1 Clarifications

1.1 Structured Light Stereo

Some relevant reiews are [1],[2] and slides. A *structured light stereometric system* is similar to a passive stereo system where one of the camera is replaced by a projector. A light source projects light a vertical plane of light that creates a narrow stripe on the scene. The intersection of an illumination plane of known spatial position (corresponds to a projector column) and a line of sight (corresponds to a camera pixel) determines a point. For dense reconstruction of the scene, many images must be taken. To speed up the scanning process, spatially modulated light projector has been suggested, in which multiple illumination planes or rays can be projected simultaneously as part of a single illumination pattern. Spatial-temporal modulation of illumination, i.e. sequentially projecting several patterns, can be used for reliable identification of light planes. To enable acquisition of dynamic scenes, the number of projected patterns used should be as small as possible. Intuitively, the projected pattern impose illusion of texture on the object, increasing the number of correspondences, which enables reconstruction. Structured light stereo is equivalent to (1) solving the correspondence problem and then (2) computing stereo using triangulation.

The Correspondence Problem The correspondence problem can be stated simply

For each point in the left image, find the corresponding point in the right image

We first note that search space for the corresponding point can be restricted to pixels lying on the epipolar line. Epipolar plane is plane formed from points p, o_1, o_2 and epipolar line is intersection of epipolar plane with the image plane. For structured light stereo systems, the art of designing robust, fast, reliable coding

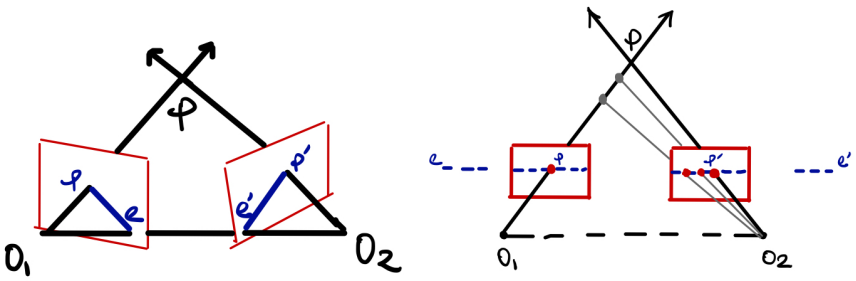

Figure 1: non parallel (left) and parallel (right) camera setup. pe and $p'e'$ are the epipolar lines

schemes serves to solve the correspondence problem.

Triangulation Given correspondence between projector and camera images, triangulation refers to the process of computing the distance of object relative to camera. Since a 3D point can be obtained by intersecting a ray (pixel of camera image) with a projector plane (a single code), it is necessary to encode a single axis in projected image to ensure unique reconstruction. *disparity* is the displacement between points of a conjugate pair p_l, p_r (points in different images of projection of same point in the scene) when the two images are superimposed. In the context of parallel calibrated cameras, disparity is inversely proportional to depth, when baseline b and focal length f are known

$$\frac{b}{z} = \frac{b + x_l - x_r}{z - f} \quad \text{implies} \quad z = \frac{bf}{x_l - x_r}$$

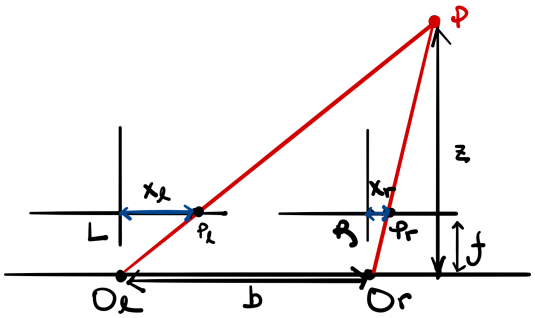

Figure 2: parallel-calibrated binocular stereo

1.2 Structured Light Coding

The goal of structured light coding is to find coding schemes (maps pattern intensity to indices of the projector light planes) that enables robust, reliable algorithms for finding correspondence.

