In 16th Portuguese Conference on Pattern Recognition (RecPad 2010)
Vila Real, Portugal, October 2010

Surveillance with Pan-Tilt Cameras: Background Modeling

Ricardo Galego Alexandre Bernardino José Gaspar
Institute for Systems and Robotics, Instituto Superior Técnico / UTL
ricardo.galego@gmail.com, {alex,jag}@isr.ist.utl.pt

***Abstract*—In this paper we propose a methodology for minimizing the variance of a cube (mosaicked) representation of a scene imaged by a pan-tilt camera. The minimization is based on the estimation of the vignetting image distortion, using the pan and tilt degrees of freedom instead of color calibrating patterns. Experiments with real images show that variance minimization is effective for improving event detection.**

I. Introduction

Surveillance with pan-tilt cameras imply finding (static) background representations. There are various ways to represent geometrically the background [1], [3]. In this paper we use the cube based representation as it allows a complete $360^o \times 360^o$ field-of-view with simple homography transformations. Defined the representation, background differencing can than be used to find intrusions (events), provided one has a good characterization of the uncertainty of the model. There are two main sources of uncertainty: inaccurate knowledge of the geometry of the camera and nonlinear-transformation of the radiometric readings. The geometric uncertainty found in the intrinsic parameters of the camera is handled by calibration. Radiometric uncertainty is mainly due to the nonlinearity of the radiometric response function and to vignetting, a decreasing gain for increasing radial distances in an image [2], [4]. A pixel based radiometric calibration is therefore required. In the following, we tackle both the uncertainty sources. First we introduce background modeling and construction.

II. Camera Model and Scene Representation

A pan-tilt camera is characterized geometrically by a perspective (pin-hole) camera surveying the scene with varying orientation while having a static projection center. The projection model is therefore represented by $m \sim PM$, where $M = [x\ y\ z\ 1]^T$ and $m = [u\ v\ 1]^T$ are 3D world and 2D image points, $\sim$ denotes equality up to a scale factor, and $P = K[R\,t]$ is a 3x4 matrix projection matrix, composed by the intrinsic parameters matrix, K, and the extrinsic parameters, R and t, representing the orientation and position of the camera with respect to the world coordinate system. We assume that the projection center is at the origin, $t = 0_3 = [0\ 0\ 0]^T$. The rotation matrix depends on the pan and tilt angles, $R = f(pan, tilt)$.

Building a cube base representation is a two steps process: (i) obtaining a back-projection for each image point and (ii) projecting the back-projection to the right face of the cube.

Fig. 1. Cube construction. Some of images used (left). Cube model and detail of a mosaicked image (right).

If the intrinsics and the orientation of the camera are known, then each image point can be back-projected to a 3D world point $[x\ y\ z]^T = (KR)^{-1}m$. This world point can be scaled to touch the lateral surfaces of a cube having edge lengths $2L$ with $[x_c\ y_c\ z_c]^T = [x\ y\ z]^T * L/max(|x|, |z|)$, from which one defines the so termed *critical latitude*, $\varphi_c(\theta)$ a latitude angle dependent on longitude, denoting the transition between the lateral faces of the cube and the top or bottom faces:

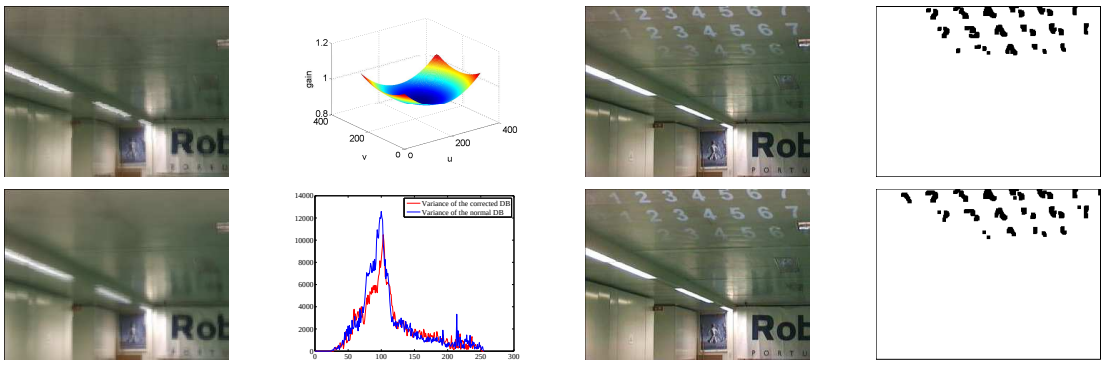
$$\varphi_c(\theta) = atan(L/\sqrt{x_c^2 + z_c^2}) = atan\left(max(|x|, |z|)/\sqrt{x^2 + z^2}\right).$$

Defined the *critical latitude*, and converting world coordinates to spherical coordinates longitude, $\theta = atan\,(x/z)$ and latitude, $\varphi = atan\left(-y/\sqrt{x^2 + z^2}\right)$ one can match the image points with the correct faces of the cube using the set of rules listed in table I.

Condition	Cube face
$\varphi \geq \varphi_c(\theta)$	Top
$\varphi \leq -\varphi_c(\theta)$	Bottom
$\lvert\varphi\rvert < \varphi_c(\theta) \wedge \lvert\theta\rvert \leq 45^o$	Front
$\lvert\varphi\rvert < \varphi_c(\theta) \wedge \lvert\theta\rvert \geq 135^o$	Rear
$\lvert\varphi\rvert < \varphi_c(\theta) \wedge 45^o < \theta < 135^o$	Right
$\lvert\varphi\rvert < \varphi_c(\theta) \wedge -135^o < \theta < -45^o$	Left

TABLE I
Matching 3D directions (φ, θ) with cube faces.

