Abstract—Pan-tilt-zoom (PTZ) cameras become popular over the last years in activities such as traffic monitoring, sports analysis and surveillance systems due to its ability to cover a wide range of views at a high definition over multiple zoom levels. Having an overall perception of a scene become significantly important in such activities and subsequent applications, thus the need for its panoramic background representation.

This dissertation work is focused on the PTZ panoramic image generation process over multiple zoom levels. The work involves studying different camera calibration algorithms, whose results are then used as input of the panoramic transformation step. Performance metrics are established to assess pixel color errors and perceived quality of the given panoramic transformations in the cube and cylinder based representations. It is further proposed, a variance minimization approach aiming at correcting calibration errors, namely the focal length parameters and consequently, improve the final mosaics.

The last part of this work consists of studying and applying different image blending techniques. This step compensates for exposure differences and reduces generated artifacts in the final images. An image blending method is implemented where the final mosaic is reconstructed from its combined gradient. The main mathematical tool used in the blending method is the Poisson partial differential equation with Dirichlet boundary conditions.

The proposed methods allowed to obtain promising results on panoramic image mosaicing.

Index Terms—PTZ camera, Panorama, Mosaicing, PTZ camera calibration, Image Blending

I. INTRODUCTION

In a world where human life and its dependency on technology grows day by day, computer vision has been providing multiple answers and solutions for existing problems. As an alternative to traditional fixed cameras, Pan-tilt-zoom (PTZ) cameras have been an active case study in activities such as traffic monitoring and surveillance [1]. Moreover, studies over cooperative mapping of multiple PTZ cameras and PTZ panoramic image generation process have been performed in [2] and [3], [4]. The orientation of Pan-tilt-zoom (PTZ) cameras can be remotely controlled as well as its level of zoom and Field of view (FOV). Moreover, these cameras are usually available at a fair price, making it a tempting resource compared to fixed cameras. Although these cameras offer a better solution, their implementation in automated surveillance systems is still restricted due to the insufficiency of automatic schemes and therefore relying on manual labour.

Panoramic images have become significantly important in surveillance tasks as it provides a compact representation of the observed scene. Having a full representation of the background is crucial in activities such as object tracking and background subtraction. In recent years, PTZ cameras revealed themselves as useful resources in panoramic image mosaicing. Compared to traditional image stitching techniques formulated as a multi-image matching problem, PTZ panoramic image generation process makes use of pan, tilt and zoom coordinates as input parameters in the panorama construction process. The building process addresses four key topics: camera calibration, background subtraction, panoramic transformation and image blending, according to [4].

Most applications require camera calibration as camera coordinate parameter estimation is a fundamental step. There are several documented methodologies on camera calibration as are usually separated in two categories: automatic or non-automatic. In [5], Bouguet proposed estimating camera parameters by changing the orientation of a chessboard positioned in front of a camera. In [6], Aziz and Karara established a calibration algorithm named Direct Linear Transformation (DLT) in which the camera projection matrix is estimated through the resolution of a linear system on the matrix entries, relying on a group of 3D points and the matching 2D points.

Later Agapito et al. [7] presented an automatic calibration for rotating and zooming cameras without non-linear optimizations. In [3], Sinha and Pollefeys suggest an automatic calibration method using non-linear optimization of re-projection errors. Their algorithm used a sequence of images taken with a PTZ camera from which its parameters were estimated.

Regarding panoramic image mosaicing, traditional image stitching methodologies [8] set the problem as multi-image matching problem, using invariant local features as Scale-invariant feature transform (SIFT) to look for matches between images. On the other hand, the PTZ panoramic image generation process makes use of camera orientation and zoom level on the panorama construction. Among the most common structures used for panoramic transformation are the cube based representation [3], sphere [9] and cylinder [4].

Image blending is an important step in panoramic image generation process. It compenses exposure differences and reduces artifacts created from small misalignments, fusing the overlapping areas within different images according to [4]. In [9] a spatially-weighting of image pixels is considered whereas in [10], a frequency-adaptive width is used by adopting a band pass (Laplacian) pyramid. The transition widths are set as a function of the pyramid level. This idea is extended in a multi-band blending approach suggested in [8], although very ineffective on large datasets. In [4], an highly efficient
A multi-band blending process is presented, in contrast to what happens in [8]. As an alternative forthright method, for cases where the scene texture is well modelled as a single band, [11] presents a gradient based approach.

This thesis focuses on the study of the PTZ panoramic image generation process at distinct zoom levels, addressing the multiple steps of the building process. For this purpose, a dataset over different zoom levels is acquired from which a sequence of mosaics are then built. Each mosaic is associated with a distinct zoom level.

Mosaicing with PTZ cameras relies on camera orientation and calibration. Thus three calibration algorithms are considered: two classified as automatic and one non-automatic. The automatic calibration algorithms also allow the estimation of the pan and tilt angles associated with each image, and therefore the direct comparison between the camera and estimated odometry. From these three, two calibration methodologies assisted the construction of the panoramas. Both were compared over pasting and visual perception metrics.

A minimization process over the mosaic construction accuracy metrics is also considered by iteratively adjusting camera focal length parameters, and consequently improve the visual perception of the final panorama.

