Learning Feature Subspaces for Appearance-Based Bundle Adjustment

Chia-Ming Cheng [1] and Hwann-Tzong Chen [2]

[1] MediaTek Inc., Taiwan
[2] National Tsing Hua University, Taiwan

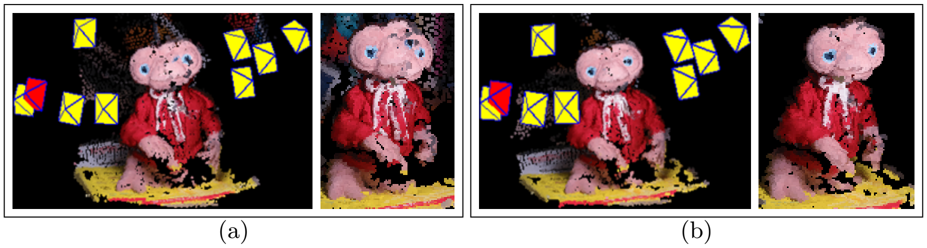
Abstract. We present an improved bundle adjustment method based on the online learned appearance subspaces of 3D points. Our method incorporates the additional information from the learned appearance models into bundle adjustment. Through the online learning of the appearance models, we are able to include more plausible observations of 2D features across diverse viewpoints. Bundle adjustment can benefit from such an increase in the number of observations. Our formulation uses the appearance information to impose additional constraints on the optimization. The detailed experiments with ground-truth data show that the proposed method is able to enhance the reliability of 2D correspondences, and more important, can improve the accuracy of camera motion estimation and the overall quality of 3D reconstruction.

1 Introduction

Recent structure from motion (SfM) systems such as [1, 3, 6, 8, 14] usually build on two key techniques: one is a distinctive-feature detector for image matching, e.g. [10, 17], and the other is an optimization process based on bundle adjustment [15]. SIFT [10] is arguably the most popular feature-extraction method for image matching. It has been successfully used in 3D modeling systems [13, 14] to extract local features for finding 2D correspondences to the same 3D point. The optimization process in an SfM system is usually based on bundle adjustment. For example, the handy SfM system *Bundler* [13, 14] uses a modified version of sparse bundle adjustment package [9] to solve the joint optimization of camera parameters and 3D point positions. More efficient algorithms on solving bundle adjustment have also been continually developed [2, 4]. The coupling of feature matching and bundle adjustment enables modern SfM systems like Bundler to model large-scale 3D structures from unordered image collections.

The sparse bundle adjustment used in Bundler requires good feature-matching results to provide reliable initial correspondences. However, local features across wide-baseline views and varied lighting conditions are not easy to be matched due to the nontrivial transformation of the feature's appearance. Havlena et al. [6] use a model-growing scheme to connect images and create new 3D points for the 3D model. More correspondences can thus be included in bundle adjustment. Our approach shares a similar notion of adding new views as [6], but we explore the use of online learning mechanisms in SfM. We seek to improve the matching quality by incorporating the online learned appearance models of 3D points into bundle adjustment. Various learning-based feature descriptors have been devised to improve image matching, e.g. [17]. Our goal is different in that we attempt to build feature representations for structure-from-motion rather than for general-purpose image matching. We incrementally update the appearance models of 3D points after each iteration of bundle adjustment, and use the appearance models to formulate a more robust bundle adjustment process.

Based on the online learning scheme for the appearance models of 3D points, we present the *appearance-based bundle adjustment* to solve the SfM problem. A feature subspace is associated with each 3D point as the appearance model, and the subspace is incrementally updated when new observations are available after each iteration of bundle adjustment. Local features in a new view are directly compared with the appearance model of each 3D point to find correspondences. Through the online learning of the appearance models, we are able to include more plausible observations of 2D features across diverse viewpoints. The experiments show that our approach is effective in improving both the visibility rates and the track lengths of correctly matched features. The appearance-based bundle adjustment is preferable to the point-based bundle adjustment in terms of the formulation of optimization problems. Relying on merely the positions of 2D points to evaluate the reprojection error might either lead to wrong estimations or make lots of points be removed as outliers. Our formulation can use the appearance information to avoid being trapped in poor local minima. Fig. 1 shows an example of using the appearance-based bundle adjustment to obtain a more consistent structure. In the experiments shown in Section 4, we use ground-truth data to show that our approach can enhance the reliability of the reconstructed 3D points, and as a result, can improve the accuracy of camera motion estimation and the overall quality of 3D reconstruction.


