Affine Covariant Features for Fisheye Distortion Local Modelling

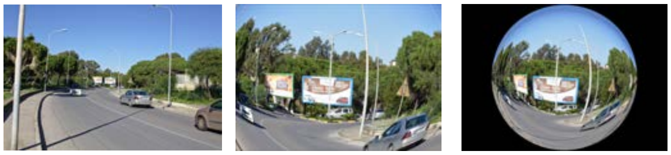
Antonino Furnari, Giovanni Maria Farinella, *Member, IEEE,*
Arcangelo Ranieri Bruna, and Sebastiano Battiato, *Senior Member, IEEE*

***Abstract*—Perspective cameras are the most popular imaging sensors used in Computer Vision. However, many application fields including automotive, surveillance and robotics, require the use of wide angle cameras (e.g., fisheye), which allow to acquire a larger portion of the scene using a single device at the cost of the introduction of noticeable radial distortion in the images. Affine covariant feature detectors have proven successful in a variety of Computer Vision applications including object recognition, image registration and visual search. Moreover, their robustness to a series of variabilities related to both the scene and the image acquisition process has been thoroughly studied in the literature. In this paper, we investigate their effectiveness on fisheye images providing both theoretical and experimental analyses. As theoretical outcome, we show that the inherently non-linear radial distortion can be locally approximated by linear functions with a reasonably small error. The experimental analysis builds on Mikolajczyk's benchmark to assess the robustness of three popular affine region detectors (i.e., Maximally Stable Extremal Regions (MSER), Harris and Hessian affine region detectors), with respect to different variabilities as well as to radial distortion. To support the evaluations, we rely on the Oxford dataset and introduce a novel benchmark dataset comprising 50 images depicting different scene categories. Experiments are carried out on rectilinear images to which radial distortion is artificially added, and on real-world images acquired using fisheye lenses. Our analysis points out that affine region detectors can be effectively employed directly on fisheye images and that the radial distortion is locally modelled as an additional affine variability.**

***Index Terms*—fisheye distortion, affine region detectors, omnidirectional vision, division model**

I. Introduction and Motivations

COMPUTER Vision algorithms are usually designed to work on images acquired using perspective cameras. The adherence to the perspective camera model ensures that straight lines in the real world are mapped to straight lines in the image, which produces a representation of the scene coherent with our perception [1]. However, many application fields such as automotive, surveillance and robotics [2], [3], [4], [5], [6], [7], require the use of wide angle cameras, which are characterized by a wide Field Of View (FOV) and are able to acquire a large part of the scene using a single device. Fig. 1 shows some examples of wide angle images, along with their perspective counterpart. Unfortunately the perspective camera model cannot be efficiently used to model the image formation process of wide angle cameras which hence require different projection models with the consequent introduction of inherent radial distortion [1], [2], [8]. Wide angle cameras can be built following two main designs: dioptric [1], [8] and catadioptric [2], [9], [10]. In particular, we consider the dioptric systems, which are built substituting the regular lens of a perspective camera with a fisheye lens able to divert the light rays on the sensor in order to achieve the desired wide Field Of View. As discussed by Miyamoto [8], the distortion introduced by such cameras should not be considered as an aberration but as the result of the projection of an hemisphere on a finite plane. The most straightforward approach to deal with wide angle images consists in explicitly removing the distortion through a rectification process. Such process however can be computationally expensive, especially in embedded settings, since it requires interpolation to account for the spatially non-uniform sampling performed by the wide angle sensor. Moreover, the interpolation process introduces artefacts in the image which can affect the feature extraction process [11]. Additionally, in order to perform the rectification process, the camera needs to be calibrated so that a mapping between the distorted points and their positions in the ideal (rectilinear) image plane can be established. Some calibration techniques require a special pattern to be present in the scene [12], [13] while others [14], [15] just require a few images of the scene and no other information. However, even when the camera can be easily calibrated, it would be advantageous to be able to work directly on the distorted images to avoid the rectification process and get rid of the artefacts due to the interpolation. Many efforts in the context of wide angle calibrated cameras already exist: the authors of [16], [17] studied how to compute the scale space of omnidirectional images, in [3], [11], [18] the Scale Invariant Feature Transform (SIFT) pipeline [19] is modified in order to be used directly on wide angle images; in [20] scale invariant features are derived



(a) rectilinear (b) full frame (c) full circle

Fig. 1. Examples of perspective (a), full frame (b) and full circle (c) images. The two fisheye images are obtained by artificially adding different amounts of radial distortion to the rectilinear image (a).

Antonino Furnari, Giovanni Maria Farinella and Sebastiano Battiato are with the University of Catania, Department of Mathematics and Computer Science, Catania, 95125, Italy (e-mail: furnari@dmi.unict.it, gfarinella@dmi.unict.it, battiato@dmi.unict.it)

Arcangelo Ranieri Bruna is with STMicroelectronics, Advanced System Technology - Computer Vision, Catania 95121, Italy (e-mail: arcangelo.bruna@st.com)

Manuscript received Month day, year; revised Month day, year.

from wide angle images mapping them to a sphere; in [21] a direct approach to detect people using omnidirectional cameras is proposed; in [22] an algorithm to extract straight edges from distorted images is presented; in [11], [22], [23] methods to estimate geometrically correct gradients of distorted images are investigated. A second category of algorithms working directly on the distorted images doesn't need the camera to be calibrated and treats the radial distortion as an additional variability present in the images. For instance, in [24] the Perspectively Invariant Normal features (PIN) are computed using a depth map in order to be independent from the acquisition point of view and from the employed camera. PIN can be successfully used to match regions between rectilinear and wide angle images as pointed out by the authors of [24]. In [25] an approach to match features between uncalibrated omnidirectional images (not rectified) and perspective images is presented. People detection and tracking are performed directly on fisheye images using a probabilistic appearance model in [26]. The authors of [27] perform feature matching on omnidirectional images through descriptor learning. This category of approaches which do not need calibration is particularly interesting for those applications in which the input images are acquired by different cameras which are not generally known in advance and hence difficult to calibrate. Examples of such applications include image retrieval (e.g., images on the web), object detection on generic cameras and registration (e.g., camera networks in surveillance applications).

In this paper we consider the context of direct approaches dealing with uncalibrated images. Specifically, we study how the detection, description and retrieval of local features can be reliably performed on uncalibrated wide angle images acquired by an unknown device. In particular, considering the amount of work already done in the field of affine covariant detectors and descriptors [28], [29], [30] in the perspective domain, we investigate whether the state-of-the-art affine detectors are suitable to be used directly on wide angle images. We support our analysis by theoretically showing that, even if the radial distortion introduced by fisheye cameras is not an affine transformation, it can be locally approximated as a linear function with a small error. Moreover, we consider three state-of-the-art affine region detectors [28], namely the Maximally Stable Extremal Regions (MSER) [31], the Harris affine region detector and the Hessian affine region detector [32], [33], and study how such detectors behave under the influence of increasing radial distortion and the variabilities included in the Oxford dataset [28], i.e., change of viewpoint angle, scale changes, blur, JPEG compression and light changes. We extend our previous work [34] on affine region detectors in the fisheye domain, moving the analysis from the theoretical fisheye projection functions to the Division Model [15] which generalizes our study to many real world fisheye cameras [35]. Moreover, we introduce a new dataset of high resolution rectilinear images depicting real world scenes belonging to different categories (see Section II-B for the details), which we use to generate artificial fisheye images with a controlled amount of distortion for testing purposes. This dataset allows us to assess the performances of the detectors on different scene types and to draw general conclusions on their robustness with respect to increasing radial distortion. Experiments show that the affine covariant region detectors under analysis achieve good performances when computed directly on fisheye images and hence that they succeed in locally modelling the radial distortion introduced by fisheye images.

The main contributions of this paper can be summarized as follows: we review the Division Model and provide meaningful interpretations for the involved distortion parameter; we perform a theoretical analysis to motivate the applicability of affine covariant region detectors directly on fisheye images; we introduce a dataset of high resolution images depicting scenes belonging to different categories artificially adding radial distortion for testing purposes; we revise the evaluation pipeline proposed by Mikolajczyk et al. [28] and define a new set of experiments to assess the performances of affine covariant feature detectors on fisheye images; by means of extensive experiments we show that the current affine detectors are robust to radial distortion and to the combination of radial distortion with other common variabilities.

The remainder of the paper is organized as follows: in Section II we discuss the fisheye domain, we briefly review the Division Model providing practical interpretations of the distortion parameter and introduce the synthetic data used for the experiments. In Section III we discuss some properties of the considered detectors and provide theoretical evidence of the applicability of affine region detectors in the fisheye domain. In Section IV the experimental setup is detailed, whereas in Section V the results are discussed. Finally in Section VI we draw the conclusions.

II. The Fisheye Domain

Fisheye lenses allow formation of an image of an hemispheric field on a finite plane [8]. This is achieved by sampling the incoming light rays in a spatially non-uniform way. Specifically, due to the properties of the lens, light is sampled densely in the central part of the image and coarsely in the peripheral areas. This phenomenon introduces a symmetric radial distortion which causes the points on the image plane to be shifted from their ideal position in the rectilinear space towards the principal point. The amount of radial distortion undergone by the acquired image is proportional to the Field of View of the camera. The Field Of View characterizing a given fisheye camera depends on the design of the lens, its focal length and the sensor size. In practice, two configurations are relevant: full frame and full circle [35]. Full frame images are characterized by a diagonal FOV equal to 180°. Such configuration is convenient since it allows to get the largest FOV which still allows to cover the full sensor (the whole sensor is illuminated and the image does not contain dark non-illuminated areas). Full circle images are characterized by a vertical FOV equal to 180°. Such configuration does not allow full coverage of the sensor (the image is formed on a circular region in the centre of the sensor), but allows to obtain the projection of the full hemispheric field on the final image. Fig. 1 shows some synthetic examples of the two configurations. While all perspective cameras follow a single

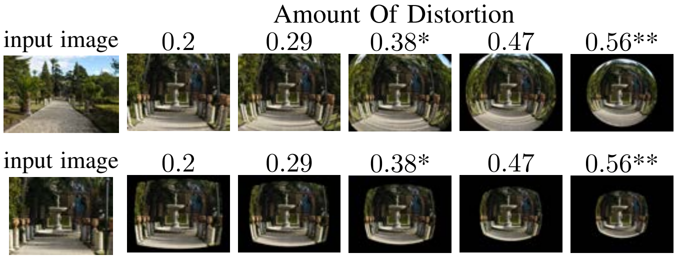


Fig. 2. Some examples of synthetic fisheye images obtained adding different amounts of radial distortion to input rectilinear images by using the Division Model. In the top row, the input image is a high resolution image (5204×3472 pixels), while in the bottom row the input image is a low resolution image (1024×768 pixels). All the output distorted images have resolution equal to 1024×768 pixels. The * and ** symbols denote the full frame and full circle distortion rates respectively.

projection model (i.e., the perspective camera model), fisheye lenses are usually manufactured to adhere to one of a set of projection functions [1], [8]. Moreover, since the design of fisheye lenses can be quite complex [36], a deviation from the ideal model is usually expected. A number of calibration models have been proposed in the literature to cope with such variability [1], [12], [15], [35]. Among these, the Division Model [15] has gained some popularity due to its ability to model real world fisheye cameras [35] with a single parameter.

