Master Thesis Summary

My master thesis was concentrated entirely on 3D Computer Vision. I summarize it in this paper. During my internship at the National University of Singapore, I researched in the CVRP lab under the supervisor, Prof.Gim Hee Lee. I worked on extrinsic parameters calibration of non-overlapping cameras. I was working on the project to obtain precise real-world experiment results. In my master thesis, I studied methods on the 3D reconstruction of non-rigid surfaces from realistic monocular video under the supervision of Prof.razvan, co-supervision of Prof.Moghadasi, and consult of Dr.Kamali. Several simplified methods based on the orthographic models have been investigated.

1 Extrinsic Parameters Calibration of Non-Overlapping Cameras

Motivation. Camera calibration is a significant problem in Computer Vision. It consists of extrinsic and intrinsic calibration. The intrinsic camera calibration is a well-known problem that researching on these concepts is beyond these research project [16]. The extrinsic parameters of cameras that have overlapping views can be calculated with linear eight-point or minimal five-point algorithms [9, 12]. In this project, we proposed a novel method for estimating extrinsic parameters of cameras without overlapping views. We formulate this problem by using light and shadow geometry and evaluate our approach with synthetic data and real-world experiments. Estimating light coordinate with respect to a reference coordinate system is crucial in photometric stereo problems. In addition to the extrinsic parameters calibration of cameras, the light position in world coordinate is computed by our proposed method.

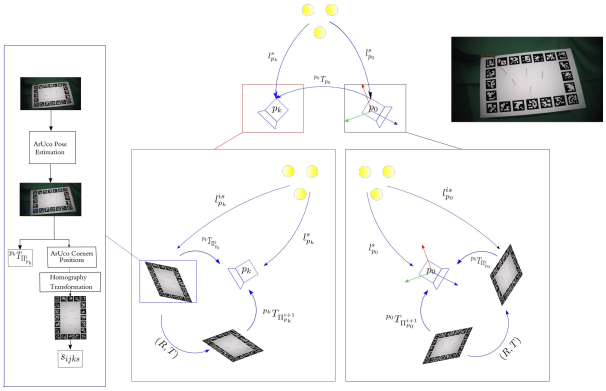


Figure 1: Overall schematic of the proposed method in two-camera extrinsic calibration, we consider camera pairs $(p_k, p_0), k = 1, \ldots, N_{cam}$ to estimate extrinsic parameters in multi-camera system.

Problem. *Suppose there exist N_{cam} cameras in multi-camera system that the views of these cameras do not have any overlap. The intrinsic parameters of these cameras are given. If we consider world coordinate system on the camera p_0, the goal of this research is to estimate homogeneous transformation matrices ${}^{p_k}T_{p_0}, k = 1, \ldots, N_{cam}$ which are 4×4 matrices that bring points defined in camera p_0 to camera p_k and they are defined as follows:*

$$ {}^{p_k}T_{p_0} = \begin{bmatrix} {}^{p_k}R_{p_0} & {}^{p_k}t_{p_0} \\ \mathbf{0} & 1 \end{bmatrix}, \quad k = 1, \ldots, N_{cam} $$

where ${}^{p_k}R_{p_0}$, ${}^{p_k}t_{p_0}$ are relative rotation and translation of camera p_k with respect to camera p_0 coordinate system.

Proposed Solution. Our goal is to calibrate cameras in multi-camera system when cameras do not have overlapping views in an indoor environment. We exploit at least three lights in the setup, and we estimate the extrinsic parameters of cameras by optimizing reprojection error. We need initialization to optimize our reprojection error. After estimating lights position in every camera coordinate system, we calculate extrinsic parameters of camera p_k by using lights position in two cameras p_k and reference camera p_0 coordinate system.
Therefore, we describe light and shadow geometry in this paper; then, we formulate our problem by using this geometry, and finally, we represent the evaluation results of the proposed solution on real-world and synthetic data.

