RESEARCH-ARTICLE

Prediction of 3D rotation and translation from 2D images

BHATTARABHORN WATTANACHEEP, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

ORACHAT CHITSOBHUK, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

Published: 27 July 2019

ICCCM 2019: The 7th International Conference on Computer and Communications Management
July 27 - 29, 2019
Bangkok, Thailand

Prediction of 3D rotation and translation from 2D images

Bhattarabhorn Wattanacheep
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang Bangkok,
Thailand 10520
nroskool2@gmail.com

Orachat Chitsobhuk
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang Bangkok,
Thailand 10520
orachat.ch@kmitl.ac.th

ABSTRACT
The prediction of three-dimensional (3D) rotation and translation can be retrieved from two-dimensional (2D) images to build 3D models from large collections of images. In this paper, the process starts by extracting the features of images via transfer learning approach from Deep Neural Network model called VGG19. Even though the features extracted from VGG19 are usually adopted in image recognition application; in this research, we apply these features to the prediction model to obtain rotation and translation parameters. Due to the large size of the feature dimensions, it is necessary to perform dimensional reduction technique called latent semantic analysis (LSA) to decrease the feature dimensions and remain only the important ones. Then, the regression estimation technique based on the idea of Support Vector Machine (SVM) is used to predict the rotation and translation parameters. The accuracy is estimated by comparing the prediction results with the corresponding ground truth set. The average errors of rotation and translation of 3D prediction from 2D images are approximately 0.2419 degrees and 1.35 meters respectively.

CCS Concepts
• **Computing methodologies→Camera calibration.**

Keywords
3D Reconstruction; Image Processing; Robotics; Deep Learning.

1. INTRODUCTION
Image transformation is a process to convert a 3D world into a 2D image using a camera model with geometric relationship between a 3D position and its 2D corresponding projection onto the image plane. The model starts by processing the color and light through the pixel sensor within the camera. Two types of parameters needed to be estimated are intrinsic camera parameters- the parameters necessary to link the pixel coordinates of an image with the corresponding coordinates in the camera reference frame- and extrinsic camera parameters- the parameters that define the translation and orientation of the camera reference frame with respect to a known world reference frame. The intrinsic camera parameters include focal length, optical center, and skew coefficient can be obtained through the camera calibration process.

ICCCM 2019, July 27–29, 2019, Bangkok, Thailand

ACM ISBN 978-1-4503-7195-7/19/07…$15.00
https://doi.org/10.1145/3348445.3348485

However, in this paper, we focus on extrinsic camera parameter estimation consisting of rotation and translation in the X-axis, Y axis and Z-axis. Several techniques have been proposed to estimate these parameters from a single camera, stereo camera and multiple cameras.

Recent well-known work on estimating extrinsic camera parameters was a structure from motion, SfM, of a series of photos on social media sites such as Flickr [2-7]. SfM operation was mostly based on feature mapping as presented in Li et al. [4]. First, the images were grouped by finding the iconic images to construct the basic spanning tree structure of the image relation. Then, the rotation and translation were estimated using the SfM technique as presented by Singha's et al [6] with the linear SfM method. Most approaches to SfM from unstructured image collections were operated iteratively, starting with a small seed reconstruction, then growing through repeated integration of additional cameras and scene points. Even though such iterative approaches have been quite successful, they initiated two significant drawbacks. First, these methods tended to require heavy calculations on repeated non-linear optimization that attempted to refine camera parameters and scene structure as well as outlier rejection to remove inconsistent measurements. Second, these methods did not consider all images equally, thus led to different results depending on the order in which photos were considered. This sometimes can cause failure due to local minima or cascades of misestimated cameras. Such methods can also make estimation of rotation and translation parameters grow over time and introduce many errors.

Afterwards, Yihui-he [14] used the same SfM technique, starting with using SIFT to find the corresponding feature point between two images, then used the MSAC technique to filter the feature points that were not consistent with the original image instead of the RANSAC. This makes the projection of the camera more accurate. Although this method provides better performance, it still continues the incremental calculation structure. As a result, the calculation load cannot be reduced, and the order still affects the estimation efficiency. Therefore, the above-mentioned problems still cannot be solved.

