Eurographics Workshop on 3D Object Retrieval (2017)
I. Pratikakis, F. Dupont, and M. Ovsjanikov (Editors)

Towards Recognizing 3D Models Using A Single Image

H. A. Rashwan[1] and S. Chambon[1] and G. Morin[1] and P. Gurdjos[1] and V. Charvillat[1]

[1]IRIT-CNRS, University of Toulouse, France

Abstract

As 3D data is getting more popular, techniques for retrieving a particular 3D model are necessary. We want to recognize a 3D model from a single photograph; as any user can easily get an image of a model he/she would like to find, requesting by an image is indeed simple and natural. However, a 2D intensity image is relative to viewpoint, texture and lighting condition and thus matching with a 3D geometric model is very challenging. This paper proposes a first step towards matching a 2D image to models, based on features repeatable in 2D images and in depth images (generated from 3D models); we show their independence to textures and lighting. Then, the detected features are matched to recognize 3D models by combining HOG (Histogram Of Gradients) descriptors and repeatability scores. The proposed methods reaches a recognition rate of 72% among 12 3D objects categories, and outperforms classical feature detection techniques for recognizing 3D models using a single image.

1. Introduction

Recognizing a 3D object consists in identifying a 3D object among a set of 3D models or categories of models given a request input data, in our case, a single 2D photograph. Capturing or indexing an image by keywords are two very easy way to describe a 3D shape. Therefore, requesting a 3D model from a photograph would be a natural way to index a repository of 3D shape. However, 2D/3D matching is a difficult task due to occlusions, view point variations, lighting and illumination changes. The type of representation used both for 2D images and 3D models is critical for the choice of the object recognition strategy. For representing and characterizing 3D models, some approaches rely on attributes of the object's 3D geometry [GBS*14], other on attributes of its 2D projection [CK01, AS13] (distinction between object- vs. image-representations).

A fist step is to choose a representation of the 3D object comparable to an image of this object. However, photographs and 3D models have very different appearance: 3D models contain only geometric information, and no color or texture, whereas the photograph (i.e., color or gray images) is the result of a combination of color, texture, lighting and shape information. Thus, the first task for recognizing a 3D model based on a single 2D image is to find an appropriate representation of 3D models in which reliable features can be extracted. 3D models can be reconstructed by a set of color images [AFS*11] or range images. For reconstruction techniques based on color images, the matching of a color image to a set of color images solves a particular 3D object recognition problem. However, these techniques rely on texture and are affected by lighting. In order to model a 3D shape independently of texture, we propose to represent the 3D model by a set of range images that express the model shape independently to color or texture information.

The second step of 2D/3D matching consists in proposing how to match entities between these two modalities in this common representation. It can be partial [IZFB09] or dense matching, based on local or global characteristics [SS02]. A key requirement on these features, as in classic 2D matching between photographs, is to be computed with a high degree of *repeatability*: the probability that key features in a photograph are found close to those extracted in a depth image must be high. Since we suppose that an individual photograph of an object of interest is acquired in a textured environment, we will focus on extracting features of photographs related to the object shape and as similar as possible to features extracted from the set of rendered images of the corresponding 3D model, more precisely, a set of depth images. Moreover, this similarity has to be the highest one when the rendered image is from the same point of view as the photograph.

Many alignment methods rely on extracting features such as points and curves of the 3D model/scene. Most of point-based methods are based on extracting SIFT, Scale invariant Features Transform [Low04], adapted from 2D to 3D [LVJ05, SLK11]. Due to the differences in appearance and the different light conditions between 2D and 3D, this kind of method can yield an inaccurate alignment process, and suffers from low precision due to the loss of 3D scene information during rendering. In turn, other methods have been proposed to extract a set of curves and lines whose points are local maxima on a surface as in [JDA04, GW11]. However, these methods often produce false edges that are not related to occluding contours, which are important for pose estimation. Recently, a rendering technique called Average Shading Gradients (ASG) was