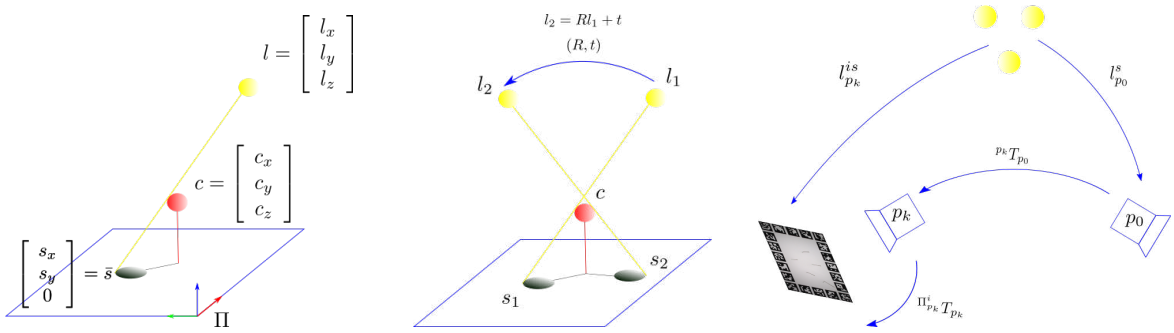


Figure 2: (a) Light and shadow geometry, a point light l casts a shadow $\bar{s}$ on the shadow receiver plane Π. (b) l_1 is transfomed to a new position l_2 with the pose (R, t). l_1, and l_2 cast shadows respectively s_1, and s_2 from the same caster c. (c) $l^s_{p_0}$ lights position in the world coordinate system are transformed to lights $l^{is}_{p_k}$ in the $\Pi^i_{p_k}$ coordinate system with the pose ${}^{\Pi^i_{p_k}}T_{p_k}{}^{p_k}T_{p_0}$.

Let us assume that the pose of a shadow receiver plane Π is fixed to the world coordinate system's xy plane, and a point light $l \in \mathbb{R}^3$ is in world coordinate. An infinitesimal caster located at $c \in \mathbb{R}^3$ casts a shadow on Π. It has a position of $s \in \mathbb{R}^2$ in Π coordinate system. As Figure 2a shows, a point light, caster, and its corresponding shadow lie on the same line, and the shadow position in world coordinate is $\bar{s} = [s^\top, 0]^\top$, so we have:

$$(c - \bar{s}) \times (l - \bar{s}) = \mathbf{0}. \tag{1}$$

By substituting $l = \begin{bmatrix} l_x & l_y & l_z \end{bmatrix}^\top, c = \begin{bmatrix} c_x & c_y & c_z \end{bmatrix}^\top, \bar{s} = \begin{bmatrix} s_x & s_y & 0 \end{bmatrix}^\top$ and their corresponding homogeneous coordinates $\hat{l}, \hat{c}, \hat{s} = \begin{bmatrix} s_x & s_y & 1 \end{bmatrix}^\top$ in the equation (1), we have equations with the constant λ as follows:

$$\begin{aligned}
\lambda\hat{s} &= \underbrace{\begin{bmatrix} -l_z & 0 & l_x & 0 \\ 0 & -l_z & l_y & 0 \\ 0 & 0 & 1 & -l_z \end{bmatrix}}_{L} \hat{c} = \underbrace{\begin{bmatrix} -l_z & 0 & l_x \\ 0 & -l_z & l_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \underbrace{\begin{bmatrix} c_x - l_x \\ c_y - l_y \\ c_z - l_z \end{bmatrix}}_{c-l} \\
&= \begin{bmatrix} -l_z & 0 & l_x \\ 0 & -l_z & l_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & -l_x \\ 0 & 1 & 0 & -l_y \\ 0 & 0 & 1 & -l_z \end{bmatrix} \hat{c} = K(I| - l)\hat{c} \\
&\Rightarrow \boxed{\lambda\hat{s} = K(I| - l)\hat{c}} \quad \texttt{and} \quad \boxed{\lambda\hat{s} = L\hat{c}} \quad \texttt{and} \quad \boxed{\lambda\hat{s} = K(c - l)}
\end{aligned} \tag{2}$$

Structure from Motion problem is analogous to light and shadow geometry. As the equation illustrates this similarity, point lights, shadow receiver plane, shadow casters, and two matrices of K, and $(I| - l)$ can be described correspondingly by pinhole cameras, the image plane, observed points, and camera intrinsic and extrinsic parameters [13].

