Using Curvilinear Features in Focus for Registering a Single Image to a 3D Object

Hatem A. Rashwan, Sylvie Chambon, Pierre Gurdjos, Géraldine Morin, Vincent Charvillat

Using Curvilinear Features in Focus for Registering a Single Image to a 3D Object

Hatem A. Rashwan, Sylvie Chambon, Pierre Gurdjos, Géraldine Morin, and Vincent Charvillat

Abstract—**In the context of 2D/3D registration, this paper introduces an approach that allows for matching features detected in two different modalities, photographs, and 3D models, by using a common 2D representation. More precisely, 2D images are matched with a set of depth images representing the 3D model. After introducing the concept of Curvilinear Saliency, which is related to curvature estimation, we propose a new ridge and valley detector for depth images rendered from 3D models. A variant of this detector is adapted to photographs, first by considering multi-scale features and second by integrating the focus curve principle. Finally, a registration algorithm determines the correct view of the 3D model and, thus, the pose of the photograph. This approach relies on the Histogram of Curvilinear Saliency (HCS), an adaptation of the Histogram of Oriented Gradients (HOG) to the proposed features in 2D and 3D. The presented results highlight both the quality of the features detected in terms of repeatability and the interest of the approach for registration and pose estimation.**

Index Terms—**2D-3D registration, pose estimation, object detection, feature extraction, curvilinear saliency.**

I. Introduction

MANY computer vision and robotic applications are used to take 2D contents as input; recently, however, 3D contents have become simultaneously available and popular. To benefit from both modalities, 2D/3D matching is necessary. For medical imaging, registration of pre-operative 3D volume data with intra-operative 2D images is increasingly necessary to assist physicians [27]. For robotics, the 2D/3D matching can be useful to determine the 3D pose of an object of interest for 3D navigation or object grasping [33]. The main goal is to find the transformation of the 3D model that defines the pose for a query 2D image. Thus, a typical 2D/3D registration problem consists of two mutually interlocked subproblems, that is, point correspondence and estimation.

To match 2D photographs directly to 3D models or point clouds, most systems rely on detecting and describing features on both 2D/3D data and subsequently on matching these features [1], [47]. Some recent approaches are based on learning by a specific supervision classifier [41], [45]. In [45], a convolutional neural network (CNN) architecture is introduced to predict a viewpoint. They combine multi-scale appearance with a viewpoint-conditioned likelihood. The objective is to predict key points to capture the finer details to correctly detect the bounding box of the objects. In [41], the authors have rendered millions of synthetic images from 3D models under varying illuminations, lighting and backgrounds and then have proceeded to use them to train a CNN model for the viewpoint estimation of real images. These methods produce very interesting results, but they require a high volume of viewpoint-annotated images to learn the classifiers. What makes it difficult to match the 3D features of an object to the 2D features of one of its photographs is the appearance of the object. Indeed, this appearance dramatically depends on the intrinsic characteristics of the object, such as texture and color/albedo, as well as the extrinsic characteristics related to the acquisition, such as the camera pose and the lighting conditions. Consequently, some approaches manually define correspondences between the query image and the 3D model, such as [10]. These manual selections can easily become difficult to apply to large image sets. Moreover, in this paper, we focus on automated approaches. Note that some systems are able to generate a simultaneous acquisition of photographs, and scanning of a 3D model; using this kind of system nevertheless induces limited applications. Other methods solve the problem by distinguishing two subproblems: choosing the data's common representation followed by finding the correspondences. More precisely, these methods transform the initial 2D/3D registration problem into a 2D/2D matching problem by rendering multiple 2D images of 3D models from different viewpoints, such as in [5], [7], [32]. The work presented in this paper focuses on this type of approach.

Consequently, the first task of 2D/3D registration is to *find an appropriate representation of 3D models such that reliable features can be extracted in 2D and 3D*. In [5], synthetic images of the 3D model are rendered, while depth images are rendered in [7]. More recently, [32] proposes Average Shading Gradients. This rendering technique for a 3D model averages the gradient normals over all lighting directions to cope with the unknown lighting of the query image. The advantage of representing the 3D model by a set of depth images lies in the fact that it can express the shape model independent of color and texture information. Therefore, we have decided to represent 3D models by sets of depth images; see Fig. 1.

The second difficulty of 2D/3D registration consists of proposing *how to match entities between the two modalities*

Manuscript received May 10, 2018; revised December 30, 2018, February 26, 2019, and March 23, 2019; accepted March 30, 2019. Date of publication April 22, 2019; date of current version July 1, 2019. This work developed within the LADIO Project was supported in part by the European Union's Horizon 2020 programme under Grant 731970 and in part by the MOBVILLE Project (Occitanie Regional Project, France). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Raja Bala. *(Corresponding author: Sylvie Chambon.)*

The authors are with the Institut de Recherche en Informatique de Toulouse, University of Toulouse, 31000 Toulouse, France (e-mail: schambon@enseeiht.fr).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

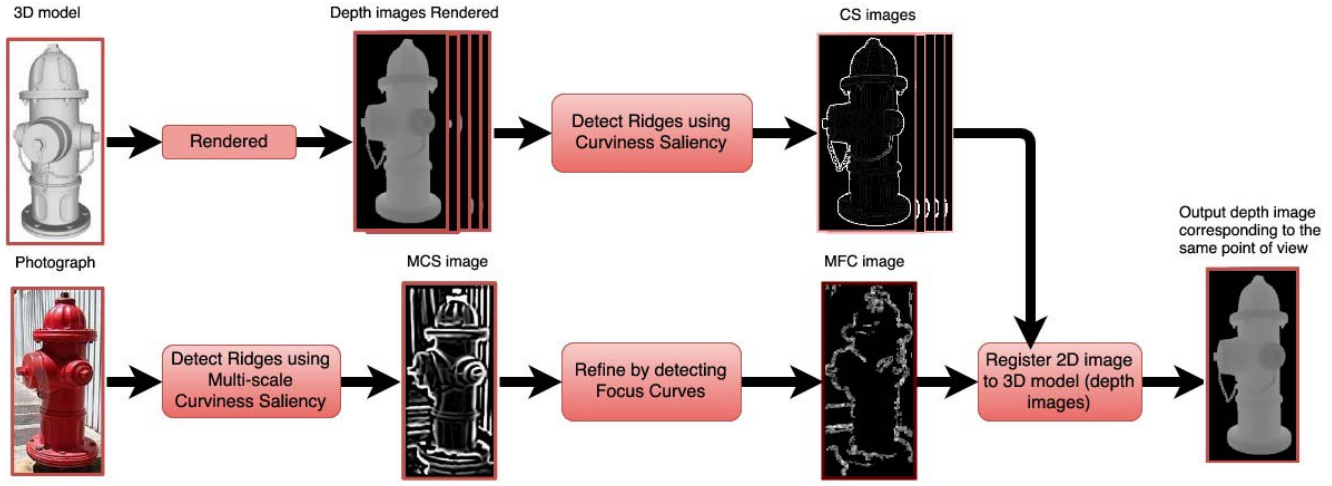

Fig. 1. To compare 2D images with 3D models, a collection of rendered images of the 3D models from different viewpoints is used, and then, points of interest (Multi-Curvilinear Saliency, MCS) are detected with common basis definitions between depth and intensity images. Each depth image is compared with the original 2D image, based on this detection of points of interest, and the proposed algorithm gives as output the depth image with the most similar point.

in this common representation. It can be partial [17] or dense matching based on local or global characteristics [38]. In [5], silhouettes extracted from synthetic images are matched to those extracted from the color images. However, this method does not have the capacity to account for most of the occluding contours useful for accurate pose estimation. In turn, in [32], Image Gradients are matched with their 3D representation, but Image Gradients are still affected by image textures and background. A key requirement for these features, as in classic 2D matching, is that the computation should be performed with a high degree of *repeatability*. Here, similar to the definition in [42], the repeatability of a feature is defined as the frequency at which an element detected in the depth image is also detected within ϵ pixels around the same location in the corresponding intensity image (if it is supposed that the features are not moving or are following a slight displacement). Subsequently, by supposing that an individual photograph of an object of interest is acquired in a textured environment, we will focus on comparing pre-processed features of color images with features extracted in a depth image set; see Fig. 1.

More precisely, the 3D object will be given by a set of 3D depth surfaces, which describe how the object surface is shortened by a perspective viewing, and the image is given by the 3D intensity surface. Since the depth and the intensity surfaces have a different order of representation, the two surfaces cannot be directly matched. Thus, bringing both rendered depth images and photographs into a common representation, such as gradient and edge representation, allows for the establishment of robust sparse 2D-to-3D matching [32]. The extraction of gradient-based features corresponding to the object's shapes in both depth and intensity images regardless of illumination and texture changes is proposed. In other words, as 2D intensity images are affected by background, textures and lighting changes, these difficulties are taken into account by reducing the influence of non-redundant information (i.e., colour and texture) on the features extracted from photographs. This means that the features in depth images that highlight the object's geometric characteristics are extracted. For photographs, a refinement step is needed, which consists of selecting salient points acquired by a camera in focus. These points depend on the degree of blurring in an image. Thus, the detected points are analyzed based on measuring the blurring volume of every feature point. Finally, what we call focus points should be able to detect the approximate shape and to discard the other components, such as textures.

To summarize, the contributions of this paper are as follows:

1) *A ridge and valley detector for depth images rendered from the 3D model.* We have called this Curvilinear Saliency (CS), as it is related to the curvature estimation. This representation directly relates to the discontinuities of the object's geometry, and the extracted features should be robust in the face of texture and light changes.
2) *A variant of this detector adapted to photographs.* This Curvilinear Saliency detector is applied at multiple scales by searching over all scales and all image locations to identify scale-invariant interest points. To reduce the influence of structures due to texture and background regions, the extraction of focus Curvilinear Saliency features is introduced. This corresponds to ridges unaffected by blurring.
3) *A registration algorithm for determining the correct view of the 3D model and thus the pose.* We have introduced the Histogram of Curvilinear Saliency (HCS), which is a descriptor computed similarly to the Histogram of Gradients (HOG) proposed in [9]. The HCS descriptor is computed on both depth images (i.e., curvilinear features extracted with Curvilinear Saliency detection) and photographs (i.e., curvilinear features in focus extracted with Multi-Scale Curvilinear Saliency detection), and it combines the Curvilinear Saliency value with the curvature orientation. The repeatability score measures

the set of repeatable points detected both in a photograph and in the rendered depth images.

After presenting the related work and notations, § II and III, the 3D model representation is introduced, § IV, followed by the image representation, § V. In addition, we describe how robustness to background and texture can be achieved by using the same principle as focus curve detection, § VI. The results obtained illustrate how this new global approach for 2D/3D matching allows for obtaining more repeatable features, compared to the state of the art, § VII. Finally, we explain how 2D/3D registration is estimated, § VIII, and how pose estimation is computed, § IX before the conclusions, § X.

II. Related Work

As mentioned earlier, a typical 2D/3D registration problem consists of two subproblems: feature correspondence and pose estimation. Thus, the related work is divided into three parts related to these subproblems: (1) Detection of features in 2D photography, (2) detection of features in a 3D model and (3) matching of 2D/3D features to estimate the 3D pose. We briefly present classical detectors in 2D and 3D and highlight that the associated points of interest, in 2D and 3D, are not comparable, i.e., cannot be directly matched.