DOI: 10.2312/3dor.20171062

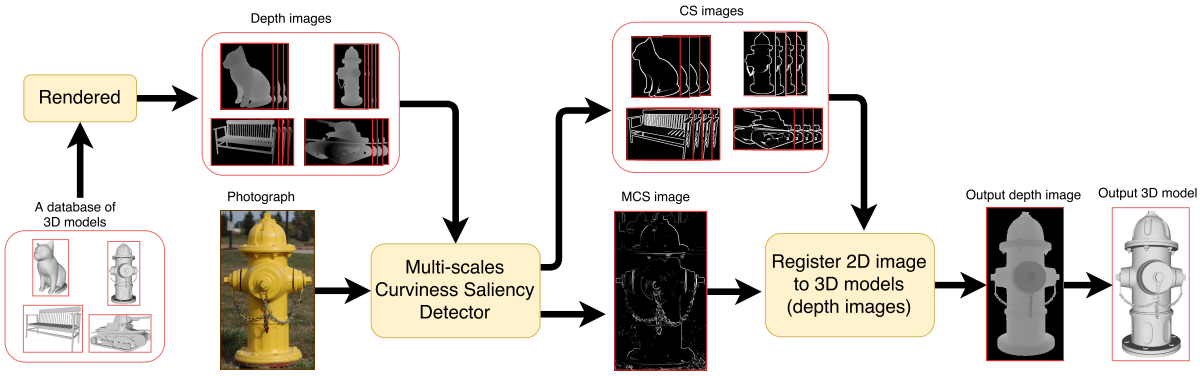

Figure 1: *To compare 2D images with 3D models, we use a collection of rendered images of the 3D models in different viewpoints, and then we detect points of interest (ridges) with common basis definitions between depth images and intensity images.*

proposed in [PR15] to register an image to its corresponding 3D model. ASG represents 3D models by averaging shading gradients over all lighting directions to cope with the unknown lighting conditions. The 2D image representation that they match with the 3D ASG representation is based on the image gradients and thus yields unreliable correspondences when the 2D image contains textures or background.

Finally, Curviness saliency (CS) [RCG*16] is related to the curvature estimation (a function of the eigenvalues derived of the Hessian matrix). This representation is directly related to the discontinuities of the object's shape, and, by nature, the extracted features should be robust to texture and light changes. Thus, in this paper, features extracted in depth images highlight geometric characteristics of an object. Regarding photographs, curviness saliency detector is able to detect the approximate shape. The proposed detector can also reduce the influence of scale changes because it is estimated over all scales and image locations in order to identify scale-invariant interest points as shown in Figure 1. In our targeted application, the user manually defines the Region Of Interest (ROI) (i.e., localizing the object projection in an image). In an image with a single object, we assume that the object is in focus, so we could detect the ROI using the bounding box of all points in focus. For object recognition, we propose an algorithm to identify a 3D model given an image and the ROI. We then compute the similarity score based both on the HOG descriptor, Histogram Of Gradients [DT05], and the repeatability scores between the features of the cropped image of the object and the features of rendering depth images from multiple 3D models.

The remainder of this paper is organized as follows: section 2 presents to the curviness saliency used for extracting the features in 2D photographs and 3D models. The recognition process of 3D models based on histogram of curviness saliency and the repeatability scores of the features between depth images and photographs is explained in section 3. A set of retrieval experiments is described and analyzed in section 4. Finally, section 5 provides a summary and perspectives for future work.

2. Feature Detection in 2D Photographs and 3D Models

The features proposed in [RCG*16] are briefly presented before using them for the recognition step, Section 3.

2.1. Curviness Saliency Detector

The principal curvatures κ_1, κ_2 of a surface S at point $\mathbf{p}$ are approximately the eigenvalues of the Hessian matrix H:

$$\mathsf{H} = \begin{bmatrix} \mathcal{Z}_{xx} & \mathcal{Z}_{xy} \\ \mathcal{Z}_{xy} & \mathcal{Z}_{yy} \end{bmatrix},$$

where $\mathcal{Z}_{xx}$, $\mathcal{Z}_{xy}$ and $\mathcal{Z}_{yy}$ are the second-order partial derivatives of a pixel of an image $\mathcal{Z}$ in x, and y directions. When $\kappa_1 \geq \kappa_2$, the pixel lies on a "curvilinear feature" in a representation common to images (e.g., photographs or depth images). The function, called *curviness saliency* (*CS*), computes the difference between principal curvatures:

$$\mathrm{CS} \triangleq \lambda_1 - \lambda_2, \tag{1}$$

where λ_1 and λ_2 are the eigenvalues of H assuming $\lambda_1 > \lambda_2$. *CS* is large when $\lambda_1 \gg \lambda_2$, that is, distant foci, or, as well, the surface is approximated (at second order) by a highly elongated ellipse or a "squashed" hyperbola. This occurs when the point is located on a ridge (either curved or straight). Otherwise, when $\lambda_1 \simeq \lambda_2$, the osculating conic approaches a circle and the distance between foci becomes very small.