In this paper, we use the Division Model to generate synthetic fisheye images by artificially adding radial distortion to input rectilinear images (as done in [11], [23], [34]). Working with these settings allows to control the exact amount of distortion present in the image and to use the source rectilinear images to build the reference ground truth for the evaluations. In general, given a rectilinear image I and the distortion function f which maps the undistorted point $\mathbf{u}$ in the rectilinear image to point $\mathbf{x} = f(\mathbf{u})$ in the distorted space, the synthetic fisheye image is defined as follows:

$$\hat{I}(f(\mathbf{u})) = I(\mathbf{u}). \tag{1}$$

It should be noted that, when the Field Of View is large, the function f is designed to project an infinite rectilinear image I to a finite fisheye image $\hat{I}$. In practice it is sufficient that the resolution of the input image I is sufficiently higher than the resolution of the output image $\hat{I}$ to achieve consistent results. Fig. 2 shows some examples of synthetic fisheye images obtained considering input rectilinear images of variable sizes. Mapping high resolution input images to low resolution ones allows to cover a larger part of the artificially distorted image, which is preferable in order to carry evaluations in the peripheral areas. However, it should be noted that distorting rectilinear images with a lens model cannot describe the image formation process with total accuracy due to the lack of depth information from the acquired scene. Nevertheless, given its synthetic nature, the proposed approach conveniently allows to study the performances of the considered detectors with respect to varying amounts of distortion. In order to complement the analysis, we also perform tests on images acquired with real lenses.

A. The Division Model

The Division Model [15] establishes a relationship between the image point $\mathbf{x}$ in the distorted space and its undistorted counterpart $\mathbf{u}$ in the rectilinear one as following:

$$\mathbf{u} = \frac{\mathbf{x}}{1+\xi||\mathbf{x}||^2} \tag{2}$$

where the parameter $\xi < 0$ regulates the amount of radial distortion in the image. It should be noted that the coordinates are referred to the principal point, which in our experiments we always consider to be coincident with the centre of the image. The relationship in (2) can be inverted in order to derive the distortion function f which maps an undistorted point $\mathbf{u}$ in the rectilinear space to the distorted point $\mathbf{x}$ in the image:

$$\mathbf{x} = f(\mathbf{u}) = \frac{2\mathbf{u}}{1+\sqrt{1-4\cdot\xi||\mathbf{u}||^2}}. \tag{3}$$

According to equations (2) and (3), a point of radial coordinate r in the undistorted space is related to a point of radial coordinate $\hat{r}$ in the distorted image by the following expressions:

$$r = \frac{\hat{r}}{1+\xi\hat{r}^2} \tag{4}$$

$$\hat{r} = g(r) = \frac{2r}{1+\sqrt{1-4\cdot\xi r^2}}. \tag{5}$$

Unfortunately, the interpretation of the values of ξ is not intuitive and the effects of setting a specific value for ξ depend on the size of the input image. Therefore we propose to express the amount of distortion as the following rate:

$$d = 1 - \frac{\hat{r}_M}{r_M} \tag{6}$$

where r_M represents the distance from the centre of the distortion (i.e., the centre of the image) to the corner of the distorted output image and $\hat{r}_M$ represents its distorted counterpart. It should be noted that such a definition is perceptually coherent and it is independent from the scale of the image size. Considering that between r_M and $\hat{r}_M$ holds relationship (5), the parameter ξ can be straightforwardly computed from a given distortion rate d using the formula:

$$\xi = -\frac{d}{[r_M(1-d)]^2}. \tag{7}$$

Even if no direct relationship between the Field Of View of a given image and parameter ξ is provided by the Division Model, the exact values of ξ can be derived for the full frame and full circle configurations discussed above. In both cases we want the distortion function (3) to project points at infinity to points on the image having a specific radius ρ. In the case of full frame images we set $\rho = r_M$ to obtain a diagonal FOV equal to 180°, while in the case of full circle images we set $\rho = h/2$ where h is equal to the image height in order to obtain a vertical FOV equal to 180°. Let us consider the limit of expression (5) as r approaches $+\infty$:

$$\lim_{r\to+\infty} \frac{2r}{1+\sqrt{1-4\cdot\xi r^2}} = \frac{2}{\sqrt{-4\xi}}. \tag{8}$$

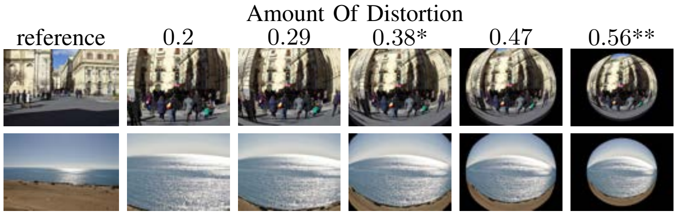


Fig. 3. Two image series from Dataset A. The * and ** symbols denote the full frame and full circle distortion rates respectively.

Equating such expression to ρ, we get:

$$\xi = -\frac{1}{\rho^2}. \tag{9}$$

Equation (9) can be used to compute the distortion parameter ξ allowing the projection of a point at infinity in the rectilinear space to a point with radial coordinate ρ in the fisheye space. Combining equations (6) and (9) and considering the values which ρ assumes in the case of full frame and full circle images, it is possible to obtain the following expressions:

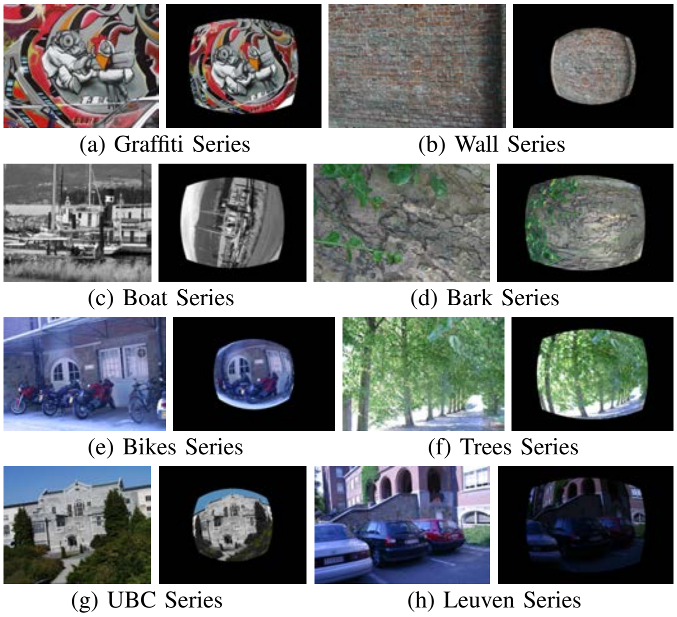
$$d_{full-frame} \approx 0.38 \tag{10}$$

$$d_{full-circle} = \frac{2\alpha^2 - \sqrt{4\alpha^2+5}+3}{2\alpha^2+2} \tag{11}$$

where $\alpha = \frac{w}{h}$ is the image aspect ratio and w is the image width. Since in our experiments we consider synthetic fisheye images of size 1024×768 pixels, then we get $\alpha \approx 1.4$ and hence $d_{full-circle} \approx 0.56$. It should be noted that, for a square image (i.e., $\alpha = 1$), the distortion rate inherent to a full circle image amounts to exactly 0.5. The deviation from such a value is entirely due to the non-rectangular aspect ratio of the output images. Fig. 3 shows some examples of synthetic fisheye images characterized by varying distortion rates, including the full frame and full circle configurations. Note that the full frame image shown in Fig. 3 still exhibits black corners. This is due to the fact that the input image is finite while an infinite image would be required in principle. For the same reason the full circle image shown in Fig. 3 is not perfectly circular and smaller than what a real full circle image should look like. Despite such considerations, the synthetic images are worth to be considered for evaluation purposes since they exhibit the amounts of radial distortion inherent to the full frame and full circle configurations and still cover most of the related Field Of View.

B. Experimental Datasets

To perform the experimental analysis, we considered three different datasets. Two of them comprise rectilinear images to which radial distortion is artificially added following the methodologies discussed in Section II. Working in these settings is convenient since it allows to control the exact amount of distortion present in the images used for the experiments. However, as discussed in Section II, distorting rectilinear images with a lens model cannot describe the image formation process with total accuracy. Hence, experiments are carried on a third dataset which comprises 39 images acquired using three different fisheye cameras.



(a) Graffiti Series (b) Wall Series

(c) Boat Series (d) Bark Series

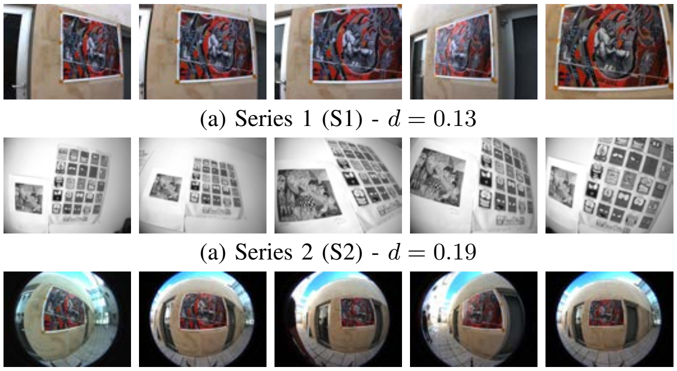
(e) Bikes Series (f) Trees Series

(g) UBC Series (h) Leuven Series

Fig. 4. Some examples from Dataset B. The leftmost image in each pair is always the reference image, while the rightmost image is one of the test images in the series characterized by a given amount of distortion. Related to Dataset B, the distortion is coupled with the variabilities considered in [28]: (a) Change of viewpoint angle for a structured scene (full frame distortion). (b) Change of viewpoint angle for a textured scene (full circle distortion). (c) Scale changes for a structured scene (full frame distortion). (d) Scale changes for a textured scene (full frame distortion). (e) Image blur for a structured scene (full circle distortion). (f) Image blur for a textured scene (full frame distortion). (g) JPEG compression (full circle distortion). (e) Light change (full frame distortion).

In order to assess the performances of local detectors with respect to varying amounts of distortion, we have built a benchmark dataset comprising 50 high resolution rectilinear images (5204×3472 pixels) to which we artificially add different radial distortion rates as we have defined in Section II. The 50 images are a random selection of the 100 images included in our previously collected dataset presented in [23]. The proposed dataset can be downloaded at the URL *http://iplab.dmi.unict.it/FisheyeAffine*. The original images have been acquired using a Canon 650D camera with a Canon EF-24mm lens and depict scenes taken by considering different categories according to the scene categorization proposed by Torralba and Oliva [37]: indoor, outdoor, natural, handmade, urban, car, pedestrian, street. To test the performances of local detectors, coherently with the evaluation procedure in [28], for each image we build a series of 6 images consisting of the reference full resolution rectilinear image, plus 5 test images affected by the following distortion rates: 0.2, 0.29, 0.38 (full frame configuration), 0.47 and 0.56 (full circle configuration). All the test distorted images have resolution 1024×768 pixels. Starting from a base value of 0.2, distortion rates have been chosen in order to be evenly spaced (at a step of 0.9) and to include the full frame and full circle configurations. We shall refer to this dataset as Dataset A. Fig. 3 shows two sample series from Dataset A.