(a)

(b)

Fig. 1. (a) The PMVS [5] reconstruction based on the result generated by a standard SfM pipeline with sparse bundle adjustment. Although the sparse bundle adjustment yields a small reprojection error, the inconsistency in the reconstructed structure is noticeable at the middle part, corresponding to the boundaries between the two clusters of views. (b) The PMVS output of our approach. Combining the geometry and the appearance helps to resolve the problem caused by insufficient matchings between the two clusters of views.

2 Learning the Subspace Representations of Local Features

In SfM, bundle adjustment is performed according to the initial pose estimation and the correspondences found by image matching. During bundle adjustment, dubious correspondences might be excluded from the optimization as outliers. A camera view that does not contain enough inlier corresponding points might thus be removed and does not contribute to the reconstruction. When more views are added into bundle adjustment, the increasing amount of information may help to identify correct matchings. Our approach to adding new views is to take account of the new information derived from the results of previous iterations of bundle adjustment. We explore the new view to find feature points that can actually fit the scene structure. To enable such an adaptive mechanism for finding 2D correspondences, we propose to learn the subspace representations for image features. The proposed subspace representations can be plugged in the appearance-based bundle adjustment optimization, which will be described in the next section.

The subspace representations are expected to model the variations of local features exhibited in former observations. We start by using SIFT to detect keypoints and extract local features. Instead of modeling 2D features image by image, we build a feature subspace associated with each 3D point. The detected local features in a new view are compared with the existing subspaces to find correspondences. The subspace representations are equipped with an incremental update scheme, such that, after bundle adjustment, local features can be used to update the subspaces.

We choose to use the $\mathcal{L}_\infty$ subspace described in [7] as the appearance model. The $\mathcal{L}_\infty$ subspace is originally presented for visual tracking. It has been shown that the $\mathcal{L}_\infty$ subspace outperforms the $\mathcal{L}_2$ (PCA-like) subspace in tracking objects under lighting changes and geometric transformations. The computation is also easier for $\mathcal{L}_\infty$ subspace since, unlike $\mathcal{L}_2$ subspace, no eigen-decomposition is involved.

Consider a set of SIFT feature vectors $\{v_1, \ldots, v_k\}$ associated with a 3D point. Our goal is to learn a subspace L that minimizes an error function given by

$$\text{Error}^{\infty}\left(L, \{v_1, \ldots, v_k\}\right) = \max_{t \in \{1, \ldots, k\}} d(L, v_t)\,, \tag{1}$$

where the function $d(\cdot,\cdot)$ measures the distance from a vector to a subspace in a least-squares sense. A subspace spanned by the entire observations of SIFT feature vectors $\{v_1, \ldots, v_k\}$ should minimize the above error function. We can find one of the subspaces that approximate to the span of $\{v_1, \ldots, v_k\}$ by applying the Gram-Schmidt process to $\{v_1, \ldots, v_k\}$, and an orthonormal basis can be obtained to represent the subspace.

The dimension of $\mathcal{L}_\infty$ subspace spanned by $\{v_1, \ldots, v_k\}$ will grow as the number k of data increases. To enable the subspace to be updated under a bounded dimension, we use a local-means method similar to the ones proposed in [12]. We keep at most s local means $\{z_1, \ldots, z_s\}$ to form the subspace. For each

3D point we learn its $\mathcal{L}_\infty$ subspace using the local means $\{z_1, \ldots, z_s\}$ rather than $\{v_1, \ldots, v_k\}$. The Gram-Schmidt process is applied to the local means $\{z_1, \ldots, z_s\}$ and yields an orthonormal basis Q for the $\mathcal{L}_\infty$ subspace. The local means are incrementally updated through the observations of $\{v_1, \ldots, v_k\}$.

In our SfM method, the orthogonal bases $\{Q_j\}$ of the learned $\mathcal{L}_\infty$ subspaces are used as the appearance models for 3D points $\{X_j\}$. Each 3D point X_j has an associated orthonormal basis Q_j. Given a detected 2D point in a new view i for camera C_i, we may find its most possible corresponding 3D point by projecting its SIFT feature vector onto the appearance subspace of each 3D point. We search for the subspace spanned by basis Q_{j^*} that has the minimum squared Euclidean distance from the SIFT feature vector to its orthogonal projection on the subspace. That 2D point is thus denoted as a 2D correspondence u_{ij^*} of the 3D point X_{j^*} in view i.

The SIFT feature vector of the 2D point is then used to update the corresponding basis Q_{j^*}. We add the SIFT vector into the closest local mean to update the set of local means. If the maximum number s of local means is not achieved and the distance from the SIFT vector to the closest mean is larger than a threshold, we create a new mean and add it into the set of local means. The updated set of local means is then used to generate a new orthonormal basis of the subspace by applying the Gram-Schmidt process. The Gram-Schmidt process is efficient. In our case we choose $s = 10$ and find that the overhead of recomputing Gram-Schmidt is negligible.