The final part of this thesis consists in improving the final panoramas over its visual perception. Hence, different image blending techniques are studied. An approach over the gradient domain is proposed based in state-of-the-art methodologies, in contrast to image pixel weighting techniques also mentioned in this work.

This paper is organized as follows, section I introduces the topic to be studied, as well as a brief discussion over the state-of-the-art and proposed approach, section II addresses the pin-hole camera model and introduces traditional panoramic image generation processes in addition to state-of-the-art image blending techniques, section III presents PTZ cameras and a sequence of camera calibration methodologies and section IV describes the PTZ panoramic image generation process, starting with the panoramic transformation step. A group of mosaic construction accuracy metrics are also introduced, followed by a minimization process over the variance of pasting. This section ends with the studied image blending techniques. Lastly, section V illustrates the experimental results on the proposed and implemented methodologies, and section VI outlines all the work developed in this work in addition to future work.

II. BACKGROUND

A. Camera Model

Among the different possible representations of a camera, the pin-hole camera model is the one selected for this work. The 3D projective space is mapped to the 2D projective plane meaning that, a 3D point $m = [X\ Y\ Z\ 1]^T$, using homogeneous coordinates, is mapped to its corresponding 2D image point $m = [u\ v\ 1]^T$. This transformation is established by a $3 \times 4$ rank-3 projection matrix $P$, denoted as camera projection matrix. Their relation is given by $m \sim PM$. $P$ can be decomposed in $P = K[R\ t]$ where $K$ represents the camera intrinsic matrix and $R$ and $t$ build the camera extrinsic matrix. The intrinsic matrix is represent as

$$K = \begin{bmatrix} k_u & s & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where $K$ is $3 \times 3$ upper triangular matrix, $k_u$ and $k_v$ are camera horizontal and vertical focal length in pixels respectively and, $u_0$ and $v_0$ are the coordinates of the principal point. $s$ is an additional parameter known as skew, which is a coefficient between the $u$ and $v$ axis often considered $s = 0$. The intrinsic matrix is fixed for a particular camera zoom level, only varying if the zoom level changes. $R$ is a $3 \times 3$ rotation matrix and it describes the orientation of the reference coordinate system with respect to camera reference frame and $t$ is a column vector, $t = [X_f\ Y_f\ Z_f]^T$, representing the reference coordinate system position in camera reference frame.

This type of representation is not very common, as representing both position and orientation of the camera with respect to the reference coordinate system is far more usual. Let $RC$ be the camera orientation with respect to the reference coordinate system and $C$ the position of the camera in the reference coordinate system. Their relationship with $R$ and $t$ can be described as $R = R_C^T$ and $t = -RC$.

Usually, the reference coordinate system is set coincident with the camera reference frame at its homing position ($t = [0\ 0\ 0]^T$). In any case, $R$ is represented as combination of three possible rotations around its three fixed axis.

The pin-hole model is a very accurate model to represent a camera however, it does not take into account radial distortion. Here, straight lines in 3D world are rendered as curves lines on the camera sensor, giving a rounded effect to the final image.

Radial distortion is often very complex to model with high precision, as true lens distortion curves are usually not easy to determine, implying the use of look-up tables or high-order models to represent camera radial distortion, as noted by Fitzgibbon [12]. However taking into account the typical computer vision applications, an approximation to the cameras true distortion functions performs as well as the preciser ones.

1) Simplified Model: Let $m_d = [x_d\ y_d]^T$ be a projection of 3D point into an image point, with normalized image coordinates ($x_d$ and $y_d$ are dimensionless). This point deviates from the real undistorted image point, $m_u = [x_u\ y_u]^T$ due to radial distortion. Their relationship can be described as

$$m_d = D(m_u, k_1, k_2) = L(r)m_u = (1 + k_1r^2 + k_2r^4)m_u$$

where $r$ is the radial distance, $r = \sqrt{x_u^2 + y_u^2}$, $L(r)$ is a radially symmetric distortion factor and the two radial distortion coefficients, $k_1$ and $k_2$. The model here described corresponds to a simplified two coefficient version of the one proposed by Heikkila [13].

B. Panoramic Image Generation Process

As studied in [8], image stitching is a multiple step procedure. The first step, feature detection, selects points of interest
within the images and finds correspondences between them. Traditional feature matching methodologies relied on Harris corners detection, whereas the techniques used today focus on invariant local features as SIFT.

Following feature detection, all matching images must be found. This process is called image matching. Here, homography estimation is performed using Random sample consensus (RANSAC). For this particular case RANSAC takes as inputs the correspondences between images, and tries to estimate image transformation parameters.

Followed by cameras parameter estimation, image straightening and gain compensation, one must deal with the remaining visible overlapping regions between images. This usually happens due to registration errors, parallax effects, vignetting and varous other causes.

Different from panoramic image stitching, PTZ panoramic generation takes advantage of PTZ cameras physical properties to assist the development of such panoramas. In [11] it is taken a four step approach: camera calibration, background subtraction, panoramic transformation and image blending.

### C. Image Blending

Here, a brief introduction to the Poisson blending approach will be stated as most content was derived from [11]. This method undergoes a minimization problem over a specified region of interest in a target image to seamless paste a common shape region of a source image. The solution of this minimization problem involves solving a Poisson partial differential equation with Dirichlet boundary conditions, which specifies the Laplacian of an unknown function over the domain of interest, along with the unknown function values over the boundary of the domain.