We compute *CS* (1) at each pixel and a pixel with high *CS* value can be considered as a potential keypoint. Note that the two eigenvalues $\lambda_1 \triangleq \lambda_+$ and $\lambda_2 \triangleq \lambda_-$ of the scaled Hessian matrix $\widetilde{\mathsf{H}}$ can be directly computed by:

$$\lambda_\pm = \frac{\alpha}{2}\left(\mathcal{Z}_{xx} + \mathcal{Z}_{yy} \pm \sqrt{(\mathcal{Z}_{xx} - \mathcal{Z}_{yy})^2 + 4\mathcal{Z}_{xy}^2}\right). \tag{2}$$

By subsititing $\lambda_{1,2}$ of (2) in (1) and $\alpha = 1$, the CS is thus given by

$$\boxed{\mathrm{CS} \cong \left(\|\mathsf{H}\|_F^2 - 2\,\det(\mathsf{H})\right)}, \tag{3}$$

where $\|\mathsf{H}\|_F$ and $\det(\mathsf{H})$ are respectively the Frobenius norm and

the determinant of the Hessian matrix H. The relation between these two terms can be used to measure how much this surface bends in different directions at any point $\mathbf{p}$. The proposed potential function *CS* can be interpreted as a scalar measurement of the curvature at a local surface patch. In addition, *CS* is a scalar curvature index, commonly used in differential geometry, which quantifies lack of flatness of the surface at a specific point [LWU13].

2.2. Multi-Scale Curviness Saliency

Computing the curviness saliency in an image at a single-scale detects points that have high curvature in a particular scale. Multi-scale helps to detect important structures as well as small details. At a coarse level, the detection of edges lacks localization precision and misses small details. At a fine level, details are preserved, but the detection suffers greatly from clutters in textured regions. In consequence, we require that a keypoint of a 2D photograph (those with high value of CS) to appear in a multi-scale space. As we expected, shape-related keypoints, which appear in all the scales, are kept whereas texture and background-related keypoints, which appear in a limited number of scales, are discarded. Indeed, the CS values of small details and textures are high in the fine level, whereas these values become lower in the coarsest levels. To combine the strengths of each scale, the CS value of each pixel over n scales is analyzed. If this value is higher than a threshold T in all scales, the maximum curviness saliency (MCS) value of this pixel over all scales is kept. This threshold is a function of n, the number of the smoothed image (here we took, $T = e^{-n}$). And, if the CS value of a pixel happens to be lower than T at a particular level, we considered this pixel as texture (or small detail) point, thus removed it from the final multi-scale curviness saliency (MCS) features. The multi-scale curviness saliency MCS if applied to the request input photograph and will be compared with features extracted on the depth image for 2D-3D matching. Our proposal for depth images is to simply use curviness saliency CS: as depth discontinuities appear at all scales, it is sufficient to consider a single scale.

3. Recognition

The goal of recognition is to identify which 3D object is represented by a (request) photograph q. For that purpose, we identify the closest model M among M_i models, $i = 1 \ldots k$, where k the number of 3D models in the database. Moreover, each model is represented by a set of depth images d_{i_n}, n the number of rendered depth images, from approximately uniformly distributed viewing angles around a sphere by changing elevation h and azimuth a angles (the choices for a and h are discussed in section 4).

To describe the features, we naturally expand the famous classical HOG, widely used in the literature (e.g. [ARS14, PR15]), to work on curviness saliency. Indeed, both in rendered depth images and in photographs, the orientation of the curvature and the value of the curviness saliency is used for building the descriptors. For that, the direction of the curviness saliency is considered as the direction of the eigenvector $\mathbf{e}_1$ corresponding to the largest eigenvalue of the Hessian matrix in depth images or the structure tensor in a photograph. Then, in the same way as HOG, the curviness saliency features are detected for an image and are binned into sparse per-pixel histograms. Given the HOG descriptor from a 2D query image $\mathbf{D}_q$ and the HOG descriptors of the rendered images $\mathbf{D}_{i_n}$, in order to compare $\mathbf{D}_q$ and every $\mathbf{D}_{i_n}$, a similarity score is computed as proposed in [ARS14]:

$$\mathbf{S}_{hog}(i,n) = (\mathbf{D}_{i_n} - \mu_{s_i})^T \Sigma_{s_i}^{-1} \mathbf{D}_q, \tag{4}$$

where, Σ_{s_i} and μ_{s_i} are, respectively, the covariance matrix and the mean over all descriptors of the n rendered depth images of the i model. At input time, evaluating S_{hog} can be done by computing the probability of the inverse of the inner product between $\mathbf{D}_q$ and a transformed set of descriptors. The $\mathbf{S}_{hog}$ probability is then maximized to identify the depth image the closest to the query image.

As already mentioned, an important property needed for pose estimation and recognition is repeatability. The repeatability score is the probability that key features in the intensity image are found close to those extracted in the depth image. So, to validate the repeatability of the detected features, we propose to also evaluate a global similarity by measuring how well detected keypoints in depth images agree with keypoints in the request photograph. More precisely, the closest view of the 3D model seen in the photograph should have a repeatability score higher than both depth images on the same object from a different viewpoint and depth images of different 3D models. If we denote $\mathbf{R}_i$ the repeatability scores of n rendered views of a i model and a given image, the similarity $\mathbf{S}_{rep}(i,n)$ is defined by:

$$\mathbf{S}_{rep}(i,n) = exp\left(\frac{-(\mathbf{B}_{i_n} - \mu_{r_i})^2}{2\,\sigma_{r_i}^2}\right), \tag{5}$$

where $\mathbf{B}_{i_n} = 1 - \mathbf{R}_{i_n}$, μ_{r_i} is the mean value of $\mathbf{B}_i$ and σ_{r_i} is the standard deviation of the repeatability scores of n views of the i model [†]. By combining HOG feature similarities and the similarity based on the repeatability, the probability of the final similarity can be computed by:

$$\mathbf{S}(i,n) = \mathbf{S}_{hog}(i,n) \odot \mathbf{S}_{rep}(i,n), \tag{6}$$

where $\odot$ is the Hadamard product. Based on the computation of S for each depth image, we select the depth image $\mathbf{d}_{i_n}$ that corresponds to the highest S to identify the answer, i.e. the corresponding 3D model M_i.

4. Experiments

For indexing a 3D model using a single photograph, we assume that an object is localized by a bounding box in the given photograph. The cropped image is then matched with a set of candidate 3D models. For recongition task, we have developed a small database of ten textureless 3D objects ([‡]), and we collected a set of 15 real images found online for each object and then used this database in our evaluation. To be more general, we also tested our approach on a classical large benchmark called *PASCAL3D+* [XMS14]. The *PASCAL3D+* dataset is used in general for object detection and pose

† In this work, $\sigma_{r_i} = 0.1$ is empirically chosen

‡ availableonlinehttp://tf3dm.com

estimation but we use it for object recognition. The *PASCAL3D+* database contains real images corresponding to 12 object categories; each category contains around 1000 real images acquired under different conditions (e.g., lighting, complex background, low contrast) and (at least) 3 reference 3D models; each color image is associated to the bounding box around the object, the 3D pose and the name of the corresponding reference 3D model. A single image is selected from the category to index the corresponding 3D model among all reference models in the database. In our context, we thus take an image and the bounding box information for *PASCAL3D+*, we output the closest view among rendering images from the 36 models (3 per categories). We use the name of the corresponding 3D model category as a ground truth for our algorithm.

We rendered the depth images from the 3D CAD models from different viewpoints by using *MATLAB 3D Model Renderer*[§]. Actually, a large number of depth images (60 in our experiments) is necessary to completely represent a 3D model. This yields a significant execution time of the matching process. Thus, we orthographically render $N = 60$ depth images per model from approximately uniformly distributed viewing angles h and a (i.e., in these experiments, h is increased by a step of 45^o, and the azimuth angle, 45^o).

In addition, in this paper, the HOG descriptor is quantized into 9 bins, as proposed in [DT05]. The photograph and each depth image are divided into a grid of square cells (i.e., in this work, the image is divided into 8×8[¶]. For each cell, the curviness saliency histograms are aggregated by weighting them with their respective magnitude.