In 2D, edge detection [6] based on the first-order derivative information is the initial technique. It can detect any kind of edge, even low contrasted edges, not due to the structure but more to texture. The second technique is to detect the points of interest [40] by, for example, analyzing the eigenvalues of the structure tensor [16]. Complementary to these methods, blob detection [46] provides a description of image structures in terms of regions. More recently, multi-scale approaches have been introduced, such as a generalization of Harris or Laplacian detectors [28] or the well-known approach of SIFT, scale-invariant feature transform [26]. In [3], SURF, speeded up robust features, a detector that is also based on Hessian matrix analysis, is introduced to be faster than these multi-scale techniques. All these techniques are robust to light changes, rotations and translations and consequently are invariant to viewpoint changes. However, they totally rely on texture and/or intensity changes. Curvature detection is one of the most important techniques of second-order derivative-based approaches. Recently, [13] has proposed a detector based on curvature κ, expressed as the change of the Image Gradient along the tangent to obtain a scalar q approximating κ. In addition, [11] presented PCBR, principal curvature-based regions, the detector using the maximum or minimum eigenvalue of the Hessian matrix in a multi-scale space.

Feature extraction of 3D models/scenes can be classified into point-based and image-based approaches. Most of the point-based methods use SIFT in 3D by proposing an adaptation of the initial SIFT [37]. In image-based approaches, the 3D model is first rendered to form images or geometric buffers. Image processing methods are then applied such as edge [23] or SIFT detection [24]. The apparent ridges, AR [19], are a set of curves with points that are local maxima on a surface; a view-dependent curvature corresponds to the variation of the surface normal with respect to a viewing screen plane. Average Shading Gradients, ASG, was proposed in [32]. This rendering technique is based on averaging gradients over all lighting directions to cope with the unknown lighting conditions.

In computer vision research, the problem of automatically aligning 2D photographs with an existing 3D model of the scene has been investigated over the past fifteen years. It can be approached through indirect and direct methods [30]. For indirect registration, these methods are implemented either by 3D-to-3D registration or by finding some appropriate registration parameters, such as the standard iterative closest point, ICP, algorithm [4]. For direct registration, in [37], correspondences are obtained by matching SIFT feature descriptors between SIFT points extracted both in 2D and 3D. However, establishing reliable correspondences may be difficult due to the fact that the set of points in 2D and 3D are not always similar. This is particularly due to the variability of the illumination conditions during the acquisitions. Methods relying on higher-level features, such as lines [48], planes [44] and building bounding boxes [25], are generally suitable only for Manhattan world scenes. Similarly, skyline-based methods [35] as well as methods relying on a predefined 3D model [8] are of limited applicability. Recently, the Histogram of Gradients, HOG, detector [2] or a fast version of HOG [7] have been used to extract the features from rendering views and real images. Finally, in [32], 3D corner points are detected using the 3D Harris detector and the rendering Average Shading Gradient images on each point. For a query image, similarly, corner points are detected on multiple scales. As the next step, the gradient computed for patches around each point is matched with the database containing Average Shading Gradient images using the HOG descriptor. This method still relies on extracting gradients of photographs affected by textures and background. Consequently, they propose a refining stage based on RANdom SAmple Consensus, RANSAC [14], to improve the pose estimation. All these approaches yield interesting results, but they do not evaluate the repeatability between the set of points detected in an intensity image and those detected in an image rendered from the 3D model.

In this paper, structural cues (e.g., curvilinear shapes) based on Curvilinear Saliency are extracted instead of only considering silhouettes since they are more robust to intensity, color, and pose variations. In fact, they have the advantage of both representing outer and inner (self-occluding) contours, which also characterize the object and are useful for estimating the pose. To merge in the same descriptor Curvilinear Saliency values and curvature orientation, the HOG descriptor, which is widely used in research and correctly describes the object shape, is employed.

III. Notations

In the rest of the paper, we use the following notations:

- x, y, Z, f, I: scalars (and scalar-valued functions), including Cartesian coordinates, are simply denoted with letters without special formatting.
- $\tilde{x}, \tilde{y}$: if needed, local Cartesian coordinates are distinguished by adding a tilde over the symbol.

- $\mathbf{P}$, $\mathbf{M}$, and $\mathbf{x}$: vectors (and vector-valued functions) are denoted by bold letters.
- $\mathtt{J}$: matrices (and matrix-valued functions) are denoted by typewriter-style letters.
- $\mathcal{S}$: regular surfaces are denoted by calligraphic mode.

We also use these special notations:

- ∇_f: the gradient vector of a scalar-valued function f.
- $\mathbf{F}_x$: the partial first-order derivative $\frac{\partial \mathbf{F}}{\partial x}$ of a vector-valued function $\mathbf{F}$ w.r.t. variable x.
- Similarly, $\mathbf{F}_{xy}$: the partial second-order derivatives $\frac{\partial^2 \mathbf{F}}{\partial x \partial y}$ of $\mathbf{F}$ w.r.t. variables x and y.

For the two last notations, if the vector function is 1D, then the scalar rule is applied.

IV. 3D Model Representation

The work most related to that proposed in this paper is the Average Shading Gradient (ASG) approach, which was proposed in [32]. After introducing how the object surface can be represented, the differences between these two approaches are highlighted. For interested readers, a reminder on differential geometry is given in the appendix *Reminder on Differential Geometry* as supplementary material.

Object Surface Representation: Let $\mathcal{M}$ be the object surface parameterized by $\mathbf{M}(\mathbf{x}) \triangleq [X(\mathbf{x}), Y(\mathbf{x}), Z(\mathbf{x})]^\top$, where $\mathbf{x} = [x, y]^\top$ varies within the restricted image domain of a given camera delimited by the occluding contour of the object. We assume that $\mathcal{M}$ is such that all of its points $\mathbf{M}(\mathbf{x})$, as seen from the camera viewpoint, are in one-to-one perspective correspondence with the image point $\mathbf{x} = [x, y]^\top$, such that $x = X(\mathbf{x})/Z(\mathbf{x})$ and $y = Y(\mathbf{x})/Z(\mathbf{x})$. As a result, we obtain

$$\mathbf{M}(\mathbf{x}) = Z(\mathbf{x})[\mathbf{x}^\top, 1]^\top \tag{1}$$

Let $\mathbf{n}$ be the unit normal of $\mathcal{M}$ at $\mathbf{P} = \mathbf{M}(\mathbf{x})$. The Gaussian map $\mathbf{N} : \mathcal{M} \to \Sigma$ of $\mathcal{M}$ at $\mathbf{P}$ is the map that assigns to $\mathbf{P}$ the vector $\mathbf{N}(\mathbf{P}) = \pm\mathbf{n}$ on the unit sphere Σ such that $\mathbf{N}$ is differentiable. Using the notation $\bar{\mathbf{N}}(\mathbf{P}) = \mathbf{M}_x(\mathbf{x}) \times \mathbf{M}_y(\mathbf{x})$ for $\mathbf{x} = \mathbf{M}^{-1}(\mathbf{P})$, it can be computed as $\mathbf{N}(\mathbf{P}) = \frac{\bar{\mathbf{N}}(\mathbf{P})}{\|\bar{\mathbf{N}}(\mathbf{P})\|}$, where

$$\bar{\mathbf{N}} = \mathbf{M}_x \times \mathbf{M}_y = Z\left[-Z_x, -Z_y, xZ_x + yZ_y + Z\right]^\top \tag{2}$$

It can be shown that the Jacobian 3×2 matrix of $\mathbf{N}$ is written as

$$\mathtt{J}_{\mathbf{N}} = \left[\mathbf{N}_x\ \mathbf{N}_y\right] = \left(\mathtt{I} - \mathbf{N}\mathbf{N}^\top\right)\mathtt{J}_{\bar{\mathbf{N}}} \tag{3}$$

where the columns of $\mathtt{J}_{\bar{\mathbf{N}}} = \left[\bar{\mathbf{N}}_x\ \bar{\mathbf{N}}_y\right]$ have the form

$$\bar{\mathbf{N}}_\star = \begin{bmatrix} Z_x Z_\star - Z_{x\star} Z \\ Z_\star Z_y - Z_{\star y} Z \\ xZ_{x\star}Z + yZ_{\star y}Z + Z_\star(xZ_x + yZ_y + 3Z) \end{bmatrix} \tag{4}$$

and $\star$ represents either x or y.

A. Average Shading Gradient (ASG) Feature [32]

Plötz *et al.* assumed that the image intensity function obeys the Lambertian shading function for parallel light source $\mathbf{s}$:

$$I(x, y) \propto \max(0, -\mathbf{N}(x, y) \cdot \mathbf{s}) \text{ with } \mathbf{s} \in \mathbb{R}^3 \tag{5}$$

Eq. (5) means that the reflectance describing the object material is assumed to be Lambertian with constant albedo.[1] In addition, the image background is assumed to be constant (e.g., as on a plane). The authors propose the magnitude of the gradient of the shading function as a feature in the intensity image. To register the intensity image to the 3D (untextured) model, the idea is to generate virtual images when viewing the object from different camera pose candidates. Nevertheless, it is clearly impossible to render any such virtual image obeying the shading function (5) without prior information about the lighting direction and therefore about $\mathbf{s}$. Thus, the authors propose to replace the gradient magnitude feature in the virtual images by a feature corresponding to the average value of the gradient magnitude computed over all light directions, which is the so-called *Average Shading Gradient* magnitude. Denoting $\|\nabla_I\|$ as the magnitude of the gradient of the shading function (5), the magnitude of the Average Shading Gradient is then

$$\overline{\|\nabla_I\|} = \int_{\mathcal{S}} \|\nabla_I\| \, \mathrm{d}\mathbf{s} \tag{6}$$

with $\|\nabla_I\|^2 = I_x^2 + I_y^2$ and where the vector $\mathbf{s}$, cf. (5), varies over the unit sphere $\mathcal{S}$ in $\mathbb{R}^3$, and $\mathrm{d}\mathbf{s}$ is the volume element. The nice contribution of Plötz *et al.* is, by applying Jensen's inequality, to derive the following closed-form bound on $\overline{\|\nabla_I\|}$

$$\begin{aligned} \overline{\|\nabla_I\|} &\leq \sqrt{\int_{\mathcal{S}} \|\nabla_I\|^2 \, \mathrm{d}\mathbf{s}} \\ &= \gamma \sqrt{\left(\|\mathbf{N}_x\|^2 + \|\mathbf{N}_y\|^2\right)} \text{ with } \gamma = \sqrt{\frac{\pi}{3}} \end{aligned} \tag{7}$$

It is reported by the authors to behave like a very good approximation of $\overline{\|\nabla_I\|}$. This is the elegant way that the authors do away with the unknown lighting direction $\mathbf{s}$.

