Introduction

April 12, 2011

1 Thesis context

This thesis concerns about some aspects of the precise 3D scene reconstruction with multiple cameras within the scope of project CALLISTO (Calibration en vision stro par mthodes statistiques) sponsored by ANR (Agence Nationale de la Recherche). This project relies on the collaboration between different universities and grandes coles: CMLA-ENS-Cachan, IMAGINE-ENPC, MAP5-Paris 6 and LCTI-Tlcom Paris. Even though this project mainly concerns about the images taken by consumer digital compact/reflex cameras, the result can be extended to the satellite images. The work thus helped to some extent the MISS (Mathmatiques de l'imageries stroscopique spatiale) project collaborated with CNES (Centre National d'Etudes Spatiales), whose aim is to reconstruct a completely controlled and reliable 3D terrain model from two images taken by the airbone satellite camera almost simultaneously.

The 3D scene reconstruction can be mainly divided into five components: camera calibration, image rectification, dense image registration, 3D scene reconstruction, 3D merging, meshing and rending (Fig. 1). Some components can be replaced by other techniques to make the chain more adapated for some specific applications. The camera calibration is the first step in the chain and thus plays a very important role. It consists in camera internal/external parameters calibration and distortion model estimation. Camera internal parameters means the intrinsic parameters of camera, like aspect ration, focal length, princiapl point and lens distortion parameters. Camera external parameters means camera orientation and position in a fixed world coordinate. And the distortion model describes the geometric deviation of real camera from a pinhole camera.

Image rectification is an auxiliary step for dense image registration. It virtually rotates two cameras about their optic center respectively such that the two cameras planes are coplane and their x-aixs are parallel to the baseline. This generates two images whose corresponding epipolar lines coincide and are parallel to the x-axis of image. A pair of rectified images is helpful for dense stereo matching algorithms. It restricts the search domain

for each matching to a line parallel to the x-axis. Due to the redundant degrees of freedom, the solution to rectification is not unique and actually can lead to undesirable distortions or be stuck in a local minimum of the distortion function. So a robust algorithm which works for any pair of images is needed.

Dense image registration is the thesis subject of Neus Sabater [3], whose aim is to find dense correspondences between two images. This is a difficult problem by considering different imaging conditions when taking images, like different geometric distortion depending on the viewing angle, non-linear lens distortion of camera, changing lighting condition, non-static scene, occlusion, etc. These problems can be to some extent reduced by using a pair of image with "low B/H" (low baseline/height), which are taken by a satellite almost simultaneously. But this raises a higher demande on the precision on the correspondences. A thorough discussion about how to obtain dense correspondences with emphasis on the control of false alarms, sub-pixel precision and the fatterning problem at the contour of image can be found in Neus Sabater's thesis [3].

Once camera is precisely calibrated and a pair of images is accurately and reliably registrated, a 3D model of scene can be easily reconstructed on the same order of precision with some classic methods up to a 3D similartiy transformation. One pair of images only permits us to reconstruct a partial 3D model. To have a more complete model, it is necessary to take photos with different angles of view around the 3D scene. The pair-wise 3D scene model can be overlapped or completely disjoint. By merging many pair-wise partial 3D models, a dense 3D point cloud can be obtained.

The chain seems complete with all the above components. But the reconstructed 3D scene is just a set of 3D points in the space. For the sake of visualization, it is better to reconstruct also the surface of objects in the scene, in particular for purpose of parts inspection and repairing in industry. Since the precision of 3D points is high, it is hoped that the precision will be kept in the surface reconstruction. This problem is largely discussed in Julie Digne's thesis [1], which treats high precision scanned raw data point sets with up to 35 million points, usually made about 300 different scan sweeps. Even though in her work, the 3D points directly come from the laser scanner instead of from 3D reconstruction of images taken by camera, the principle of problem remains the same: how to reconstrut the surface reliably by keeping the high precision without smoothing and re-sampling ?

2 Thesis summary

The thesis focus on the precision aspect in 3D reconstruction without reviewing all the components in the chain. The origin of imprecision can lie at any step of the chain. The imprecision caused in a certain step will

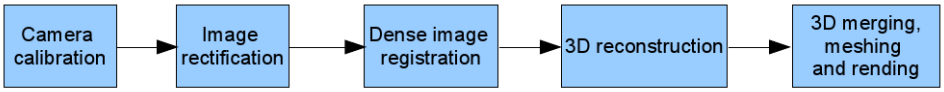


Figure 1: The 3D reconstruction chain.

compensate the precision gained in the previous steps, then be propagated, amplified or mixed with the error in the following steps and finally leads to an imprecision 3D reconstruction. So it is difficult to directly improve a lot the precision from the final imprecise 3D data. The approriate approach to obtain a precise 3D model is to study the precision component by component. We pay more attention to the camera calibration for three reasons. Fisrt, it is the first component in the chain. Second, it is by itself already a complicated system containing many unknown parameters. Third, the (intrinsic) parameters of camera only need to be calibrated one time depending on the camera configuration. In addition, the camera calibration is a subject supposed to have been resolved since years. But the result is still not satisfying if high-precision is required. With a brief review of the developpement of 3D reconstruction techniques in section 2.1, we realize that it is still worth the effort to improve the precision. Then in section 2.2, we try to explain the key problem in traditional camera calibration methods with a new camera concept proposed in section 2.3. The thesis orgnaization is in section 2.4.

