorithm. Notice that our algorithm can detect many different styles of chairs in different poses. For many cases, the predicted chair matches closely the input depicted chair style and pose. In many other cases a similar style is returned, often retrieving an accurate partial match to the depicted chair. Moreover, our approach shows some robustness to background clutter, partial occlusion, and cropping.

In Figure 3 we compare the output of our algorithm with the Deformable Parts Model (DPM) [14]. While the DPM correctly predicts the 2D location of the depicted chairs, along with the 2D locations of its parts, our algorithm produces a more informative result. The aligned 3D chair pose and style allows for true 3D reasoning about the input scene.

For a given chair detection, often there is a set of related, highly confident 3D chair alignments having the same pose and similar style. We visualize these similar chair alignments in Figure 4. Notice that the styles are similar, often sharing one or more 3D parts. This suggests that when there is not an exact style match in the database a composite representation could be used to explain the entire input chair by composing well-matching 3D parts from different 3D chairs. Results for the entire dataset are available on the project webpage [2].

Quantitative evaluation.

We evaluate the detection accuracy of our algorithm on the PASCAL VOC 2012 dataset [1]. We report detection precision-recall on images marked as non-occluded, non-truncated, and not-difficult in the chairs validation set. While this is an easier set compared to the full validation set, nonetheless it is very challenging due to the large intraclass variation, chair poses, and background clutter. Note that removing these difficult examples nonetheless yields some partially-occluded and truncated chairs, as seen in Figure 2. The resulting set contains 179 images with 247 annotated chairs.

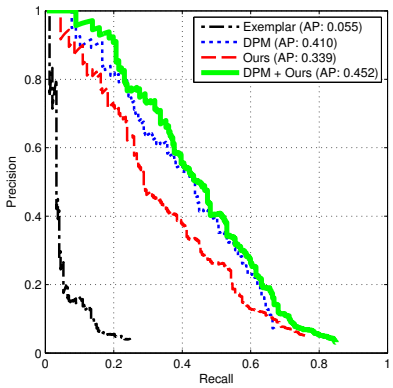


Figure 5: Precision-recall

(a) Input images (b) DPM [14] output (c) Our aligned outputs (d) Retrieved 3D chairs

Figure 3: Comparison of our algorithm output with the deformable parts model (DPM) [14]. While the DPM correctly predicts the 2D location of the depicted chairs, along with the 2D location of its parts, our algorithm is able to predict the 3D pose and style of the chair.

Figure 4: For each image (left), we show the most confident matches (right). Even if we do not have an exact match in our dataset, our algorithm returns meaningful matches.

In Figure 5 we report full precision-recall curves for our algorithm and compare it against two baselines: (i) DPM [14] and (ii) a root template detector using the LDA version of Exemplar-SVM [28]. We train the root template exemplar detector using the whitened HOG formulation in Section 3.1 assuming a single root template that covers the entire 3D chair for the given viewpoint. We calibrate the template as described in Section 3.2. During detection we run the template in a sliding-window fashion across the input image.

Our approach achieves an average precision (AP) of 0.339 on this task. The DPM and root template exemplar detector baselines achieve AP 0.410 and 0.055, respectively. Our performance is noteworthy as it does not use any of the PASCAL VOC training images. We investigated combining our algorithm with DPM for the detection task. For this we estimated an affine transformation for the DPM scores to calibrate it in the range of our returned scores. For overlapping detected windows for the two methods, we give one twice the confidence and discard the other. Combining our approach with DPM yields an AP of 0.452, which significantly out-performs the state-of-the-art DPM baseline.

	Alignment		Style		
	Good	Bad	Good	Ok	Bad
Exemplar-LDA	52%	48%	3%	31%	66%
Ours	90%	10%	21%	64%	15%

Table 1: Results of user study to evaluate the goodness of the alignment and style recovery. The data used in the user study appears on the project webpage [2].

We performed a user study to evaluate the quality of the output alignment and returned chair style. For correct detections at 25% recall, users were asked to label the alignment as "Good" (the returned alignment has very similar pose as the depicted chair) or "Bad" (the alignment is incorrect) and to label the returned chair style as "Good" (the returned chair style is an accurate match), "Ok" (part of the returned chair matches the depicted chair style), or "Bad" (no style match).