B. Proposed Curvilinear Saliency Features (CS)

As already mentioned, our goal is to find a common representation between the 3D model and the 2D image to match them. For that purpose, we first show how the 3D model can be represented from different points of view and how these different viewpoints can be compared to a 2D image. The observed 3D object is represented by a set of synthetic depth maps generated from camera locations distributed on concentric spheres encapsulating it, by sampling elevation and azimuth angles, as well as distances from the camera to the object. A depth map $Z(x, y)$ associates to every image point (x, y) the Z-coordinate, w.r.t. the camera frame, of the object 3D point (1) that projects at image location (x, y). Let $\mathcal{D}$ denote the depth surface that is the 3D surface with graph parameterization is[2] $\mathbf{D}(x, y) = [x, y, Z(x, y)]^\top$. It is worth noting that any two depth surfaces (from two different views) are *not* equal to some Euclidean transformation.

Which features should be extracted in the depth map? We aim at detecting depth discontinuities by searching points on

[1] A general shading function is $I(x, y) = \rho(\mathbf{M}(x, y)) \max(0, -\mathbf{N}(x, y) \cdot \mathbf{s})$, where $\rho(\mathbf{M}(x, y))$ is the albedo at the object point $\mathbf{M}(x, y)$.

[2] Note the difference with (1).

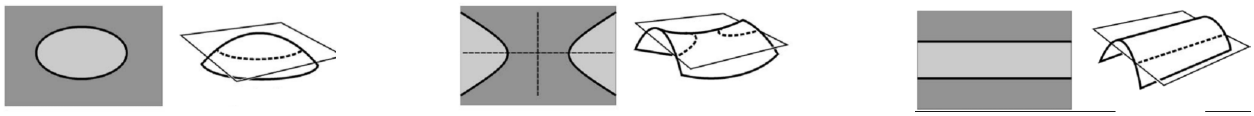

Fig. 2. The real conics of the Dupin indicatrix.

$\mathcal{D}$ having high principal curvature in one direction and low principal curvature in the orthogonal direction. We denote this as the *curviness saliency* features of surface loci of such points that correspond to the *ridges* and *valleys* of this surface. Here, the difference of the principal curvatures $\kappa_1 - \kappa_2$ is used to describe the ridges and valleys, and we explain why.

1) Principal Curvatures and Directions: Consider a point $\mathbf{P} = \mathbf{D}(x, y)$. Let $\mathbf{N}'(x, y)$ denote the *Gaussian map* of $\mathcal{D}$, assigning to $\mathbf{P}$ the unit normal of $\mathcal{D}$:

$$\mathbf{N}' = \frac{\bar{\mathbf{N}}'}{\left\|\bar{\mathbf{N}}'\right\|} \text{ where } \bar{\mathbf{N}}' = \mathbf{D}_x \times \mathbf{D}_y = \alpha \begin{bmatrix} -\nabla_Z \\ 1 \end{bmatrix} \tag{8}$$

with $\nabla_Z = [Z_x, Z_y]^\top$ and $\alpha = 1/\sqrt{1 + \|\nabla_Z\|^2}$.

As the two columns of the Jacobian matrix $\mathtt{J}_\mathbf{D}$ of $\mathcal{D}$ are $\mathbf{D}_x = [1, 0, Z_x]^\top$ and $\mathbf{D}_y = [0, 1, Z_y]^\top$, the *first fundamental form of* $\mathcal{D}$ can be computed as

$$\mathtt{I}_\mathbf{P} = \mathsf{I}_3 + \nabla_Z \nabla_Z^\top$$

and the *second fundamental form of* $\mathcal{D}$ can be computed as

$$\mathtt{II}_\mathbf{P} = \alpha \mathtt{H}_Z$$

where $\mathtt{H}_Z$ is the Hessian matrix of Z, i.e., with the second-order partial derivatives of Z w.r.t. x and y as elements.

The *principal curvatures* of $\mathcal{D}$ at $\mathbf{P}$ coincide with the eigenvalues κ_α $(\alpha = 1, 2)$ of $\mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P}$, which are always real. In the tangent plane $T_\mathbf{P}(\mathcal{D})$, the local coordinates of the *principal directions* of $\mathcal{D}$ at $\mathbf{P}$ are given by the eigenvectors $\mathbf{e}_\alpha$ of $\mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P}$, so the 3D principal directions in 3D are written as $\mathtt{J}_\mathbf{D}\mathbf{e}_\alpha$. As Koenderink wrote in [22], "*it is perhaps not superfluous to remark here that the simple (eigen-)interpretation in terms*[3] *of* $\mathtt{II}_\mathbf{P} = \alpha \mathtt{H}_Z$ *is only valid in representations where* $\nabla_Z = \mathbf{0}$", which is the condition for the point to be a local extremum.

Thanks to proposition 1, presented on page 3 of the supplementary material, we know that that the principal curvature κ_α at $\mathbf{P}$ associated to the principal 3D direction $\mathbf{T}_\alpha = \mathtt{J}_\mathbf{D}\mathbf{e}_\alpha$ is equal to the absolute magnitude of the change in the normal

$$|\kappa_\alpha| = \left\|\mathrm{d}\mathbf{N}'_\mathbf{P}(\mathbf{T}_\alpha)\right\| \tag{9}$$

where $\mathrm{d}\mathbf{N}'_\mathbf{P}(\mathbf{T})$ denotes the differential of $\mathbf{N}'$ at $\mathbf{P}$ in direction $\mathbf{T}$. We will make use of this result for the image representation, cf. §V. Now, let us explain the difference $\kappa_1 - \kappa_2$, where $\kappa_1 \geq \kappa_2$ is proposed as a feature.

2) Curvilinear Feature: Without loss of generality, let κ_1 and κ_2 be the principal curvatures computed as ordered eigenvalues of $\mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P}$ so that $\kappa_1 \geq \kappa_2$. We aim at detecting points lying on "elongated" surface parts. In this work, we detect points at which this difference is high:

$$CS(x, y) = \kappa_1(x, y) - \kappa_2(x, y) \tag{10}$$

We call (10) the *Curvilinear Saliency (CS)* feature. Curvilinear means a feature that belongs to a curved line. The rest of this paragraph justifies such a choice.

Given a point $\mathbf{P}$ on $\mathcal{D}$, let $(\tilde{x}, \tilde{y})$ be the Cartesian coordinates on the tangent plane $T_\mathbf{P}(\mathcal{D})$) w.r.t. the 2D frame whose origin is $\mathbf{P}$, and the orthonormal basis is formed by the principal directions $\{\mathbf{e}_1, \mathbf{e}_2\}$. As a result, $\mathcal{D}$ can now locally be associated to the new parameterization $\mathbf{F}(\tilde{x}, \tilde{y}) = \left[\tilde{x}, \tilde{y}, F(\tilde{x}, \tilde{y})\right]^\top$, for some height function F. In that case, it can be readily seen that $\mathtt{I}_\mathbf{P}$ is the identity matrix, and so, $\mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P} = \mathtt{II}_\mathbf{P} = \mathrm{diag}(\kappa_1, \kappa_2)$ is exactly the Hessian matrix of F. For some sufficiently small $\epsilon > 0$, consider, on the two planes parallel to $T_\mathbf{P}(\mathcal{D})$ at distances $\pm\epsilon$ from $T_\mathbf{P}(\mathcal{D})$, the curves $\mathscr{C}_\pm = \{(\tilde{x}, \tilde{y}), \mathbf{F}(\tilde{x}, \tilde{y}) \in T_\mathbf{P}(\mathcal{D}) \mid F(\tilde{x}, \tilde{y}) = \pm\epsilon\}$. It can be shown [15, p500] that the first-order approximation of the intersections of $\mathcal{D}$ with the two parallel planes is the union of two conics (one real and one virtual) with equations $\mathtt{II}_\mathbf{P}(\tilde{x}, \tilde{y}) = \pm 2\epsilon$. This union is known as the *Dupin indicatrix* when written in canonical form (i.e., by replacing 2ϵ by 1).

The real Dupin conic characterizes the local shape of $\mathcal{D}$ and provides local information on the first-order geometry of the surface, at least at points where the conic is non-degenerate. It specializes as a parabola if the Gauss curvature vanishes, i.e., $\kappa_1\kappa_2 = 0$, to an ellipse if $\kappa_1\kappa_2 > 0$ and to a hyperbola if $\kappa_1\kappa_2 < 0$; see Fig. 2. Points are said to be elliptical, hyperbolic or parabolic; more details are given in the appendix *Analysis of the Dupin central conics* of the supplemental materials. The Curvilinear Saliency CS is significant when $\kappa_1 \gg \kappa_2$, which is in the presence of distant foci and therefore a highly elongated ellipse or a "squashed" hyperbola; see Fig. 2. This occurs, for example, when the point is located on a depth "discontinuity". In turn, when $\kappa_1 \simeq \kappa_2$, the conic approaches a circle, and the distance between foci becomes very small.

3) A Simple Way to Compute the Curvilinear Feature: After algebraic manipulations, it can be shown that $\mathtt{I}_\mathbf{P}^{-1}\mathtt{II}_\mathbf{P} = \frac{1}{\alpha}\mathtt{M}$ where

$$\mathtt{M} \triangleq \begin{bmatrix} (Z_y^2 + 1)\ Z_{xx} - Z_x Z_y Z_{xy} & (Z_y^2 + 1) Z_{xy} - Z_x Z_y Z_{yy} \\ (Z_x^2 + 1)\ Z_{xy} - Z_x Z_y Z_{xx} & (Z_x^2 + 1) Z_{yy} - Z_x Z_y Z_{xy} \end{bmatrix}$$

Proposition 1: The squared curviness feature can be computed as

$$\mathrm{CS}^2 \triangleq \|\nabla_Z\|^2 \left((\mathrm{trace}\,\mathtt{M})^2 - 4 \det \mathtt{M}\right) \tag{11}$$

$$= 4 \|\nabla_Z\|^2 (\bar{\kappa}^2 - K) \tag{12}$$

where $\bar{\kappa}$ is the mean curvature of $\mathcal{D}$, and K is its Gaussian curvature (a proof is available in the appendix).

[3] i.e., by neglecting $\mathtt{I}_\mathbf{P}$.

Fig. 3. Curvilinear saliency of two shapes (columns 1, 5) with minimum (2, 6), maximum (3, 7) and the difference between maximum and minimum eigenvalues (4, 8).

The reliance on the highest or smallest principal curvature alone is not adequate for defining accurate ridges [36]. In Fig. 3, we show the different detections obtained using the minimum or the maximum principal curvature. The maximum provides a high response only for dark lines on a light background, while the minimum gives the higher answers for the light lines on a dark background. The difference in the principal curvatures, $\kappa_1 - \kappa_2$, improves robustness as it responds in both settings.

V. Image Representation

We recall these notations:

- $I(x, y)$ denotes the value of the image intensity function $I : U \subset \mathbb{R}^2 \to \mathbb{R}$ at image point (x, y).
- The intensity image can also be treated as an intensity surface $\mathcal{I}$ defined by the vector function

$$\mathbf{I}(x, y) = [x, y, I(x, y)]^\top \tag{13}$$

Proposed Curvilinear Features for Images: Similar to the work of [32], the Lambertian shading model (5) is assumed, i.e., $I(x, y) \propto \max(0, -\mathbf{N}(x, y) \cdot \mathbf{S})$. Recall that the unit normal is $\mathbf{N}(x, y) = \tilde{\mathbf{N}}(x, y)/\left\|\tilde{\mathbf{N}}(x, y)\right\|$ where $\tilde{\mathbf{N}}$ is defined in (2) and so only depends on the depth $Z(x, y)$ and its derivatives up to order-1.