Let \( S \) and \( \Omega \) be groups of finite points defined on an infinite discrete grid, where \( S \) can include all pixels of an image or only a sub-group of them. For each pixel \( p \) in \( S \), let \( N_p \) be the group of its four adjacent neighbors which belong to \( S \), let \((p,q)\) denote a pixel pair such that \( q \in N_p \) and let \( f_p \) be the value of \( f \) at \( p \). The interpolant \( f \) of \( f^* \) over \( \Omega \) is the membrane interpolant defined as the solution of the following discrete, quadratic optimization problem

\[
\min_{f^*} \sum_{(p,q) \in \Omega} (f_p - f_q - v_{pq})^2, \text{ with } f_p = f^*_p, \text{ for all } p \in \partial \Omega (3)
\]

where \( v_{pq} \) is is the projection of \( v(\frac{p+q}{2}) \) on the oriented edge \( [p,q] \). Considering the gradient field of a source image \( g \) as the chosen guidance field \( v \), the interpolation is performed under the guidance of

\[
\text{for all } (p,q), v_{pq} = g_p - g_q. (4)
\]

The solution of equation 3 is therefore given by the following set of linear equations

\[
\text{for all } p \in \Omega, |N_p| f_p = \sum_{q \in N_p, q \in \Omega} f_q = \sum_{q \in N_p, q \in \partial \Omega} f^*_q + \sum_{q \in N_p} v_{pq}. (5)
\]

Equation 5 forms a classical sparse, symmetric, positive-definite system. Setting it under the form of \( Ax = b \), it can be solved separately for the different Red, Green and Blue (RGB) channels using iterative solvers as Gauss-Seidel iteration with successive overrelaxation, V-cycle multigrid, among others.

### III. Pan-Tilt-Zoom Camera Modeling

#### A. Pan-Tilt-Zoom Camera

PTZ cameras (see figure 1) can change its zoom level and orientation, through its pan and tilt angles, contrarily to fixed cameras orientation restriction.

Let \( wM = \begin{bmatrix} wX_M & wY_M & wZ_M \end{bmatrix}^T \) be a 3D point in represented in the world reference frame and let \( cM = \begin{bmatrix} cX_M & cY_M & cZ_M \end{bmatrix}^T \) be the same point represented in camera reference frame. \( wM \) can be directly map to \( cM \) through \( cM = RMw \), considering both frames coincident at camera homing position. As stated in section II-A, \( R \) can be written as \( R = R_C^{-1} = R_M^T \). In the particular case of PTZ cameras, \( R_C \) will be a function of the pan and tilt angles. Thus it can be written as

\[
R_C = \begin{bmatrix}
\cos(\alpha) & 0 & \sin(\alpha) \\
\sin(\beta) \sin(\alpha) & \cos(\beta) & -\sin(\beta) \cos(\alpha) \\
-\sin(\alpha) \cos(\beta) & \sin(\beta) & \cos(\beta) \cos(\alpha)
\end{bmatrix} (6)
\]

where pan (\( \alpha \)) is associated with a rotation over the y axis and tilt (\( \beta \)) is associated with a rotation over the x axis, according to figure 1(b).

#### B. Camera Calibration

It was assumed that PTZ cameras could be represented by the pin-hole model, where the extrinsic matrix is represented by the pan and tilt coordinates, as illustrated in equation 5. At its homing position, both the camera and world reference frame are considered coincident (\( t = [0 \ 0 \ 0]^T \)).

1) Matlab Calibration Toolbox: Inspired in the work of Jean Ives Bouguet [5], the calibration toolbox for Matlab estimates the intrinsic and extrinsic parameters of a camera, as well as its radial distortion coefficients, using a planar calibration pattern (chessboard) as the key element of the approach. This calibration method takes advantage of the planar surface of the chessboard and thus characterizes it as \( z = 0 \). Placing the reference frame at one of the corners, all 3D points characterizing the chessboard can be generated knowing the size of each square. Setting the board at \( z = 0 \), any 3D point within the chessboard can be described as
\[ M = [X_M \ Y_M \ 0 \ 1]^T \] in homogeneous coordinates. Looking at the camera model in section II-A the relationship between the 2D points \((m)\) and the 3D points \((M)\) is expressed as
\[ m \sim K[r_1 \ r_2 \ r_3 \ 0] [X_M \ Y_M \ 0 \ 1]^T \]
where \(r_i\) represents the column \(i\) of the rotation matrix. \(r_3\) can be removed, considering its null contribution due to \(Z_M = 0\). This simplification leads to an homography transformation between \(m\) and \(M\) described as \(H = [h_1 \ h_2 \ h_3] \sim K[r_1 \ r_2 \ t]\).
From this equation is written \(r_1 = K^{-1}h_1\) and \(r_2 = K^{-1}h_2\), from which it is derived two equations on the columns of \(H\), \(h_1\) and \(h_2\), considering the properties of rotation matrices \((r_1\) and \(r_2\) are orthonormal meaning that \(r_1^T \cdot r_2 = 0\) and that \(||r_1|| = ||r_2|| = 1\)).

\[ h_1^T K^{-T} K^{-1} h_2 = 0 \]
\[ h_1^T K^{-T} K^{-1} h_1 = h_2^T K^{-T} K^{-1} h_2. \]

Taking equations (8) and one homography \(H\), it is possible to write two constraints on the camera intrinsic parameters. Considering that \(K\) has six parameters, it is necessary at least three homographies to compute \(K\). Both equations (8) and (9) do not depend on \(t\), allowing different board poses within image acquisitions. Multiple homographies can be then generated, where each \(H_j\) is associated with a particular image \(j\) in the dataset.