We report the results of the user study evaluating the quality of the returned alignments and chair styles in Table 1. We compare against the root template exemplar baseline and out-perform it on both tasks. The fact that the number of exact matches is only 3% for the baseline exemplar detector suggests that there are just too many variations within the chair category. This also motivates using a part-based approach since we are able to obtain high quality partial matches, allowing us to find a close-by chair, whereas the root template detector (Exemplar-LDA baseline) must match the entire chair at once. We found that it was somewhat difficult to judge the returned styles from the exemplar-LDA baseline since the matches were not as good.

Finally, we quantitatively evaluated the accuracy of the estimated chair orientation. For this, we manually annotated the azimuth angle for the same set of detections as used in the user study. Our algorithm returns an azimuth angle within 20° of the ground truth for 87% of the examples. The complete distribution of the orientation errors is shown in figure 6.

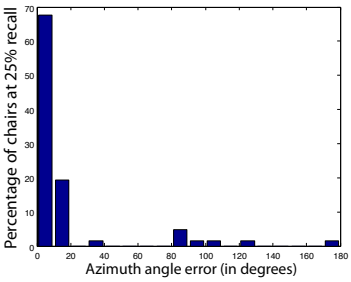


Figure 6: Azimuth angle error.

Failure cases. In Figure 7 we show common failure modes of our algorithm. We observed two main causes for false positives. First, some of the database chairs have a particular texture that can produce confident detections on textured regions, such as the rear view of the chair matching the checkered window pattern in Figure 7(a). Second, there exist many regions in images that, through the eyes of a HOG descriptor, appear as a chair, as seen in Figure 7(b). We also observed this effect in estimating the pose of the chair, which resulted in confusing the front/back of the chair, as seen in Figure 7(c). The chairs our algorithm miss are mainly special chair types, such as sofa-like chairs, for which we do not have training examples, and chairs strongly occluded by other objects and people.

Computational cost. Computing the discriminative elements takes 6 seconds per rendered view (1 second for extraction and 5 seconds for calibration) on a single machine, which can be performed offline and parallelized. The main bottleneck is having to match and align an image to the large number of rendered views from the 3D models, which involves detecting over 800K discriminative visual elements. Our Matlab implementation matches and aligns an image in 2 minutes on an 80-core cluster. While this computational complexity is high, there is room for optimism that our approach can be made significantly more efficient. For instance, the recent part-based hashing technique of [10] could be directly applied to our algorithm by applying winner-take-all hashing [40] on the discriminative visual elements. As we have a similar number of part detectors as considered in their work, we believe we can process an image in less than 20 seconds on a single multicore processor.

5. Conclusion

We have demonstrated successful detection and alignment of 3D CAD chair models to chairs depicted in 2D imagery. Our approach relied on matching spatial configurations of mid-level discriminative visual elements extracted from a large database of CAD models having a large number of rendered views. Our algorithm is able to recover the chair pose, in addition to its style. We evaluated our approach on the challenging PASCAL VOC dataset and showed that, when combined with the output of the deformable parts model detector [14], we are able to achieve higher detection accuracy than using either method alone. We also demonstrated that our algorithm is able to reliably recover the chair pose and style, as shown in our user study and orientation error analysis. The output alignments produced by our system open up the possibility of joint 3D reasoning of the depicted objects in a scene toward the larger goal of full 3D scene understanding.

Acknowledgments. We are grateful to the anonymous reviewers for their constructive comments. This work was partly supported by the EIT ICT Labs, MSR-INRIA laboratory, ANR project SEMAPOLIS (ANR-13-CORD-0003), ONR-MURI Grant N000141010934 and Google Research Grants.