To properly assess the performances of the local descrip-



(a) Series 1 (S1) - $d = 0.13$

(a) Series 2 (S2) - $d = 0.19$

(a) Series 3 (S3) - $d = 0.54$

Fig. 5. Some sample images from the three image series in Dataset C.

tors, we also consider the popular Ofxord dataset proposed in [28], which provides 8 image series each affected by one of the following variabilities: change of viewpoint angle, scale changes, image blur, JPEG compression, light changes. The dataset comprises both structured and textured scenes. Each series consists of a reference image, containing the least amount of the specified variability (i.e., the zero-variability) and 5 test images characterized by increasing amounts of the specified variability. To assess the influence of the combination of radial distortion with the aforementioned variabilities, we artificially add radial distortion to each test image in the dataset. It should be noted that no distortion is added to the reference images. Specifically, for each series in the Oxford dataset we generate two additional series characterized by the amounts of distortion inherent to the full frame and full circle configurations. The exact distortion rates are computed using equations (10) and (11) in order to account for the different aspect ratios characterizing the input images. As discussed in Section II, in order to avoid black borders in the target distorted images, high resolution rectilinear images should be used for reference. Since the resolutions of the images in the Oxford dataset are not high (640×480 pixels), the approach proposed in Section II is not a viable option. Hence, we keep the resolution of the output distorted image equal to the one of the input image. Note that, even if the distorted images generated in this way are not able to cover all the target distorted image, they are affected by the amounts of distortion inherent to the full frame and full circle configurations. We refer to this second dataset as Dataset B. Fig. 4 shows some samples from the considered series.

To perform tests with real fisheye images, we consider the benchmark dataset introduced in [11]. It comprises three image series acquired using fisheye lenses characterized by different amounts of distortion. Calibration images and division model parameters for each camera are included in the dataset. Fig. 5 shows some sample images from the considered dataset. For each image series, we report the distortion rates computed according to our model: 0.13, 0.19 and 0.54. Each series consists of 13 images related by different transformations including viewpoint change, rotation and scale. Images within a series represent a scene containing the same planar object acquired from different positions. All image pairs within a series are provided with an homography relating their undistorted counterparts. Differently from the Oxford dataset (Dataset B), the amount of variability present in each image is not quantified with respect to a given reference image. Hence, instead of considering only reference-test image pairs, all possible 78 image pairs within a series are considered in the experiments. We refer to this dataset as Dataset C.

III. Affine Region Detectors on The Fisheye Domain

We consider three state-of-the-art affine region detectors for our analysis: the Maximally Stable Extremal Regions (MSER) [31] detector, the Harris and the Hessian affine region detectors [32], [33]. Such detectors have shown top performances in the benchmark by Mikolajczyk et al. [28] with respect to two evaluation criteria (repeatability and matching ability, which are discussed thoroughly in Section IV) on both structured and textured images under the influence of different geometric and photometric transformations. All the considered region extractors are based on a region detection step followed by an affine covariant construction which allows to obtain elliptical features. The reader is referred to [28] for a review of affine region detectors. To provide evidence supporting the applicability of affine covariant region detectors directly on fisheye images, in Section III-A, we perform a theoretical analysis to study under what conditions radial distortion can be locally approximated by linear transformations. In Section III-B, we show that affine regions produced by the detectors under analysis are characterized by strong locality, and hence they are likely to yield a reasonably small error.

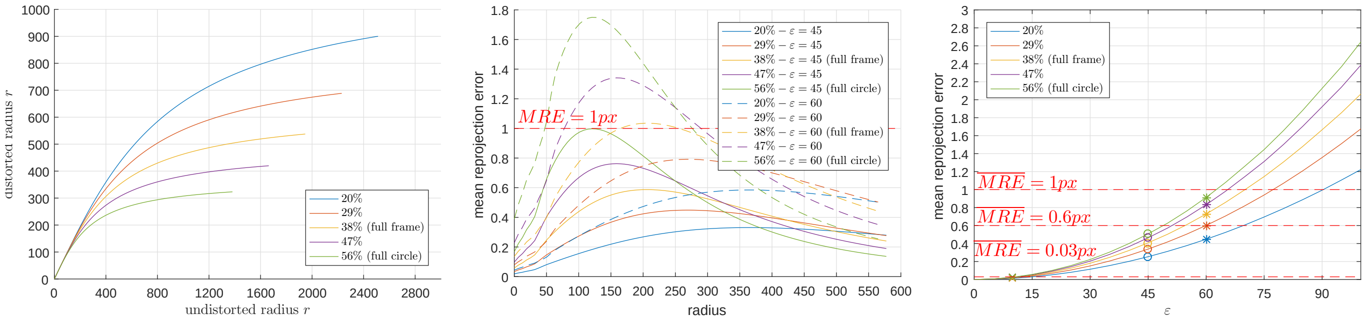
A. Local Linearity of the Division Model

In order to provide theoretical evidence to support the applicability of the affine covariant region detectors on fisheye images, in this Section we show that, even if the radial distortion introduced by fisheye cameras is not an affine transformation, it can be modelled as a linear function in small local neighbourhoods. For sake of generality, we base our analysis on the Division Model which has proven successful in modelling real fisheye cameras [35]. Specifically, we show that the radial distortion function of Equation (5) can be linearly approximated locally and that if the neighbourhood is sufficiently small, the approximation error is negligible. It should be noted that, while we base our theoretical analysis on the division model, in Section V, we report experiments on images acquired using real fisheye lenses to ensure the validity of our analysis when real-world lenses are considered.

Let us consider the first order Taylor polynomial approximation of the mapping function shown in Equation (5), centred at an arbitrary point $\hat{r}_0$ and restricted to the local neighbourhood of radius ε centred at $\hat{r}_0$ denoted by $\mathcal{N}(r_0, \varepsilon) = (\hat{r}_0 - \varepsilon, \hat{r}_0 + \varepsilon)$:

$$g(\hat{r})_{|\mathcal{N}(r_0,\varepsilon)} \approx \tilde{g}(\hat{r}, \hat{r}_0) = g(r_0) + (r - r_0)g'(r0). \qquad (12)$$

We expect the error given by such an approximation to be proportional to the extent of the chosen radius ε. To measure such error, we define the Mean Reprojection Error



(a) division model projections (b) MRE curves (c) $\overline{MRE}$ curves

Fig. 6. (a) The plot of the function in Eq. (5) for different distortion rates. (b) The Mean Reprojection Error curves for fixed values of ε. (c) The average Mean Reprojection Error for varying neighbourhood radii ε.

of expression (12) in a given point r_0 and for a chosen radius ε as follows:

$$MRE(r_0,\varepsilon) = \frac{\int_{r\in\mathcal{N}(r_0,\varepsilon)} |g(r) - \tilde{g}(r,r_0)| dr}{\int_{r\in\mathcal{N}(r_0,\varepsilon)} dr}. \tag{13}$$

Moreover, for a fixed value of ε, we define the average MRE value as follows:

$$\overline{MRE}(\varepsilon) = \frac{\int_{r=0}^{r_{max}} MRE(r,\varepsilon)dr}{\int_{r=0}^{r_{max}} dr} \tag{14}$$

where r_{max} is introduced to avoid to carry the integration up to infinity, where the curves related to Eq. (5) tend to become rectilinear as shown in Fig. 6 (a) and the MRE value would be close to zero. In particular we set r_{max} to the half diagonal of the distorted images of resolution 1024×768 pixels considered in the experiments, i.e., $r_{max} = \frac{1}{2}\sqrt{(1024^2 + 768^2)} = 640$ pixels. Fig. 6 (b) shows the MRE curves for two selected values of ε (i.e., 45 and 60 pixels) and different amounts of distortion, while Fig. 6 (c) shows the average MRE for varying values of ε and different amounts of distortion. In particular Fig. 6 (c) shows that the fisheye distortion of local regions having radii smaller than $\varepsilon = 60$ pixels can be approximated as a linear function with average subpixel precision (see points marked with the symbol "*" in Fig. 6 (c)). The average error drops to about 0.6 pixels for radii smaller than 45 pixels (see points marked with the symbol "$\circ$" in Fig. 6 (c)) and to about 0.03 pixels for radii smaller than 10 pixels (see points marked with the symbol "+" in Fig. 6 (c)). Fig. 6 (b) shows how the MRE values vary in the different parts of the image. Specifically, the error is small in the central and peripheral areas of the image and higher in between. It is worth noting that for regions with radii smaller than 45 pixels, the MRE is always under 1 pixel for all distortion rates.