For each category of objects, we compute the average recognition rate of finding the correct model for the input single image. Only non-occluded and non-truncated objects in the real images of the PASCAL+3D were used for the evaluation (i.e., around 600 images per category). We compare CS and two concurrent 3D representations: Average Shading Gradients (ASG) [PR15] and Apparent Ridges (AR) [JDA04] against MCS and four classical feature detectors on intensity image (i.e., Principle Curvature Image (PCI) [DZM*07], MinEig [ST94], Canny Edge detection, and SIFT [Low04]) by testing the matching on every possible pairs.

As shown in Table 1, with the PASCAL+3D database, the correct recognition rate achieved between the proposed CS representation of the 3D models with MCS of photograph representation outperforms all other variations of the different methods. This confirms that curviness saliency representation computed on the depth images of a 3D model can properly capture the discontinuities of surfaces. In addition, MCS can reduce the influence of texture and background components and it can also approximately extract the edges related to the object shape in intensity images. Furthermore, the recognition rate drops by more than 50% when combining AR on 3D depth images and Canny edge detection on intensity images. Apparent Ridges rendering yields the smallest recognition rate accuracy with the three image representations among 3D models representation techniques. However, using ASG on textureless 3D models against MCS yields acceptable recognition rate. Which indicates that ASG computed from the normal map of an untextured geometry is a good rendering technique. However, since image gradients are affected by texture, they do not perform well.

2D Representation / 3D Representation	MCS	PCI	MinEig	SIFT	Canny
CS	0.72	0.61	0.45	0.41	0.40
ASG	0.63	0.52	0.43	0.36	0.35
AR	0.59	0.50	0.41	0.34	0.33

Table 1: *Average Correct Recognition rate of* 12 *categories of the Pascal+3D dataset with all the combination of representations. Only non-occluded and non-truncated objects in the real images of the PASCAL+3D (i.e., around* 600 *images per category) were used for the evaluation.*

Figure 2 shows three examples (photographs and the identified 3D models) from our developed database, in turn Figure 3 shows two examples from the PASCAL3D+ dataset. As shown in the two figures, we see that both MCS and PCI methods are able to detect the edges belonging to occluded contours of the 3D objects. However, MCS better filters edges due to texture and background regions; thus, MCS properly identifies the object shape in intensity images and so, yields a better matching between the request intensity and the 3D representation (i.e., the rendered depth images of the 3D model). In turn, regarding 3D model representation, CS and ASG properly detect the surface variations in depth images. However, CS properly detects the silhouette of the object, as well as the contours related to the surface discontinuities. The detected contours with CS are very important to get accurate matching with photographs. In turn, ASG is affected with the normal map discontinuities, but it also detects a lot of edges that are not related to the shape of the objects.

In the second experiment, Figure 4 shows the average recognition rate over the 12 object categories of the PASCAL+3D of images among the top r similarities (i.e., ranks 1, 3, 5, 10 and 20). That similarity is ranked based on the highest r correspondences of S The corresponding 3D object is searched within this set of rendered images. Of course, the average recognition rate is improved when the number of images/ranks increases with all three methods used for 3D models and MCS for image representation. In fact, AR/MCS yields the smallest precision value, because AR often produces false edges that are not related to occluding contours.

5. Conclusion and Future Work

We proposed a first step towards a simple indexation of 3D repositories by a request using a single photograph, that can be captured in arbitrary conditions, and the object may be textured. 3D models are represented by a set of depth images invariant to lighting and texture. Curviness saliency estimation is used and we estimate features of the photograph by using multi-scale. The similarity between the features of a photograph and depth images is computed by both the HOG descriptor and the repeatability scores. The results show high recognition rate on a set of natural request photographs for 12 models and thus highlight the feasibility of indexing 3D by a natural image thanks to the fact that the extracted keypoints are

§ http://www.openu.ac.il/home/hassner/projects/poses/

¶ Different grids (4×4, 8×8 and 16×16) were tested, and a 8×8 grid yields the best precision rate.