Consequently, this paper proposes the prediction of 3D rotation and translation from 2D images of multiple cameras - every photo is used to calculate the rotation and translation simultaneously - using the machine learning of the features learned from deep learning model. The process starts by training the VGG19 to extract features from the sample images. These features are then used to predict the rotation and translation in 3D of image. In order to reduce large computation, the dimensional reduction technique called Latent Semantic Analysis (LSA) technique is adopted to remain just the dimension of important features. Finally, the features are regressed with the principle of Support Vector Machine, in which all images are considered equally. Therefore, the results will not depend on the sequence of photos

that have been processed. In addition, the proposed method introduces the noniterative process for parameter estimation, which help to improve computational efficiency.

The proposed algorithm is detailed in section 2 while section 3 presents the experimental results. Finally, summarize and conclusion are presented in section 4.

2. PROPOSED ALGORITHM

This research presents the prediction of 3D rotation and translation from multiple views of 2D images simultaneously. The overview of the proposed prediction system is shown in Figure 2.1. Image features are analyzed using Deep feature extraction; then the extrinsic camera parameters in terms of rotation and translation are further estimated by the principle of Support Vector Machine (SVM). The number of 4096 features of each image are extracted from the 7th fully connected layers of the VGG19. Since feature dimension is quite large, it would require excessive computational complexity. It is necessary to reduce the feature dimensions in order to preserve only the necessary ones. In this paper, we adopt the Latent Semantic Analysis (LSA) technique for the dimensional reduction task. Finally, rotation and translation parameters are estimated using support vector regression technique. We obtain the means of the root mean square error (RMSE) to determine prediction errors.

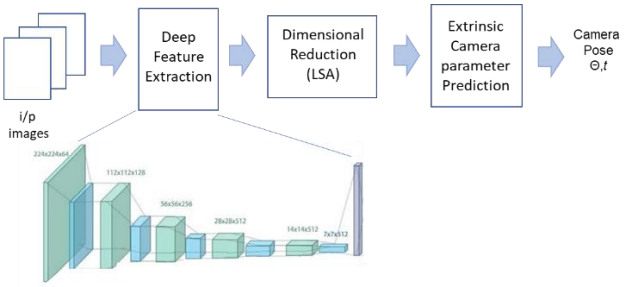

Figure 1. The overview of the proposed prediction of 3D rotation and translation from 2D images.

2.1 Deep Feature Extraction using VGG19 [8]

The input image will be randomly divided into training dataset and test dataset. Then, we apply deep feature extraction adopted from VGG19 deep learning model to estimate the important image features. Even though the model was originally used to extract features from the two-dimensional images for the object recognition applications; in this research, the features will be applied for learning and predicting rotation and translation in 3D from images.

Typically, the size of the input image submitted to the VGG19 is 224×224 pixels of the RGB color image obtained from ImageNet. The image is passed through a stack of convolutional (conv.) layers; where filters are employed with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations, the model also utilizes 1×1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution. Spatial pooling is carried out by five max-pooling layers, which is performed over a 2×2 pixel window with a stride of 2. A stack of convolutional layers is followed by three fully connected (fc) layers: the first two layers consist of 4096 channels each while the third one performs 1000-way ImageNet Large Scale Visual Recognition Challenge (ILSVRC) classification. The final layer is the soft-max layer, which is classification output layer. Since the objective of our proposed system is to perform regression not classification, we choose to retrieve the image features from the second fc layer (fc7) and used them for estimating the designated parameters.

2.2 Dimensional Reduction [9]

Features extracted from previous process contain large dimensions and can introduce excessively high complexity in prediction of rotation and translation. As a result, feature dimension must be reduced in order to accelerate the prediction while preserve only the necessary meaningful ones. In this paper, the Latent Semantic Analysis (LSA) technique is adopted. It is constructed from a mathematical technique called Singular Value Decomposition (SVD). In this way, the number of rows will be reduced, but still retaining a similar image feature structure.