Our main objective is to estimate the pose of cameras with respect to the world coordinate system. Since camera views do not have any overlaps, we use light and shadow geometry for constraining the problem. We denote the shadow receiver plane in front of camera p_k as Π_{p_k}. This plane undergoes multiple poses $\{{}^{\Pi^i_{p_k}}R_{p_k}, {}^{\Pi^i_{p_k}}t_{p_k}\}$ because single shadow observation does not provide sufficient information to solve the problem, so $\Pi^i_{p_k}$ is the plane Π_{p_k} in the ith pose. According to Figure 2c, the position of lights in the $\Pi^i_{p_k}$ coordinate system $l^{is}_{p_k}$ are related to the lights position in the world coordinate system $l^s_{p_0}$ respectively as follows:

$$\hat{l}^{is}_{p_k} = {}^{\Pi^i_{p_k}}T_{p_k}{}^{p_k}T_{p_0}\hat{l}^s_{p_0}, \quad {}^{\Pi^i_{p_k}}T_{p_k} = \begin{bmatrix} {}^{\Pi^i_{p_k}}R_{p_k} & {}^{\Pi^i_{p_k}}t_{p_k} \\ \mathbf{0} & 1 \end{bmatrix}, \quad s = 1, \ldots, N_l,\ k = 1, \ldots, N_{cam},$$
$$i = 1, \ldots, N_p$$

where ${}^{\Pi^i_{p_k}}T_{p_k}$ is a homogeneous transformation that transforms camera p_k to shadow receiver plane $\Pi^i_{p_k}$ coordinate systems, and $\hat{l}^{is}_{p_k}, \hat{l}^{s}_{p_0}$ are homogeneous coordinates of $l^{is}_{p_k}, l^{s}_{p_0}$. With the index of iks the matrix L_{iks} becomes

$$L_{iks} = \begin{bmatrix} -l^{iks}_z & 0 & l^{iks}_x & 0 \\ 0 & -l^{iks}_z & l^{iks}_y & 0 \\ 0 & 0 & 1 & -l^{iks}_z \end{bmatrix}$$

This problem can be formulated from the $\lambda\hat{s} = L\hat{c}$ as a reprojection error:

$$\begin{aligned} &\min_{c_j, l^s_{p_0}, {}^{p_k}R_{p_0}, {}^{p_k}t_{p_0}, \lambda_{ijks}} \sum_{k=1}^{N_{cam}} \sum_{s=1}^{N_l} \sum_{j=1}^{N_{cast}} \sum_{i=1}^{N_p} ||\lambda_{ijks}\hat{s}_{ijks} - L_{iks}\hat{c}_j||_{\mathcal{H}} \\ &s.t \quad \hat{l}^{is}_{p_k} = {}^{\Pi^i_{p_k}}T_{p_k}\, {}^{p_k}T_{p_0}\hat{l}^{s}_{p_0}, \end{aligned} \tag{3}$$

where N_{cam}, N_l are respectively the number of cameras and lights in the multi-camera system, N_{cast} is the number of casters on the plane Π_{p_k}, and N_p indicates the number of poses that the shadow receiver plane Π_{p_k} have in front of the camera p_k. A point light $l^{is}_{p_k}$ lights on the caster c_j, and a shadow with the position s_{ijks} on the shadow receiver plane $\Pi^i_{p_k}$ is casted. Huber cost function $||.||_{\mathcal{H}}$, which is hybrid between L_1 and least-square cost function, is used to reduce the effects of outliers with the parameter $b = 1$. We solve this nonlinear problem with the Levenberg-Marquardt algorithm, and it is implemented with Ceres Solver [1]. It is required to calculate the initial $c_j, l^s_{p_0}, {}^{p_k}R_{p_0}, {}^{p_k}t_{p_0}, j = 1, \ldots, N_{cast}; k = 1, \ldots, N_{cam}; s = 1, \ldots, N_l$ to optimize the problem.