2.1 History

The 3D scene reconstruction is not a new subject in computer vision. Before the "computer" and digital camera came into history at the end of year 1970s, the scene reconstruction was already a classic problem in photogrammetry, where it has a different name "stereophotogrammetry". Its aim is always to determine the geometric properties of objects from photographes. At that time, more attention was paid to the methods in optics and metrology due to the lack of computational power. The distances and angles are directly measured by hand from photographes, objects in scene and cameras separated by a fixed baseline. The precision is not ensured in the measurement, which can lead to inaccurate final result. A lot of imprecise manual work limits the practical application of early photogrammetry techniques.

With the advent of the digital camera and high-performance computers, the 3D reconstruction techniques become more mature. Fewer or no manual measurement is required. Many algorithms can work directly on the images

to find the necessary information to deduce 3D geometry. There exist some algorithms to reconstruct 3D scene from images automatiquely with few human intervention. Nonetheless, 3D reconstruction is still a complicated chain including several components. Any error in a certain component will make the whole system unreliable.

Another possibility to obtain 3D reconstruction is to use 3D laser or LIDAR (LIght Detection And Ranging). Not like the above multi-view geometric method, this kind of methods is active. The position of object in the space is measured by the time delay between transmission of a pulse and detection of the reflected signal. Thus a dense 3D points can be directly obtained by some post-processing like filtering and merging. The traditional idea is that 3D stereo reconstruction is not as precise as the result obtained by 3D laser scanner. A high quality 3D scanner can provide 3D point cloud with the precision about 20 μm. This is cannot be achieved by state-of-art image-based 3D stereo reconstruction algorithm. Yet, 3D laser scanner is usually a big, expensive and sophisticated machine, which have to be set up carefully and cannot be easily transported for onsite 3D reconstruction tasks, while a camera, even a reflex camera of good quality, costs almost nothing compared to a scanner machine and can be flexibly transported anywhere to take photos. In addition, it is not feasible to install a huge laser scanner on satellite to do 3D reconstruction. So it is still of interest to use camera photos to reconstruct 3D scene if the precision can achieve or surpass that of 3D scanner.

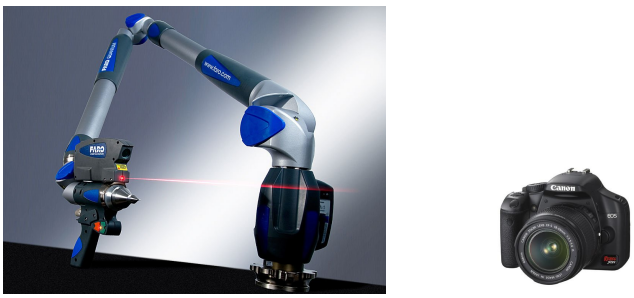


Figure 2: Left: 3D laser scanner. Right: Canon EOS Reflex camera.

2.2 Challenge

To have a precise final 3D reconstruction from 2D photos, each component should produce a precise result. This gives the first importance to the camera calibration because its imprecision will be inevitably propagated to the other components and usually cannot be corrected later. The camera calibration is also the most error-prone component due to the sophisticated camera optic structure. With a precise camera calibration, it is already sufficient

to determine the relative position of two cameras and reconstruct a sparse model of 3D scene from the correspondences between two images. A typical camera model has the form:

$$C = \mathcal{D}KR[I|-T] \tag{1}$$

with $\mathcal{D}$ the lens distortion, K calibration matrix, R camera orientation and T the camera optici center in a fixed world coordinate. According to this model, camera calibration consists of two parts: internal parameters calibration ($\mathcal{D}$ and K) and external parameters calibration (R and T). The internal calibration is to retrieve the intrisic parameters of camera, which should be constant once the camera has fixed its configuration. But in the experiments, it shows that the internal parameters varies from one experiement to the other even if the same camera was used with fixed configuration. This makes the internal calibration unreusable for the other data sets. Similarly, for the external calibration, the camera position and orientation also varies by using two data sets containing a common camera position. This leads us to think over again the whole camera calibration system. The conjecture about the phenomen can be two-fold:

1. the distortion model cannot capture the real physical aspect of real distoriton

2. the error in external calibration for R and T compensates the error in internal calibration for $\mathcal{D}$ and K. So the whole camera model can be precise in the sense that the observed point is close to the point computed by the camera model. But none of the components ($\mathcal{D}$, K, R, T) is precise.