References

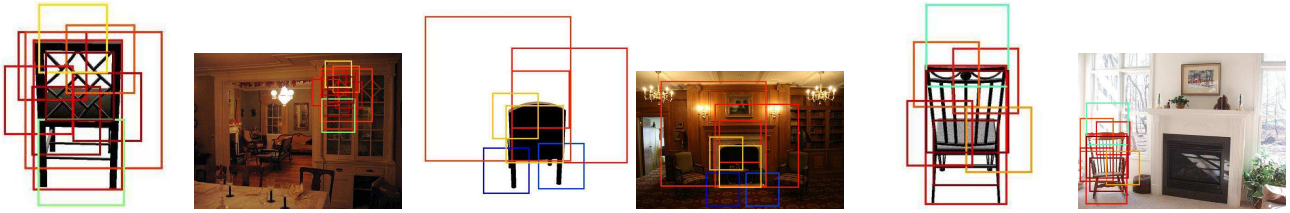
[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/, 2012. 2, 5

[2] http://www.di.ens.fr/willow/research/seeing3Dchairs/, 2012. 3, 5, 6

[3] R. Arandjelović and A. Zisserman. Smooth object retrieval using a bag of boundaries. In *ICCV*, 2011. 2

[4] M. Aubry, B. Russell, and J. Sivic. Painting-to-3D model alignment via discriminative visual elements. *ACM Transactions on Graphics*, 2014. 2, 3, 4

[5] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *ECCV*, 2012. 2



(a) Confusing discriminative patterns (b) Chair-like structures in the background (c) Errors in pose estimation

Figure 7: Common failures of our algorithm.

[6] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 2

[7] W. Choi, Y. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3D geometric phrases. In *CVPR*, 2013. 2

[8] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total Recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007. 2

[9] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 1, 3

[10] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *CVPR*, 2013. 7

[11] L. Del Pero, J. Bowdish, B. Kermgard, E. Hartley, and K. Barnard. Understanding Bayesian rooms using composite 3D object models. In *CVPR*, 2013. 2

[12] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013. 4

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 1

[14] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9), 2010. 1, 2, 4, 5, 6, 7

[15] S. Fidler, S. Dickinson, and R. Urtasun. 3D object detection and viewpoint estimation with a deformable 3D cuboid model. In *NIPS*, 2012. 2

[16] M. Gharbi, T. Malisiewicz, S. Paris, and F. Durand. A Gaussian approximation of feature space for fast image similarity. Technical report, MIT, 2012. 3

[17] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In *ICCV*, 2011. 2

[18] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010. 2

[19] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012. 3

[20] M. Hejrati and D. Ramanan. Analyzing 3D objects in cluttered images. In *NIPS*, 2012. 2

[21] D. P. Huttenlocher and S. Ullman. Object recognition using alignment. In *ICCV*, 1987. 1

[22] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis. Representing videos using mid-level discriminative patches. In *CVPR*, 2013. 2

[23] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013. 2

[24] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3D point clouds. In *ECCV*, 2012. 2

[25] J. Lim, H. Pirsiavash, and A. Torralba. Parsing IKEA objects: Fine pose estimation. In *ICCV*, 2013. 2

[26] D. Lowe. The viewpoint consistency constraint. *IJCV*, 1(1):57–72, 1987. 1

[27] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[28] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 2, 3, 4, 6

[29] J. L. Mundy. Object recognition in the geometric era: A retrospective. In *Toward Category-Level Object Recognition, volume 4170 of Lecture Notes in Computer Science*, pages 3–29. Springer, 2006. 1

[30] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3D geometry to deformable part models. In *CVPR*, 2012. 2

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 2

[32] L. Roberts. Machine perception of 3-D solids. In *PhD. Thesis*, 1965. 1

[33] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *CVPR*, 2003. 2

[34] S. Satkin, J. Lin, and M. Hebert. Data-driven scene understanding from 3D models. In *BMVC*, 2012. 2

[35] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, 2011. 2

[36] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2

[37] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple classifiers. In *CVPR*, 2001. 1

[38] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012. 2

[39] J. Xiao, B. Russell, and A. Torralba. Localizing 3D cuboids in single-view images. In *NIPS*, 2012. 2

[40] J. Yagnik, D. Strelow, D. Ross, and R.-S. Lin. The power of comparative reasoning. In *ICCV*, 2011. 7

[41] M. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3D representations for object recognition and modeling. *IEEE PAMI*, 2013. 2