Note that fixed pose of the shadow receiver plane and moving light position with (R, t) is equivalent to fixed light position and moving shadow receiver plane with (R, t). Therefore, let us assume that the pose of the shadow receiver plane is fixed, and point lights are moving in computations while the position of point lights are fixed and we move shadow receiver plane $\Pi_{p_k}, k = 0, \ldots, N_{cam}$ in practical experiments. Suppose that a point light l_1 is moved with the rotation R and translation t to a new position l_2, so there exist two shadows s_1 and s_2 from caster c (Figure 2b). We use equation (2) for estimating lights position in the world coordinate $l^s_{p_0}, s = 1, \ldots, N_l$, and the camera coordinate systems $l^s_{p_k}, k = 1, \ldots, N_{cam}$, so we have:

$$\begin{cases} \lambda_1\hat{s}_1 = K_1(c - l_1) & \rightarrow \quad \lambda_1 K_1^{-1}\hat{s}_1 = c - l_1, \\ \lambda_2\hat{s}_2 = K_2(c - l_2) & \rightarrow \quad \lambda_2 K_2^{-1}\hat{s}_2 = c - l_2, \end{cases} \implies \quad \hat{s}_2^\top K_2^{-\top}[l_1 - l_2]_\times K_1^{-1}\hat{s}_1 = 0, \tag{4}$$

where $[l_1 - l_2]_\times$ denotes 3×3 skew-symmetric matrix. We expand the equation (4) by substituting $\hat{s}_1 = \begin{bmatrix} s_{1x} \\ s_{1y} \\ 1 \end{bmatrix}, \hat{s}_2 = \begin{bmatrix} s_{2x} \\ s_{2y} \\ 1 \end{bmatrix}, l_1 = \begin{bmatrix} l_x \\ l_y \\ l_z \end{bmatrix}, R = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{bmatrix}, t = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$, so the equation becomes

$$\begin{aligned} &(a_0)l_x^2 + (a_1)l_x l_y + (a_2)l_x l_z + (a_3)l_x + (a_4)l_y^2 + (a_5)l_y l_z + (a_6)l_y + (a_7)l_z^2 + (a_8)l_z + a_9 = 0 \\ &a_0 = r_7 s_{2y} - r_7 s_{1y} \\ &a_1 = r_7 s_{1x} - r_7 s_{2x} - r_8 s_{1y} + r_8 s_{2y} \\ &a_2 = r_1 s_{1y} - r_1 s_{2y} - r_4 s_{1x} + r_4 s_{2x} - r_9 s_{1y} + r_9 s_{2y} \\ &a_3 = s_{2y} t_3 - s_{1y} t_3 - r_7 s_{1x} s_{2y} + r_7 s_{1y} s_{2x} \\ &a_4 = r_8 s_{1x} - r_8 s_{2x} \\ &a_5 = r_2 s_{1y} - r_2 s_{2y} - r_5 s_{1x} + r_5 s_{2x} + r_9 s_{1x} - r_9 s_{2x} \\ &a_6 = s_{1x} t_3 - s_{2x} t_3 - r_8 s_{1x} s_{2y} + r_8 s_{1y} s_{2x} \\ &a_7 = r_3 s_{1y} - r_3 s_{2y} - r_6 s_{1x} + r_6 s_{2x} \\ &a_8 = s_{1x} s_{2y} - s_{1y} s_{2x} - s_{1x} t_2 + s_{1y} t_1 + s_{2x} t_2 - s_{2y} t_1 - r_9 s_{1x} s_{2y} + r_9 s_{1y} s_{2x} \\ &a_9 = -s_{1x} s_{2y} t_3 + s_{1y} s_{2x} t_3 \end{aligned}$$

The unknown variables are l_x, l_y, l_z in the above nonlinear polynomial. Since this equation is a quadratic polynomial with three variables, we use Gröbner basis to solve this equation. This polynomial has three variables, so we need to have three equations to provide enough constraints for Gröbner Solver. Polyjam is used as a Gröbner Solver in implementations [10]. Two cases are considered to find the minimal solutions for equation (4):

- **Case one**: One caster is placed on the shadow receiver plane, and four positions for point lights are assumed.
- **Case two**: Two casters are placed on the shadow receiver plane, and three positions for point lights are assumed.