This is in fact the common drawback of many global camera calibration methods, which perform the internal and external calibration together. Typical global calibration algorithms (Lavest *et al.* algorithm and Cognitech) were tested to verify the problem. In particular, Lavest *et al.* algorithm shows a very small residual error about 0.02 pixels, which confirms the precision of the global camera model. But the residual distortion error can be 10 times bigger. Since Lavest *et al.* algorithm considers also the non-flatness of the pattern and estimate its shape, the factor of 10 can only be explained by the error compensation.

One possible solution to resolve the problem is to do the 3D reconstruction and the calibration at the same time. This can be done by putting the calibration pattern in the scene or by directly using the scene as the pattern. For Lavest *et al.* algorithm, even though the calibration result changes from time to time, the estimated 3D positions of pattern points are more stable. So if our aim is just to reconstruct the 3D scene, it is feasible to adapt Lavest *et al.* algorithm to do calibration and reconstruction at

the same time. One difficulty lies on the incompatibility between the wide viewing angles of cameras and precise feature detections. On the one hand, to reconstruct 3D scene more completely, the camera should change a wide angle of view to capture different aspects of the object. On the other hand, the geometric distortion caused by viewpoint change raises difficulty to detect the feature points precisely even if they are visible in all the images. That is the reason why traditional calibration methods use a flat pattern containing regular geometric shape as the feature points. This difficulty can be overcome by dividing all the feature points into several subsets which are visible among two or three images. Since these images are taken by the camera with similar viewing angles, the distortion will be limited. But there is still an obstacle ahead: unless some prior relationship between these 3D feature points is known, the 3D position of the points must be measured manually in high precision, which is not an easy task.

2.3 Mathematic camera

People are always looking for the absolute external and internal parameters of camera without realizing that there can be some error compensation between them. The measured re-projection error, which is defined as the difference between the observed position and the predicted position by camera model on the selected feature points, cannot reflect the absolute error in external or internal parameters. In the global optimization process, errors in the external and internal camera parameter can be compensated by opposite errors in the distortion model. Thus, an inaccurate distortion model can pass undetected.

For the above reason, it is risky to use global calibration method to correct lens distortion. It is preferable to separating the distortion correction from the global calibration. So the camera calibration is decomposed into two steps. The first step is to correct lens distortion to obtain a pinhole camera; the second step is to calibrate the pinhole camera. The following fundamental theorem gives the definition of pinhole camera and also some hint on how to perform and evaluate the distortion correction.

Theorem 1 *A camera follows the pinhole model if and only if the projection of every line in space onto the camera is a line.*

This theorem implies that the distortion can be corrected by rectifying the distorted lines in images. Different methods to fabricate straight lines can be tried. The only requirement is that the lines must be straight. The correction can be finally verified by the the straightness of the rectified lines. It is important to remark that a straight line in image remains to be straight under a 2D homogrpahy. This means the distortion is corrected up to an arbitrary homography. This distortion is noted by $\tilde{\mathcal{D}}$ to tell from the absolute

distoriton $\mathcal{D}$. Assume the arbitrary homography is H, then the estimated distortion is $\tilde{\mathcal{D}} = DH$. By applying the inverse of $\tilde{\mathcal{D}}$ on camera, we obtain a new camera noted by $\tilde{C}$:

$$\tilde{C} = \tilde{\mathcal{D}}^{-1}C = H^{-1}\mathcal{D}^{-1}\mathcal{D}KR[I| \ -T] = H^{-1}KR[I| \ -T]. \tag{2}$$

H, K being invertible, the decomposition $H^{-1}K = \tilde{K}R'$ is unique by QR decomposition with the constraint that $\tilde{K}$ is an upper-triangle 3×3 matrix and R' is a 3×3 rotation matrix. This new camera becomes $\tilde{C} = \tilde{K}R'R[I| \ -T] = \tilde{K}\tilde{R}[I| \ -T]$. We call it *mathematic (or virtual) camera after distortion correction* because the calibration matrix $\tilde{K}$ and rotation matrix $\tilde{R}$ do not match the physics of the actual camera, but yield a virtual pinhole camera that can be used to the very same purposes. Indeed, consider several positions of the physical camera inducing as many camera models $C_i = DKR_i[I| \ -T_i]$. Applying the correction $\mathcal{D}^{-1}$ to all images obtained from these camera positions yields virtual pinhole cameras $\tilde{C}_i = \tilde{K}\tilde{R}_i[I| \ -T_i]$, which maintains the same relative orientations: $\tilde{R}_i^{-1}\tilde{R}_j = R_i^{-1}R_j$. From these cameras the whole 3D scene can be reconstructed by standard methods, up to a 3D similarity.