3 Appearance-Based Bundle Adjustment

Bundle adjustment is formulated as a process of simultaneously refining 'the sparse 3D points of the scene structure' and 'the parameters of cameras capturing the images'. The underlying optimization problem often involves minimizing the reprojection error of 3D points according to their 2D correspondences across images. Assume that we have m cameras $\mathbf{C} = (C_1, \ldots, C_m)$ observing n points $\mathbf{X} = (X_1, \ldots, X_n)$ in 3D space. An observation of 2D point is denoted by u_{ij}, which is derived from the observation model $f(C_i, X_j)$ that yields the 2D image coordinates of the 3D point X_j projected into the view of camera C_i plus some unknown noise. The visibility of point X_j in the view of camera C_i is indicated by an index set $\mathcal{I}$, such that $(i, j) \in \mathcal{I}$ if and only if point X_j is observed in the ith image.

We present an appearance-based formulation of bundle adjustment in which the learned appearance subspaces of 3D points can be used to provide additional evidence for the measurement of the reprojection error. Instead of estimating the parameters $\{\mathbf{C}, \mathbf{X}\}$ through minimizing the reprojection error of 3D points, we incorporate the appearance into the optimization problem defined by

$$\{\mathbf{C}^*, \mathbf{X}^*\} = \arg\min_{\mathbf{C},\mathbf{X}} \sum_{(i,j)\in\mathcal{I}} \phi_{ij} \left\| f(C_i, X_j) - u_{ij} \right\|^2 , \tag{2}$$

where we multiply the reprojection error $\|f(C_i, X_j) - u_{ij}\|^2$ by an appearance weight ϕ_{ij}. For a camera C_i that has been considered in previous bundle adjustment iterations, the appearance weight ϕ_{ij} is defined by

$$\phi_{ij} = \exp\left\{-\frac{d(Q_j, v_{ij})^2}{2\sigma_a^2}\right\}, \tag{3}$$

where v_{ij} is the SIFT feature vector for the unknown 2D correspondence u_{ij} of X_j in view i, and $d(Q_j, v_{ij})$ is the distance from v_{ij} to its matched appearance subspace spanned by basis Q_j.

On the other hand, for a new camera view i', we select the 2D correspondence $\hat{u}_{i'j}$ whose feature $\hat{v}_{i'j}$ best fits the subspace Q_j, that is, yields the smallest value $d(Q_j, \hat{v}_{i'j})$ among the candidates within a radius r from the initial reprojection coordinates $f(\bar{C}_{i'}, \bar{X}_j)$ before the current iteration of bundle adjustment, where $\bar{C}_{i'}$ and $\bar{X}_j$ are previous estimations. The new view is then associated with an appearance weight

$$\phi_{i'j} = \exp\left\{-\frac{d(Q_j, \hat{v}_{i'j})^2}{2\sigma_a^2} - \frac{\|\hat{u}_{i'j} - f(\bar{C}_{i'}, \bar{X}_j)\|^2}{2\sigma_s^2}\right\}, \tag{4}$$

where we lessen the weight according to how far $\hat{u}_{i'j}$ diverges from the initial reprojection coordinates. We set $\sigma_s = 0.4r$ as a spatial scale factor based on the radius r. In our experiments we set $r = 5.0$, $\sigma_s = 2.0$ and $\sigma_a = 0.6$. Note that we use the factor of re-projection error in (4) because we would like to introduce a soft decision boundary for the inclusion of $\hat{u}_{i'j}$. If we use only the factor of $d(Q_j, \hat{v}_{i'j})$ in (4), we actually adopt a hard boundary to decide whether we should include $\hat{u}_{i'j}$. Such a hard decision boundary would be more sensitive to the parameter setting for the search radius r.

The optimization can be expressed in matrix form:

$$\{\mathbf{C}^*, \mathbf{X}^*\} = \arg\min_{\mathbf{C},\mathbf{X}} \left\|\boldsymbol{\Phi}\left(f(\mathbf{C}, \mathbf{X}) - \hat{\mathbf{U}}\right)\right\|^2, \tag{5}$$

where $\|\cdot\|$ is the Frobenius norm, $\boldsymbol{\Phi}$ contains the appearance weights in the corresponding matrix elements, and $\hat{\mathbf{U}}$ consists of the 2D correspondences. Let $\mathbf{J} = [\partial f/\partial \mathbf{C}\ \partial f/\partial \mathbf{X}]^\mathsf{T}$. By the first order Taylor approximation we may write the solution as

$$\begin{bmatrix} \Delta\mathbf{C} \\ \Delta\mathbf{X} \end{bmatrix} = (\mathbf{J}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\mathbf{J})^{-1}\mathbf{J}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\left(\hat{\mathbf{U}} - f(\bar{\mathbf{C}}, \bar{\mathbf{X}})\right). \tag{6}$$