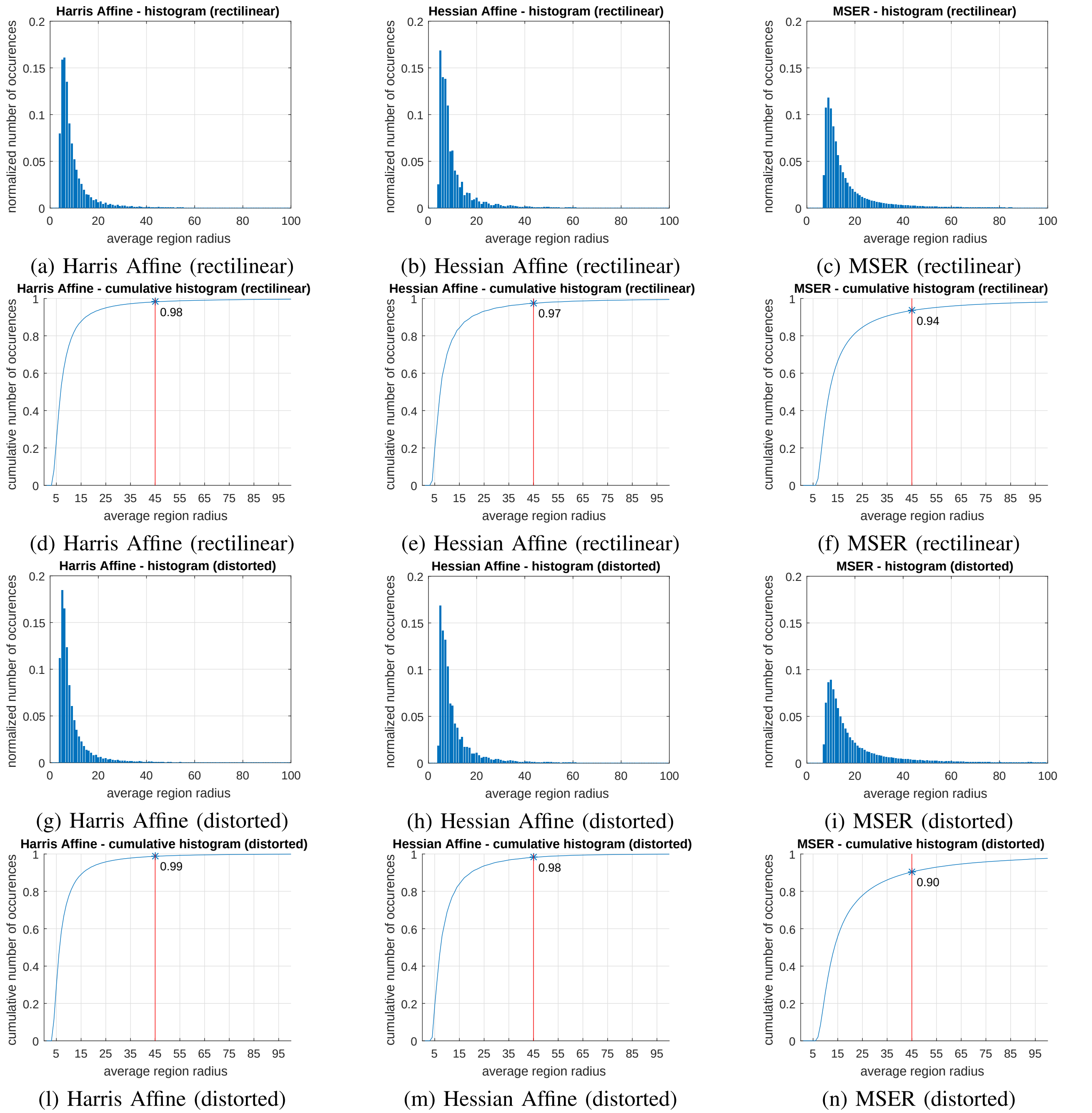
Our analysis points out that, up to a given extent, circular regions can be mapped from a reference non-distorted space to its distorted counterpart in the fisheye image using an appropriate linear function with a small projection error. If the error is low enough, an affine covariant region detector should be able to correctly extract both the reference and distorted regions modelling the latter as an affine transformation of the former. Moreover, in the description stage, the distorted region will be mapped to its undistorted counterpart with a small error using the inverse of the affine transformation estimated by the region detector. Hence we expect small linear approximation errors to be beneficial for both the feature detection and description steps.

B. Region Size

To assess the applicability of affine covariant region detectors on distorted images, we have performed an analysis of the distribution of sizes of regions extracted using the detectors under analysis. In particular, we extracted regions using the considered three detectors on all images present in Dataset A. An average radius is computed for each elliptical region as the average between the lengths of semi-major and semi-minor axes. Interestingly, all the considered detectors tend to extract regions characterized by a strong locality. This is summarized in Fig. 7, which shows the normalized histograms of average radii for all rectilinear and distorted images in Dataset A. In particular, normalized histograms reported in Fig. 7 (a) to (c) and Fig. 7 (g) to (i) show how the vast majority of regions have average radii around 10 pixels. Moreover, the cumulative histograms reported in Fig. 7 (d) to (f) and Fig. 7 (l) to (n), show how in any case more than 90% of the detected regions have an average radius smaller than 45 pixels. As it has been pointed out in Section III-A, the linear approximation error is low for regions with average radii under 45 pixels and negligible for regions with average radii around 10 pixels. These results suggest that affine covariant features are able to model the radial distortion introduced by fisheye images as a local variability.

IV. Evaluation Protocol

Following the protocol in [28], all experiments are performed on series of 6 images $S = \{I_0, I_1, \ldots, I_5\}$ affected by a specific variability. The first image in the series I_0 is affected by the least amount of the considered variability (the zero-variability) and is referred to as the reference image, while the remaining five images $\{I_i\}_{1\leq i\leq 5}$ are affected by increasing amounts of the considered variability and are referred to as test images. Given an image series S, we assess the performances of the detectors on each of the 5 image pairs $\{(I_0, I_i)\}_{1\leq i\leq 5}$ using the reference image to define the ground truth. We assume that for each image pair it is possible to establish a



(a) Harris Affine (rectilinear) (b) Hessian Affine (rectilinear) (c) MSER (rectilinear)

(d) Harris Affine (rectilinear) (e) Hessian Affine (rectilinear) (f) MSER (rectilinear)

(g) Harris Affine (distorted) (h) Hessian Affine (distorted) (i) MSER (distorted)

(l) Harris Affine (distorted) (m) Hessian Affine (distorted) (n) MSER (distorted)

Fig. 7. (a) to (c) Normalized histograms of average radii of regions extracted by the three detectors on the rectilinear images of Dataset A. (d) to (f) Normalized cumulative histograms of average radii of regions extracted by the three detectors on the rectilinear images of Dataset A. (g) to (i) Normalized histograms of the average radii of the regions extracted by the three detectors on the distorted images of Dataset A. (l) to (n) Normalized cumulative histograms of the average radii of the regions extracted by the three detectors on the distorted images of Dataset A.

mapping ψ_{i0} between the points of the test image I_i and the ones of the reference image I_0. Specifically, for Dataset A, such mapping is given by the inverse of the distortion function f_{0i} (3) used to generate the test image from the reference one:

$$\psi_{i0}^{A} = f_{0i}^{-1}. \tag{15}$$

The Oxford dataset provides homographies h_{0i} relating the reference image I_0 to the test images I_i. Hence, for the undistorted series contained in Dataset B, we define:

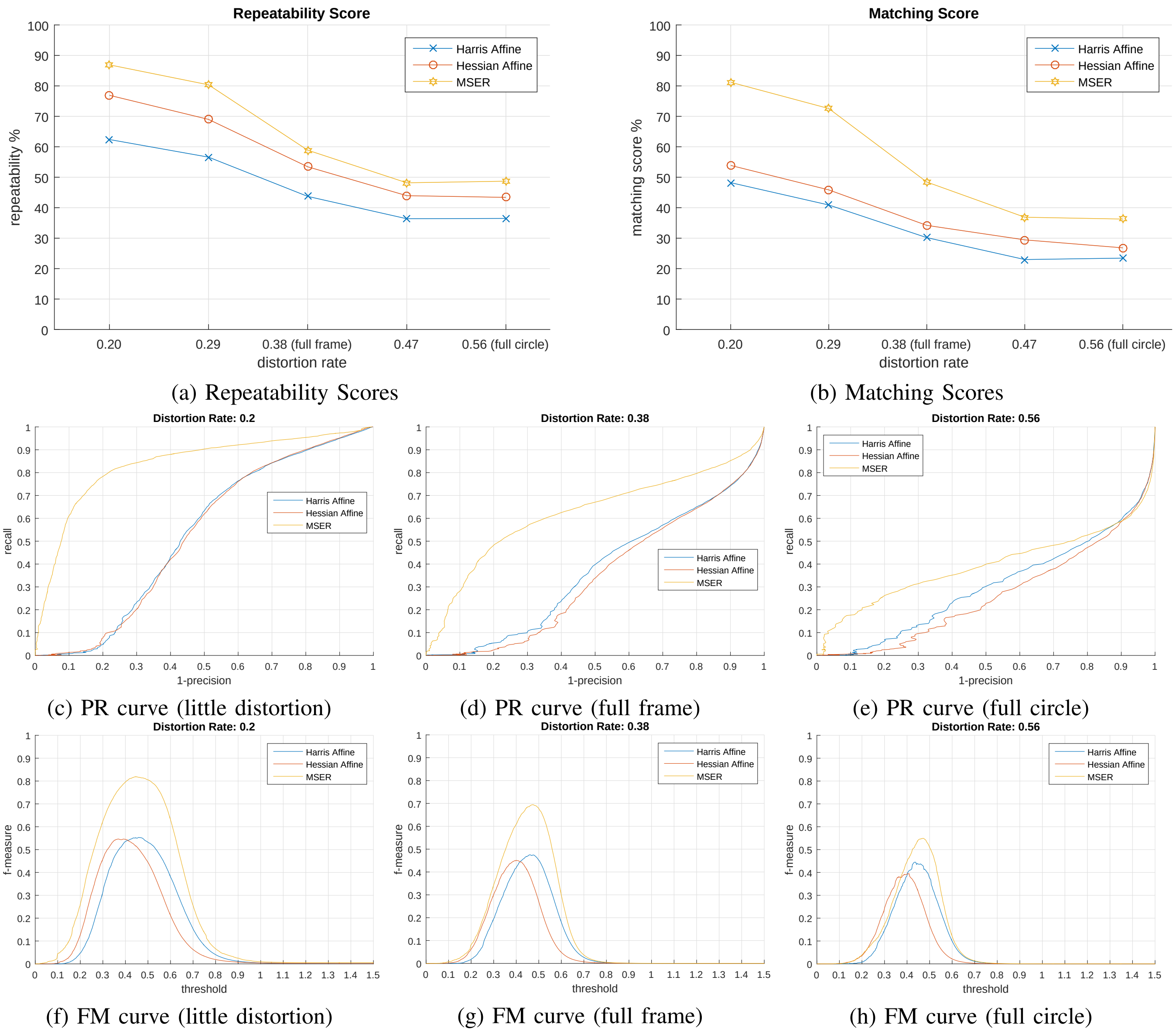
$$\psi_{i0}^{B1} = h_{0i}^{-1}. \tag{16}$$

In the case of the distorted series of Dataset B, instead, the projection from the distorted test image I_i to the undistorted reference image I_0 is carried through the following composition:

$$\psi_{i0}^{B2} = f_i^{-1} \circ h_{0i}^{-1} \tag{17}$$

where f_i is the distortion function used to generate the distorted test image I_i.

As proposed by Mikolajczyk [28], we measure two important properties of the affine detectors under analysis: the repeatability, i.e., the ability to extract regions which corre-



(a) Repeatability Scores

(b) Matching Scores

(c) PR curve (little distortion)

(d) PR curve (full frame)

(e) PR curve (full circle)

(f) FM curve (little distortion)

(g) FM curve (full frame)

(h) FM curve (full circle)

Fig. 8. Results related to Dataset A - increasing radial distortion. All numbers are obtained averaging the results for the 50 image series of Dataset A. (a) Repeatability scores for different amounts of radial distortion. (b) Matching scores for different amounts of radial distortion. (c) to (e) 1-precision vs recall (PR) curves for different amounts of radial distortion. (f) to (h) threshold vs F-measure (FM) curves for different amounts of radial distortion.