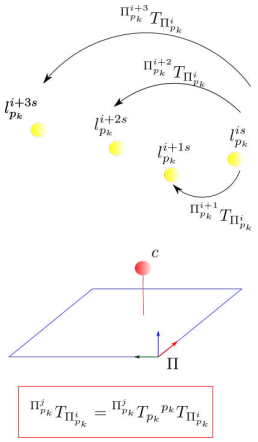


Figure 3: Case one

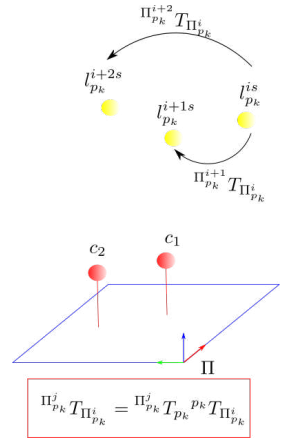


Figure 4: Case two

In both of these cases, we have more than one solution. Therefore, we use an additional pose of the shadow receiver plane and check this pose in constrain (1) to determine the unique solution. After estimating $l^s_{p_k}, s = 1, \ldots, N_l, k = 0, \ldots, N_{cam}$, we need to find the good initial estimation for ${}^{p_k}R_{p_0}, {}^{p_k}t_{p_0}$. As rotation matrix and translation vector have six degrees of freedom, we have to exploit at least three lights to estimate ${}^{p_k}R_{p_0}, {}^{p_k}t_{p_0}, k = 1, \ldots, N_{cam}$ uniquely. The lemma below explains the algorithm for computing the initial homogeneous transformation ${}^{p_k}T_{p_0}$. The proof of this can be found in [14].

Lemma. *Suppose two sets $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ and $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$ are corresponding point sets in $\mathbb{R}^d$, then our goal is to estimate rotation matrix R and translation vector t that transforms point set $\mathcal{P}$ to point set $\mathcal{Q}$ which holds below:*

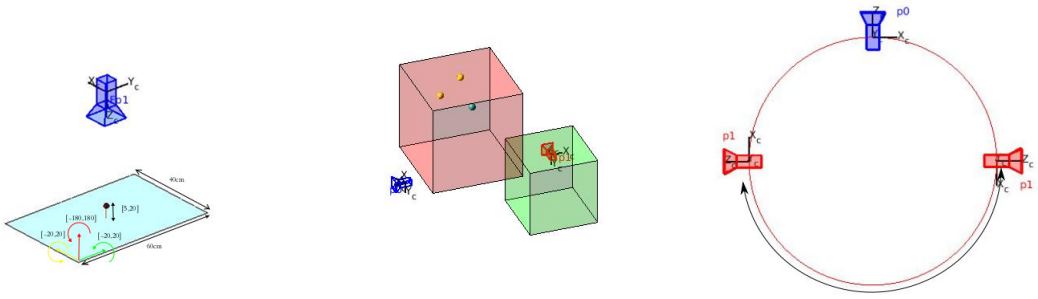
$$(R, t) = \underset{R \in SO(3), t \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} ||(Rp_i + t) - q_i||^2$$

Then optimal rotation and translation are as follow:

$$R = V \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \ldots & & \\ & & & 1 & \\ & & & & det(VU^\top) \end{pmatrix} U^\top,$$
$$t = \bar{q} - R\bar{p},$$

where centroid of point sets are $\bar{p} = \frac{\sum_{i=1}^n p_i}{n}, \bar{q} = \frac{\sum_{i=1}^n q_i}{n}$, and centered vectors are estimated as follows: $x_i = p_i - \bar{p}, y_i = q_i - \bar{q}; i = 1, \ldots, n$. The singular value decomposition of the $d \times d$ covariance matrix $S = XY^\top$, such that X and Y are the $d \times n$ matrices that have x_i and y_i as their columns respectively, should be computed as $S = U\Sigma V^\top$. U, V are orthogonal matrices and Σ is a diagonal matrix with non-negative real numbers on the diagonal.

Implementation. We evaluate the proposed method with synthetic data and real-world experiments. We assume two cameras in both of the simulation and real-world experiments. We simulate our proposed method with synthetic data. We uniformly distribute board poses, casters, and lights position. As Figure 5b illustrates, at least three lights are sampled uniformly in a cube with a length of 80 cm. Pins are placed



(a) (b) (c)

Figure 5: (a) Plane Π undergoes different poses randomly. (b) At least three lights are distributed uniformly in a red cube with a length 80 cm, 10 cm away from the world coordinate center, and the second camera is selected uniformly in the green cube with the length 60 cm, 80 cm away from world coordinate center. (c) angle around y-axis is selected randomly within the range $[-270°, -90°]$.