We would like to detect features in the intensity surface $\mathcal{I}$ and check whether they are good candidates to be matched to detected curvilinear features in the depth surface $\mathcal{D}$ w.r.t. a given camera pose. The key issue here is that detected features in $\mathcal{I}$ can be matched to detected features in $\mathcal{D}$ on the condition that both are based on measurements with the same order of derivation in $Z(x, y)$ to yield a "compatible" matching that ensures repeatability. The fact that I depends on $Z(x, y)$ and its derivatives up to order-1 entails that the detection of features in $\mathcal{I}$ must rely on order-1 variations of the surface $\mathbf{I}(x, y)$, e.g., on its differential along some adequate direction. Consider a point $\mathbf{Q} = \mathbf{I}(x, y)$ on the image surface. Let $\mathrm{d}\mathbf{I}_\mathbf{Q} : U \to \mathbb{R}^3$ be the differential of $\mathbf{I}$ at $\mathbf{Q}$. Given a unit direction $\mathbf{v} = [a, b]^\top$ in the image xy-plane, we have that $\mathrm{d}\mathbf{I}_\mathbf{Q}(\mathbf{v}) = a\mathbf{I}_x + b\mathbf{I}_y = \jmath_\mathbf{I}\mathbf{v}$ is the Jacobian matrix of $\mathbf{I}$, $\mathbf{I}_x = [1, 0, I_x]^\top$ and $\mathbf{I}_y = [0, 1, I_y]^\top$, where

$$I_\star = \frac{1}{2}\left(\mathrm{sign}\left(\mathbf{N} \cdot \mathbf{s}\right) - 1\right)\left(\mathbf{N}_\star \cdot \mathbf{s}\right) \tag{14}$$

$\star$ represents either x or y. It is an order-1 measurement of the image surface variation at $\mathbf{Q}$ and is compatible with our curvilinear measurements of the depth surface (i.e., with same order of the derivatives of Z).

To a obtain a scalar measurement, define the unit vectors $\mathbf{T}_1 = \jmath_\mathbf{I}\frac{\nabla I}{\|\nabla I\|}$ and $\mathbf{T}_2$ by rotating $\mathbf{T}_1$ by $\frac{\pi}{2}$. For $\alpha = 1, 2$, define

$$|\mu_\alpha| = \left\|\mathrm{d}\mathbf{I}_\mathbf{Q}\left(\mathbf{T}_\alpha\right)\right\| \tag{15}$$

which is the differential of $\mathbf{I}$ along unit direction $\mathbf{T}_\alpha$ in the image plane. It can be readily seen that $\nabla_I/\|\nabla_I\|$ is the eigenvector of

$$\jmath_\mathbf{I}^\top\jmath_\mathbf{I} = \begin{bmatrix}\mathbf{I}_x \cdot \mathbf{I}_x & \mathbf{I}_x \cdot \mathbf{I}_y \\ \mathbf{I}_x \cdot \mathbf{I}_y & \mathbf{I}_y \cdot \mathbf{I}_y\end{bmatrix} = \begin{bmatrix}1 + (I_x)^2 & I_x I_y \\ I_x I_y & 1 + (I_y)^2\end{bmatrix} = \mathtt{I} + \nabla_I\nabla_I^\top \tag{16}$$

associated with the largest eigenvalue μ_α. The similarity between the expression of the principal curvature computed for the depth surface is noteworthy, cf. (15). In addition, note that the matrix (16) is that of the first fundamental form of $\mathcal{I}$. Clearly,[4] the maximum and minimum values of the quadratic form $\left\|\mathrm{d}\mathbf{I}_\mathbf{Q}(\mathbf{v})\right\|^2$ correspond the two eigenvalues of the first fundamental form matrix given in (16). By a similar approach to §IV-B, we can propose a feature $\mu_1 - \mu_2$, where $\mu_1 \geq \mu_2$.

Proposition 2: Let μ_1, μ_2 be the two eigenvalues of the first fundamental form matrix $\jmath_\mathbf{I}^\top\jmath_\mathbf{I}$ of $\mathcal{I}$, in descending order. Then, we have

$$\mu_1 - \mu_2 = \|\nabla_I\|^2 \tag{17}$$

Proof: The ordered eigenvalues of $\mathtt{I}_\mathbf{P}^\mathcal{I} = \jmath_\mathbf{I}^\top\jmath_\mathbf{I}$ can be deduced from those of $\nabla_I\nabla_I^\top$, i.e., $\|\nabla_I\|^2$ and 0, so $\mu_1 = \|\nabla_I\|^2 + 1$ and $\mu_2 = 1$. This concludes the proof stage. ■

The local shape of $\mathcal{I}$ at $\mathbf{Q}$ can be described by means of the eccentricity of a conic, here given by the quadratic form $\mathbf{v}^\top\jmath_\mathbf{I}^\top\jmath_\mathbf{I}\mathbf{v} = \pm 1$. How can this conic be interpreted? The first-order Taylor expansion for infinitesimal changes $(\mathrm{d}x, \mathrm{d}y)$ in the vicinity of $\mathbf{Q} = \mathbf{I}(x, y)$ yields

$$\mathbf{I}(x + \mathrm{d}x, y + \mathrm{d}y) - \mathbf{I}(x, y) \approx \jmath_\mathbf{I}[\mathrm{d}x, \mathrm{d}y]^\top \tag{18}$$

For any unit direction $\mathbf{v} = [a, b]^\top$ in the xy-plane, the quadratic form $\mathbf{v}^\top\jmath_\mathbf{I}^\top\jmath_\mathbf{I}\mathbf{v}$ returns the linear part g of growth in arc length from $\mathbf{I}(x, y)$ to $\mathbf{I}(x + a, y + b)$. In addition,

$$g^2 = \left\|\mathrm{d}\mathbf{I}_\mathbf{Q}((\mathrm{d}x, \mathrm{d}y)\right\|^2 = \mathbf{v}^\top\jmath_\mathbf{I}^\top\jmath_\mathbf{I}\mathbf{v} \tag{19}$$

The following is an important remark that we highlight here and is not mentioned in [32]. The AVG feature defined in (7) is actually the Frobenius norm of the Jacobian matrix $\jmath_\mathbf{N}$ of the map $\mathbf{N}(x, y)$, see (3), up to constant γ. Clearly, this describes the second-order behavior of the surface $\mathcal{M}$ relative to the normal at one of its points in the immediate vicinity of

[4] If $\mathbf{A}^\top\mathbf{A}$ is full rank, then the maximum (resp. minimum) of $\|\mathbf{A}\mathbf{x}\|^2 = \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}$ under the constraints $\|\mathbf{x}\|^2 = 1$ is given by the largest (resp. smallest) eigenvalue of $\mathbf{A}^\top\mathbf{A}$. Here, $\mathbf{A} = \jmath_\mathbf{I}^\top\jmath_\mathbf{I}$ is 2×2 and generally full rank.

this point. Using the results in (2), (3) and (4), we can claim that the extracted feature in the virtual image only depends on X, Y, Z and their derivatives up to order-2. This is consistent (regarding the considered orders of the derivatives of X, Y, Z) with the feature $\|\nabla_I\| = \sqrt{I_x^2 + I_y^2}$ detected in the intensity image, where $I_\star$, with $\star \in \{1, 2\}$, is given in (14).

We have presented how some information relative to Curvilinear Saliency can be extracted, both in 3D and 2D. In the next section, this Curvilinear Saliency measurement is improved in 2D to be robust to texture and to background.

VI. Robustness to Texture and Background

A. Multi-Curvilinear Saliency (MCS)

Contrary to depth images that represent textureless 3D shapes, intensity images are composed of shape and texture components. Consequently, the Curvilinear Saliency (CS) estimated from intensity images is affected by the textured regions. Our idea is to put forward the assumption that multi-scale analysis can discriminate between key points (those with high CS value in the image) due to shape and key points due to texture.[5] At a coarse level, edges detected are reliable but with a poor localization, and they miss small details. At a fine level, details are preserved, but detection suffers greatly from clutter in textured regions. In addition, the CS values of small details and textures are high at the coarse level, whereas these values decrease in the finest levels. To combine the strengths of each scale, the CS value of each pixel over n scales is analyzed. If this value at all scales is higher than a threshold T, the maximum Curvilinear Saliency (MCS) value of this pixel over all scales is then kept. This threshold is a function of the number of the smoothed images, n, (i.e., $T = e^{-n}$: when n is small, then T is high, and vice versa). However, if the CS value is lower than T in one level, it is considered a point that belongs to a texture (or a small detail) point; thus, it is removed from the final Multi-Scale Curvilinear Saliency image. Adding this multi-scale step should help reduce the impact of the texture; however, in the next section, we propose introducing the principle used for estimating focus maps to increase the robustness to the background and to the presence of the texture. Before introducing the proposed improvement, we briefly present existing works concerning texture detection and, in particular, those concerning focus curve estimation.

B. Extraction of Texture: State of the Art

Various methods, such as [20], [34], have been proposed for extracting the texture from a natural image. In these approaches, the image is separated into two components while preserving edges by first smoothing the intensity image, as a pre-processing stage, and then extracting the shape/the structure from that image relying on prior knowledge. These methods are analogous to the classical signal processing low pass-high pass filter decomposition. However, even if it is correct to consider that the structure part of an image contains strong edges, the texture can also contain medium and high frequencies. Another possibility is to consider focusness.

Usually, focusness, which is related to the degree of focus, is defined as being inversely proportional to the degree of blur (blurriness) [18]. It is a valuable tool for depth recovery [50] and also for blur magnification or image quality assessment. Blurring is usually measured in regions containing edges since edges would appear in images as blurred luminance transitions of unknown blurring scale [12]. Then, the estimation of the blur can be propagated to the rest of the image. Since blur occurs for many different reasons, this task is challenging, and in research, many methods have been proposed [43]. Interested readers can find details about techniques that take into account penumbra blur or shading blur [50], in particular, with multiple scales [18].

Finally, most of the existing algorithms [18], [50] depend on measuring the blur amount using the ratio between the edges at two different scale levels (i.e., the original image and the re-blurred image). Consequently, we propose using the ratio between the two Curvilinear Saliency images that contain robust edges at different scales to determine the blur amount based on the methods developed in [50]. Concerning the multi-scale aspect, our approach is inspired by the principles explained in [18].

C. Removing Background With Focus Curves: State of the Art

Based on the mapping between the depth of a point light source and the focus level of its image, shape from defocus (SFD) approaches recover the 3D shape of a scene from focused images that represent the focus level of each point in the scene [31]. Consequently, it seems interesting to introduce what is called the detection of "focus curves". More precisely, these curves mean that the scale of blurring is estimated at the Curvilinear Saliency feature of the 2D image and that these features are supposedly related to discontinuities.