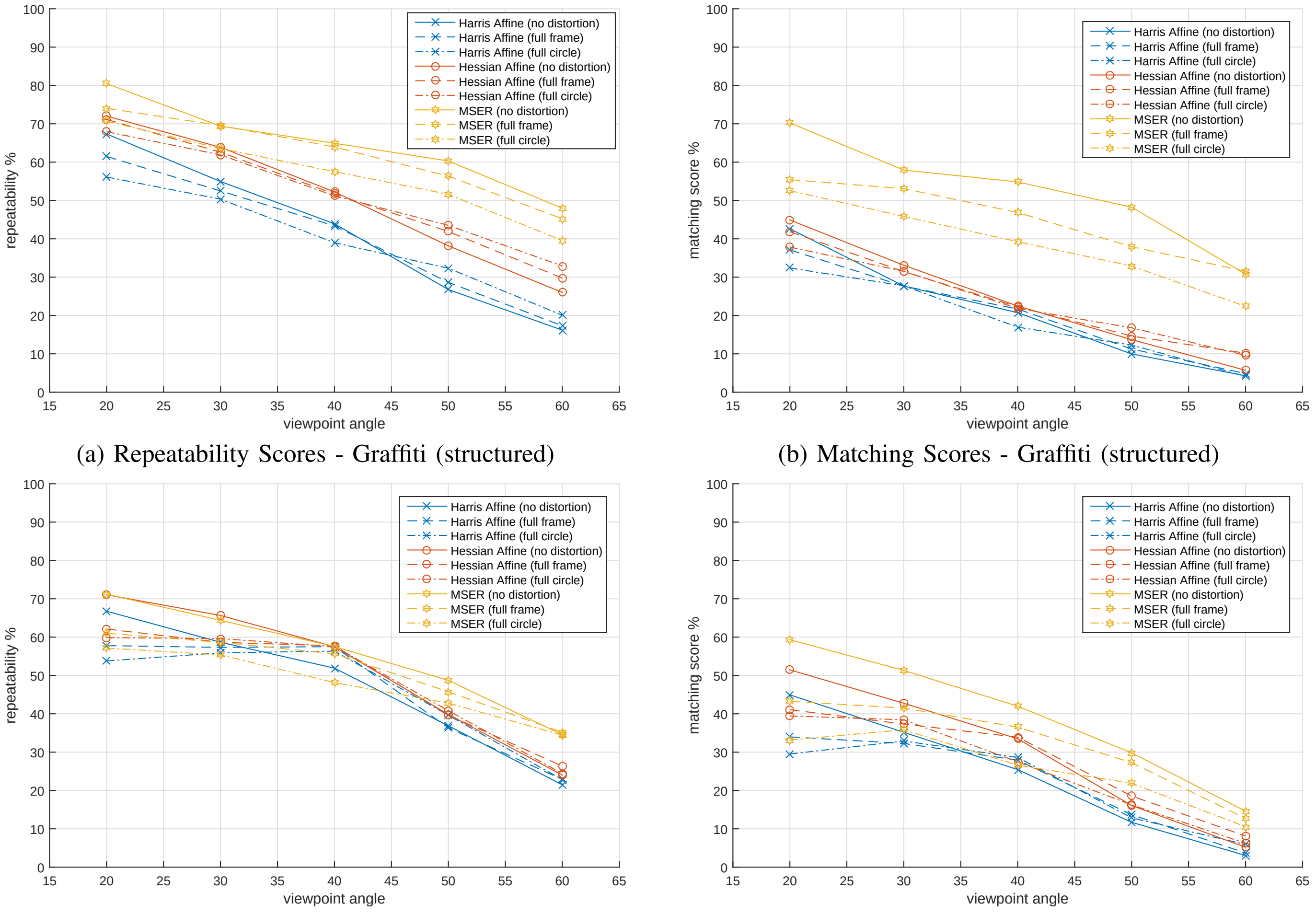
spond to the same geometrical areas under the considered variabilities, and the matching ability, which is the ability to extract distinctive regions that, given a suitable descriptor, can be matched reliably under the considered variabilities.

A. Repeatability

Let be $\mathcal{D}$ the affine region detector under analysis and let be $\mathcal{F}_i = \mathcal{D}(I_i)$ the set of elliptical features extracted from the generic image I_i using detector $\mathcal{D}$. Since the projection of an ellipse using a distortion function in the form of equation (3) is not an ellipse in general, we sample the elliptical features at an angular step of $\frac{\pi}{30}$ in order to obtain the set of polygonal regions $\mathcal{R}_i$. The repeatability of detector $\mathcal{D}$ is assessed counting how many test regions in $\mathcal{R}_i$ overlap significantly with the reference regions in $\mathcal{R}_0$. In order to measure the overlap, the test regions are first mapped to the reference space using the mapping function ψ_{i0}:

$$\mathcal{R}_{i0} = \{r' = \psi_{i0}(r), \forall r \in \mathcal{R}_i\} \tag{18}$$

where r is a polygon and $\psi_{i0}(r)$ is the point-wise projection of r through the mapping function ψ_{i0}. It should be noted that, even if the reference and test images I_0 and I_i are related by the mapping ψ_{i0}, in general they don't cover the same physical areas and hence not all the regions in the sets $\mathcal{R}_0$ and $\mathcal{R}_{i0}$ are guaranteed to lay in the part of the scene present in both images. Given the generic set of regions $\mathcal{R}$, we denote with the notation $\mathcal{R}^{(0,i)}$ the subset of regions of $\mathcal{R}$ entirely contained in the common part of the scene of images I_0 and I_i. For each pair of regions $(r_h, r_k) : r_h \in \mathcal{R}_0^{(0,i)}, r_k \in \mathcal{R}_{i0}^{(0,i)}$,



(a) Repeatability Scores - Graffiti (structured)

(b) Matching Scores - Graffiti (structured)

(c) Repeatability Scores - Wall (textured)

(d) Matching Scores - Wall (textured)

Fig. 9. Results related to Dataset B - change of viewpoint angle for both a structured (“graffiti”) and a textured (“wall”) scenes. (a) Repeatability scores for the “graffiti” image series (structured scene). (b) Matching scores for the “graffiti” image series (structured scene). (c) Repeatability scores for the “wall” image series (textured scene). (d) Matching scores for the “wall” image series (textured scene).

we compute the overlap error as following:

$$err_{hk} = 1 - \frac{area(\alpha \cdot r_h \cap \alpha \cdot r_k)}{area(\alpha \cdot r_h \cup \alpha \cdot r_k)} \qquad (19)$$

where α is a scaling factor such that $area(\alpha \cdot r_h) = \pi r^2$ and r is a normalized radius. Following the protocol of [28], we set $r = 30$ pixels. Unions, intersections and areas are computed numerically. In order to compute the set of most likely correspondences x_{hk} between regions $r_h \in \mathcal{R}_0^{(0,i)}$ and $r_k \in \mathcal{R}_i^{(0,i)}$, such that the overlap error in Eq. (19) between r_h and r_k is under a given overlap threshold o_t, we solve the following assignment problem using the Hungarian algorithm [38]:

$$\begin{cases} \min(\sum_{hk} e_{hk} x_{hk}) & \\ \sum_k x_{hk} \leq 1 & \forall h : 1 \leq h \leq |\mathcal{R}_0^{(0,i)}| \\ \sum_h x_{hk} \leq 1 & \forall k : 1 \leq k \leq |\mathcal{R}_{i0}^{(0,i)}| \\ x_{hk} \; \in \{0,1\} & \forall h, \; \forall k : 1 \leq h \leq |\mathcal{R}_0^{(0,i)}| \\ & \qquad \wedge 1 \leq k \leq |\mathcal{R}_{i0}^{(0,i)}| \end{cases} \qquad (20)$$

where:

$$e_{hk} = \begin{cases} err_{hk} & \text{if } err_{hk} \leq o_t \\ +\infty & \text{otherwise} \end{cases}. \qquad (21)$$

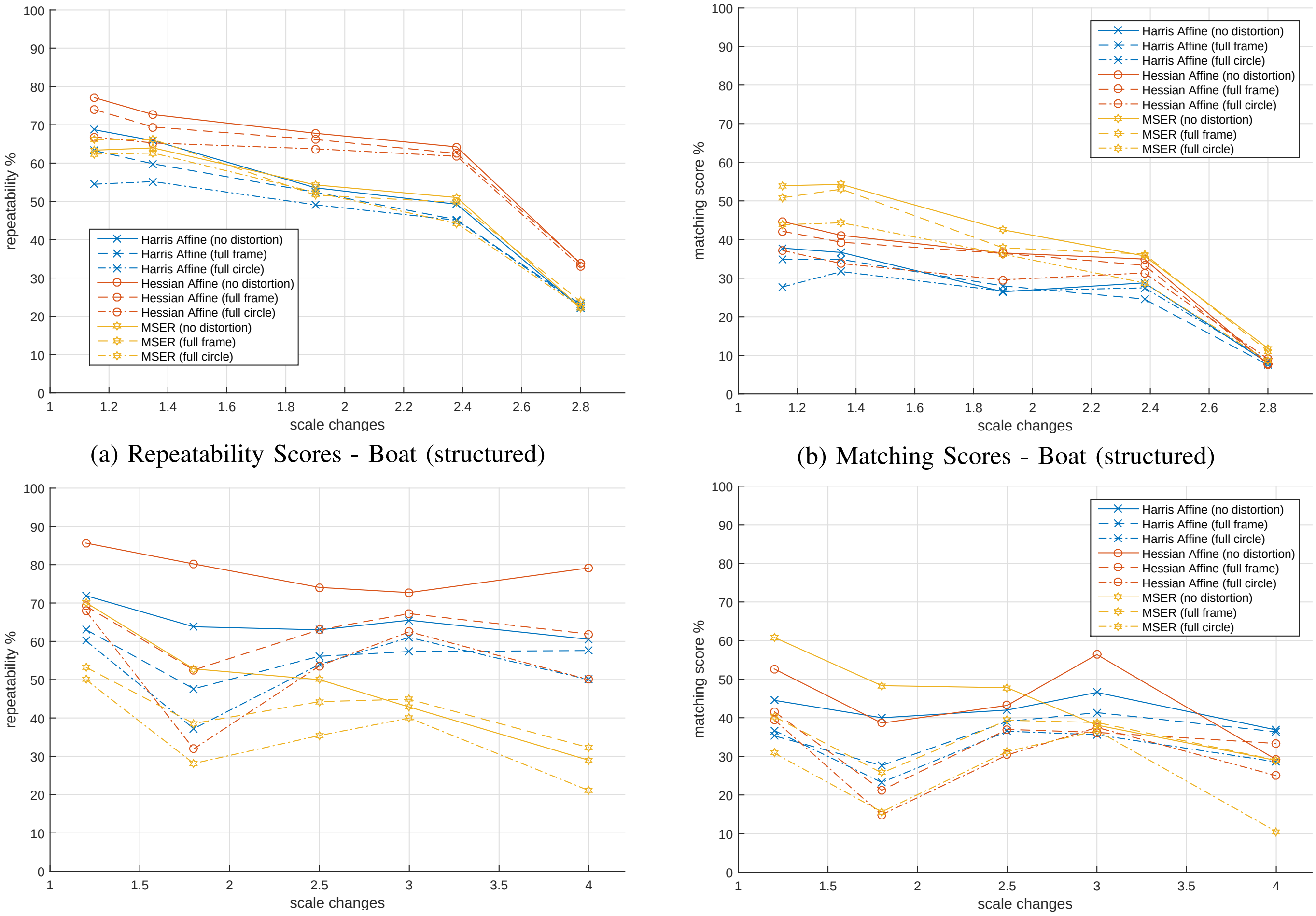
Threshold o_t is set to $o_t = 0.4$ as discussed and justified in [28]. The repeatability score is defined as the number of correspondences normalized by the minimum number of regions detected in the two images (excluding the regions not entirely contained in the common part):

$$repeatability\ score = \frac{\sum_{hk} x_{hk}}{\min(|\mathcal{R}(I_0)^{(i,0)}|, |\mathcal{R}(I_{i0})^{(i,0)}|)}. \qquad (22)$$

The repeatability score measures the ability of the detector to extract features corresponding to the same geometrical regions under varying amounts of a given variability.