SVD is robust and reliable orthogonal matrix decomposition method. It is one of the most powerful computational tools in numerical linear algebra. In particular, SVD is commonly used to solve i) the unconstrained linear least squares problems, ii) matrix rank estimation and iii) canonical correlation analysis. Further, SVD tells that any matrix A with arbitrary dimensions $m \times n$ can be represented as orthogonal matrices U and V and a diagonal matrix D as followed.

$$A = UDV^T \tag{1}$$

where the columns of U and V are the left and right singular vectors, respectively, and D is a diagonal matrix whose diagonal entries are the singular values of A. Since matrix V is orthogonal, V^T can instead be V^{-1}. The diagonal of the matrix D consists of singular values ordered from the largest value in the first column to the smallest value in the last column of the matrix. Since the matrix A consists of larger number of rows than those of columns $(m \geq n)$, it results in orthogonal matrix U of size $m \times n$, D is the diagonal matrix of size $n \times n$and V is the $n \times n$ orthogonal matrix. Therefore, the matrix product gives the relationship between the image and the feature, which can be written as: $A^TA = VDU^TUDV^T = VD^2V^{-1}$. D^2 is the eigen value of A^TA and the column of V is the eigen vector of A^TA. The singular value of A is the square root of the eigen value of A^TA, Therefore, the singular value is a real number and $AA^T = UD^2U^{-1}$. In order to understand the SVD, the rows of an $m \times n$ matrix A are represented as m images in a n-dimensional space and it is considered as the problem of finding the best k-dimensional subspace with respect to the set of images as shown in equation 2.

$$X_k = U_k\Sigma_kV_k^T \tag{2}$$

After finding best k-dimensional subspace of the features, they will be derived to obtain reduced dimension features and used for prediction of 3D rotation and translation. The prediction is performed using the Support Vector Regression method as discussed in the next section.

2.3 Support Vector Regression [10]

Support Vector Machine is one of the most popular machine learning algorithms. The concepts are relatively simple. Suppose $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset \mathcal{X} \times \mathbb{R}$ is a training data when $\mathcal{X}$ denotes the space of input format. The goal is to find a function $f(x)$ with $\mathcal{E}$ deviation from the actual target, y_i are received for all the

training data. Similarly, the error value will be ignored as long as it is less than the deviation. Equation 3 descries the case of linear relationship, where $\langle \cdot,\cdot \rangle$ is the dot product of $\mathcal{X}$, w is the small seek value.

$$f(x) = \langle w, x \rangle + b \qquad \text{where } w \in \mathcal{X}, b \in \mathbb{R} \quad (3)$$

The linear functions f estimates precision of all pairs (x_i, y_i) with $\mathcal{E}$; which is a possible convex optimization solution. However, some error values may be excluded. In practice, data may not be able to be linearly divided. In this case, the "soft margin" loss function may be used instead of slack variable $\xi_i, \xi_i^* \geq 0$ for training data samples that violate the support vector conditions. The conditions of the specified Slack value are shown in Figure 2

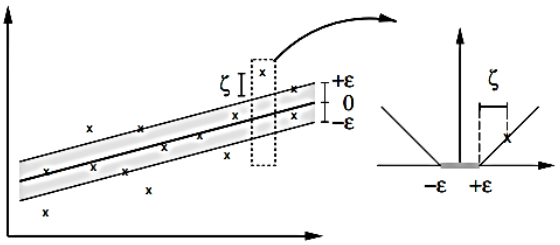

Figure 2. The soft margin loss setting for a linear SVM (from Scholkopf and Smola, 2002)

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$\text{subject to} \quad \begin{cases} y_i - \langle w, x \rangle - b \leq \mathcal{E} + \xi_i \\ \langle w, x \rangle + b - y_i \leq \mathcal{E} + \xi_i^* \end{cases} \quad (4)$$

The constant C determines the importance of the scope and the need of the Slack variable. In other words, the low value C makes the method focus on the Soft-margin SVM, while the large value C makes the method focus on Hard-margin SVM.