3.1 Comparison with the original bundle adjustment [15]

Although we formulate the optimization in a form of weighted least squares as in [15] so that stable numerical solutions can be more easily obtained, the notion of our formulation is quite different from [15], where the weight matrix is just an inverse covariance matrix modeling the uncertainty. Our formulation includes the additional information provided by the learned appearance models, and we perform the optimization and learning in an EM-like manner that is embedded in the iterations of bundle adjustment. At each iteration of bundle adjustment we search among the candidate appearance models to associate individual 2D points in the new view with the 3D points. After an iteration of bundle adjustment, we can update the appearance models using the current results of 2D to 3D correspondences.

Our appearance-based formulation is also different from the intensity-based model which solves for the transformations between image patches, as is mentioned in [15]. For the problem of SfM, the transformations between image patches on surfaces are not fully dependent on the parameters of the camera poses and the scene structures of interest. To include extra parameters of patch transformations might burden the optimization rather than alleviate the adjustment computation. Our formulation does not include the extra parameters but make use of the appearance information to avoid infeasible solutions found by point-based bundle adjustment.

4 Experiments

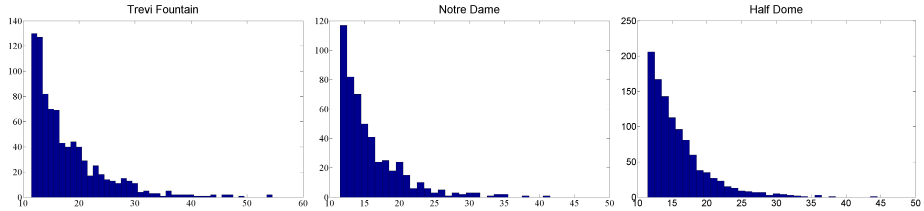
In the first part of the experiments, we evaluate the performance of learning the subspace representations for local features. We show that the proposed learning method can be applied to large datasets and can achieve very good precision-recall rates, significantly better than the baseline strategy of descriptor averaging. Our learning method performs comparably well as the direct matching strategy (nearest-neighbor criterion), in which all descriptors are kept for matching without any learning. However, our learning method is much more efficient than the direct matching, especially for large datasets.

In the second part of the experiments, we evaluate the structure-from-motion results using the appearance-based bundle adjustment. We use three datasets that provide calibrated cameras and ground-truth correspondences for evaluation. Our method shows the advantages of increasing the track length and the number of observations per view. More important, the accuracy of camera motion estimation and 3D reconstruction is also improved, in comparison with the point-based sparse bundle adjustment.

4.1 Evaluation of Learning Subspace Representations

We use the datasets provided by Winder and Brown in [17] to evaluate the effectiveness of learning the subspace representations. The image data are taken from *photo tourism* [13] reconstructions of Trevi Fountain, Notre Dame, and Half Dome. Each dataset consists of $100,000$ grayscale patches, which are obtained by projecting 3D points from *photo tourism* reconstructions back into the original images. Due to the mechanism of deciding the scales and orientations of the 2D projected points, many of the correspondences identified in the datasets may not have been matched using SIFT descriptors. The patches might also have some local occlusion due to parallax.

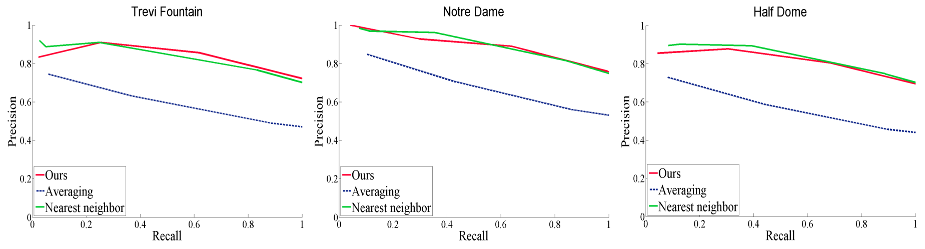
For each dataset, we select the 3D points that have at least twelve 2D correspondences (twelve corresponding patches), since we would like to see how effectively the subspace representations can perform for modeling longer tracks of matched 2D correspondences. As a result, the number of selected 3D points is 852, 515, and 1,071 for Trevi Fountain, Notre Dame, and Half Dome. Totally there are 15,267, 8,164, and 17,050 patches selected from the three datasets. The average number of patches of a selected 3D point for Trevi Fountain, Notre Dame, and Half Dome is 18, 16, and 16, respectively, and the histograms regarding the number of patches of selected 3D points are shown in Fig. 2. Some of the selected 3D points may have more than 30 corresponding patches. We separate the patches of each dataset into a training set and a test set, with a ratio of 4 : 1. The size of a patch is 64 × 64 pixels. We extract the SIFT descriptor from each patch for subspace learning.