B. Matching Ability

In order to measure the matching ability of the detectors, we count how many test features in $\mathcal{F}_i$ are correctly matched to the reference features in $\mathcal{F}_0$ given a suitable descriptor. The ground truth matchings are given by the correspondences x_{hk} computed solving the assignment problem in (20). Each ellip-



(a) Repeatability Scores - Boat (structured)

(b) Matching Scores - Boat (structured)

(c) Repeatability Scores - Bark (textured)

(d) Matching Scores - Bark (textured)

Fig. 10. Results related to Dataset B - scale changes for both a structured (“boat”) and a textured (“bark”) scenes. (a) Repeatability scores for the “boat” image series (structured scene). (b) Matching scores for the “boat” image series (structured scene). (c) Repeatability scores for the “bark” image series (textured scene). (d) Matching scores for the “bark” image series (textured scene). The legend of (d) applies to (c) as well.

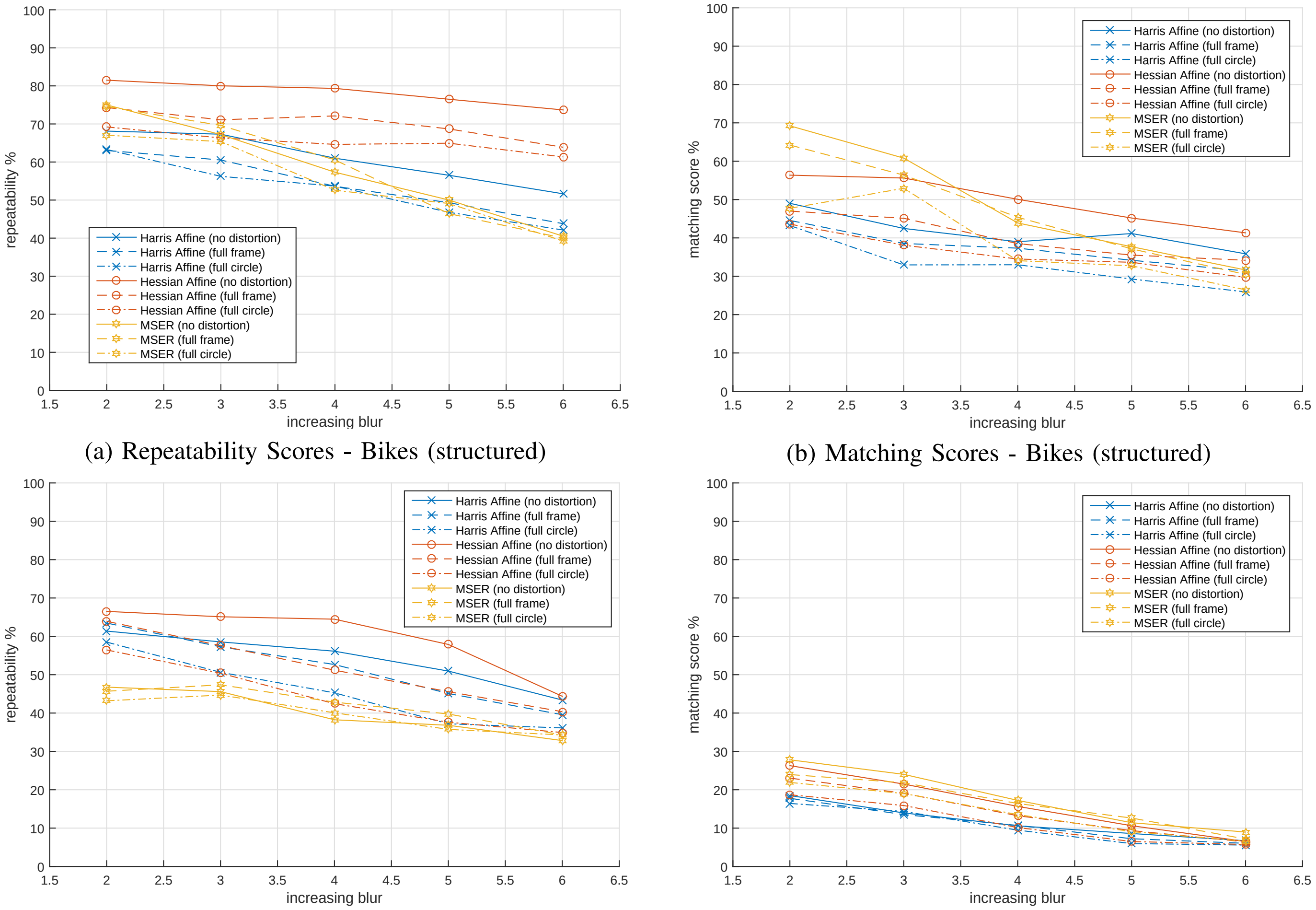
tical feature is normalized to a circular region of dimensions 20×20 pixels and the Local Intensity Order Pattern (LIOP) descriptor is computed over that region [30]. We compute the nearest neighbour matchings between the reference and test descriptors and denote them by m_{hk}, where $m_{hk} = 1$ if f_h matches f_k in the nearest neighbour sense and $m_{hk} = 0$ otherwise. The matching ability is defined as the number of correct nearest neighbour matchings normalized by the minimum number of regions detected in the two reference and test images (excluding the regions not entirely contained in the common part):

$$matching\ score = \frac{\sum_{hk}(m_{hk} \cdot x_{hk})}{\min(|\mathcal{R}(I_0)^{(i,0)}|, |\mathcal{R}(I_{i0})^{(i,0)}|)}. \quad (23)$$

The matching score measures the ability of the detector to extract distinctive features, i.e., regions which can be reliably described and matched under different variabilities. As pointed out in [28], the matching results should follow the repeatability scores if the regions extracted are distinctive. It should be noted that we use the LIOP descriptor to compute the matching ability instead of using the standard SIFT algorithm as proposed by Mikolajczyk et al. in [28]. Our choice is motivated by recent studies [30] in which the LIOP descriptor outperforms SIFT on the Oxford dataset (corresponding to Dataset B in this paper) and supplementary image pairs with complex illumination changes. Since we are benchmarking the ability of the detectors to extract highly distinctive features and we are not interested in assessing the performances of the descriptors themselves, we choose LIOP as the best performing algorithm up-to-date for our evaluations.

C. *Precision-Recall Curves*

To better assess the matching ability of the detectors with respect to increasing radial distortion (Dataset A), we also compute 1-precision vs recall (PR) curves following the scheme proposed in [29]. According to such scheme, two descriptors match if their euclidean distance is smaller than a given threshold t. Each test descriptor is compared with each reference descriptor and the number of false and correct matchings is counted in order to compute the precision and recall values corresponding to threshold t using the following



(a) Repeatability Scores - Bikes (structured)

(b) Matching Scores - Bikes (structured)

(c) Repeatability Scores - Trees (textured)

(d) Matching Scores - Trees (textured)

Fig. 11. Results related to Dataset B - increasing blur for both a structured (“bikes”) and a textured (“trees”) scenes. (a) Repeatability scores for the “bikes” image series (structured scene). (b) Matching scores for the “bikes” image series (structured scene). (c) Repeatability scores for the “trees” image series (textured scene). (d) Matching scores for the “trees” image series (textured scene).

formulas:

$$precision = \frac{\#correct\ matchings}{\#matchings} \quad (24)$$

$$recall = \frac{\#correct\ matchings}{\#correspondences}. \quad (25)$$

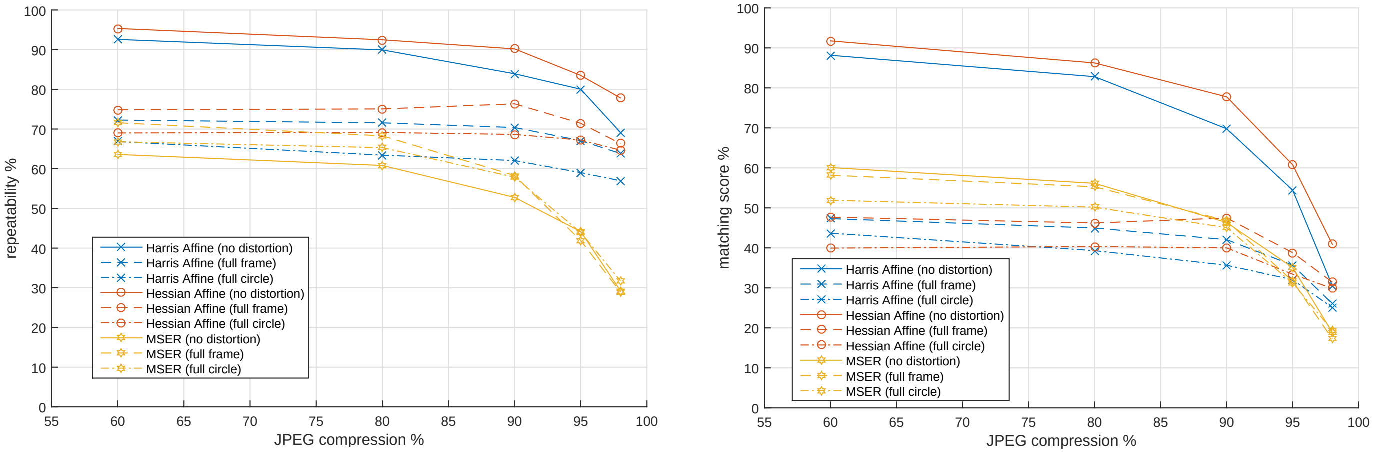
The curves are obtained varying the threshold t. An ideal 1-precision vs recall curve would have recall equal to 1 for any precision, while in practice the recall increases as the precision decreases. A steep curve denotes a detector able to produce distinctive regions with a reduced amount of non-distinctive regions. We also report the threshold vs F-measure (FM) curves, where the F-measure is computed as follows [39]:

$$F_\beta = \frac{(1+\beta^2) precision \times recall}{\beta^2 \times precision + recall} \quad (26)$$

where $\beta^2 = 0.3$ to weigh precision more than recall. The threshold vs F-measure curves have a retrieval-based interpretation: a good curve would have a high peak for a small threshold, indicating that a high number of regions can be retrieved with little noise.