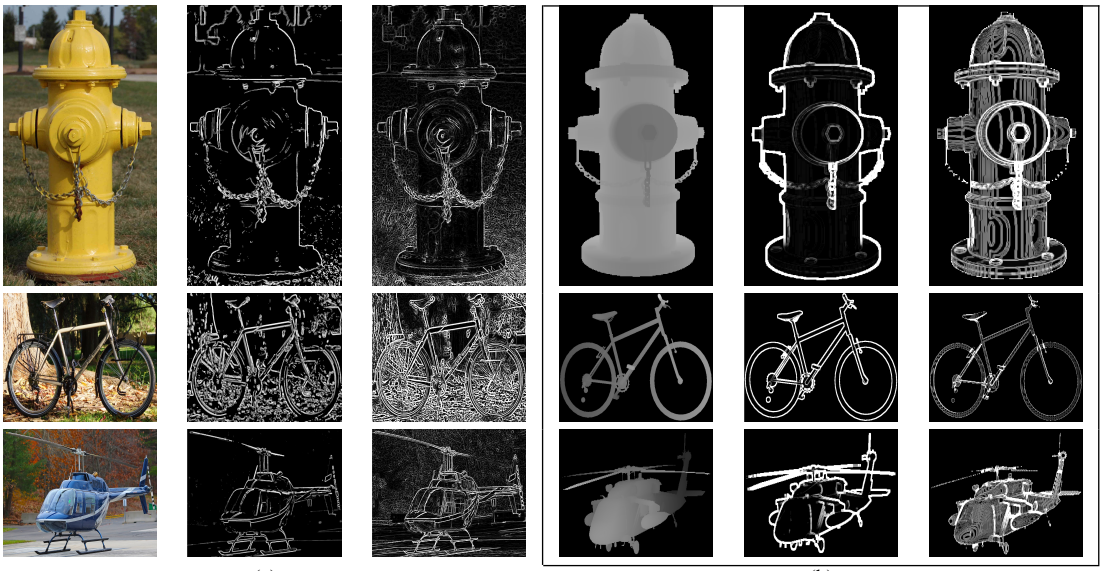

Figure 2: *Three examples of our developed database: (a) (Col 1) Request photograph and extracted features with (Col 2) MCS, (Col 3) PCI [DZM*07], both with 5 scales, and (b) (Col 1) depth images (3D representation) and extracted features with (Col 2) CS and (Col 3) ASG [PR15].*

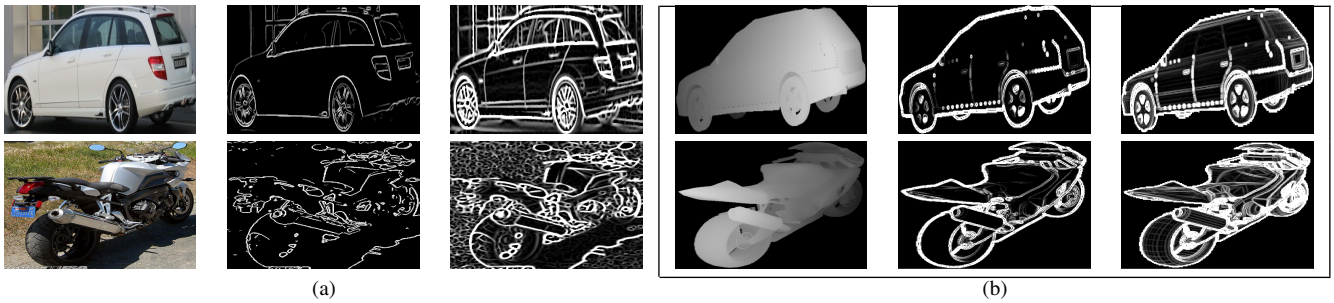

Figure 3: *Two examples of the PASCAL3D+ dataset: (a) (Col 1) Request photograph and extracted features with (Col 2) MCS, (Col 3) PCI [DZM*07], both with 5 scales, and (b) (Col 1) depth images (3D representation) and extracted features with (Col 2) CS and (Col 3) ASG [PR15].*

more repeatable than classical detectors. Future work will first deal with more models and second, to introduce deep learning system based on the proposed curviness saliency estimation to get more accurate recognition rate.