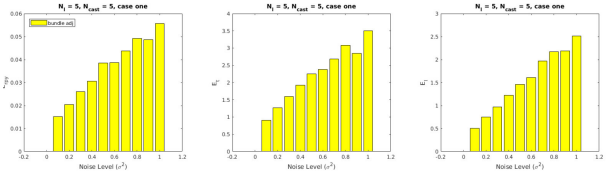
uniformly on the plane 10 cm away from edge of the plane. Pins with the height uniformly within the $[5, 20]$ cm are selected. The board size is 40×60 cm, and its poses are randomly sampled within the range shown in Figure 5a. The second camera is placed uniformly in the green cube shown in Figure 5b, and the camera is rotated randomly around x-axis and z-axis within the range $[-10°, 10°]$, and y-axis within the range $[-270°, -90°]$. As Figure 5c shows, we want to avoid overlapping views in cameras. For both cases mentioned previously, we compute shadows with the synthetic data, and we compute the errors as follow:

$$E_l = \frac{\sum_{s=1}^{N_l} ||l^s_{p_0} - \{l^s_{p_0}\}_{GT}||_2}{N_l},$$

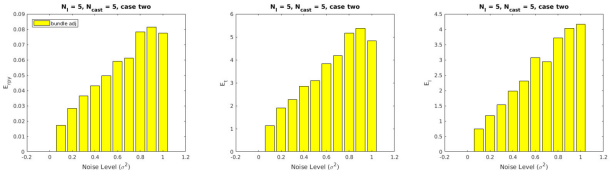
$$E_t = ||\{^{p_1}R_{p_0}\}^{\top}_{GT} {}^{p_1}t_{p_0} - \{^{p_1}R_{p_0}\}^{\top}_{GT}\{^{p_1}t_{p_0}\}_{GT}||_2,$$

$$E_{rpy} = f(\{^{p_1}R_{p_0}\}^{\top}_{GT} {}^{p_1}R_{p_0})$$

where E_l, E_t, E_{rpy} are mean lights error, translation, and rotation error respectively. Note that translation and rotation error are relative pose error between computed pose ${}^{p_1}T_{p_0}$ and ground truth pose $\{^{p_1}T_{p_0}\}_{GT}$. The relation $f : R \in SO(3) \longrightarrow f(R) = (\psi, \theta, \phi)$ transforms rotation matrix R to angles around axis (ϕ, θ, ψ), where $R = R_z(\phi)R_y(\theta)R_x(\psi)$. We perturbed the shadow positions with Gaussian noise with mean zero and variance σ^2 to analyze the influence of shadow localization. The unit specified in the simulation results is centimeter. As Figure 6 shows, three errors increase when the noise variance increases, and it is shown that the method is more robust in case one in comparison to case two. Hence, we evaluate our method on real-world experiments in case one



(a) Noise effects on shadow localization for case one. Five casters and five lights are used in this simulation with Gaussian noise of variance σ^2 and mean zero. Every data is the median of 500 simulated iterations.



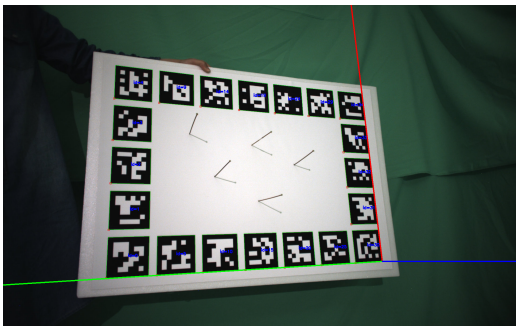
(b) Noise effects on shadow localization for case two. Five casters and five lights are used in this simulation with Gaussian noise of variance σ^2 and mean zero. Every data is the median of 500 simulated iterations.