2.4 Thesis by chapter

The thesis will follow the above track to try to correct the distortion in high precision. With some basic introductions to the camera model and epipolar geometry in Chapter **??**, we first propose a non-parametric pattern-based distortion correction method in Chapter **??**. The precision of the non-parametric distortion correction being limited by the non-flatness of the pattern, a plumb-line based method is also tried to improve the precision in Chapter **??** by using a polynomial model, which is shown to be more universal than the other distortion models in Chapter **??**. After the distortion problem, we go back to the basic correspondences precision between two images by analysing and improving SIFT method in Chapter **??**, which will be used in image rectification in Chapter **??** and burst denoising in Chapter **??**.

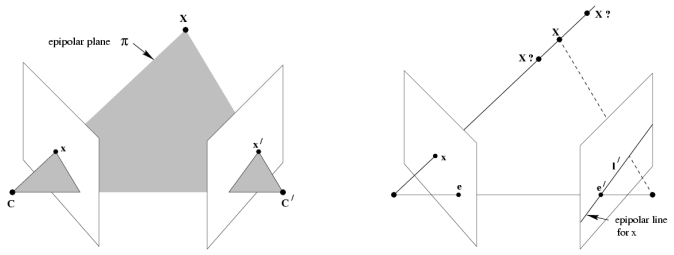
2.4.1 Chapter ??: Camera model and epipolar geometry

Many algorithms in multi-view geometry are based on the assumption that the camera is ideal pinhole. But in practice, the camera is deviated from a pinhole model by lens distortion. In this sense, distortion is the only gap between the theory and the practice. That is the reason why we want to correct it in high-precision in Chapter **??**, **??** and **??**. In this chapter, some basic concepts about pinhole camera model and distortion model are introduced. A typical bundle adjustment like camera calibration method is

also explained in detail to show the problems we shall meet in distortion correction.

Once the distortion is removed and the camera becomes pinhole, the projective geometry is a useful tool to solve different problems in multi-view geometry, like image rectification and mosacing. The geometric constraint becomes more complicated when the number of views increases. Here we concentrate on the simplest two-view geometry because it is difficult for three or more images to share enough stable feature points in practice. The relation between corresponding points in two views is described by epipolar geometry. In algebraic viewpoint, the epipolar geometry is coded by a 3×3 matrix, called fundamental matrix, F. Fundamental matrix only depends on the relative position of two cameras and the intrinsic parameters of cameras, but not on the 3D scene. Two important observations of epipolar geometry is:

- Given one point x in the left image, its corresponding point in the right image must be on the line called epipolar line, which can be explicitly computed as Fx. This provides a necessary condition to test whether two points correspond to the same 3D point, see Fig. 3.
- Given a set of corresponding points in two images, the 3D scene can be reconstructed up to a 3D projective transformation, as well as the camera position and orientation.



(a) (b)

Figure 3: (a) The optic center of two cameras and a 3D point form a epipolar plane which intersects two image planes at the epipolar line. (b) Given one point x in the left image, its correspondence in the right image must be on the corresponding epipolar line.

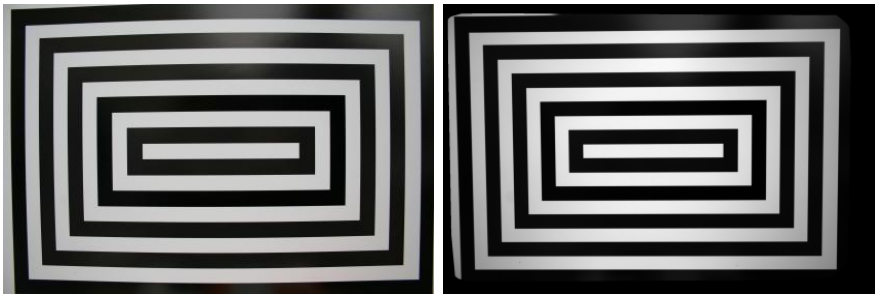
2.4.2 Chapter ??: Non-parametric pattern-based distortion correction

Pinhole camera model is widely used in computer vision applications because of its simplicity and its linearity in terms of projective geometry. But a

real camera is deviated from the ideal pinhole model by lens distortion. Distortion correction is the first important step in many multi-view geometry applications. The error caused in distortion correction will be inevitably propagated to the final result and usually cannot be correct again. But a lot of available methods just assume that the camera is brought back to be ideal pinhole camera after removing the dsitortion without paying much attention to the precision.