(a) (b) (c)

Fig. 2. The histogram of the number of patches corresponding to the selected 3D points ($\geq$ 12 patches) for (a) Trevi Fountain, (b) Notre Dame, and (c) Half Dome datasets. Some of the selected 3D points may have more than 30 corresponding patches.

Precision-recall We apply the proposed learning method to each of the three training sets and build the feature subspaces for the corresponding 3D points. The maximum number s of local means is 10, as described in Section 2. For the test data, the correspondences to the 3D points are decided by finding the closest subspaces. We can verify the ground-truth correspondences to evaluate the quality of matching results. If we set a threshold for the distance between a test feature and its closest subspace, we may remove some incorrect correspondences. By modulating the threshold value, we can derive a precision-recall curve. Precision is the number of 'true positives' divided by the sum of 'true positives' and 'false positives'; recall is the number of 'true positives' divided by the sum of 'true positives' and 'false negatives'. If we set a larger threshold value, then the recall rate will be higher but the precision might decrease. The precision-recall curves for the three test sets are shown in Fig. 3. The subspace learning method is compared with two strategies: The first one is to average all the SIFT descriptors that belong to the same 3D point, and use the mean descriptor as the feature representation of the 3D point. To find the correspondence for a test descriptor, we measure the similarity between the test descriptor and each of the mean descriptors using the Euclidean distance. The second strategy is to keep all SIFT descriptors of the training data and use the nearest-neighbor criterion to find 2D correspondence for the test descriptor, where the Euclidean distance is also used as the measurement for SIFT descriptors. As shown in Fig. 3, our subspace method can achieve comparable performances as the nearest-neighbor strategy. The averaging strategy does not perform very well because the mean descriptors might not be distinctive enough for large datasets.



(a) (b) (c)

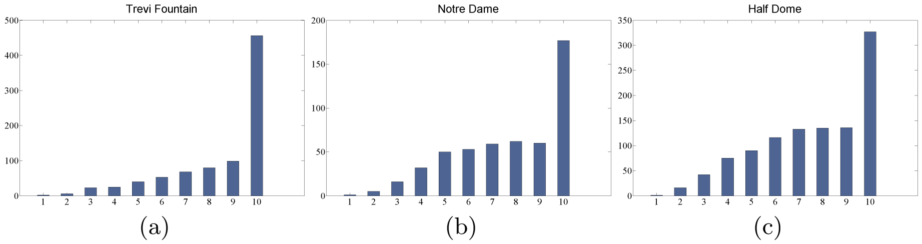
Fig. 3. The precision-recall curves for (a) Trevi Fountain, (b) Notre Dame, and (c) Half Dome datasets. Our subspace representations can achieve comparable performances as the nearest-neighbor strategy. The averaging strategy does not perform very well, probably because the mean descriptors are not distinctive enough for large datasets.

Timing Learning the subspace representations using our method is very fast. For example, the subspaces for the 13,640 training descriptors of the Half Dome data can be learned in less than 2 seconds, in MATLAB on a PC with quad-core 2.8GHz CPU and 12GB memory. The training time for the averaging strategy is close to our method. The nearest-neighbor strategy does not require training, and only some overhead processing time is involved. Regarding the matching between the test data and the training data for finding correspondences, our method and the averaging strategy are faster. The nearest-neighbor strategy, as expected, is very slow. The timing results for matching are shown in Table 1.

Table 1. The timing results of feature matching using different strategies.

	# of test patches	# of training patches	Timing for matching		
			Nearest neighbor	Averaging	Subspace
Trevi Fountain	3,053	12,214	729s	46s	51s
Notre Dame	1,632	6,532	252s	17s	18s
Half Dome	3,410	13,640	989s	65s	71s

Further discussions The evaluation shows that the learned appearance subspaces provide effective representations for finding correspondences to 3D points. By using the learned subspaces, we can have similar precision-recall rates without keeping all the descriptors of 2D features, and therefore greatly reduce the time required for matching. Since we set the maximum number s of local means to be 10, the dimension of a learned subspace is at most 10. We find that the average dimension of the learned subspaces is 9, 8, and 8 for Trevi Fountain, Notre Dame, and Half Dome. The distributions of the subspace dimensions are shown in Fig. 4. We may choose a larger value of s to allow higher dimensional subspaces to be built, particularly when the dataset is very large, but the training and matching time might also increase. The trade-off of descriptiveness and efficiency would be dependent on the data. For a dataset with a scale about 1,000 3D points and 15,000 2D features, our current setting seems suitable.