Identified the cube faces for mapping the image points, the mapping process consists simply in projecting the back-projections of the image points using the projection matrix $P_{WF} = K_F[R_{WF}\ 0_3]$, where K_F is an intrinsics matrix characterizing the resolution (size) of the cube faces, and R_{WF} are rotation matrices defining optical axis orthogonal to the cube faces. Figure 1 shows a representation of a laboratory given images acquired by a pan-tilt camera, sweeping 180^o longitude and 30^o latitude, in 10^o steps.



(a) Database (b) Gain and Variance (c) Run time (d) Detections

Fig. 2. Event detection experiment. Vignetting corrected images are in the bottom row, columns a and c. Notice the digits superimposed by a video-projector on the ceiling (column c). Vignetting correction allowed to increase the number of detections without raising the false alarms (column d, bottom row).

In order to map an image into the background (cube), one has to know precisely the camera orientation, R and the intrinsic parameters, K. In this work we assume that R is given by the camera control system and we obtain a first estimate of K using Bouguet's calibration toolbox. Then, we optimize K, a function of $\vartheta = [u_o\ v_o\ f_u\ f_v]^T$, i.e. the principal point and the focal lengths, by minimizing the back-projections of corresponding image points in two images, m_{1i} and m_{2i} obtained by matching SIFT features:

$$\vartheta^* = arg_\vartheta min \sum_i \|R_1^{-1} h({K_{(\vartheta)}}^{-1} m_{1i}) - R_2^{-1} h({K_{(\vartheta)}}^{-1} m_{2i})\|^2$$

where R_1 e R_2 denote the (known) rotation matrices representing the poses of the camera for acquiring the images and $h(.)$ denotes normalization to unit norm.

III. Uncertainty and Event Detection

In order to decrease the variance of a background representation, we estimate from (and apply to) all images defining the background, a vignetting correction function. The estimation process starts by choosing one image point, computing its back-projection to a 3D point, and then moving the camera and re-projecting the 3D point. Vignetting implies that the re-projected point-images will, in general, be different. Collecting all re-projections in an image V_{uv}, one can fit a correction function $g(\Delta u, \Delta v; a) = cosh(a_1 \Delta u) cosh(a_2 \Delta v) + a_3$:

$$a^* = arg_a min \sum_{uv} \|max(V_{uv})/V_{uv} - g(\Delta u, \Delta v; a)\|^2$$

where $a = [a_1\ a_2\ a_3]^T$, $\Delta u = u - u_0$ and $\Delta v = v - v_0$. Given g, we can now correct all acquired images I_{0uv} as $I_{uv} = g(u - u_0, v - v_0; a) \cdot I_{0uv}$.

Event detection is done by comparing the currently captured image, vignetting-corrected, I_{uv} with the corresponding image retrieved from the background database, B_{uv}, using the log likelihood function, L_{uv}:

$$L_{uv} = -0.5(I_{uv} - B_{uv})^2 / \Sigma_{uv}^2 - 0.5 \ln(\Sigma_{uv}^2) - 0.5 \ln(2\pi)$$

where Σ_{uv}^2 denotes the background variance. A pixel (u, v) is considered active, foreground, if L_{uv} is larger than a threshold in at least two of the three (RGB) components of I_{uv}.

IV. Experiments

In our experiments we used a Sony EVI D30 to scan a room and create two background representations: one lacking and the other one having vignetting-correction (see Fig. 2). The images with events to be detected, were created afterwards using a video projector superimposing text (digits) towards the ceiling of the room. Figure 2 shows correct detections of digits, in both cases, despite the variance of the background motivated by the imaging distortions (e.g. radiometric response function and vignetting). In addition, the results show that vignetting correction implies a lower variance in the background representation (Fig. 2, column b). The average pixel-wise standard deviation gain is about $-9.5dB$. This reduction of the variance (standard deviation) improves the detection when using a fixed, proportional to variance, threshold for the background subtraction algorithm.

V. Final notes and future work

In this paper we proposed a vignetting correction method for pan-tilt cameras. Experiments have shown that the correction allows building (mosaicked) scene representations with less variance and therefore more effective for event detection. Future work will focus on maintaining minimized variance representations accompanying the daylight change.

Acknowledgements

This work has been partially supported by the Portuguese Government - Fundação para a Ciência e Tecnologia (ISR/IST pluriannual funding) through the PIDDAC program funds, and by the project DCCAL, PTDC / EEA-CRO / 105413 / 2008.

References

[1] M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *Int. Journal of Comp. Vision*, 74(1):59–73, 2007.

[2] Seon Joo Kim and M. Pollefeys. Robust radiometric calibration and vignetting correction. *IEEE T-PAMI*, 30(4):562 –576, 2008.

[3] S. N. Sinha and M. Pollefeys. Towards calibrating a pan-tilt-zoom camera network. In *Department of Computer Science, University of North Carolina at Chapel Hill*, pages 91–110, 2006.

[4] Wonpil Yu. Practical anti-vignetting methods for digital cameras. *IEEE T. Consumer Electronics*, 50(4):975 – 983, 2004.