Horn & Kiryati This paper generalizes Gray code to n-ary code in order to reduce the number of patterns that needs to be projected, (L^K instead of 2^K code words) [3]. The authors draw inspirations from communication theory, where the projector projects unique temporal codes, received at each image plane through a noisy channel and subsequently decoded. Let there be K patterns and L code words (distinct planes of light), we want to encode the indices of vertical light planes $x \in [L]$ using some encoding scheme $f : [L] \to \mathbb{R}^K$ such that the nearest neighbor decoding $\hat{x}(y) = \arg\min_{x\in[L]}(f(x)-y)$ of a normalized noisy observation $y \in \mathbb{R}^K$ minimizes the probability of depth estimation error. Given the forwarding model,

$$\mathsf{y} = f(\mathsf{x}) + \mathsf{n} \qquad \text{where} \qquad \mathsf{x} \sim \mathrm{Cat}(1/L) \quad \mathsf{n} \sim p_{\mathsf{n}}$$

implying $\mathsf{y}|\mathsf{x} = x \sim p_{\mathsf{n}}(y - f(x))$. We want to minimize the probability of depth estimation error, which roughly proportional to difference between true index x of the plane of light and the estimated index $\hat{x}$,

$$\text{minimize}_f \quad \left[\mathbb{E}_{\mathsf{x},\mathsf{y}}\left[(x-\hat{x}(y))^2\right] = \sum_{x=1}^{L} p_{\mathsf{x}}(x) \int p_{\mathsf{y}|\mathsf{x}}(y|x)(x-\hat{x}(y))^2 dy \propto \sum_{x=1}^{L} \int (x-\hat{x}(y))^2 p_{\mathsf{n}}(y-f(x))dy\right]$$

This optimization problem is hard. The paper suggest the use of space filling curves as the encoding function and established that Gray code is a special limiting case of the space filling curve. Note here we assume there is no *mutual illumination*, i.e. there is no interval reflection and so the projected codes $f(x)$ is proportional to observation y.

Phase Shifting Phase shifting is a particular coding method whereby the projector column coordinates are encoded as the (absolute) phase of a spatial sinusoidal pattern. Note we want to encode column coordinates because we only need to search along the horizontal epipolar lines. Let N be number of columns to be encoded. When the scene is projected with a cosine pattern of period T (measured in number of pixels) and therefore frequency $f = \frac{1}{T}$, an idealized image formation model for any pixel is

$$I = I_0 + A\cos(\Phi) = I_0 + A\cos(\phi) \tag{1}$$

where I_0 is the pixel intensity when no projection is used; I is the pixel intensity measured; A is amplitude (albedo/reflectance) of the signal; $\Phi = 2\pi fx = 2\pi n + \phi \in [0, 2\pi fN]$ for some number of period $n \in \mathbb{N}$

is the absolute phase, $\phi \in [0, 2\pi]$ is the relative phase. When measured w.r.t. number of pixels, x is the corresponding absolute phase and $\tilde{x}$ the corresponding relative phase, satisfying

$$x = Tn + \tilde{x} \qquad \text{or} \qquad x \equiv \tilde{x} \mod T \tag{2}$$

where $\tilde{x} = \frac{T\phi}{2\pi}$. In (1), I_0, A, ϕ are unknown and so ϕ cannot be determined. To solve for ϕ, phase shifting method projects K sinusoidal patterns of same frequency, each shifted by $\varphi_k = \frac{2\pi(k-1)}{K}$ for $k = 1, \cdots, K$. Thereby obtaining a system of K equations in 3 unknowns.

$$I_k = I_0 + A\cos(\phi + \varphi_k) \qquad \text{for} \qquad k = 1, \cdots, K \tag{3}$$

Although $K = 3$ suffices, larger values of K makes determination of relative phase ϕ more robust to noise. We determine ϕ using least squares

$$\text{minimize}_\phi \quad \left[\epsilon(\phi, I_0, A) := \sum_{k=1}^{K} \left[I_k - (I_0 + A\cos(\phi + \varphi_k))\right]^2\right] \tag{4}$$

Similar to appendix in [4] and results shown in [5, 4], we can show the following

$$0 = \frac{\partial \epsilon}{\partial \phi} = 2A \sum_{k=1}^{K} I_i \sin(\phi + \varphi_k) \propto \cos(\phi) \sum_{k=1}^{K} I_k \sin(\varphi_k) + \sin(\phi) \sum_{k=1}^{K} I_k \cos(\varphi_k)$$

$$\phi = \tan^{-1}\left[-\frac{\sum_{k=1}^{K} I_k \sin(\varphi_k)}{\sum_{k=1}^{K} I_k \cos(\varphi_k)}\right]$$

phase unwrapping aims to recover absolute phase x from relative phase $\tilde{x}$, which is non-trivial unless $T \geq N$. One particular choice of phase unwrapping method relies results in number theory [6]. The idea is to project patterns whose periods $T_1, \cdots, T_F$ are relative co-prime, each shifted by K times such that relative phase $\tilde{x}_1, \cdots, \tilde{x}_F$ can be solved using (4), from which we can use the Chinese Reminder Theorem to solve the following system of congruences

$$\begin{aligned} x &= \tilde{x}_1 \mod T_1 \\ &\vdots \\ x &= \tilde{x}_F \mod T_F \end{aligned} \tag{5}$$