Figure 6

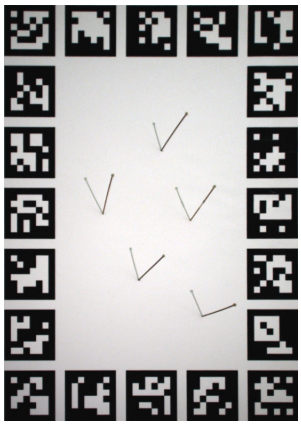
Table 1: Result of real-world experiments in three laboratory environments

Environment	**Case**	N_l	N_{cast}	N_p	E_{rpy}	E_t
Env1	one	3	4	5	0.0532576	4.18041
Env2	one	5	4	5	0.0928035	8.61596
Env2	one	3	6	5	0.163707	24.207

We evaluate the proposed method in three laboratory environments Env1, Env2, and Env3. We exploit two Point Grey cameras and three lights in the laboratory environment (Figure 7c).
ArUco markers [7] are printed on a A2 paper, and we use OpenCV 3D pose estimation[4] to obtain the shadow receiver plane pose with respect to camera coordinate system ${}^{\Pi^i_{p_k}}T_{p_k}$. Pins with length $\sim$ 10 cm and head diameters of $\sim$ 3 mm, which are small enough to localize them accurately and big enough to detect, are used as shadow casters.
For shadow detection, we compute homography transformation by using four corners of ArUco markers, Figure 7a is transformed by homography transformation, and the result is Figure 7b.
After computing homography transformation, we manually label shadow positions and estimate s_{ijks} in this transformed image, and we report the evaluation result in Table 1. Two errors E_{rpy}, and E_t are reported in this table.
According to this table, translation errors are big enough to use this method for calibration, so the future objective is to recognize the mistake that caused this significant error.



(a) A photo taken from ArUco markers, pose of the board with respect to camera coordinate is computed.



(b) Homography transformation of Figure 7a

(c) Laboratory environment

Figure 7

2 3D Reconstruction of Non-rigid Surfaces

Motivation. Structure from Motion (SfM) is the process of estimating 3D structure from a set of 2D images and the pose of a camera in the world coordinate system. Recovering 3D structures of a scene observed by moving cameras is a classical problem in Computer Vision. Rigid Structures have made significant progress toward achieving this goal, especially in multiple-view approaches, the problem will be complicated when givens are monocular realistic 2D correspondences. The difference between rigid and non-rigid structures is that the deformable objects generally vary their shapes over time, so the number of unknown parameters increases dramatically in comparison to rigid SfM problems. Non-rigid 3D shape recovery has broad applications in different domains, such as the entertainment, media industry, and the medical field. It is a challenging and ambiguous problem when the sensor data is noisy, which is typically the case when dealing with real images; in this case, the camera model is a perspective view, and images are uncalibrated.

Problem. *Consider the non-rigid structure has n points, and there are its homogeneous coordinates in F frames. The jth point is projected in the ith frame as below:*

$$d_{ij}U_{ij} = K_i[R_i|T_i]X_{ij},$$

where $U_{ij} = (u_{ij}, v_{ij}, 1)^\top$ and $X_{ij} = (x_{ij}, y_{ij}, z_{ij}, 1)^\top$ are respectively the homogeneous image coordinate and 3D world coordinate of the jth point in the ith frame. K_i is the 3×3 calibration matrix that is just related to the focal length of f_i in every frame. This assumption is practical in real-world sequences. R_i and T_i are respectively rotation matrix and translation vector in the ith frame, and $d_{ij}$is known as a projective depth. Now if we consider n points and F frames in one equation, the problem will be formulated as follows:

$$W = \begin{bmatrix} d_{11}U_{11} & \dots & d_{1n}U_{1n} \\ \vdots & \vdots & \vdots \\ d_{F1}U_{F1} & \dots & d_{Fn}U_{Fn} \end{bmatrix} = \begin{bmatrix} K_1[R_1|T_1]S_1 \\ \vdots \\ K_F[R_F|T_F]S_F \end{bmatrix},$$

where the $4 \times n$ matrix S_i represents the 3D shape observed in the frame i. W is $4F \times n$ input measurement matrix that is known in this problem and all other parameters are unknown.
Under the orthographic model, unknown parameters are just R_i and S_i in F frames and formulation of the problem is more straightforward in comparison to the equations above, and we have:

$$\begin{aligned} W &= \begin{bmatrix} W_1 \\ \vdots \\ W_F \end{bmatrix} = \begin{bmatrix} R_1 & & \bigcirc \\ & \ddots & \\ \bigcirc & & R_F \end{bmatrix} \begin{bmatrix} S_1 \\ \vdots \\ S_F \end{bmatrix} + \begin{bmatrix} t_1 \\ \vdots \\ t_F \end{bmatrix} [1, \dots, 1] \\ &= RS + T\mathbf{1}^\top \end{aligned}$$