The dot product of the pair of samples can be viewed as similarity within the samples. Therefore, it is possible to use SVM to divide the group of similar data without the need to use the actual feature values which are replaced by the similarity of data. These similarities can be formed as a Kernel. Function $K(\bar{X}, \bar{Y})$ measures the similarity between point $\bar{X}$ and $\bar{Y}$. The concept is that the Kernel function may be a dot product between pairs of points in the newly converted area (shown by mapping function $\Phi(\cdot)$)

The SVM algorithm only depends on dot products between patterns $\bar{X}$. Hence, it suffices to know $K(\bar{X}, \bar{Y}) = \Phi(\bar{X}) \cdot \Phi(\bar{Y})$ rather than Φ explicitly which allows us to restate the Support Vector optimization problem.

Several kernels can be chosen such as Linear, Radial Basis, or Polynomial kernels. In this research, the Gaussian Radial Basis Kernel is used as followed

$$K(\bar{X}_i, \bar{X}_j) = e^{-\left|\left|\bar{X}_i - \bar{X}_j\right|\right|^2 / 2\sigma^2} \quad (5)$$

Where $\left|\left|\bar{X}_i - \bar{X}_j\right|\right|^2$ may be recognized as the squared Euclidean distance between the two feature vectors, σ is the standard derivation

After predicting the rotation and translation of the test set, the model accuracy is computed using Root Mean Square Error (RMSE) measurement.

2.4 The Root Mean Square Error (RMSE)

The RMSE has been used as a standard statistical parameter to measure model performance in several natural sciences. The parameter indicates the standard deviation of the residuals or how far the points are from the regression or modelled line as presented in equation 6.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \quad (6)$$

Where: N is number of samples, x_i is the observed values, $\hat{x}_i$ is the forecasts.

3. EXPERIMENTAL RESULTS

In this paper, we adopt the 2D image Dataset from the CMU Panoptic Studio [1]. The images are taken by different rotation and translation parameters in X, Y and Z axis around the objects. We conduct the model performance evaluation on a variety of challenging scenes such as in the presence of significant occlusion (Circular Movement: Three people rotate around a person at the center (Figure 3(a)) and large displacement (Bat Swing: A person swings a baseball bat. (Figure 3(b)). Sequence frames of Circular Movement were captured for 250 samples and Bat Swing was captured 200 samples for each of 480 cameras. The cameras are extrinsically and intrinsically calibrated and synchronized via an external clock. The dataset contains 209,934 images, which are randomly divided into train data 146,957 images and test data 62,977 images. The ground truth of each image consists of observed rotation (R) and translation (T) parameters defined as extrinsic camera parameters. The total number of R and T ground truth combinations were 480 values.

(a) The examples of postures from Bat Swing Dataset

(b) The examples of postures from Circular Movement Dataset

Figure 3. The examples of Dataset from different challenging scenes.

This research presents the prediction of 3D rotation and translation from 2D images from multiple cameras estimated simultaneously. The deep features of each image are extracted from the second fully connected (fc7); with the number of 4096 features, from the VGG19 and reduced to 1000 features using LSA. Finally, the rotation and translation parameters are estimated with support vector regression (SVR). The prediction performance of the SVR with several adjusted C and gramma parameters is

presented in Table 1, where C is in [0.001,0.01,0.1,1,10,100 and 1000] and the gamma is in [0.0001,0.001,0.01,0.1,1.0,10,100], respectively. From table 1, the best C and gamma parameters are C = 10 and gamma = 1 for rotation estimation with the RMSE of 0.24 degree and C = 1000 and gamma = 1 for translation prediction with the RMSE of 1.35 meters.

Table 1. The prediction accuracy of rotation (R) and translation (T) for fine tune model parameters