MATLAB automatically detects the 2D points associated with the chessboard pattern, contrary to what happens in J.Bouquet toolbox. It also automatically generates the 3D points assuming \(z = 0\) for the chessboard. MATLAB calibrates in two separate steps. First solves for the intrinsic and extrinsic parameters in closed form assuming zero radial distortion and then, estimates all parameters including radial distortion coefficients using nonlinear least-squares minimization (Levenberg-Marquardt algorithm).

2) Image Based Calibration: The calibration procedure mentioned hereafter, was developed by Agapito et al. [14]. It focuses on a self-calibration methodology for rotating cameras relying on image texture. Let \(M\) be a 3D world point represented in multiple images as \(m_i = P_i M\). The relationship between the images points associated with \(M\) can be written as
\[ m_i = K_i R_i R_j^{-1} K_j^{-1} m_j = K_i R_{ij} K_j^{-1} m_j, \]
where \(R_{ij} = R_i R_j^{-1}\). Looking at equation (10) it is derived \(H_{ij} = K_i R_{ij} K_j^{-2}\), where \(H_{ij}\) represents an homography transformation. Using SIFT features to find matches \((m_i, m_j)\) between pairs of images \((i, j)\), \(H_{ij}\) can be automatically computed in a least square sense. Considering the rotation matrix properties \((R_{ij} = R_{ij}^{-T})\) one writes

\[ K_i^{-T} K_j^{-1} = H_{ij}^{-T} K_j^{-T} K_j^{-1} H_{ij}^{-1} \]
which can be simplified considering that all images were taken at a fixed zoom level \((K_i = K_j)\). Therefore, equation (11) reads \(w = H_{ij}^{-T} w H_{ij}^{-1}\), where \(w = K^{-T} K^{-1}\) is a symmetric matrix and thus it can be parameterized as a single vector \(\vartheta = [w_1 \ w_2 \ w_3 \ w_4 \ w_5 \ w_6]\). With one homography are written six equations, which can be rewritten under the form of a linear system \(A \vartheta = 0\). Having six equations and six unknowns, the solution can be found using singular value decomposition where \(||A\vartheta||\) is minimized under the constraint \(||\vartheta|| = 1\). With \(w\), one can define the matrix \(w\) and consequently compute \(K\). Multiple constraints can be considered on \(w\) to help the estimation of \(K\), e.g., a null skew parameter. Having computed \(K\), \(R_{ij}\) can directly estimated by

\[ R_{ij} = K^{-1} H_{ij} K \]

3) Automatic Calibration Method: Inspired on the work of Diogo Leite et al. [15], this methodology for PTZ camera calibration relies on non-linear optimization of re-projection errors. The proposed method stated in [15] is a two step calibration procedure where first both intrinsic and radial distortion coefficients are iteratively estimated at minimum zoom level and then computed for an increasing zoom sequence. The first step starts by capturing images with intersecting FOVs, with constant pan and tilt steps.

All images then go through key feature detection and matching using SIFT algorithm, and later excluded all matches not forming closed loops. RANSAC based homography estimation and non-linear optimization is used to compute homographies between every adjacent vertical images \(V_i\), and between every adjacent horizontal images \(H_i\). From composition of both \(H_i\) and \(V_i\), \(T_i\) homographies are calculated mapping all \(I_i\) images to a chosen reference image \(I_r\). Considering eventual residual errors accumulated in the homographies computation process, a non-linear optimization of the re-projection errors is applied using Levenberg-Marquardt algorithm. Initializing this minimization problem with the \(K\) matrix obtained from Agapito method and null radial distortion coefficients, \(k_1 = 0\) and \(k_2 = 0\), the cost function reads

\[ (K^* \vartheta, D^* \vartheta) = \arg \min_{K^{*\theta}, D^{*\theta}} \sum_{j=1}^{n} \sum_{i=1}^{m} \left\| u_i^j - h((K^* \vartheta)D^*(R_i X^j, k_1, k_2)) \right\|^2 \]

where \(D\) is a homogenization of \(D\) defined on equation 2, i.e. transforming a 2-vector into a 3-vector output by adding a third unitary coordinate, \(h([a \ b \ c]^T) = [a/c \ b/c]^T\), \(u_i^j\) are the selected features for the given image \(I_i\), \(R_i\) the rotation matrices directly obtained from the pan and tilt coordinates and, \(m\), \(n\) and \(X^j\) are the feature count, image count and global feature list of back-projected points respectively, for the given images.

This work uses the first step of the calibration procedure developed in [15] and generalizes it to all levels of zoom.