(a) (b) (c)

Fig. 4. The histogram of subspace dimensions for (a) Trevi Fountain, (b) Notre Dame, and (c) Half Dome datasets.

4.2 Evaluation of Appearance-Based Bundle Adjustment Using Ground-Truth Data

We use the datasets created by Moreels and Perona [11] to evaluate the performance of the appearance-based bundle adjustment. The images in the datasets are captured by a calibrated stereo system with a turntable. The advantage of using these datasets is that we are able to verify the correctness of correspondences based on the ground-truth geometric constraints. We choose three of the datasets, BallSander, Standing, and StorageBin, as shown in Figs. 5a– 5c. The 'ground-truth' camera poses are shown in Fig. 5d. The world center is set at $(0, 0, 0)$, and the average distance between each camera and the world center is 1.0. The proposed appearance-based bundle adjustment is compared with the sparse bundle adjustment in respect of several evaluation metrics which we will describe later in this section. For fair comparison, the numbers of initial 2D features extracted by SIFT are the same for both methods.

Evaluation metrics We focus on the comparisons between the point-based sparse bundle adjustment [9] and our online-learned appearance-based bundle

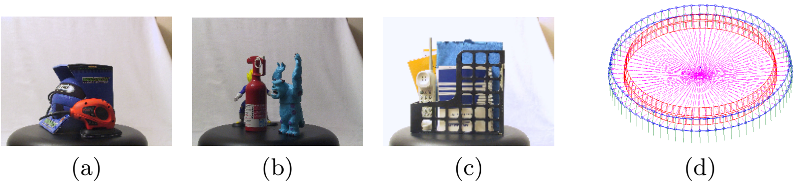


Fig. 5. Three of the datasets created by Moreels and Perona [11]: (a) the BallSander dataset, (b) the Standing dataset, and (c) the StorageBin dataset. (d) The camera poses for those datasets are derived from the calibrated stereo system with a turntable. We set the world center at $(0, 0, 0)$, and the average distance between each camera and the world center is 1.0. The evaluations of the 3D errors are based on the scale after this normalization.

adjustment. The pipeline of incremental SfM is not taken into consideration for the evaluation. Several metrics are used to evaluate the performances: *i)* the visibility rate, *ii)* the outlier rate, *iii)* the false 3D-point rate, *iv)* the camera motion estimation error (the average rotation and translation errors), and *v)* the average 3D reconstruction error.

The visibility rate is computed by (# of observations)/(# of views × # of 3D points). By 'outlier' we mean that a 2D feature within a track does not satisfy the ground-truth geometry constraint. The outlier rate is defined by (# of outliers)/(# of observations). Furthermore, we can use the ground-truth geometry constraints to verify the correctness of a reconstructed 3D point. We compute the false 3D-point rate by (# of false 3D points)/(# of all reconstructed 3D points).

Incorrect matching results would induce outliers into the minimization of the reprojection error. Outliers might bias the solution due to overemphasizing the errors. Equipping the point-based bundle adjustment with an outlier-removal mechanism might increase the robustness, but would also make bundle adjustment prone to be trapped in trivial local minima. Ideally, the reprojection error should be minimized under the assumption that all 3D points can be observed in all views. A higher visibility rate and a lower outlier rate are preferable in a sense that they imply the ideal case of the original objective of bundle adjustment.

To further evaluate the quality of camera motion estimation and 3D reconstruction, we use the ground-truth camera poses and geometry constraints derived from the datasets of Moreels and Perona. As mentioned earlier, we measure the errors of camera motion estimation and 3D reconstruction based on a normalized scale: the average distance between each camera and the world center $(0, 0, 0)$ is 1.0. The quality of camera motion estimation is evaluated by the translation error and the rotation error of camera pose. We align all of the estimated camera poses to the normalized ground-truth coordinates shown in Fig. 5d. The translation error is computed as the distance between the estimated camera center and the ground-truth camera center. The rotation error is measured by the geometric mean of the Euler angles of $\mathbf{R}_{\text{est}}\mathbf{R}_{\text{gt}}^{\mathsf{T}}$, where $\mathbf{R}_{\text{est}}$ is an estimated rotation matrix and $\mathbf{R}_{\text{gt}}$ is the ground-truth rotation matrix.