D. Note on the Normalization Scheme

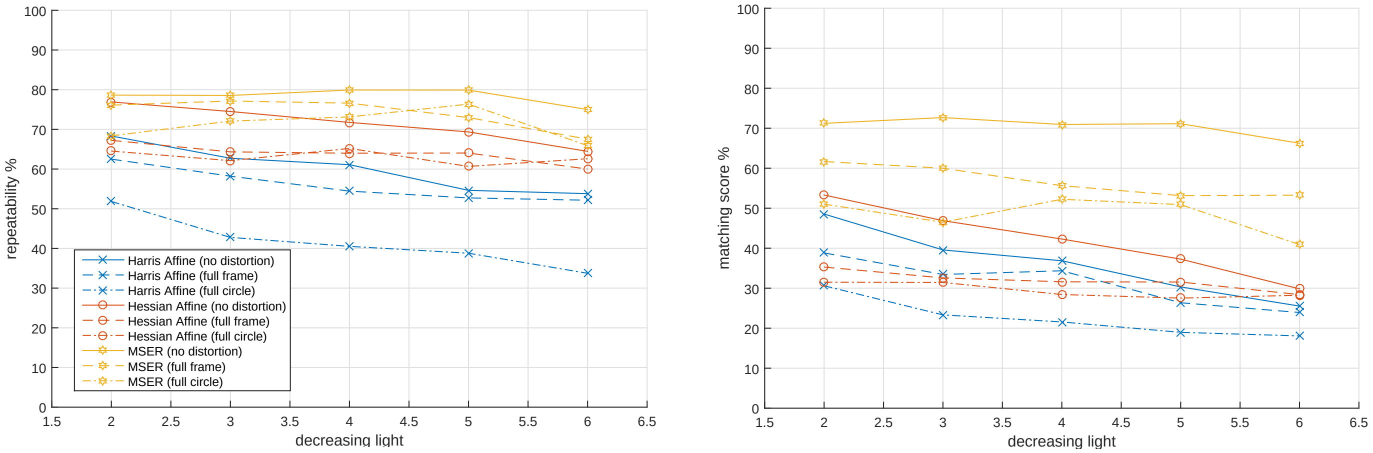
The repeatability and matching scores reported in Eq. (22) and (23) are defined normalizing the number of correspondences and matchings by the minimum number of regions detected in the test and reference images. Such normalization scheme, proposed in [28] and fully recognized by the Computer Vision community, is based on the observation that the chosen normalization value is the maximum number of correspondences or matchings which it is possible to achieve. This normalization scheme accounts for those situations in which, due to an extreme amount of the considered variability (e.g., increasing radial distortion, change of viewpoint angle), most of the regions extracted from the reference image are unlikely to be detected by any algorithm in the test image since they are represented by just a few pixels. Nevertheless, it should be noted that, according to such definitions, the scores referring to the same image series but different test images are not in general normalized by the same number. As it shall be clearer later on in our analysis, for this reason, the scores related to different test images of the same series are not directly comparable in a quantitatively fashion and the reported results should be considered indicative instead as pointed out



(a) Repeatability Scores - UBC

(b) Matching Scores - UBC

Fig. 12. Results related to Dataset B - increasing JPEG compression. (a) Repeatability scores for the “UBC” image series. (b) Matching scores for the “UBC” image series.



(a) Repeatability Scores - Leuven

(b) Matching Scores - Leuven

Fig. 13. Results related to Dataset B - decreasing light. (a) Repeatability scores for the “leuven” image series. (b) Matching scores for the “leuven” image series. The legend of (a) applies to (b) as well.

in [28].

V. Experimental Results and Discussion

We have performed three sets of experiments using the datasets described in Section II-B. The first set of experiments is aimed at assessing the robustness of the detectors to increasing amounts of radial distortion and is performed on dataset A. We take advantage of the repetition of the tests on a number of images depicting scenes belonging to different categories, to draw general conclusions on the performances of the detectors. The second set of experiments is aimed at assessing the performances of the considered detectors when the variabilities present in the Oxford dataset are combined to the radial distortion and it is performed on Dataset B. The third set of experiments is performed on Dataset C and is intended to complement the analysis on images acquired using real fisheye lenses. Fig. 8 to 13 report the results of the experiments described in Section IV. It should be noted that, due to the normalization scheme discussed in Section IV-D, the curves related to the repeatability and matching scores are not guaranteed to be strictly monotonically decreasing with respect to increasing amounts of a considered variability as the reader could expect. As pointed out earlier and in [28], such results have an indicative rather than quantitative value and the reader is advised to focus more on the general trends of the presented curves rather than on local configurations. In the following, results related to the three sets of experiments are discussed.

Fig. 8 reports the results related to Dataset A. All the numbers have been obtained by averaging the results for the 50 image series of Dataset A depicting different scene types. This allows us to draw general conclusions on the performances of the detectors under analysis. Fig. 8 (b) and (c) show that all detectors retain good performances for increasing fisheye distortion. Interestingly, MSER clearly outperforms the other detectors on both the repeatability and matching tests. In particular, the superior performances of MSER in the matching test highlight that the regions extracted by MSER tend to be more distinctive than the ones extracted by the competitors under the influence of radial distortion. This observation is

TABLE I
RESULTS RELATED TO DATASET C - IMAGES ACQUIRED USING REAL FISHEYE LENSES.

Series	**Repeatability %**			**Matching ability %**		
Affine Detector	Harris	Hessian	MSER	Harris	Hessian	MSER
S1 ($d = 0.13$)	61.94	69.34	74.73	36.09	39.14	59.20
S2 ($d = 0.19$)	60.14	71.00	72.24	32.32	37.33	54.23
S3 ($d = 0.54$)	23.54	27.97	32.88	12.80	13.65	25.68
S1 ($d = 0.13$), rect	68.00	75.47	77.22	40.28	41.73	62.29
S2 ($d = 0.19$), rect	63.10	73.88	73.97	33.55	38.74	53.53
S3 ($d = 0.54$), rect	43.41	52.91	57.32	26.87	24.99	44.37

strengthen by the 1-precision vs recall and threshold vs F-measure curves shown in Fig. 8 (d) to (i). Moreover, the decays of the curves shown in Fig. 8 (b) and (c) are reminiscent of the results related to the robustness of the detectors with respect to affine variabilities such as the change of viewpoint angle (solid lines in Fig. 9). This observation supports our premise that affine covariant region detectors can locally model the radial distortion introduced by a fisheye camera as an affine variability (Section III-A).

Fig. 9 to 13 show the results related to Dataset B. Each figure reports the repeatability and matching scores related to a specific variability (i.e., change of viewpoint angle, scale change, increasing blur, JPEG compression, decreasing light). Specifically, each figure reports the results related to the original series (no radial distortion is introduced) of the Oxford dataset (solid lines), the results related to the series to which a full frame distortion is added (dashed lines) and the results related to the series to which a full circle distortion is added (dot-dashed lines). It should be noted that, since the reference image is never distorted in Dataset B, in each plot all the data series are related to the same reference image. The results are in line with [28] also when radial distortion is added; no detector performs systematically better than the competitors on all the image series and the relative ordering of the curves tends to change for the structured and textured scenes even when the variability under analysis is the same. However, some general considerations are possible. The combination of radial distortion and the variabilities present in the Dataset B, i.e., Oxford dataset (dashed and dot-dashed lines) degrades the performances of the detectors. Nevertheless, the curves related to the distorted series are often characterized by decays and relative ordering similar to the ones of the original series not affected by distortion (solid lines). This is especially true for the structured scenes both for the repeatability and matching scores (Fig. 9 to 13 (c) - (d)). This observation is a further evidence of how the introduction of the fisheye distortion is in most of the cases handled by the detectors as an additional variability to cope with. As general remarks, moreover, the Hessian Affine detector achieves the best repeatability results in most of the configurations, while the MSER detector extracts highly distinctive regions in all the cases (i.e., the matching results follow the repeatability results).

TABLE I reports the results related to dataset C. For each image series and considered feature detector, we report the average repeatability and matching ability scores over the 78 image pairs. The last three rows report results obtained performing a rectification step prior to extracting affine covariant features from the images. The results reported in TABLE I confirm the general findings discussed earlier in this section. In particular, repeatability and matching computed on real fisheye images are generally lower, but still consistent with the ones reported in Fig. 9 (viewpoint change + radial distortion) and Fig. 10 (scale and rotation transformations + radial distortion). As observed in the previous experiments, regions extracted by the MSER detector are highly distinctive (matching scores follow the trend of repeatability scores). In agreement with the experiments performed on Dataset A, the MSER detector systematically outperforms the competitors both in terms of repeatability and matching ability. Moreover, when the distortion rate is low (i.e., S1 and S2 in TABLE I), affine covariant feature detectors perform reasonably well directly on fisheye images as compared to employing rectification. In the case of low distortion, in fact, using affine covariant region detectors directly implies an average performance drop under the 3% with respect to both repeatability and matching ability scores, which suggests that radial distortion is successfully modelled as an additional affine variability. When distortion is severe (i.e., S3 in TABLE I), performing rectification allows to improve both repeatability and matching ability by a good margin, leading to average gains of about 23% for repeatability and 15% for matching ability. It should be noted that, even in the case of severe distortion, results obtained on Dataset C are still coherent with those obtained on Dataset B, suggesting that direct employment of affine covariant region detectors on fisheye images is able to produce usable results. This can be particularly useful when rectification is not a viable option, e.g., when the camera is not known (and hence cannot be calibrated) in advance.

VI. CONCLUSION

We have studied the applicability of affine covariant region detectors on fisheye images. Relying on the Division Model for modelling the radial distortion introduced by fisheye cameras, we have provided both theoretical and experimental evidence that affine region detectors can successfully deal with radial distortion as a local affine transformation. Specifically, inspired by the work by Mikolajczyk et al. [28], we have designed a series of experiments aimed at assessing the performances of three popular region detectors, i.e., MSER, Hessian Affine, Harris Affine, with respect to increasing radial distortion. We have also tested the combination of the variabilities included in the Oxford dataset with two different degrees of radial distortion and performed testes on images acquired using three real fish-eye lenses. Interestingly, MSER outperformed the Hessian and Harris affine region detectors in both the repeatability and matching tests in the experiments related to the increasing radial distortion and on images acquired using real fisheye lenses. The evaluations carried on the Oxford dataset have shown that the detectors behave consistently when the scene variability is combined with the radial distortion, providing further evidence that radial distortion is effectively modelled as an additional affine variability by the detectors. Tests on images acquired using real fisheye lenses show that affine region detectors are able to handle low levels of radial distortion making rectification avoidable. When distortion is severe, affine region detectors yield results consistent with the ones obtained in the presence of strong scale and rotation transformation with artificially distorted images. Our analysis can be exploited in all the application domains where the input images are acquired by unknown, non-calibrated cameras (both fisheye or rectilinear).

References

[1] C. Hughes, P. Denny, E. Jones, and M. Glavin. Accuracy of fish-eye lens models. *Applied Optics*, 49(17):3338–3347, 2010.

[2] L. Puig and J. J. Guerrero. *Omnidirectional Vision Systems*. Springer, 2013.