Focal blurring occurs when a point is out of focus. When the point is at the focus distance d_f from the lens, all the rays from it converge to a sharp single sensor point. Otherwise, when $d \neq d_f$, these rays generate a blurred region in the sensor area. The blurred pattern generated in this way is called the circle of confusion (CoC), the diameter of which is denoted c.

In [18], [50], the defocus blur can be modelled as a convolution of a sharp image with the point spread function (PSF). The PSF is usually approximated by a Gaussian function $g(x, \sigma)$, where the standard deviation $\sigma \propto c$ measures the blurring amount and is proportional to the diameter of the CoC:

$$c = \frac{|d - d_f|}{d} \frac{f}{d - f},$$

where d, d_f, f are the focus distance, defocus distance and focal length, respectively. A blurred edge $i(x)$ is then given by

$$i(x) = f(x) \otimes g(x, \sigma) \tag{20}$$

where $f(x) = Au(x) + B$ is an ideal edge, and $u(x)$ is the step function. The terms A and B correspond to the amplitude

[5] To build the scale pyramid, an edge-preserving smoothing approach, denoted as an anisotropic diffusion filter [29], is used. It tries to separate the low-frequency components (i.e., sharp edges) from the high-frequency components (i.e., textures) by preserving the largest edges in an image.

and the offset of the edge, respectively. Note that the edge is located at $x = 0$. In [50], the blur estimation method was described for a 1D case. The gradient of the re-blurred edge is

$$\begin{aligned}\nabla i_1(x) &= \nabla(i(x) \otimes g(x, \sigma_0)) \\ &= \nabla((Au(x) + B) \otimes g(x, \sigma) \otimes g(x, \sigma_0)) \\ &= \frac{A}{\sqrt{2\pi(\sigma + \sigma_0)}} \exp\left(-\frac{x^2}{2(\sigma^2 + \sigma_0^2)}\right) \end{aligned} \quad (21)$$

where σ_0 is the standard deviation of the re-blur Gaussian kernel. Thus, the gradient magnitude ratio between the original and the re-blurred edges is

$$\begin{aligned} R &= \frac{|\nabla i(x)|}{|\nabla i_1(x)|} \\ &= \sqrt{\frac{\sigma^2 + \sigma_0^2}{\sigma^2}} \exp - \left(\frac{x^2}{2(\sigma^2)} - \frac{x^2}{2(\sigma^2 + \sigma_0^2)}\right). \end{aligned} \quad (22)$$

It can be proven that the ratio is a maximum at the edge location ($x = 0$), and the maximum value is given by

$$R = \sqrt{\frac{\sigma^2 + \sigma_0^2}{\sigma^2}} \quad (23)$$

Finally, given the maximum value R at the edge locations, the unknown blurring amount σ can be estimated using

$$\sigma = \frac{\sigma_0}{\sqrt{R^2 - 1}} \quad (24)$$

D. Multi-Focus Curves (MFC) Based on Curvilinear Saliency

We propose using the Curvilinear Saliency computation instead of the edge response to estimate the focus curves of an input image. In addition, focus curves are estimated at multiple scales rather than at one scale as proposed in [50]. All the information obtained from different blurring scales is combined. In consequence, the Curvilinear Saliency is given by

$$CS = \alpha((I_x^2 + I_y^2)) \otimes g(x, y, \sigma) \quad (25)$$

Then, the re-blurred Curvilinear Saliency image, denoted CS_i, at multiple scales can be defined as

$$CS_i = \alpha((I_x^2 + I_y^2)) \otimes g(x, y, \sigma) \otimes g(x, y, \sigma_i) \quad (26)$$

with n being the number of scales, and $i = 1, 2, \ldots, n$. Hence, the ratio between the original and the re-blurred Curvilinear Saliency is

$$R_i = \frac{CS_i}{CS} = \frac{\sigma^2 + \sigma_i^2}{\sigma^2} \exp - \left(\frac{x^2 + y^2}{2(\sigma^2)} - \frac{x^2 + y^2}{2(\sigma^2 + \sigma_i^2)}\right)$$

Within the neighborhood of a pixel, the response reaches its maximum when $x = 0$ and $y = 0$; thus,

$$R_i \mid_0 = \frac{CS_i}{CS} = \frac{\sigma^2 + \sigma_i^2}{\sigma^2} = 1 + \frac{\sigma_i^2}{\sigma^2}$$

Finally, given the maximum value R_i at each scale level, the unknown blur amount σ_i can be estimated using

$$\sigma_i = \frac{\sigma_i}{\sqrt{R_i \mid_0 - 1}}, \quad (27)$$

For n scales, $n - 1$ focus curve scales are computed by using the ratio between the Curvilinear Saliency of the coarse level (i.e., the original image) and the next scale levels. By following the same remarks as in section VI-A, we define Multi-Focus Curves (MFC) that correspond to the fusion of all the focus curves into one map by keeping only the pixels that have a high focus value in all the $n-1$ scales (i.e., a high value means a value larger than $T = e^{-n}$, chosen in the same way as in section VI-A). If the pixel has a high value at all scales, the maximum value of the scale of blur is taken into account to build the final multi-scale curve map:

$$MFC = \frac{1}{\arg\max_i\ (s_i)}. \quad (28)$$

In conclusion, the highest values of the estimated MFC indicate edges that have low blurring (i.e., sharp edges). On the contrary, low values indicate ones that have a high level of blurring. Consequently, we expect that focus curves highlight salient Curvilinear Saliency in images that are approximately similar to the detected Curvilinear Saliency features in depth images.

VII. Experiments for Feature Detection

A. Comparison With Existing Methods

One of our most important objectives in this work was to introduce a detector that is more repeatable between 2D images and 3D models than classical detectors. Consequently, we compare the features detected on 3D models with the proposed Curvilinear Saliency detector with features detected on real images with these three 2D detectors: Image Gradient (IG), Multi-Scale Curvilinear Saliency (MCS) and multi-scale focus curves (MFC). In addition, the repeatability is measured between the two other 3D model detectors, i.e., Average Shading Gradient (ASG) [32] and Hessian Frobenius Norm (HFN), and the same three 2D detectors. In addition, MFC and MCS are then compared with nine classical 2D detectors:

(1) **Edge detectors**: (*i*) Sobel, (*ii*) Laplacian of Gaussian (LoG), (*iii*) Canny [6] and (*iv*) Fuzzy logic technique [21];

(2) **Corner detectors**: (*v*) Harris detector based on auto-correlation analysis and (*vi*) Minimum Eigenvalues detector based on analysis of the Hessian matrix [39];

(3)**Multi-scale detectors**: (*vii*) SIFT [26], (*viii*) SURF, speeded up robust features, a multi-scale technique based on the Hessian matrix [3] and (*ix*) A multi-scale principal curvature image (PCI) detector [11].

B. Evaluation Criteria

The eleven 2D detectors are evaluated with two **measures**:

(1) Intersection percentage (IP): the probability that a 2D intensity-based key feature can be found close to those extracted in a depth image [36].

(2) Hausdorff distance (HD): the classical measurement is defined for two point sets A and B by

$$HD(A, B) = \max\left(h(A, B), h(B, A)\right),$$

TABLE I

MEAN INTERSECTION PERCENTAGE (IP) (*Higher Is Better*) OF ALL DEPTH IMAGES RENDERED FROM DIFFERENT VIEWPOINTS AND ALL REAL IMAGES CAPTURED UNDER DIFFERENT TEXTURES AND LIGHTING FOR THE *Web Collection*

Methods	MFC	MCS	PCI	[39]	Harris	SIFT	SURF	Sobel	Canny	LoG	[21]
Car	59	50	46	08	04	03	03	10	18	11	05
Shoe	38	31	31	02	03	10	01	04	04	05	02
Plane	58	55	38	06	04	10	03	18	21	21	14
T-Rex	66	64	59	09	06	02	05	16	18	20	12
Elephant	37	32	32	03	03	05	03	06	08	06	04
Fhydrant	56	51	42	06	04	02	09	09	14	13	06
Jeep	69	62	58	05	05	05	06	09	15	11	06
Mug	57	54	50	02	03	04	03	08	12	07	08
Teddy	44	39	32	04	05	09	04	07	14	08	07
Pistol	69	67	61	09	09	09	04	13	23	14	07

where $h(A, B) = \max_{a \in A} \min_{b \in B} \| a - b \|$. The lowest distance corresponds to the most similar two sets.

C. Datasets

Two datasets are evaluated:

(1) Web collection: we have collected 10 objects and 15 real images of each object on the web by choosing views as close as possible to the views used for the generation of the depth images. Moreover, to highlight the robustness of the approach to different acquisition conditions, many real images of a similar model are taken.

(2) PASCAL3D+ dataset [49]: it is used to assess scalability. It contains real images corresponding to 12 rigid objects categories. We have computed average results for all non-occluded objects in each category, i.e., approximately 1000 real images and 3 or more reference models per category. The real images are acquired under different acquisition conditions (e.g., lighting, complex background, and low contrast). We have rendered the depth images of the corresponding 3D CAD model using the viewpoint information from the dataset. Only non-occluded and non-truncated objects in the real images were used. Furthermore, we choose 3D textureless objects (available online: http://tf3dm.com/),

For all the tested 3D models, depth images have been rendered using MATLAB 3D Model Renderer: http://www.openu.ac.il/home/hassner/projects/poses/.

D. Analysis of the Results

As shown in tables I and II, and as expected, the proposed approach using focus curves based on Curvilinear Saliency, named MFC, is able to find the highest number of features in the intersection with the features detected on real images captured under different textures and lighting conditions. More precisely, MFC obtains an average mean intersection percentage greater than 56%, whereas for MCS and PCI, it is, respectively, greater than 50% and 44% for the web collection dataset. With the PASCAL+3D dataset, MFC also yields the highest mean average IP among all the tested detectors: 46%.

In addition, as shown in tables III and IV, the average Hausdorff distance (HD) with MFC is less than 35 and, with MCS, is less than 52. For all the presented results, the two proposed approaches always give the lowest HD.