To compute the 3D reconstruction error, we exclude the false 3D points from the reconstructed 3D points. We then aligned the reconstructed 3D structure with the ground-truth structure by applying absolute pose estimation [16]. The average 3D reconstruction error is measured by the average distance from each aligned 3D point to its corresponding ground-truth 3D point.

Results We summarize all of the evaluation results in Tables 2, 3, & 4. The results show that the appearance-based bundle adjustment achieves better performance than the point-based sparse bundle adjustment on all of the evaluation metrics. The average track length and the visibility rate of 2D features both significantly increase. The improved outlier rate means that the appearance-based bundle adjustment is capable of removing more incorrect correspondences. The appearance-based bundle adjustment can also achieve a very low false 3D-point rate, which means that its reconstruction of 3D points is quite reliable. Most important, the appearance-based bundle adjustment indeed improves the accuracy and quality of camera motion estimation and 3D structure reconstruction, as explicitly shown in the evaluation results.

Table 2. Evaluations with the BallSander dataset. We compare the proposed appearance-based bundle adjustment with the sparse bundle adjustment (SBA).

	SBA	Appearance-based
# of 3D points	943	494
average track length	4.11	9.87
visibility rate (%)	10.81	25.97
outlier rate (%)	1.29	0.72
false 3D-point rate (%)	1.70	0.20
average camera rotation error	2.061	1.793
average camera translation error	0.0073	0.0070
average 3D reconstruction error	0.0074	0.0059

Further discussions After learning the subspaces and applying the learned representations to the appearance-based bundle adjustment, we can find more 2D features that can be modeled by the learned subspaces. From the results shown in Figs. 6, 7, & 8, we observe that the online learned appearance representations can help to increase the track length as well as the number of registered 2D features in each view. These newly-included 2D correspondences will contribute to solving the 3D points in later iterations. Overall, the integrated mechanism of subspace learning and appearance-based bundle adjustment provides a plausible way of computing structure and motion.

Although the reliability of the 3D points is enhanced, a limitation of our approach is that it would merge short tracks into longer ones, and as a result, the number of reconstructed 3D points might greatly decrease. The number of 3D

Table 3. Evaluations with the Standing dataset. We compare the proposed appearance-based bundle adjustment with the sparse bundle adjustment (SBA).

	SBA	Appearance-based
# of 3D points	1,226	621
average track length	4.86	12.50
visibility rate (%)	12.15	31.25
outlier rate (%)	1.16	0.98
false 3D-point rate (%)	1.47	0.00
average camera rotation error	1.603	1.402
average camera translation error	0.0065	0.0059
average 3D reconstruction error	0.0056	0.0055

Table 4. Evaluations with the StorageBin dataset. We compare the proposed appearance-based bundle adjustment with the sparse bundle adjustment (SBA).

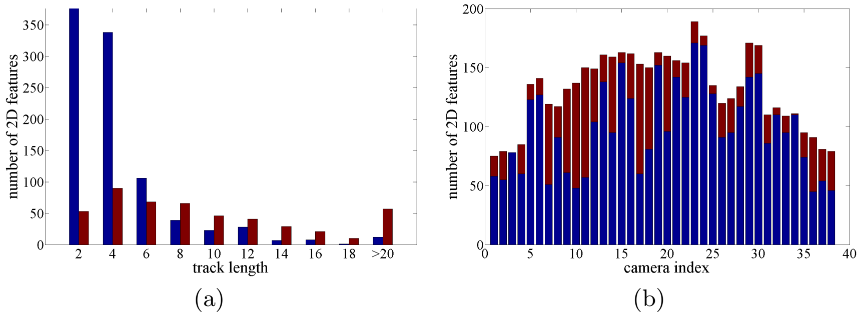
	SBA	Appearance-based
# of 3D points	1,741	697
average track length	3.82	10.85
visibility rate (%)	8.88	25.22
outlier rate (%)	5.67	1.48
false 3D-point rate (%)	6.03	0.01
average camera rotation error	1.923	1.646
average camera translation error	0.0100	0.0076
average 3D reconstruction error	0.0108	0.0074

points reconstructed by our approach is about half of the number of 3D points obtained by the point-based sparse bundle adjustment, as can be observed in Tables 2, 3, & 4. This is a trade-off between ensuring a more consistent structure and reconstructing as more 3D points as possible.

About the time complexity, the additional computational cost of the appearance based bundle adjustment is due to the computation of the appearance-weight matrix, of which the size is the number of views times the number of 3D points. We also need to compute the appearance weights and multiply the appearance-weight matrix by the Jacobian matrix, but the computation of Jacobian matrix is efficient owing to the the longer tracks and the reduced number of redundant points. In practice the computation time of solving the appearance-based bundle adjustment is close to solving the sparse bundle adjustment if the optimization involves similar numbers of views and 3D points.