C	AVG RMSE of R							AVG RMSE of T						
	gamma							Gamma						
	0.0001	**0.001**	**0.01**	**0.1**	**1**	**10**	**100**	**0.0001**	**0.001**	**0.01**	**0.1**	**1**	**10**	**100**
0.001	0.34	0.34	0.34	0.34	0.34	0.34	0.34	1.78	1.78	1.78	1.78	1.78	1.78	1.78
0.01	0.34	0.34	0.34	0.32	0.28	0.29	0.34	1.78	1.78	1.78	1.78	1.77	1.77	1.78
0.1	0.34	0.34	0.32	0.28	0.26	0.26	0.33	1.78	1.78	1.78	1.77	1.76	1.76	1.78
1	0.34	0.32	0.28	0.27	0.24	0.26	0.32	1.78	1.78	1.77	1.75	1.66	1.67	1.77
10	0.32	0.28	0.27	0.25	0.24	0.26	0.32	1.78	1.77	1.75	1.65	1.55	1.47	1.74
100	0.28	0.27	0.25	0.24	0.24	0.26	0.32	1.77	1.75	1.65	1.59	1.46	1.37	1.66
1000	0.27	0.25	0.25	0.24	0.24	0.26	0.32	1.75	1.65	1.61	1.56	1.35	1.36	1.64

Once we obtain the optimized parameters of C and gamma, they are adopted as SVR model parameters. The model performance evaluation is performed on test data using the proposed model and that of [14] as shown in Table 2. From the experimental results, it can be seen that the performance of the proposed technique in term of the RMSE of rotation and translation is decreased approximately 2.69 degree and 3.11 meters, respectively compared to that of [14].

Table 2. The performance evaluation of the rotation and translation of the proposed with [14]

Method	AVG RMSE of R (Degree)	AVG RMSE of T (Meter)
SfM [14]	2.93	4.47
Propose	0.24	1.36
Performance comparison	2.69	3.11

4. CONCLUSION

This article presents the prediction of 3D rotation and translation from 2D multi-view images. The prediction model has been trained from large amount of 2D image dataset with a variety of challenging scenes. Deep feature extraction is proposed to construct corresponding features among images. The reduced features from LSA are then learned by Support Vector Regression to predict rotation and translation of the images. The model performance comparison is conducted. The experimental results show that the RMSE of rotation and translation prediction of the propose method is 0.24 degree and 1.35 meters, respectively. It can be seen that the average errors in term of RMSE is decreased by 2.69 degree and 3.11 meters respectively compared to the reference. This can demonstrate the significant performance improvement of the proposed algorithm.

5. REFERENCES

[1] Hanbyul J., Hyun S. P., Yaser Sh., "MAP Visibility Estimation for Large-Scale Dynamic 3D Reconstruction," In Proceedings of CVPR, pp. 4321-4328, 2014.

[2] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski, "Building Rome in a Day," Proc. 12th IEEE Int'l Conf. Computer Vision, 2009.

[3] J. - M. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, and S. Lazebnik, "Building Rome on a Cloudless Day," Proc. 11th European Conf. Computer Vision, 2010.

[4] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm, "Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs," Proc. 10th European Conf. Computer Vision, pp. 427-440, 2008.

[5] N. Snavely, S. Seitz, and R. Szeliski, "Photo Tourism: Exploring Photo Collections in 3D," ACM Trans. Graphics, vol. 25, no. 3, pp. 835-846, 2006.

[6] S. Sinha, D. Steedly, and R. Szeliski, "A Multi - Stage Linear Approach to Structure from Motion," Proc. 11th European Conf. Computer Vision, 2010.

[7] David J. C., Andrew O., Noah S. and Daniel P. H., "SfM with MRFs: Discrete-Continuous Optimization for Large-Scale Structure from Motion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2841 - 2853, 2013.

[8] V.M. Govindu, "Combining Two - View Constraints for Motion Estimation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 218-225, 2001.

[9] C. Rother, "Linear Multi-View Reconstruction of Points, Lines, Planes and Cameras Using a Reference Plane," Proc. Ninth IEEE Int'l Conf. Computer Vision, pp. 1210-1217, 2003.

[10] D. Martinec and T. Pajdla, "Robust Rotation and Translation Estimation in Multiview Reconstruction," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2007.

[11] K. Sim and R. Hartley, "Recovering Camera Motion Using l1 Minimization," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1230-1237, 2006.

[12] F. Kahl and R. Hartley, "Multiple - View Geometry under the l - Infinity - Norm, " IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 9, pp. 1603-1617, Sept. 2008.*Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

[13] Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[14] Github (2016) at: https://github.com/yihui-he/3D-reconstruction