IV. PTZ PANORAMIC IMAGE GENERATION PROCESS

A. Panoramic Transformation

1) Cube based representation: Let the camera reference frame at its homing position be coincident with both the world and the cube frame as represented in figure 2(a)
Given the camera intrinsic parameters, \( K \), and the camera orientation, \( R \), every 2D point can be directly back-projected to a 3D point through \( [X_M \ Y_M \ Z_M]^T = (KR)^{-1}m \). Computing \( \chi = \max(|X_M|, |Y_M|, |Z_M|) \), the rules determining which cube face does an image point belong to are: if \( \chi = -Y_M \) top face, if \( \chi = Y_M \) bottom face, if \( \chi = Z_M \) front, if \( \chi = -Z_M \) back face, if \( \chi = X_M \) right face or if \( \chi = -X_M \) left face.

![Fig. 2. Cube and cylinder based representations.](image)

Once established the correct cube faces, one can map the back-projected image points through a projection matrix \( P_{WF} = K_F[R_{WF} \ 0_{3 \times 1}] \), where \( K_F \) consists in an intrinsic parameters matrix characterizing the resolution (size) of the cube faces, and \( R_{WF} \) are rotation matrices setting the optical axis orthogonal to the cube faces. Considering a cube face with a resolution of \( N \times N \) pixels, \( K_F \) is represented as

\[
K_F = \begin{bmatrix}
(N + 1)/2 & 0 & (N - 1)/2 \\
0 & (N + 1)/2 & (N - 1)/2 \\
0 & 0 & 1
\end{bmatrix}
\]

which stands for a perspective camera with an image coordinate system where the top-left pixel corresponds to \((1,1)\), and a 90° × 90° FOV. \( R_{WF} \) is responsible for rotating the 3D points from world’s reference frame to the correct cube face. The rotations for top, bottom, front, back, right and left faces are given by \( R_x(-90°) \), \( R_x(90°) \), \( I_{3 \times 3} \), \( I_{3 \times 3} \), \( R_y(90°) \), and \( R_y(-90°) \), respectively.

In brief, an image point \( m_i \) is re-projected to a point on a cube face \( m_{Fi} \) through \( m_{Fi} = K_F R_{WF} R_{Fi}^{-1} m_i \).

2) Cylinder based representation: Let the camera reference frame at its homing position be coincident with both world and cylinder reference frame as illustrated in figure 2(b). To successfully map an image into a cylinder based representation, first, the 2D points must be back-projected into 3D points. This is achieved with the help of the camera intrinsic and extrinsic parameters, as performed in the cube based representation. Then, the 3D points must be projected onto a cylindrical surface with unitary radius given by

\[
[X_L \ Y_L \ Z_L]^T = [X_M \ Y_M \ Z_M]^T \frac{1}{\sqrt{(X_L^2 + Z_L^2)}}
\]

as illustrated in figure 2(b). Any point in the cylindrical surface can be parameterized with a height \( h \) and an angle \( \theta \), known as 3D cylindrical coordinates \((h, \theta)\). The transformation between coordinates is given by \( [\sin \theta \ h \ \cos \theta]^T = [X_L \ Y_L \ Z_L]^T \), from which it is derived \( \theta = \arctan(2(X_L, Z_L)) \) and \( h = Y_L \).

The final mapped coordinates in the final panorama are given by \( (u, v) = (f_L \theta + u_L, f_L h + v_L) \), where \( f_L \) is a scaling factor in pixels that directly influences the panorama resolution. This value must be carefully adjusted according to the size of the panorama. \( u_L \) and \( v_L \) are simply offsets associated with the panorama coordinate system.

B. Panorama Construction Assessment Metrics

1) Variance and Standard Deviation of Pasting: Let \( \mu_i \) be the average pixel intensity of pixel \( i \) in the final mosaic, where \( i \) is the linear indexing format of pixel \((u, v)\). \( \mu_i \) is given by \( \mu_i = \frac{1}{K_i} \sum_{j=1}^{K_i} x_{ij} \), where \( K_i \) is the total number of images superimposed under pixel \( i \), and \( x_{ij} \) are the intensity values associated with these images resampled in panoramic coordinates. \( \mu_i \) has been computed using an iterative approach where each \( x_{ij} \) is stored and sum to the next \( x_{ij+1} \), as the images are pasted in the final panorama. The total number of images pasted in pixel \( i \) is also iteratively obtained as the pasting process takes place. Both values are kept in auxiliary cumulative matrices, having the size of the mosaic, which are later used in a pixel to pixel division for non-zero values of \( K_i \). Worth mentioning that the computation of \( \mu_i \) is only valid when \( K_i > 0 \), i.e. when at least one image has been pasted over \( i \).

The sum of squared intensities \( \sigma_i^2 \) is also computed within the pasting process, using a similar approach as the one described for \( x_{ij} \). The variance of pixel \( i \) can be then calculated as \( \sigma_i^2 = \left( \frac{1}{K_i} \sum_{j=1}^{K_i} x_{ij}^2 \right) - \mu_i^2 \), where \( K_i \) is the total number of images pasted in pixel \( i \). The variance of pixel \( i \) is divided by the auxiliary matrix associated with \( K_i \), in a pixel to pixel division for non-zero elements of \( K_i \). These results are later subtracted to the corresponding \( \mu_i^2 \).