[3] Z. Arican and P. Frossard. OmniSIFT: Scale invariant features in omnidirectional images. In *International Conference on Image Processing*, pages 3505–3508, 2010.

[4] C. Hughes, M. Glavin, E. Jones, and P. Denny. Wideangle camera technology for automotive applications: a review. *IET Intelligent Transport Systems*, 3(1):19–31, 2009.

[5] S. Battiato, G. M. Farinella, A. Furnari, G. Puglisi, A. Snijders, and J. Spiekstra. Vehicle tracking based on customized template matching. In *VISAPP International Workshop on Ultra Wide Context and Content Aware Imaging*, 2014.

[6] S. Battiato, G. M. Farinella, A. Furnari, G. Puglisi, A. Snijders, and J. Spiekstra. An integrated system for vehicle tracking and classification. *Expert Systems with Applications*, 2015.

[7] D. Scaramuzza, N. Criblez, A. Martinelli, and R. Siegwart. Robust Feature Extraction and Matching for Omnidirectional Images. In *International Conference on Field and Service Robotics*, pages 1–11, 2007.

[8] K. Miyamoto. Fish eye lens. *Journal of the Optical Society of America*, pages 2–3, 1964.

[9] S. Baker and S. K. Nayar. A theory of catadioptric image formation. In *International Conference on Computer Vision*, pages 35–42, 1998.

[10] C. Geyer and K. Daniilidis. A unifying theory for central panoramic systems and practical implications. In *European Conference on Computer Vision*, pages 445–461, 2000.

[11] M. Lourenço, J. P. Barreto, and F. Vasconcelos. sRD-SIFT: Keypoint detection and matching in images with radial distortion. *IEEE Transactions on Robotics*, 28(3):752–760, 2012.

[12] J. Kannala and S. Brandt. A generic camera calibration method for fisheye lenses. In *International Conference on Pattern Recognition*, pages 10–13, 2004.

[13] D. Scaramuzza, A. Martinelli, and R. Siegwart. A Toolbox for Easily Calibrating Omnidirectional Cameras. In *International Conference on Intelligent Robots and Systems*, pages 5695–5701, 2006.

[14] F. Devernay and O. Faugeras. Straight lines have to be straight. *Machine Vision and Applications*, 1:14–24, 2001.

[15] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Computer Vision and Pattern Recognition*, volume 1, 2004.

[16] L. Puig and J. J. Guerrero. Scale space for central catadioptric systems: Towards a generic camera feature extractor. In *International Conference on Computer Vision*, pages 1599–1606, 2011.

[17] I. Bogdanova, X. Bresson, J. Thiran, and P. Vandergheynst. Scale space analysis and active contours for omnidirectional images. *IEEE transactions on image processing*, 16(7):1888–901, 2007.

[18] J. Cruz-Mota, I. Bogdanova, B. Paquier, M. Bierlaire, and J. Thiran. Scale Invariant Feature Transform on the Sphere: Theory and Applications. *International Journal of Computer Vision*, 98(2):217–241, 2011.

[19] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[20] P. Hansen, P. Corke, W. Boles, and K. Daniilidis. Scale-invariant features on the sphere. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[21] I. Cinaroglu and Y. Bastanlar. A direct approach for human detection with catadioptric omnidirectional cameras. In *Signal Processing and Communications Applications Conference*, pages 2275–2279, 2014.

[22] M. S. Islam and L. J. Kitchen. Straight-edge extraction in distorted images using gradient correction. In *Digital Image Computing: Techniques and Applications*, 2009.

[23] A. Furnari, G. M. Farinella, A. R. Bruna, and S. Battiato. Generalized Sobel filters for gradient estimation of distorted images. In *IEEE International Conference on Image Processing*, pages 3250–3254, 2015.

[24] K. Koser and R. Koch. Perspectively Invariant Normal Features. In *International Conference on Computer Vision*, pages 1–8, 2007.

[25] L. Puig, J. J. Guerrero, and P. Sturm. Hybrid matching of uncalibrated omnidirectional and perspective images. In *Informatics in Control, Automation and Robotics*, 2002.

[26] M. Saito, K. Kitaguchi, G. Kimura, and M. Hashimoto. People Detection and Tracking from Fish-eye Image Based on Probabilistic Appearance Model. *Society of Instrument and Control Engineers Annual Conference*, pages 435–440, 2011.

[27] J. Masci, D. Migliore, M. M. Bronstein, and J. Schmidhuber. Descriptor learning for omnidirectional image matching. In *Registration and Recognition in Images and Videos*, volume 532 of *Studies in Computational Intelligence*, pages 49–62. Springer Berlin Heidelberg, 2014.

[28] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision (IJCV)*, 65:43–72, 2005.

[29] K. Mikolajczyk and C. Schmid. Performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–30, 2005.

[30] Z. Wang, B. Fan, and F. Wu. Local Intensity Order Pattern for Feature Description. In *International Conference on Computer Vision*, pages 603–610, 2011.

[31] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2002.

[32] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, 2002.

[33] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision (IJCV)*, 60(1):63–86, 2004.

[34] A. Furnari, G. M. Farinella, G. Puglisi, A. R. Bruna, and S. Battiato. Affine region detectors on the fisheye domain. In *IEEE International Conference on Image Processing*, 2014.

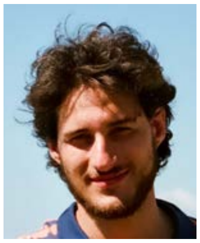
[35] C. Hughes, E. Jones, M. Glavin, and P. Denny. Validation of Polynomial-based Equidistance Fish-Eye Models. In *Signals and Systems Conference*, 2009.

[36] J. J. Kumler and M. L. Bauer. Fish-eye lens designs and their relative performance. In *International Symposium on Optical Science and Technology*, pages 360–369, 2000.

[37] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391–412, 2003.

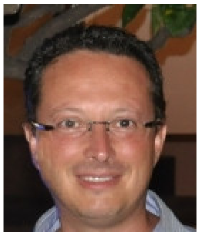
[38] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly (NRL)*, 2(1-2):83–97, 1955.

[39] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.



Antonino Furnari received his bachelor degree and his master degree (both summa cum laude) from the University of Catania in 2010 and 2013 respectively. In 2012 he joined the IPLab and since 2013 he is a Computer Science PhD student at the University of Catania under the advisorship of Prof. Sebastiano Battiato and Dr. Giovanni Maria Farinella. His main research interests concern Computer Vision and Pattern Recognition and include First Person Vision, Computer Vision algorithms for Fisheye Cameras and Visual Saliency. Since 2013 he is involved in the ENIAC Joint Undertaking Panorama European Project and in a collaboration with ST Microelectronics (IPLab-STMicroelectronics joint laboratory). In 2013 he has been awarded the ICVSS Reading Group Competition prize sponsored by the University of California, Los Angeles. In 2014 he has joined a collaboration between UNICT-DMI and INGV Catania on the analysis of SAR images acquired from the Mt. Etna. In 2015 he has been Technical Program Committee of the third workshop on Assistive Computer Vision and Robotics (ACVR) held in conjunction with the International Conference of Computer Vision (ICCV). In 2015 he held the Web Application II course at the ITS Steve Jobs institute. In 2016 he has been Local Arrangments Chair of Variational Inequalities, Nash Equilibrium Problems and Applications (VINEPA) 2016.

Giovanni Maria Farinella (M'11–SM'16) is Assistant Professor at the Department of Mathematics and Computer Science, University of Catania, Italy. He received the (egregia cum laude) Master of Science degree in Computer Science from the University of Catania in April 2004. He was awarded the Ph.D. in Computer Science from the University of Catania in October 2008. From 2008 he serves as a Contract Professor of Computer Science for undergraduate courses at the University of Catania. He is also an Adjunct Professor at the School of the Art of Catania in the field of Computer Vision for Artists and Designers (Since 2004). From 2007 he is a research member of the Joint Laboratory STMicroelectronics - University of Catania, Italy. His research interests lie in the field of Computer Vision, Pattern Recognition and Machine Learning. He is author of one book (monograph), editor of 4 international volumes, editor of 2 international journals, co-author of 90 papers in international book chapters, international journals and international conference proceedings, and co-author of 18 papers in national book chapters, national journals and national conference proceedings. He is co-inventor of 6 patents involving industrial partners. Dr. Farinella serves as a reviewer and on the board programme committee for major international journals and international conferences. He has been Video Proceedings Chair for the International Conferences ECCV 2012 and ACM MM 2013, General Chair of the International Workshop on Assistive Computer Vision and Robotics (ACVR) held in conjunction ECCV 2014 and ICCV 2015, and chair of the International Workshop on Multimedia Assisted Dietary Management (MADiMa) 2015. He has been Speaker at international events, as well as invited lecturer at industrial institutions. Giovanni Maria Farinella founded (in 2006) and currently directs the International Computer Vision Summer School (ICVSS). He also founded (in 2014) and currently directs the Medical Imaging Summer School (MISS). Dr. Farinella is an IEEE/CVF/IAPR/GIRPR/AIxIA/BMVA member.



Arcangelo Ranieri Bruna received the Italian degree in electronic engineer (summa cum laude) from the University of Palermo, Palermo, Italy, in 1998, and the Ph.D. degree in applied mathematics from the University of Catania, Catania, Italy. He worked for a telecommunication company in Rome. He then joined STMicroelectronics, Catania, in 1999, where he is manager of advanced R&D projects with the Advanced System Technology Catania Laboratory. His current research interests are in the areas of computer vision, digital image processing from the physical digital acquisition to the final image compression, and image forensics. He has authored several papers and book chapters, and he is co-author of 30 international patents on these activities. He serves as a reviewer and program committee for several international journals and international conferences.

Sebastiano Battiato (M'04–SM'06) received his degree in computer science (summa cum laude) in 1995 from University of Catania and his Ph.D. in computer science and applied mathematics from University of Naples in 1999. From 1999 to 2003 he was the leader of the Imaging team at STMicroelectronics in Catania. He joined the Department of Mathematics and Computer Science at the University of Catania as assistant professor in 2004 and became associate professor in the same department in 2011. His research interests include image enhancement and processing, image coding, camera imaging technology and multimedia forensics. He has edited 6 books and co-authored about 200 papers in international journals, conference proceedings and book chapters. Guest editor of several special issues published on International Journals. He is a co-inventor of 22 international patents, reviewer for several international journals, and he has been regularly a member of numerous international conference committees. Prof. Battiato has participated as principal investigator in many international and national research projects. Chair of several international events (ACIVS 2015, VAAM2014-2015, VISAPP2012-2015, IWCV2012, ECCV2012, ICIAP 2011, ACM MiFor 2010-2011, SPIE EI Digital Photography 2011-2012-2013, etc.). He is an associate editor of the SPIE Journal of Electronic Imaging. He is the recipient of the 2011 Best Associate Editor Award of the IEEE Transactions on Circuits and Systems for Video Technology. He is director (and co-founder) of the International Computer Vision Summer School (ICVSS), Sicily, Italy. He is a senior member of the IEEE.