References

[AFS*11] AGARWAL S., FURUKAWA Y., SNAVELY N., SIMON I., CURLESS B., SEITZ S., SZELISKI R.: Building rome in a day. *Communications of the ACM 54*, 10 (2011). 1

[ARS14] AUBRY M., RUSSELL B., SIVIC J.: Painting-to-3d model alignment via discriminative visual elements. *ACM Trans. on Graphics 33*, 2 (2014). 3

[AS13] ATMOSUKARTO I., SHAPIRO L.: 3d object retrieval using salient views. *Int. journal of multimedia information retrieval 2*, 2 (2013). 1

[CK01] CYR C. M., KIMIA B. B.: 3d object recognition using shape similiarity-based aspect graph. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (2001), vol. 1, IEEE, pp. 254–261. 1

[DT05] DALAL N., TRIGGS B.: Histograms of oriented gradients for human detection. In *CVPR* (2005), vol. 1. 2, 4

[DZM*07] DENG H., ZHANG W., MORTENSEN E., DIETTERICH T., SHAPIRO L.: Principal curvature-based region detector for object recognition. In *CVPR* (2007). 4, 5

[GBS*14] GUO Y., BENNAMOUN M., SOHEL F., LU M., WAN J.: 3D object recognition in cluttered scenes with local surface features: A survey. *PAMI 36*, 11 (2014). 1

[GW11] GODIL A., WAGAN A. I.: Salient local 3d features for 3d shape retrieval. In *IS&T/SPIE Electronic Imaging* (2011), International Society for Optics and Photonics, pp. 78640S–78640S. 1

[IZFB09] IRSCHARA A., ZACH C., FRAHM J., BISCHOF H.: From structure-from-motion point clouds to fast location recognition. In *CVPR* (2009), IEEE. 1

[JDA04] JUDD T., DURAND F., ADELSON E.: Apparent ridges for line drawing. *ACM T. Graphics 26*, 3 (2004). 1, 4

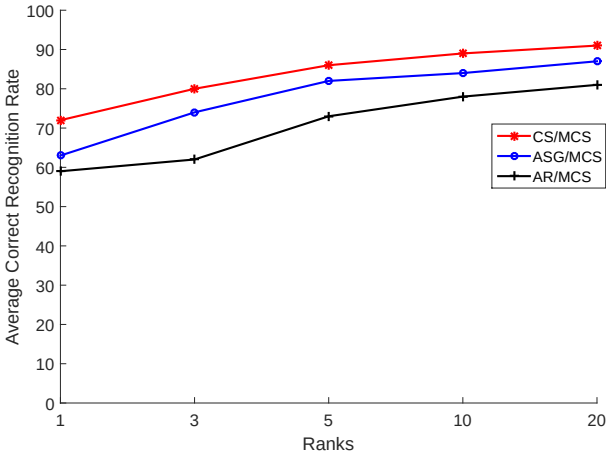

Figure 4: *Average precision values with different ranks (i.e., 1, 3, 5, 10 and 20) with 3D models and photographs representation using: CS representation for 3D and MCS for photographs (CS/MCS), ASG representation for 3D and MCS for photographs (ASG/MCS), and AR representation for 3D and MCS for photographs (AR/MCS).*

[Low04] LOWE D.: Distinctive image features from scale-invariant keypoints. *IJCV 60*, 2 (2004). 1, 4

[LVJ05] LEE C. H., VARSHNEY A., JACOBS D. W.: Mesh saliency. In *ACM transactions on graphics (TOG)* (2005), vol. 24, ACM, pp. 659–666. 1

[LWU13] LEFKIMMIATIS S., WARD J. P., UNSER M.: Hessian schatten-norm regularization for linear inverse problems. *IEEE transactions on image processing 22*, 5 (2013), 1873–1888. 3

[PR15] PLOTZ T., ROTH S.: Registering images to untextured geometry using average shading gradients. In *ICCV* (2015). 2, 3, 4, 5

[RCG*16] RASHWAN H., CHAMBON S., GURDJOS P., MORIN G., CHARVILLAT V.: Towards multi-scale feature detection repeatable over intensity and depth images. In *ICIP* (2016). 2

[SLK11] SATTLER T., LEIBE B., KOBBELT L.: Fast image-based localization using direct 2d-to-3d matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (2011), IEEE, pp. 667–674. 1

[SS02] SCHARSTEIN D., SZELISKI R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV 47*, 1-3 (2002). 1

[ST94] SHI J., TOMASI C.: Good features to track. In *CVPR* (1994). 4

[XMS14] XIANG Y., MOTTAGHI R., SAVARESE S.: Beyond pascal: A benchmark for 3d object detection in the wild. In *Applications of Computer Vision (WACV)* (2014). 3