Traditionally, the lens distortion is estimated with the other camera parameters together (camera internal and external parameters). So we call these methods global camera calibration methods. In these global calibration methods all parameters are estimated by minimizing the error between the camera and its numerical model on feature points identified in several views, all in a single non-linear optimization. The advantage is that any distortion model can be incorportated into global camera calibration. But the result will be precise if (and only if) the model captures the correct physical property of cameras and if the minimization algorithm finds a global minimum. To avoid the local minimum in non-linear minimization, a two-step strategy is often used. The closed-form solution is first found with linear method by ignoring the lens distortion and is refined by non-linear optimization by adding the lens distortion parameters. But the global camera calibration suffers a common drawback: errors in the external and internal camera parameter are being compensated by opposite errors in the distortion model. Thus the residual error can be small but the distortion model is not that precise.

The error compensation can be avoided by only correcting the distoriton without treating other camera parameters. These methods can be classified into two categories: enlarged epipolar geometry methods and plumb-line based method. Enlarged epipolar geometry methods incorporate the distortion model into the epipolar geometry and use a set of corresponding points in two images suffereing the same distortion. The distortion can be estimated by linear or non-linear methods. Plumb-line based method is based on the famous fact that a 3D line remains to be straight in 2D image if the camera is a pinhole camera (no lens distortion). Yet, these methods are all parametric and depend on the a priori choice of a distortion model with a fixed number of parameters. This per se is a drawback: such calibration methods require several trials and a manual model selection. In this chapter, a non-parametric, non-iterative and model-free method is proposed. This method requires a flat and textured pattern. By using the dense matchings between the pattern and its photo, a distortion field can be obtained by triangulation and local affine interpolation. The obtained precision compares favorably to the distortion given by state of the art global calibration and reaches a RMSE of 0.08 pixels (see Fig. 4 for a correction example). Nonetheless, we also show that this accuracy can still be improved in the next two Chapters.



(a) (b)

Figure 4: A correction example. (f) distorted image. (g) corrected image.

2.4.3 Chapter ??: Self-consistency and universality of camera lens distortion models

This Chapter is a preparation for the next chapter. Due to the difficulty to control the flatness of a pattern, we will go back to the parametric method in the next chapter to obtain a higher correction precision. For any parametric method, an approriate distortion model is necessary. Even though there exists many distortion models in litterature, it is not clear which one is more approriate than the others. In addition, the role of distorted points and undistorted points seems to be interchangeable in litterature, which makes the distortion model more ambiguous. In this chapter, the concepts of "self-consistency" and "universality" are introduced to evaluate the validity and precision of camera lens distortion models. Self-consistency is evaluated by the residual error when distortion generated with a certain model is corrected (using the model in reverse way) by the best parameters for the same model. Analogously, universality is measured by the residual error when a model is used to correct distortions generated by a family of other models. Five classic camera lens distortion models are reviewed and compared for their degree of self-consistency and universality. The realistic synthetic experiments show that the polynomial model is self-consistent and more universal than the other models. The polynomial model, with order from 8 to 19, permits to approximate any other four models, and the inverse of any other four models including itself, at the precision about 1/100 pixels. This high order is more than compensated by its linearity and its translation invariance, which makes it independent of the distortion center. A real experiment shows that the polynomial model of degree 6 can approximate a real distortion field between a textured pattern and its photo at 1/100 pixel precision (see Fig. 5). So the polynomial model will be chosen in the next chapter as distortion model to improve the correction precision.