TABLE II

MEAN INTERSECTION PERCENTAGE (IP) (*Higher Is Better*) OF ALL DEPTH IMAGES RENDERED FROM DIFFERENT VIEWPOINTS AND ALL REAL IMAGES CAPTURED UNDER DIFFERENT TEXTURES AND LIGHTING FOR THE *PASCAL3D+*

Methods	MFC	MCS	PCI	[39]	Harris	SIFT	SURF	Sobel	Canny	LoG	[21]
Plane	55	50	37	15	09	08	13	10	13	11	10
Bicycle	69	61	57	25	08	16	24	13	15	18	14
Boat	42	36	28	09	10	06	10	09	14	11	09
Bus	31	24	17	05	06	02	04	04	06	04	04
Car	44	41	24	08	08	03	06	16	18	14	13
Diningtable	40	38	19	06	05	04	08	11	12	11	07
Motorbike	59	53	48	07	09	06	14	18	11	14	08
Sofa	67	66	60	10	11	18	20	19	16	15	11
Train	31	28	14	06	07	03	05	08	07	04	06
Tvmonitor	44	40	37	05	06	12	11	14	17	11	10
Chair	58	54	42	18	10	09	18	20	19	21	14
Bottle	56	54	51	18	14	18	21	17	19	13	12

TABLE III

MEAN HAUSDORFF DISTANCE (HD) (*Lower Is Better*) OF ALL DEPTH IMAGES RENDERED FROM DIFFERENT VIEWPOINTS AND ALL REAL IMAGES CAPTURED UNDER DIFFERENT TEXTURES AND LIGHTING FOR THE *Web Collection*

Methods	MFC	MCS	PCI	[39]	Harris	SIFT	SURF	Sobel	Canny	LoG	[21]
Car	21	29	40	57	77	85	71	48	46	47	49
Shoe	34	52	67	102	106	111	108	71	71	71	71
Plane	26	23	19	37	43	46	47	26	26	24	24
T-Rex	20	17	25	41	100	143	46	28	28	32	22
Elephant	21	41	55	80	91	114	74	57	58	57	57
Fhydrant	15	23	35	62	86	74	67	38	37	36	42
Jeep	29	31	42	70	67	74	89	47	47	46	47
Mug	35	56	65	129	133	134	145	72	76	75	75
Teddy	19	24	31	72	69	77	101	47	44	47	47
Pistol	18	16	26	34	96	44	73	30	65	29	26

TABLE IV

MEAN HAUSDORFF DISTANCE (HD) (*Lower Is Better*) OF ALL DEPTH IMAGES RENDERED FROM DIFFERENT VIEWPOINTS AND ALL REAL IMAGES CAPTURED UNDER DIFFERENT TEXTURES AND LIGHTING FOR THE *PASCAL3D+*

Method	MFC	MCS	PCI	[39]	Harris	SIFT	SURF	Sobel	Canny	LoG	[21]
Plane	47	48	59	61	63	68	73	68	65	69	71
Bicycle	71	75	79	90	101	93	100	83	84	82	87
Boat	62	68	75	79	77	87	76	75	71	78	76
Bus	106	110	117	128	123	131	127	121	118	122	123
Car	80	85	98	102	100	113	108	89	88	94	97
Diningtable	84	85	96	117	118	118	111	117	114	116	120
Motorbike	62	64	78	84	96	94	86	88	91	86	92
Sofa	70	75	77	86	98	93	92	95	99	89	96
Train	101	108	121	126	123	133	127	125	129	129	122
Tvmonitor	96	102	104	109	104	111	105	116	112	114	106
Chair	92	105	115	119	108	107	112	98	97	112	106
Bottle	78	84	87	89	90	92	97	82	86	79	88

All these quantitative results support that MFC is able to detect Curvilinear Saliency features that are more repeatable between an intensity image and its corresponding depth image than the state of the art.

In the rest of this section, we illustrate the results for the most significant dataset, PASCAL3D+ [49]. In Fig. 4, the repeatability percentage between the three comparable 3D detectors, i.e., MFC, MSC and Image Gradient (IG), and the three comparable 2D detectors, Hessian Frobenius Norm, Average Shading Gradient and CS, is presented. These results highlight that Image Gradients are affected by texture. Moreover, MCS improves the repeatability between depth and

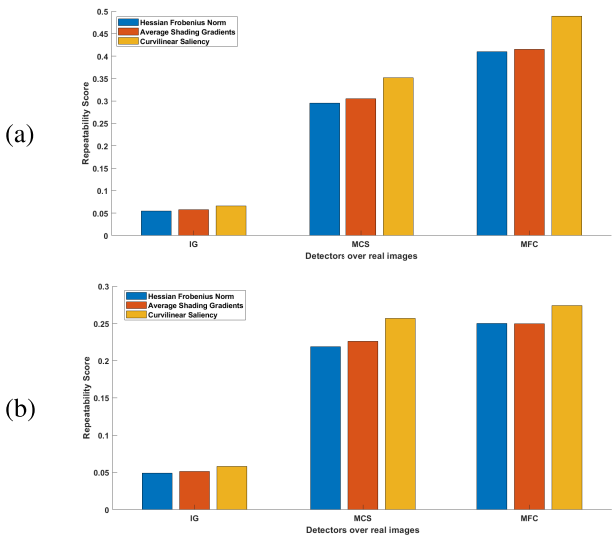

Fig. 4. Average repeatability percentages for two examples of 3D models of *PASCAL3D+ dataset*: car (a) and sofa (b).

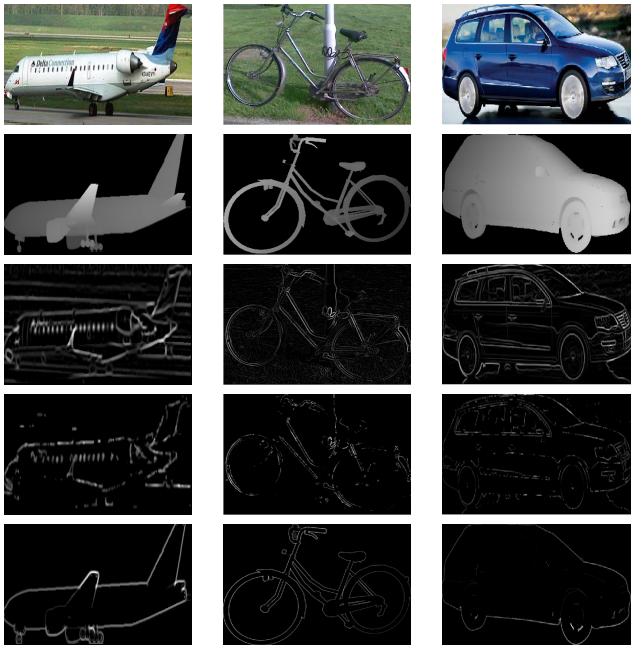

Fig. 5. Real images (row 1), depth images (row 2), and Curvilinear Saliency resulting with 5 scales with MCS (row 3), MFC (row 4) and CS (row 5).

real images, compared to IG, and as expected, MFC still yields the best repeatability scores. Among the detectors used for depth images, the Curvilinear Saliency detector yields the best repeatability scores between the three intensity-based 2D detectors. In conclusion, using CS with MFC gives the best repeatability among all the other possible combinations. In Fig. 5, some visual results show that MFC can reduce a high number of edges belonging to texture information.

E. Robustness to Illumination Changes

The MCS and MFC methods have been tested with sequences of the web collection database by changing the global illumination of the image depending on $I_o = 255(I_i/255)^{\gamma}$, where I_i and I_o are the input and output

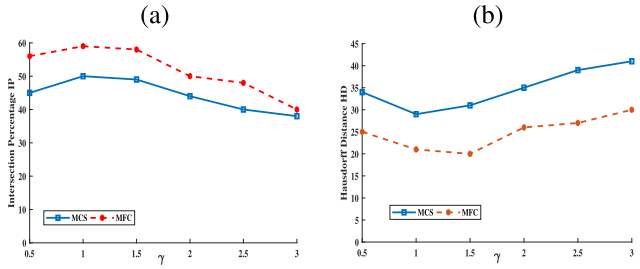

Fig. 6. Robustness against γ correction changes (illumination changes on the x-axis) illustrated with (a) the intersection percentage (y-axis) and (b) the Hausdorff distance (y-axis).

TABLE V
MEAN EXECUTION TIMES IN SECONDS OF MATLAB CODES EXECUTING ON A 2.9 GHZ INTEL CORE(I7)

MFC	MCS	PCI	MinEig	Harris	SIFT	SURF	Sobel	Canny	LOG
0.023	0.018	0.041	0.022	0.057	0.121	0.088	0.023	0.024	0.062

images, respectively, and $\gamma > 0$ is the gamma correction. Fig. 6 shows a qualitative comparison of the intersection percentage (IP) and the Hausdorff distance (HD). Both MCS and MFC are robust against small and significant changes in γ.

F. Execution Time for Detection

The proposed approaches obtain good results without a substantial impact on the execution time. As shown in table V, where the mean execution times are given, both MFC and MCS execution time is compared with the 8 tested detectors. The proposed approaches are finally less time-consuming than SIFT or even SURF approaches. Moreover, MFC and MCS are also twice as fast as PCI, which also works with curvatures.

VIII. REGISTRATION OF 2D IMAGES TO 3D MODELS

In this section, a 2D query image is registered to a 3D model by finding the closest view d between all the rendered images of the 3D model d_k, $k = 1 \ldots N$, with N being the number of rendered views (i.e., depth images). The object to recognize is supposed to be contained in a bounding box, and we would like to estimate the 3D pose. Estimating the pose consists in estimating the elevation and the azimuth angles, (h) and (a), respectively, and the distance between the model and the camera, (v). For each 3D model, depth images are generated from almost uniformly distributed viewing angles around a sphere by changing h, a and v to have N views per model. The choices for these terms are explained in paragraph IX-A.

To describe an object in a photograph and in all the rendered depth images, we naturally expand the famous classical descriptor HOG, Histogram of gradient, which is presented in [9] and widely used [2], [32], to work on Curvilinear Saliency by generating Histogram of Curvilinear Saliency, HCS. A sliding window is used to generate dense features based on binning the gradient orientation over a region. Indeed, both in rendered depth images and in photographs, the curvature orientation and the magnitude of the Curvilinear Saliency are used for building the descriptors. For depth images, CS is

multiplied by the eigenvector e_{H_1} corresponding to the largest eigenvalue of the matrix M in (11):

$$\overrightarrow{CS} = CS.\overrightarrow{e_{H_1}}.$$

For photographs, MCS values are multiplied by the eigenvector e_{S_1} corresponding to the Curvilinear Saliency $\lambda_1 - \lambda_2$:

$$\overrightarrow{MCS} = MCS.\overrightarrow{e_{H_1}}.$$

Moreover, MFC values are also multiplied by the eigenvector e_{S_1}:

$$\overrightarrow{MFC} = MFC.\overrightarrow{e_{S_1}}.$$

Using the HOG principle, we propose a descriptor that contains the curvature orientation and the magnitude of CS, MCS and MFC, binned into sparse Histograms.

Given the HCS descriptor from a 2D query image D_q, the HCS descriptors of the rendered images D_{d_N}, with N rendering depth images, are computed. To compare D_q to every D_{d_N}, the similarity scores are computed as in [2]:

$$\mathbf{S}_{hcs}(k, h, a, \upsilon) = (\mathbf{D}_{d_N} - \boldsymbol{\mu}_s)^T \Sigma^{-1} \mathbf{D}_{\mathbf{q}}, \tag{29}$$

where $k = 1 \ldots N$, and Σ and $\boldsymbol{\mu}_s$ are, respectively, the covariance matrix and the mean over all descriptors of the rendered images. For the registration process, evaluating $\mathbf{S}_{hcs}(k, h, a, \upsilon)$ can be carried out by computing the probability of the inverse of the inner product between D_q and a transformed set of descriptors. The $\mathbf{S}_{hcs}(k, h, a, \upsilon)$ probability is then maximized to find the closest corresponding views of the query image.