Given the variance of pasting \( \sigma_i^2 \) of each pixel \( i \), an average value of all variances is consider in order to get an estimate of the variance of pasting associated with a particular mosaic. This metric is given by \( M_{\sigma^2} = \frac{1}{M} \sum_{i=1}^{M} \sigma_i^2 \), where \( M \) and \( T \) are the total number of non-zero \( K_i \) elements and the total number of pixels of the final panorama, respectively.

The standard deviation of pasting is also consider as an assessment metric. Considering the variance \( \sigma_i^2 \), the standard deviation of each pixel \( i \) can be computed as \( \rho_i = \sqrt{\sigma_i^2} \). An average value of all \( \rho_i \) is then computed, following the approach adopted in the average variance computation. This metric is therefore given by \( M_{\rho} = \frac{1}{M} \sum_{i=1}^{M} \rho_i \).

2) Assessment based on the SSIM: Presented in the work of [19], the Structural Similarity Index (SSIM) metric is an image quality metric taken between two images, \( I_1 \) and \( I_2 \), where the \( I_1 \) is chosen to be the reference image. This metric addresses the visual impact of image luminance, contrast and structure. The SSIM is given by

\[
SSIM(I_1, I_2) = \frac{(2\mu_{I_1} \mu_{I_2} + C_1)(2\sigma_{I_1 I_2} + C_2)}{(\mu^2_{I_1} + \mu^2_{I_2} + C_1)(\sigma^2_{I_1} + \sigma^2_{I_2} + C_2)}
\]

where \( \mu_{I_1}, \mu_{I_2}, \sigma_{I_1}, \sigma_{I_2} \) and \( \sigma_{I_1 I_2} \) are the local means, standard deviation and cross-covariance for images \( I_1 \) and \( I_2 \).
$C_1$, $C_2$ and $C_3$ are regularization constants for the luminance, contrast and structural terms respectively, used to avoid instability for image areas where the standard deviation or local mean are close to zero. The SSIM function returns a value between 0 and 1 which can also be represented in $[0\% - 100\%]$ interval.

The extraction of images from the final panorama according to its original pan and tilt coordinates, i.e. the inverse process of the panoramic transformation procedure, allows a direct assessment to be made with the original images. Thus, the SSIM metric is used as an assessment metric between the original and the extracted images. An average value over the SSIM metric of all pairs of images that are part of the final panorama is considered. This visual perception metric is termed mosaic quality.

C. Optimize Mosaic Alignment

This section presents an optimization approach where the camera focal length parameters are iteratively adjusted within the panoramic transformation step in order to minimize the average variance of pasting. This minimization was performed under Matlab optimization function `fminsearch`, and its cost function reads

$$(k_u^*, k_v^*) = \arg \min_{k_u, k_v} \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{K_i} \frac{x_i^2(k_u; k_v)}{K_i} - \mu_i^2(k_u; k_v)$$ (17)

where $x_i^2(k_u; k_v)$ stands for the high dependability of the pixel to pixel transformation on the camera focal length parameters $k_u$ and $k_v$, during the panoramic transformation.

D. Seamless Mosaicing by Image Blending

1) Average and Median: A pixel within the cube or cylinder based representations can be overlapped by multiple images according to the image positioning in the final panorama. Hence, an intuitive idea of a blending approach is to consider the average pixel intensity value. This computation is given by $P(u,v) = \sum_{j=1}^{L} \frac{I_j(u,v)}{K(u,v)}$, where $P(u,v)$ is the average pixel intensity value of pixel $(u,v)$ in the panoramic image $P$, $I_j(u,v)$ is the pixel intensity value of $(u,v)$ set by the resampled image $I_j$ in panoramic coordinates, $K(u,v)$ is the number of overlapping images in pixel $(u,v)$ and $L$ is the total number of images that are part of the panorama.

In practise, the average is computed in two parts. First, the sum of $I_j$ and $K$ variables for each $(u,v)$ pixels are stored in two independent auxiliary matrices with the same size of the final panorama. Then both matrices undergo a pixel to pixel division for non-zero elements.

Another simple approach, is the median pixel intensity computation. Here, every non-zero $I_j(u,v)$ is saved in a vector $P(u,v)[n]$ with size $K(u,v)$ from smallest to greatest, and from which the median pixel intensity is later computed.

2) Linear and Hard Blending: Both linear and hard blending approaches rely on spatial-weighting of image pixels i.e, each pixel is associated with a weight according to its position in the image with respect to the image borders.

The linear blending approach uses the grassfire transform to perform a weighted pixel average intensity computation. In this method, all pixel intensities are multiplied by its corresponding weight and then summed accordingly to the number of overlapping images in each pixel. These values are later divided by the sum of weights for each pixel. The linear blending computation can be described as

$$P(u,v) = \frac{\sum_{j=1}^{L} I_j(u,v) w_j(u,v)}{\sum_{j=1}^{L} w_j(u,v)}$$ (18)

where $P(u,v)$, $I_j(u,v)$, $w_j(u,v)$ and $L$ are the final pixel intensity value of pixel $(u,v)$ within the panoramic image $P$, the pixel intensity value of the resampled image $I_j$ in panoramic coordinates, the pixel weight of $I_j(u,v)$ set by the weight image $w_j$ also resampled in panoramic coordinates and the total number of images, respectively.