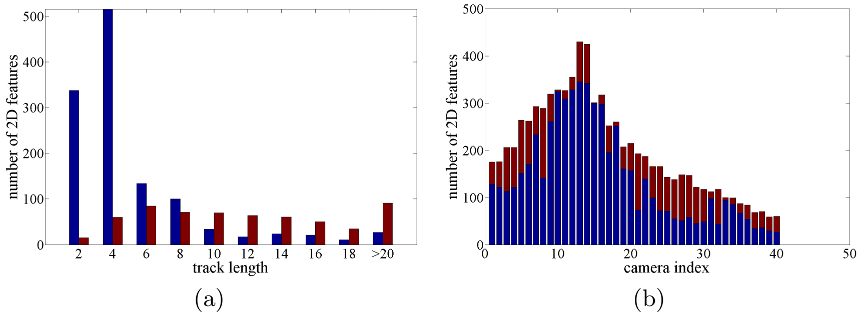
5 Conclusion

We have presented a new bundle adjustment method based on an online-learned appearance model associated with each 3D point. The proposed appearance-based bundle adjustment is able to include more 2D observations into the op-



(a) (b)

Fig. 6. The BallSander dataset. (a) The distribution of the track length. (b) The number of registered 2D points in each view. Blue bars: before subspace learning. Red bars: after subspace learning.



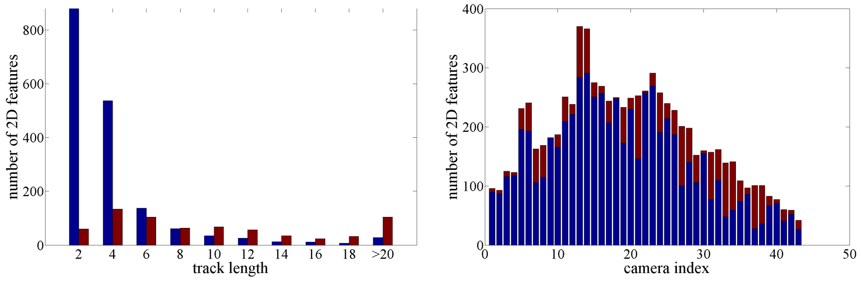
(a) (b)

Fig. 7. The Standing dataset. (a) The distribution of the track length. (b) The number of registered 2D points in each view. Blue bars: before subspace learning. Red bars: after subspace learning.

timization. As shown in our experiments, the lengths of most tracks in conventional sparse bundle adjustment are usually quite small. The appearance-based bundle adjustment is able to achieve a significant increase in the number of long tracks and the number of correctly matched features. The visibility rates of 2D correspondences and the outlier rates are greatly improved by appearance-based bundle adjustment. Through the detailed evaluations on the ground-truth datasets, we show that our method can improve the accuracy of camera motion estimation and the quality of 3D reconstruction, in comparison with the point-based sparse bundle adjustment.

References

1. S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing rome. *IEEE Computer*, 43(6):40–47, 2010.
2. S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski. Bundle adjustment in the large. In *ECCV (2)*, pp. 29–42, 2010.



(a) (b)

Fig. 8. The StorageBin dataset. (a) The distribution of the track length. (b) The number of registered 2D points in each view. Blue bars: before subspace learning. Red bars: after subspace learning.

3. S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, 2009.
4. M. Byröd and K. Åström. Conjugate gradient bundle adjustment. In *ECCV (2)*, pp. 114–127, 2010.
5. Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *CVPR*, 2007.
6. M. Havlena, A. Torii, and T. Pajdla. Efficient structure from motion by graph optimization. In *ECCV (2)*, pp. 100–113, 2010.
7. J. Ho, K.-C. Lee, M.-H. Yang, and D. J. Kriegman. Visual tracking using learned linear subspaces. In *CVPR (1)*, pp. 782–789, 2004.
8. X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV (1)*, pp. 427–440, 2008.
9. M. I. A. Lourakis and A. A. Argyros. Sba: A software package for generic sparse bundle adjustment. *ACM Trans. Math. Softw.*, 36(1), 2009.
10. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
11. P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73(3):263–284, 2007.
12. H. Park, M. Jeon, and J. B. Rosen. Lower dimensional representation of text data based on centroids and least squares. *BIT*, 43, 2003.
13. N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006.
14. N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.
15. B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Workshop on Vision Algorithms*, pp. 298–372, 1999.
16. S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4):376–380, 1991.
17. S. A. J. Winder and M. Brown. Learning local image descriptors. In *CVPR*, 2007.