Moreover, a global similarity is evaluated by measuring how well each individual detected point in an image can be matched with a corresponding detected point in the depth map, i.e., how well each image's detected points are repeatable. More precisely, this repeatability score, Rep, normalized between 0 and 1, is the probability that key features in the intensity image are found close to those extracted in the depth image $Rep_{d_N \dashrightarrow q}$. Since the closest view should have high repeatability scores in comparison to other views, the dissimilarity based on repeatability scores is defined by $R_{d_i} = 1 - Rep_{d_i \dashrightarrow q}$. If $\mathbf{R}_{d_i}$ is the repeatability scores of N rendered views of a model and an image, the similarity $\mathbf{S}_{rep}$ is defined by

$$\mathbf{S}_{rep}(k, h, a, \upsilon) = exp\left(\frac{-(\mathbf{R}_{d_i} - \mu_r)^2}{2\,\sigma_r^2}\right). \tag{30}$$

where μ_r is the mean value of $\mathbf{R}_{d_N}$, and σ_r is the standard deviation (i.e., in this work $\sigma_r = 0.1$). Finally, by combining all HCS feature similarities and the similarity based on the repeatability, the probability of the final similarity is given by

$$\mathbf{S}(m, h, a, \upsilon) = \mathbf{S}_{hcs}(k, h, a, \upsilon) \odot \mathbf{S}_{rep}(k, h, a, \upsilon). \tag{31}$$

where $\odot$ is the Hadamard product. Based on calculating $\mathbf{S}(k, h, a, \upsilon)$, we select at least the highest three correspondences to estimate the full pose. From the selected three views, the logically ordered or connected views (i.e., coherent views) are first selected. As a following step, minimum and maximum values of h, a and υ of the corresponding views are estimated. Subsequently, additional views are generated in the vicinity of the selected views that is between the minimum and the maximum values of the three parameters with small steps (e.g., $\delta h = 5°$, $\delta a = 5°$ and $\delta \upsilon = 5\ cm$). The process is repeated for these ranges to find the closet view to the object in a query image until the differences between the minimum and maximum values of h, a, υ of the selected coherent views are as small as possible; more precisely, $|dh| = 5$, $|da| = 5$, $|d\upsilon| = 1$ are used to stop the algorithm repetition.

IX. Pose Estimation Experiments

A. 3D Model Representation and Alignment

Matching photographs and rendered depth images requires a 3D model representation. Each depth image represents a 3D model from different viewpoints. Hence, we need to have a significant number of depth images to completely represent a 3D model, which yields a high execution time. Consequently, N depth images (approximately 700 in our experiments) have been orthographically rendered from approximately uniformly distributed viewing angles h and a and the distance υ (i.e., in these experiments, h is empirically chosen and is increased by a step of 50°, the azimuth angle, 20°, and the distance, 0.3 m, for a range between 0 and 2 m).

Moreover, we need to parameterize the model's view alignment between the depth image and the object detected in a color image. For comparing two models, the optimal measure of similarity, over all possible poses, has to be computed. To do so, each model is placed into a canonical coordinate frame normalized for translation and rotation. Since the model centroids are known, the models are normalized for translation by shifting them to align the center of mass with the origin. Subsequently, the two models are normalized for rotation by aligning the model's principal axes with the x- and y-axes. This defines the ellipsoid that best fits the model. By rotating the two point sets so that the ellipsoid's major axis is aligned with the x-axis and the second major axis is aligned with the y-axes, the model is obtained in a normalized coordinate frame. Then, principal component analysis, PCA, is used to find the orientation of the major axis of the ellipse. The model's point set is rotated by the difference in the direction of the two major axes. After normalization, the two models are (almost) optimally aligned and can be directly compared in their normalized poses.

In addition, the HCS descriptor is quantized into 9 bins, exactly as proposed in [9]. The photograph and each depth image are divided into a grid of square cells (we have empirically chosen that the image is divided into 8×8).[6] For each cell, Histograms are aggregated by weighting them with their respective magnitudes.

B. Analysis of the Results

For pose estimation or even for object recognition, the probability that photograph key features are found close to depth key features must be high when the photograph and the depth image come from the same viewpoint. This aspect is illustrated

[6]Different grids were tested: 4×4, 8×8 and 16×16. The grid with 8×8 size yields the best precision rate.

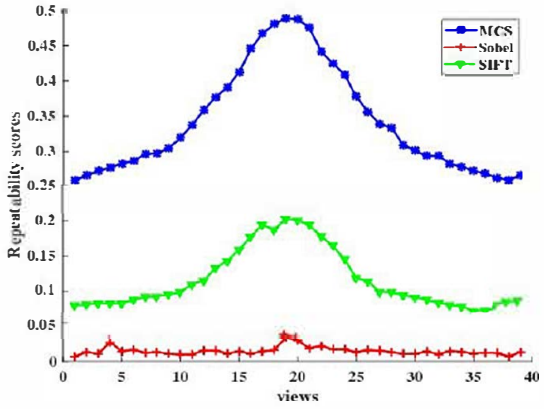

Fig. 7. Repeatability scores of an image with 40 rendered views around the correct view. The correct view is view#20 in the middle of the x-axis.

TABLE VI

PRECISION OF POSE ESTIMATION FOR CS, ASG AND AR AGAINST MFC, MCS AND IG

Methods 3D / 2D	CS			ASG			AR		
	MFC	MCS	IG	MFC	MCS	IG	MFC	MCS	IG
Plane	0.85	0.83	0.62	0.84	0.80	0.59	0.78	0.70	0.50
Bicycle	0.81	0.76	0.60	0.80	0.78	0.61	0.74	0.73	0.49
Boat	0.78	0.71	0.58	0.75	0.70	0.57	0.71	0.68	0.52
Bus	0.87	0.82	0.56	0.82	0.80	0.52	0.75	0.74	0.51
Car	0.86	0.85	0.58	0.86	0.83	0.51	0.76	0.72	0.47
Diningtable	0.86	0.83	0.61	0.81	0.81	0.60	0.79	0.77	0.54
Motorbike	0.79	0.78	0.60	0.78	0.75	0.58	0.69	0.62	0.52
Sofa	0.85	0.81	0.64	0.80	0.72	0.61	0.68	0.61	0.53
Train	0.87	0.86	0.70	0.81	0.82	0.71	0.74	0.67	0.58
Tvmonitor	0.83	0.80	0.55	0.80	0.79	0.54	0.66	0.64	0.52
Chair	0.80	0.76	0.62	0.78	0.73	0.56	0.68	0.66	0.54
Bottle	0.77	0.74	0.61	0.79	0.71	0.59	0.62	0.64	0.50

in Fig. 7. As expected, the three tested detectors yield the highest repeatability score with the correct viewpoint (even if the difference between views is slight, as with Sobel). In addition, as expected, the score is gradually diminished whenever it is at a distance from the correct viewpoint. The most important remark is that MCS results in the highest differences between the correct view and all the other views. Consequently, this illustrates that it is the most adapted detector for pose estimation based on 2D/3D registration. This result is quite coherent because SIFT was designed to be robust in the face of numerous change difficulties. Hence, it induces that the differences should be lower than MCS, which is designed to be efficient in the case of 2D/3D matching.

In addition, the other experiment was performed with the Pascal+3D dataset. For each category of objects, we compute the precision rate for detecting the correct view. This is done subsequent to using the three aforementioned methods for 3D model representations, i.e., Curvilinear Saliency (CS), Average Shading Gradient (ASG) and apparent ridges (AR) [19], against the three techniques for intensity image representation, i.e., Image Gradient (IG), Multi-Curvilinear Saliency (MCS) and Multi-Focus Curves (MFC). As shown in table VI, the registration between our Curvilinear Saliency (CS) representation of the 3D model and the multi-scale focus curves (MFC) extracted on corresponding images outperforms all other variations of the tested methods. This confirms the fact that

TABLE VII

AVERAGE ERROR OF THE ESTIMATED POSE (EST.) (a) ELEVATION, (b) AZIMUTH AND (c) YAW ANGLES AND (d) DISTANCE, IN CENTIMETRES, OF THE POSE OF THE CAMERA. THE TERM CLO. INDICATES THE CLOSEST VIEW TO THE CORRECT POSE. THESE QUANTITATIVE RESULTS DEMONSTRATE THAT THE BEST COMBINATION IS MFC/CS

Methods	(a)		(b)		(c)		(d)	
	Est.	Clo.	Est.	Clo.	Est.	Clo.	Est.	Clo.
CS/MFC	16.5°	4.8°	08.8°	1.2°	5.6°	0.8°	18	7
CS/MCS	16.0°	5.2°	11.4°	1.5°	6.1°	1.1°	21	8
ASG/MFC	19.2°	5.3°	10.1°	1.3°	5.1°	0.8°	22	9
ASG/MCS	19.6°	5.9°	13.6°	1.9°	6.2°	1.2°	23	11
AR/MFC	28.7°	7.1°	16.5°	2.5°	8.5°	1.8°	36	13
AR/MCS	29.5°	8.0°	17.3°	3.1°	9.2°	2.0°	39	17

TABLE VIII

$Acc_{\pi/6}$ MEASURES POSE ESTIMATION ACCURACY (THE HIGHER THE BETTER), AND $MedErr$ MEASURES THE VIEWPOINT ERROR (THE SMALLER THE BETTER) FOR THE PROPOSED MODEL BASED ON MFC FEATURES AND TWO DEEP LEARNING MODELS [41], [45]

Models	mean $Acc_{\pi/6}$	mean $MedErr$
Render [41]	0.82	13.6
ONet [45]	0.81	11.7
Our Model with MFC	0.80	09.5

TABLE IX

CONTRIBUTION OF EACH STEP OF THE PROPOSED ALGORITHM TO THE ENTIRE EXECUTION

Rendering	**Depth Feature Extraction**	**Image Feature Extraction**	**Registration**
25.25%	28.28%	6.06%	40.40%

Curvilinear Saliency representation computed from the depth images of a 3D model can capture the surface discontinuities. In addition, MFC can reduce the influence of texture and background components by extracting the edges related to the object shape in intensity images rather than MCS. Furthermore, the precision rate is reduced by more than 25% compared to ASG and IG. Apparent ridge rendering yields the lowest registration accuracy with the three image representations among all the 3D model representation techniques. Moreover, using ASG with untextured 3D models against MFC and MCS increases the correct pose estimation rate. All these results indicate that Average Shading Gradients computed from the normal map of an untextured geometry are a good rendering technique for the untextured geometry. However, Image Gradients are not the appropriate representation of intensity images to match rendering images and real images since they are affected by image textures. All these results are confirmed in table VII, where the details regarding the pose estimation precision in terms of elevation, azimuth, yaw angles and distance are given.