Intuitively, projecting F spatial sinusoids with period $T_1, \cdots, T_F$ with different frequency emulates the projection of a low frequency spatial sinusoid with period $T = T_1 \times \cdots \times T_F$, shown in Figure (3). As

Figure 3: Two sinusoids with $T_1 = 17, T_2 = 31$ emulates a low frequency sinusoid with $T = 527$

specified in [7] and experimented in Figure (4), phase unwrapping is unstable when the relatives phases $\tilde{x}_1, \cdots, \tilde{x}_F$ are noisy. Simple application of medium filtering of the relative phase does not seem to help with phase unwrapping. Wavelet denoising with `wdenoise` seems to help to a limited extend, as shown in Figure (5)

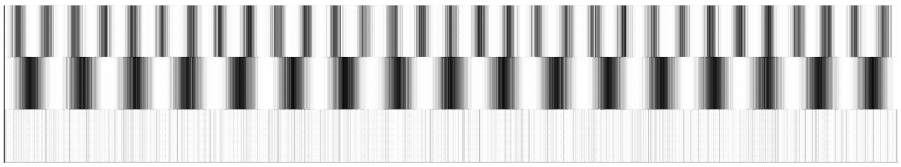

Figure 4: When relative phase $\tilde{x}_1, \tilde{x}_2$ are corrupted with i.i.d. additive Gaussian noise $\mathcal{N}(\mathbf{0}, 0.1 \cdot \mathsf{range}(\tilde{x}) \cdot \mathbf{I})$, phase unwrapping is unable to recover the absolute phase x robustly.

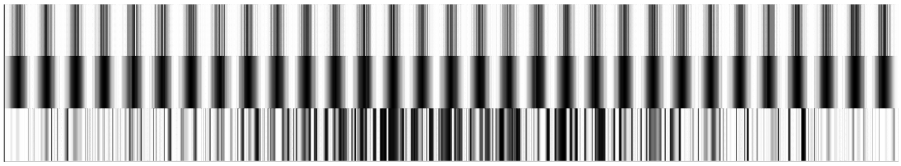

Figure 5: $\tilde{x}_1$ and wavelet denoised $\tilde{x}_1$ ($x_2, \tilde{x}_2$ not shown) helps with phase unwrapping to some extend

1.3 Image Priors

The choice of regularization has been an important research topic in image processing. Handcrafted priors have been successful in a number of different image recovery tasks. For example, we can choose to enforce task-specific priors: (1) the sparsity of $\mathbf{x}$ with ℓ-1 norm in image deblurring [8] (2) total variation in image denoising [9] (3) cross-channel correlation in color image demosaicing [10] (4) dark channel prior in image dehazing [11], etc. More exhotically, randomly initialized neural network can inject inductive bias to the optimization and act as image priors. [12]

In addition to hand-crafted priors, there has been interest in algorithm induced priors. Alternating direction method of multipliers (ADMM) is a common convex optimization method for inverse problem where the objective function is separable with respect to the *data term* and the *regularizer*. Each primal update involves an evaluation of a proximal operator, which can be interpreted as performing denoising on some iterate. [13, 14, 15] proposed plug-and-play priors where the choice of regularization is implicitly specified by the denoiser used. [16] proposed an explicit laplacian-based expression for the regularizer and generalizes the method to a number of different iterative optimization algorithms.

The convergence of plug-and-play ADMM is studied by a number of papers. [17] showed fixed point convergence of plug-and-play ADMM. [16] showed convergence of the algorithm under some mild conditions of the denoiser, which are satisfied by some state of the art denoisers like *Block-matching and 3D filtering (BM3D)* [18] and *Trainable Nonlienar Reaction Diffusion (TNRD)* [19]. Most recently, [20] established convergence given that the denoising network satisfy certain Lipschitz condition.

Some proposed to learn proximal operator from data. [21] used a CNN denoiser [22]. Instead of substituting the proximal operator with a denoiser, [23] learns a projection mapping to the space of natural images by training a single neural network and showed impressive results on a number of different linear inverse problems.