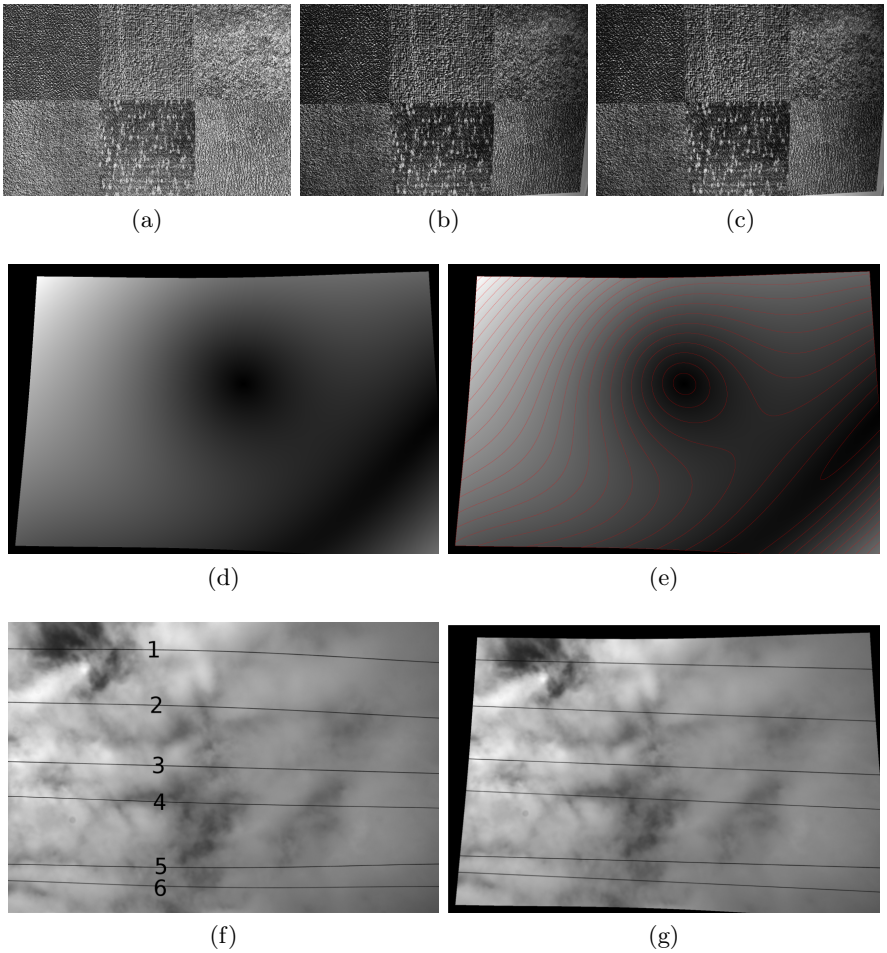


Figure 5: (a) digital pattern. (b) and (c) two photos of the digital pattern. (d) the distortion field constructed by the estimated parameters of the polynomial model. (e) level lines of (d) with quantization step of 20. (f) distorted images of tightly stretched lines. (g) corrected image by polynomial model.

2.4.4 Chapter ??: Distortion correction with a calibration harp

This chapter is a continuation of the last chapter to improve the precision of lens distortion correction. Even though non-parametric pattern based method do not depend on the a priori choice of a distortion model with a fixed number of parameters, to achieve a high precision, they require a very flat non deformable plate with highly accurate patterns printed on it. It is shown that $100\mu m$ flatness error can introduce 0.3 pixels error in distortion estimation. Either to fabricate such pattern or to estimate its shape is very difficult in practice. Perphas the thing we can easily ensure the quality is the straightness of strings. That is the reason why we resort to the plumb-line method to correct the distortion. Based on the famous fact that a camera follows the pinhole model if and only if the projection of every line in space onto the camera is a line, it is sufficient to correct the distorted lines to obtain the pinhole camera. The "calibration harpe" can be easily obtained by tightly stretching good quality sewing strings on a frame. It seems that the hardware problem of plumb-line methdo is sovled. But we still need a good distortion model to integrate into the plumb-line method to treat differet types of realistic lens distortion. By testing the self-consistency and universality of different models, we find that the polynomial model is more adpated to correct real distortion. In addition, it is invariant to the translation to the distoriton center. So the distortion center can be fixed anywhere without being estimated. This is a big advantage compared to other models. Real experiments show that no artificial bias is created on the corrected strings and higher precision is attained compared to non-parametric pattern based method. Further study shows that the residual oscillation after correction is due to the twisted structure of sewing strings used to build the harp. With more smooth fishing strings, we do gain a factor about 2 and achieve the correction precision about 0.02 pixels, much better than the result given the global camera calibration, which is not stable and varies with the parameters used in the distortion model.

2.4.5 Chapter ??: SIFT

SIFT (Scale-Invariant Feature Transform) is one of the most successful feature detection and matching algorithm since years. It is totally invariant to a similarity transformation and also robust to illumination change and partial change of view point. Precise feature points are really the basis of 3D reconstruction chain. SIFT features are a very good candidate if we need some sparse features before a dense image registration. In spite of its importance, the precision of SIFT features has not been extensively studied before.

In this chapter, we first review SIFT method and understand why it is completely scale invariant. The SIFT matching precision under differ-

ent geometric transformation is tested to show its average precision. By studying the structure of SIFT scale space, we realize that its convolution-subsampling structure will decrease the precision through octaves. We propose to cancel the subsampling between octaves to keep the precision. This change leads to other modification in order to keep the scale invariance of SIFT. We call the new SIFT schema precise SIFT.

Synthetic test verifies the improvement of precision compared to Lowe's SIFT. But in case of scale change, precise SIFT does not gain much with respect to Lowe's SIFT. This is due to the inconsistency between fractional scale change and the dyadic structure of SIFT scale space. Two schema are proposed to furthur refine the precision: one is to apply the estimated transformation on one of the image and launch precise again on two image; the other is to refine the points by a local homography estimated by some neighbouring matchings. The second method shows better performance and will also be used in our non-parametric distortion correction to remove the oscillation in the estimated distortion field in Chapter **??**. By assuming that the distortion follows a polynomial model, the precise SIFT is also verified in real images. Precise SIFT is directly usable in three directions of applications:

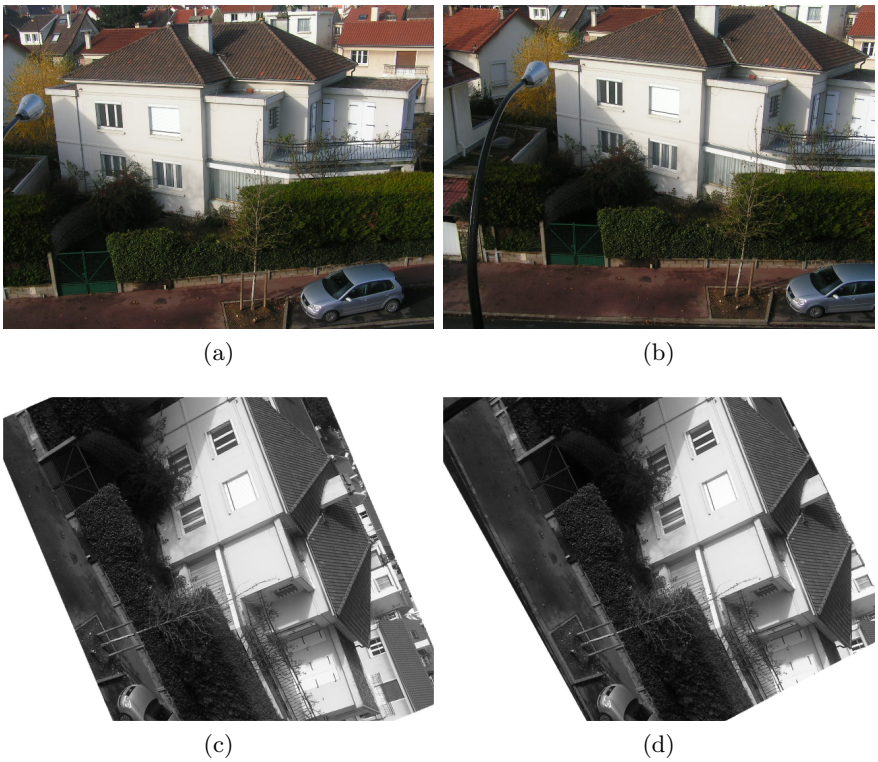
- panorama: the aim is to perform a photo montage seamlessly from several photos. These photos are taken at large distance from the object and the overlapping between two adjacent photos is important.
- super-resolution: the aim is to obtain an image in higher resolution from a burst of low resolution images of a flat objetct.
- global camera calibraion: the aim is to calibrate camera model parameters by several photos of a flat pattern.

2.4.6 Chapter ??: Rectification

The epipolar geometry enters into a special situation when the two camera planes are co-plane and parallel with the line connecting the optic center of two cameras. In this case, the corresponding epipolar lines also coincide and align with the x-axis of the image. This means that one point has the same y-coordinate as its corresponding point. This special geometry can be achieved by rotating two camera without changing their optic center. This is equivalent to apply a homography on each image respectively. This process is called image rectification in two-view geometry (Fig. 6 shows a pair of images before and after rectification). A pair of stereo-rectified images is helpful for dense stereo matching algorithms. It restricts the search domain for each match to a line parallel to the x-axis. Due to the redundant degrees of freedom, the solution to stereo-rectification is not unique and actually can lead to undesirable projective distortions or be stuck in a local

minimum of the distortion function. Many rectification methods reduce the distortion by different explicit measure. But it is not clear which measure is the most approriate. We propose a rectification method by three steps of camera rotation. In each step, the distortion is explicitly reduced by minimizing the rotation angle. For un-calibrated cameras, this method can be formulated as an efficient minimization algorithm by optimizing only one natural parameter, the focal length. This is in contrast with many methods which optimize between 3 and 6 parameters.

Finally, we should not forget the assumption for the epipolar geometry that the camera is considered ideal pinhole without lens distortion. If it is not true, the rectification will not be very precise. So a preliminay precise distortion correction is required before the rectification.



(a) (b)

(c) (d)

Figure 6: Rectification example. (a) and (b): pair of original images. (c) and (d): the rectified images.

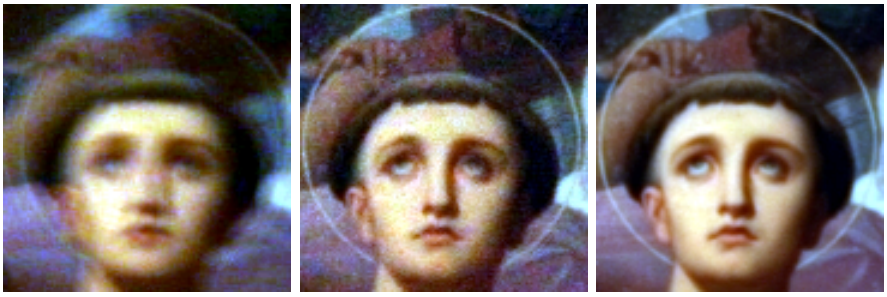
2.4.7 Chapter ??: Burst denoising

Denoising is one of the most important image enhancement techniques. Even though denoising algorithms have been largely developed since years, most of them concentrate on the single image denoising. However, with the increase