In the hard blending method every $P(u,v)$ is determined by the maximum weight of pixel $(u,v)$ in all weight images $w_j(u,v)$. Its computation is given by

$$P(u,v) = \sum_{j=1}^{L} I_j(u,v) W_j(u,v)$$ (19)

$$W_j(u,v) = \begin{cases} 1 & \text{if } w_j(u,v) = \arg \max_k w_k(u,v), \\ 0 & \text{otherwise}, \end{cases}$$ (20)

where $W_j$ is a binary weight image for each $I_j$.

This technique is computationally more expensive compared to the linear blending approach since all weights and image
3) Blending based on a Gradient Approach: The image blending approach described hereafter is a gradient method from which the final panoramic image is reconstructed by solving a Poisson equation similarly to the Poisson blending methodology described in section III-C. The developed approach starts by computing the discrete image gradient for all \( I_j(u, v) \) represented in panoramic coordinates. \( g_u \) and \( g_v \) stand for the image gradient in the \( u \) and \( v \) directions of image \( I_j \) resampled in panoramic coordinates. They can be computed using the discrete gradient computation (see equation 21).

With the resulting image gradients, all \( g_u \) and \( g_v \) must be combined using a weight average approach in a similar way to the linear blending method described in the previous section. This computation is given by

\[
G_u(u, v) = \frac{\sum_{j=1}^{L} g_u(u, v)w_j(u, v)}{\sum_{j=1}^{L} w_j(u, v)}
\]

where \( G_u(u, v) \) represents the combined gradient values of the final panoramic image in the \( u \) direction. \( G_v(u, v) \) is computed in a similar way, replacing \( g_u \) by \( g_v \). \( w_j \) is obtained with the grassfire transform as previously described. Taking \( G = (G_u, G_v) \), the divergence of \( G \) is given by \( \text{div} \ G = \frac{\delta G_u}{\delta u} + \frac{\delta G_v}{\delta v} \), from which \( \frac{\delta G_u}{\delta u} \) and \( \frac{\delta G_v}{\delta v} \) can be computed in a discrete form in a similar way as the gradient computation.

To successfully solve equation 5, one must determine the \( \Omega \) and \( \delta \Omega \) regions. Considering that the main goal is to recover an image from its combined gradient, the whole image was set as \( \Omega \). At this point, \( \delta \Omega \) although undefined, is still required to solve such equation. Hence, \( \delta \Omega \) was extracted from the result obtained in the linear blending approach due to the expected similarity between both approaches final image.

A variable \( b \) determined the number of boundary pixels from the linear blending result that were used for \( \delta \Omega \), as the same amount of boundary pixels that were removed in \( \Omega \) (combined gradient). From both \( \Omega \) and \( \delta \Omega \), equation 5 is written in the form of \( Ax = b \), in which \( A \) was obtained using matlab \texttt{delsq} function. The linear system was solved under matlab \texttt{mldivide} function, where a sparse LU decomposition algorithm is considered due to the shape of \( A \). This method has a closed formed solution unlike the iterative approaches suggested in [11]. Reshaping the solution \( x \) into an image shape, one gets the final panoramic image. Equation 5 was solved separately for each colour channel, as the final panoramic image was obtained as composition of these three.

V. EXPERIMENTAL RESULTS
A. Dataset
A sequence of images were taken with a fixed range of pan, tilt and zoom coordinates. The images followed a grid structure, with the pan ranging from 20° to 24° and tilt from −5° to −7° with steps of 1° for both angles, performing a total of fifteen images per zoom. The zoom ranged from 1% to 60% with steps of 10%, approximately. These images were taken under a controlled environment where the background remains still from image to image. The sequence of images were acquired in the Institute for Systems and Robotics (ISR), room 8.18 of north tower in Instituto Superior Técnico, using an AXIS P5512 PTZ Camera.

B. Feature-Points based Calibration
Section III-B3 introduced an automatic calibration technique based on the image calibration method described in section III-B2. Taking as input a grid of images acquired at a fixed range of pan and tilt orientations, the algorithm calculates an initial estimate of the camera intrinsic parameters through SIFT feature detection and matching between images. From this initial estimate, it is also possible to estimate camera odometry with the given homography transformations between images \( H_{ij} \).

Selecting camera odometry or estimated odometry together with the initial estimate, the camera intrinsic parameters go over a bundle adjustment process with the aim of minimizing the re-projection errors. As output, this method returns the final intrinsic matrix, as well as the camera pan and tilt coordinates used to run the optimization process.

The given camera, Axis P5512, has also been calibrated in the work of Luís Alves et al. [17] with Jean-Yves Bouguet camera calibration toolbox for Matlab. At this study, the calibration was performed for zoom levels of 1%, 20%, 40% and 60% and an interpolation method was used to find the intrinsic parameters for the remaining zoom levels. Figures 3(a) and 3(b) display the focal length parameters \( k_u \) and \( k_v \) respectively, and figures 3(c) and 3(d) the principal point
coordinates \( u_0 \) and \( v_0 \), regarding the multiple results obtained with the different calibration methodologies.