In the following experiment, in Fig. 8, the precision of image registration is shown among the top r similarities, i.e., we sort all the similarity scores obtained for all views, and the r first highest similarities are analyzed (more precisely, the 1, 3, 5, 10 and 20 first ranks). The correct pose is searched for within this view set. As shown, the precision rate is increased when the number of views is increased for any

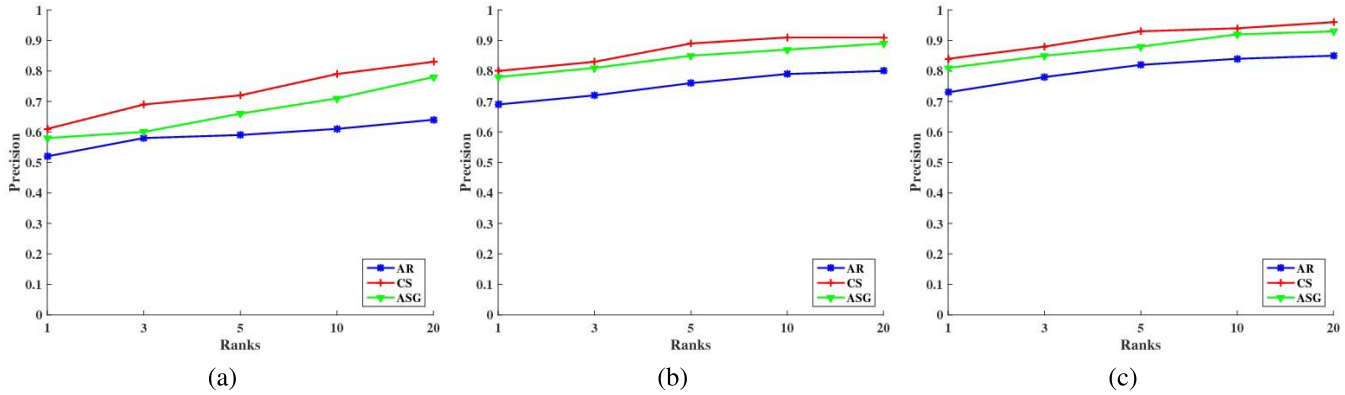

Fig. 8. Precision values with different ranks with image representation using (a) Image Gradient (IG), (b) Multi-Scale Curvilinear Saliency (MCS) and (c) multi-scale focus curves (MFC).

Fig. 9. Some correct registration examples with the Pascal+3D dataset. We show the query image (column 1), the corresponding 3D model (column 2) and the first ranked pose estimation (column 3). This illustrates that even if the 3D model does not have the same detailed shape, the registration can be correctly executed.

combination of 3D model representation and image representation. However, MFC yields the highest precision rate with the three tested methods for representing 3D models. In addition, MCS yields good precision values. In fact, IG yields the lowest precision values due to the fact that the edges detected with texture information have a negative influence on estimating the successful registration.

Finally, Fig. 9 shows some examples of correct registrations with the top-ranked pose estimation. It can be seen that our system is able to register an image with a wide variety of textures and viewing angles. In addition, the proposed algorithm can register images regardless of light changes.

C. Comparison With a CNN Model

The proposed model based on MFC features is compared with two deep pose estimation models [41], [45] using the same dataset, PASCAL3D+. We used the same metrics $Acc_{\pi/6}$ and $MedErr$ as in [45]. The quantitative results are shown in table VIII. As indicated, our model based on MFC yields an average $Acc_{\pi/6}$ of 80%, which is comparable with the work put forward [41], [45] with accuracies of 82% and 81%, respectively, although these methods have rendered millions of synthetic images to train their deep models. For $MedErr$, the proposed method yields the smallest error among the two tested methods [41], [45], achieving $MedErr$ of 9.5°, while [41], [45] achieved 13.6° and 11.7°.

D. Execution Time for Registration

In Table IX, we show how each step of the proposed approach, i.e., rendering, depth feature extraction (CS), image feature extraction (MFC/ MCS) and registration, contributes to the total execution time. It is shown as a pie

chart: approximately 53% of the execution time of the entire algorithm's total time is spent on rendering and depth feature extraction. In addition, the time requested to extract the input color feature is only 6%. Finally, to determine the final viewpoint, the time spent corresponds to approximately 40% of the entire operation. Optimizing the code was not the priority, and we can imagine that this execution time can be improved.

X. Conclusions and Perspectives

After an analysis of existing tools for 2D/3D registration, the major goal of this paper was to propose a more adapted approach for 2D/3D matching, and, in particular, more justified than existing approaches. For that purpose, we also put forward an evaluation protocol based on the repeatability study. More precisely, to carry out this matching process, we have studied these two important aspects: how to represent the data in 2D and 3D and, subsequently, how to compare them. In this context, we introduce a 3D detector based on Curvilinear Saliency and a 2D detector based on the same principle but adapted on multiple scales and combined with the principle of focus curves. The interest in this method is illustrated by quantitative evaluation on pose estimation and 2D/3D registration. All the results are encouraging, and the next step of this work is to use this registration to identify object defaults. For this purpose, we need to study the robustness of this work with regard to missing parts of objects and to adapt the registration process.

Acknowledgment

The authors would like to thank Anne Brittain for the careful reading of the paper.

References

[1] S. Agarwal *et al.*, "Building rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[2] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models," in *Proc. CVPR*, Jun. 2014, pp. 3762–3769.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[4] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[5] R. J. Campbell and P. J. Flynn, "A survey of free-form object representation and recognition techniques," *Comput. Vis. Image Understand.*, vol. 81, no. 2, pp. 166–210, 2001.

[6] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[7] C. B. Choy, M. Stark, S. Corbett-Davies, and S. Savarese, "Enriching object detection with 2D-3D registration and continuous viewpoint estimation," in *Proc. CVPR*, Jun. 2015, pp. 2512–2520.

[8] M. J. Clarkson, D. Rueckert, D. L. G. Hill, and D. J. Hawkes, "Using photo-consistency to register 2D optical images of the human face to a 3D surface model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1266–1280, Nov. 2001.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, Jun. 2005, pp. 886–893.

[10] M. Dellepiane, M. Callieri, F. Ponchio, and R. Scopigno, "Mapping highly detailed colour information on extremely dense 3D models: the case of David's restoration," in *Comput. Graph. Forum*, vol. 27, no. 8, pp. 2178–2187, 2007.

[11] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, and L. Shapiro, "Principal curvature-based region detector for object recognition," in *Proc. CVPR*, Jun. 2007, pp. 1–8.

[12] J. H. Elder and S. W. Zucker, "Local scale control for edge detection and blur estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 7, pp. 699–716, Jul. 1998.

[13] P. Fischer and T. Brox, "Image descriptors based on curvature histograms," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 239–249.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[15] J.-H. Gallier, *Geometric Methods and Applications: For Computer Science and Engineering*. New York, NY, USA: Springer-Verlag, 2001.

[16] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 1–5.

[17] A. Irschara, C. Zach, J. M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *Proc. CVPR*, Jun. 2009, pp. 2599–2606.

[18] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by ufo: Uniqueness, focusness and objectness," in *Proc. ICCV*, Dec. 2013, pp. 1976–1983.

[19] T. Judd, F. Durand, and E. Adelson, "Apparent ridges for line drawing," *ACM Trans. Graph.*, vol. 26, no. 3, p. 19, 2007.

[20] L. Karacan, E. Erdem, and A. Erdem, "Structure-preserving image smoothing via region covariances," *ACM Trans. Graph.*, vol. 32, no. 6, p. 176, 2013.

[21] K. Kiranpreet, V. Mutenja, and E. I. S. Gill, "Fuzzy logic based image edge detection algorithm in MATLAB," *Int. J. Comput. Appl.*, vol. 1, no. 22, pp. 55–58, 2010.

[22] J. J. Koenderink and A. J. Van Doorn, "Surface shape and curvature scales," *Image Vis. Comput.*, vol. 10, no. 8, pp. 557–564, 1992.

[23] Y. Lee, L. Markosian, S. Lee, and J. F. Hughes, "Line drawings via abstracted shading," *ACM Trans. Graph.*, vol. 26, no. 3, p. 18, 2007.

[24] Y. Y. Lee, K. Park, J. D. Yoo, and K. H. Lee, "Multi-scale feature matching between 2D image and 3D model," in *Proc. SIGGRAPH Asia*, 2013, p. 19.

[25] L. Liu and I. Stamos, "Automatic 3D to 2D registration for the photorealistic rendering of urban scenes," in *Proc. CVPR*, Jun. 2005, pp. 137–143.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[27] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3D/2D registration methods for image-guided interventions," *Med. Image Anal.*, vol. 16, no. 3, pp. 642–661, 2012.

[28] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[29] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 24–52, 2009.

[30] D. P. Paudel, C. Demonceaux, A. Habed, and P. Vasseur, "Localization of 2D cameras in a known environment using direct 2D-3D registration," in *Proc. ICPR*, Aug. 2014, pp. 196–201.

[31] A. P. Pentland, "A new sense for depth of field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 4, pp. 523–531, Jul. 1987.

[32] T. Plötz and S. Roth, "Automatic registration of images to untextured geometry using average shading gradients," *Int. J. Comput. Vis.*, vol. 125, no. 1, pp. 65–81, 2017.

[33] F. Pomerleau, F. Colas, and S. Roland, "A review of point cloud registration algorithms for mobile robotics," *Found. Trends Robot*, vol. 4, no. 1, pp. 1–104, 2015.

[34] L. Q. Xu Zhang, X. Shen, and J. Jia, "Rolling guidance filter," in *Proc. ECCV*, 2014, pp. 815–830.

[35] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand, "Geolocalization using skylines from omni-images," in *Proc. ICCV Workshops*, Sep./Oct. 2009, pp. 23–30.

[36] H. A. Rashwan, S. Chambon, P. Gurdjos, G. Morin, and V. Charvillat, "Towards multi-scale feature detection repeatable over intensity and depth images," in *Proc. ICIP*, Sep. 2016, pp. 36–40.

[37] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *Proc. ICCV*, Nov. 2011, pp. 667–674.

[38] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

[39] J. Shi and C. Tomasi, "Good features to track," in *Proc. CVPR*, Jun. 1994, pp. 593–600.

[40] S. M. Smith and J. M. Brady, "SUSAN—A new approach to low level image processing," *Int. J. Comput. Vis.*, vol. 23, no. 1, pp. 45–78, May 1997.
[41] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2686–2694.
[42] R. Szeliski, *Computer Vision–Algorithms and Applications*. London, U.K.: Springer-Verlag, 2011.
[43] Y.-W. Tai and M. S. Brown, "Single image defocus map estimation using local contrast prior," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 1797–1800.
[44] M. Tamaazousti, V. Gay-Bellile, S. N. Collette, S. Bourgeois, and M. Dhome, "Nonlinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment," in *Proc. CVPR*, Jun. 2011, pp. 3073–3080.
[45] S. Tulsiani and J. Malik, "Viewpoints and keypoints," in *Proc. CVPR*, Jun. 2015, pp. 1510–1519.
[46] T. Tuytelaars and L. van Gool, "Matching widely separated views based on affine invariant regions," *Int. J. Comput. Vis.*, vol. 59, no. 1, pp. 61–85, 2004.
[47] C. Wu, B. Clipp, X. Li, J. M. Frahm, and M. Pollefeys, "3D model matching with viewpoint-invariant patches (VIP)," in *Proc. CVPR*, Jun. 2008, pp. 1–8.
[48] C. Xu, L. Zhang, L. Cheng, and R. Koch, "Pose estimation from line correspondences: A complete analysis and a series of solutions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1209–1222, Jun. 2017.
[49] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond pascal: A benchmark for 3D object detection in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2014, pp. 75–82.
[50] S. Zhuo and T. Sim, "Defocus map estimation from a single image," *Pattern Recognit.*, vol. 44, no. 9, pp. 1852–1858, 2011.