of memory size and data storage speed, today it is possibile for cameras to take a burst of images in a few seconds. This opens a new possibility to do image denoising, in particular with dim light. Perphas all of us have the frustrated experience to take photos in musem under low light conditions, where the flash of camera and tripod are forbiden. In such situation, taking photographs with a handheld camera is problematic. If the camera is set to a long exposure time, the photograph gets motion blur. If it is taken with short exposure, the image is noisy. This dilemma can be solved by taking a burst of images, each with short-exposure time, as shown in Fig. 7. But then, as classical in video processing, an accurate registration technique is required to align the images. Denote by $u(x)$ the ideal non noisy image color at a pixel x. Such an image can be obtained from a still scene by a camera in a fixed position with a long exposure time. The observed value for a short exposure time τ is a random Poisson variable with mean $\tau u(x)$ and the standard variation proportional to $\tau u(x)$. Thus the SNR increases with the exposure time proportionally to τ. The core idea of the burst denoising method is a slight extension of the same law. The only assumption is that the various values at a cross-registered pixel obtained by a burst are i.i.d.. Thus, averaging the registered images amounts to averaging several realizations of these random variables. An easy calculation shows that this increases the SNR by a factor proportional to $\sqrt{n}$ where n is the number of shots in the burst. (We call SNR of a given pixel the ratio of its temporal standard deviation to its temporal mean). Fig. 7 summarizes the possibilities offered by an image burst. A long exposure image is exposed to motion blur. The short exposure image is noisy, but sharp. Finally, the image obtained by averaging the images of the burst after registration is both sharp and noiseless. In this real example the burst taken in a gallery had 16 images. The noise should therefore be divided 4.

Even though the denoising power of burst denoising is eventually hemmed by the low growth of the square root, dividing the noise by the mentioned factors and getting an artifact free image is in no way a negligible ambition. Indeed, even the best state of the art denoising methods can create slightly annoying artifacts. If a fine non-periodic texture is present in an image, it is virtually indistinguishable from noise, and actually contains a flat spectrum part which has the same Fourier spectrum as the white noise. Such fine textures can be distinguished from noise only if several samples of the same texture are present in other frames and can be accurately registered.

Yet, this method rises serious technical objections. The main technical objection is: how to register globally the images of a burst? Fortunately, there are several situations where the series of snapshots are indeed related to each other by a homography, and we shall explore these situations first. The homography assumption is actually valid if one of the assumptions is satisfied:



(a) (b) (c)

Figure 7: (a) one long-exposure image (time = 0.4 sec, ISO=100). (b) one of 16 short-exposure images (time = 1/40 sec, ISO = 1600). (c) the average after registration. All images have been color-balanced to show the same contrast. The long exposure image is blurry due to camera motion. The middle short-exposure image is noisy, and the third one is some 5.6 times less noisy, being the result of averaging 32 short-exposure images.

- the only motion of the camera is an arbitrary rotation around its optic center;
- the photographed objects share the same plane in the 3D scene;
- the whole scene is far away from the camera.

In those cases, image registration is equivalent to computing the underlying image homography. But this registration should be sub-pixel accurate. To this aim we will use the precise SIFT in Chapter **??** and a generalization of ORSA (Optimized Random Sampling Algorithm, [2]) to register all the images together. Yet, in general, the images of 3D scene are not related by a homography, but by an epipolar geometry. Even if the camera is well-calibrated, 3D point-to-point correspondence is impossible to obtain without knowing the depth of the 3D scene. Therefore, we should not expect that a simple homography will work everywhere in the image, but only on a significant part. On this part, we shall say that we have a dominant homography. At each pixel that is well-registered, the registered samples are i.i.d. samples of the same underlying Poisson model. As a result, a signal dependent noise model will be accurately estimated for each colour channel. This model simply is a curve of image intensity versus the standard deviation of the noise.

Averaging does not work at the mis-registered pixels, and block matching metods are at risk on the fine image structures. Thus they will be combined. The simple combination used here will be a convex combination of them, the weight function being based on the noise curve and on the observed standard deviation of the values for the accumulation at a certain pixel.

If this standard deviation is compatible with the noise model, the denoised value will be the mean of the samples. Otherwise, the standard deviation test will imply that the registration at this point is inaccurate and a conservative denoising will be applied.

References

[1] J. Digne. *Inverse Geometry: From the raw point cloud to the 3D surface - Theory and Algorithms*. PhD thesis, École Normale Supérieur de Cachan, 2010.

[2] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. Of the ACM*, 24:381395, 1981.

[3] N. Sabater. *Reliability and Accuracy in Stereovision. Application to Aerial and Satellite High Resolution Images*. PhD thesis, École Normale Supérieur de Cachan, 2009.