C. Colour based Calibration and Panoramic Imaging

The cube based representation presented in section [IV-A] is a six face structure where each face has a resolution of \( N \times N \) pixels. In this work, \( N \) was set as \( N = 2000 \) equally for all zoom levels in order to preserve image resolution.

Considering the multiple calibration results given in the previous section, the panoramic transformation step for the cube based representation was studied for the (i) toolbox calibration results with camera odometry, and (ii) automatic calibration results considering camera odometry or estimated odometry. The results are shown in figures 5(a,b,c) and 5(g,h,i).

All three performance measures indicate that the best performance is obtained with the automatic calibration encompassing odometry estimation. In fact, the lowest variance and standard deviation of pasting and, higher mosaic quality (SSIM based metric) are observed in the obtained mosaics. Figure 6(a) shows a particular case where this behavior is illustrated.

The mosaics based on the toolbox calibration and camera odometry show much distortion. The automatic calibration together with camera odometry provides more accurate results but still more distorted than using estimated odometry.

Figures 5(d,e,f) and 5(g,h,i) show the variance, standard deviation and mosaic quality results for the variance minimization approach, respectively. On the other hand, figure 4 illustrates the optimized focal length parameters and figure 6(b) illustrates the final images for zoom level of 40%.

For the cylinder it was assumed a resolution of \( N \times N \) pixels with \( N = 2000 \) and scale factor of \( f_L = 500 \), fixed for all zoom levels. Under these constraints, the cylinder went under a similar case study of the one described in the previous section regarding the cube structure. The results were identical. The automatic calibration results together with estimated odometry showed the best performance.

Due to the similar behaviour illustrated in both cube and cylinder results, this section only illustrates the best overall performance associated with the automatic calibration results with estimated odometry (see figures 5(g,h) and 5(i)).
Fig. 6. Final mosaics for zoom level 40%. (a) illustrates the non-minimized results. The toolbox result with camera odometry correspond to the image on the left, and the automatic calibration result with camera and estimated odometry are associated with images at the center and in the right, respectively. (b) illustrates the minimized variance results. (c) shows the results for the automatic calibration results with estimated odometry regarding the cylinder structure.

Fig. 7. Image blending results. (a) illustrates the final mosaics for zoom level 40%. These results are associated with the cube representation. (b) and (c) demonstrate the mosaic quality results for the cube and cylinder representations, respectively.
D. Image Blending Results

Section [IV-D] introduced the average and median techniques, the linear and hard blending methods and the proposed gradient-based approach. Figures [7(b)] and [7(c)] illustrate the cube and cylinder results regarding the optimized automatic calibration results with estimated odometry, respectively. Figure [7(a)] shows the final mosaics for zoom level of 40% using the cube based representation.

VI. Conclusion

The work described within this work focused on studying the PTZ panoramic image generation process. Starting with the acquisition of a sequence of images taken at a fixed range of pan and tilt coordinates for multiple zoom levels, the given PTZ camera was calibrated with an automatic calibration method relying on an image-based calibration technique. This texture-based approach provided an initial estimate on camera intrinsic parameters as well as camera odometry, which were later used as inputs on the automatic calibration bundle adjustment process. Along with the camera calibration results obtained in [17] with Jean-Yves Bouguet camera calibration toolbox, the automatic calibration results with camera odometry illustrated a continuous increase in camera focal length parameters for an increasing level of zoom. On the other hand, the automatic calibration results together with estimated odometry illustrated a variable behaviour possibly associated with camera decreasing image resolution for an increasing level of digital zoom. This constraint results in calibration and odometry estimation errors, considering its impact on the detection and matching of features.

The multiple calibration results were later tested as input parameters of the panoramic transformation step for the cube and cylinder based representations. The automatic calibration results with estimated odometry presented the lowest variance and standard deviation of pasting as well as the highest mosaic quality results. The final mosaics were also less distorted and more appealing than the ones presented by the other inputs.

With the variance minimization approach, the automatic and toolbox calibration results with camera odometry were effectively adjusted, leading to a major improvement on all performance measures and final mosaics. The automatic calibration results with estimated odometry, were barely affected. This behaviour suggests that these results were already close to its best case scenario.

On the image blending step the average, median and linear blending techniques improved the mosaic quality measures, although the image overlapping regions were still visible on the average and median results. The gradient-based approach successfully recovered most images from its combined gradient compensating exposure differences in image overlapping regions. However, it presented the lowest mosaic quality measures. Given these results, the mosaic quality metric was somewhat ineffective on evaluating the image blending results and therefore can be recorded as an inconclusive performance measure regarding this topic.

Considering future work, the gradient based approach presented in this work can be optimized. In this method, the selected boundary assumed a rectangular shape in which existed, for some cases, pixels with undefined intensities. Although it has not affected the quality of the final mosaics, this boundary can be more carefully selected, i.e. avoiding undefined pixels, and thus assume an arbitrary shape.

The memory requirements of PTZ panoramic image generation process for multiple zoom levels should also be taken into account. As most panoramic images demand high resolution images, having an efficient way of saving a sequence of multi zoom level panoramas preserving image quality would be especially useful in any panoramic image generation process.

REFERENCES