Note that we can eliminate the translation component from the equation above by registering the image coordinates to the centroid in each frame i. As a result, the equation becomes

$$W = RS$$

Therefore, the main goal is to estimate R and S under the orthographic model.
Literature review. Tomasi and Kanade proposed a factorization method for recovering 3D shapes of rigid objects from 2D correspondences with the assumption of the orthographic camera model [15]. Bregler [5] pioneered the first solution to non-rigid structure from motion by extending Tomasi and Kanade rigid factorization approach, he assumed that 3D deformable shape in every frame is a linear combination of a set of shape basis. Therefore, we need to estimate shape basis and its coefficients instead of shapes in every frame. Akhter [3] attacked the problem in another direction and showed that 3D shape in any frame in the sequence could be expressed as a linear combination of trajectory basis, which are predefined and Discrete Cosine Transform (DCT) basis can be used to describe most real motions compactly. Bregler, the same as Tomasi and Kanade, has enforced the orthonormality constraint of rotation matrices to solve ambiguities because matrix factorization is not unique. Xiao [17] asserted that using only the rotation constraints results

in ambiguous and invalid solutions. The ambiguity arises from the fact that the shape bases are not unique. An arbitrary linear transformation of the bases produces another set of eligible bases. To eliminate the ambiguity, a set of new constraints, basis constraints, which uniquely determine the shape bases, has been proposed. Akhter [2] proved that orthonormality constraints are sufficient to recover the 3D structure from image observations alone. Some papers have discussed non-rigid SfM under weak perspective camera model, While this camera model can be sufficient when the variation in depth over the whole object is relatively small, it is known that the full perspective model is often more accurate to what is observed in real images. Therefore, several authors have proposed non-rigid SfM formulations for the full perspective case [18, 8]. One of the weaknesses of non-rigid structure-from-motion techniques is their sensitivity to missing data and mismatches. Garg [6] offered the first variational approach to the problem of dense 3D reconstruction of non-rigid surfaces from monocular video sequences. The problem is formulated with calibrated images under the orthographic camera model. However, this method reconstructs highly deforming smooth surfaces densely and accurately directly from video without the need for any prior models. Kumar [11] proposed a new approach for dense non-rigid SfM under the orthographic model by modeling the problem on a Grassmann manifold. Specifically, they assumed the complex non-rigid deformations lie on a union of local linear subspaces both spatially and temporally. This naturally allows for a compact representation of the complex non-rigid deformation over frames that previous methods are unable to show.

References

[1] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. `http://ceres-solver.org`.

[2] Ijaz Akhter, Yaser Sheikh, and Sohaib Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1541. IEEE, 2009.

[3] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009.

[4] Gary Bradski and Adrian Kaehler. Opencv. *Dr. Dobb's journal of software tools*, 3, 2000.

[5] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *cvpr*, volume 2, page 2690. Citeseer, 2000.

[6] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1272–1279, 2013.

[7] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[8] Richard Hartley and René Vidal. Perspective nonrigid shape and motion recovery. In *European Conference on Computer Vision*, pages 276–289. Springer, 2008.

[9] Richard I Hartley. In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 19(6):580–593, 1997.

[10] Laurent Kneip. Grobner solver. `https://github.com/laurentkneip/polyjam`.

[11] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2018.

[12] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):0756–777, 2004.

[13] Hiroaki Santo, Michael Waechter, Masaki Samejima, Yusuke Sugano, and Yasuyuki Matsushita. Light structure from pin motion: Simple and accurate point light calibration for physics-based modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[14] Olga Sorkine-Hornung and Michael Rabinovich. Least-squares rigid motion using svd. *Computing*, 1(1), 2017.

[15] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2):137–154, 1992.

[16] Roger Y Tsai. An efficient and accurate camera calibration technique for 3d machine vision. *Proc. of Comp. Vis. Patt. Recog.*, pages 364–374, 1986.

[17] Jing Xiao, Jinxiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67(2):233–246, 2006.

[18] Jing Xiao and Takeo Kanade. Uncalibrated perspective reconstruction of deformable structures. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1075–1082. IEEE, 2005.