Content Selection Using Frontalness Evaluation of Multiple Frames

Sungmin Eum* and David Doermann†
Institute of Advanced Computer Studies
University of Maryland, College Park, MD, USA
*smeum@umiacs.umd.edu, †doermann@umiacs.umd.edu

***Abstract*—This paper addresses the problem of selecting instances of a planar object in a video or from a set of images based on an evaluation of its "frontalness". We introduce the idea of "evaluating the frontalness" by computing how close the object's surface normal aligns with the optical axis of a camera. The unique and novel aspect of our method is that unlike previous planar object pose estimation methods, our method does not require the true frontal image as a reference. The intuition is that a true frontal image can be used to produce other non-frontal images by perspective projection, while the non-frontal images have limited ability to produce other non-frontal images. We show that this intuition of comparing 'frontal' and 'non-frontal' can be extended to comparing 'more frontal' and 'less frontal' images. Based on this observation, our method estimates the relative frontalness of an image by exploiting the objective space error. We also propose the usage of K-invariant space to evaluate the frontalness even when the camera intrinsic parameters are unknown (e.g., images/videos from the web). We show that our method outperforms the homography decomposition-based method which also does not require reference images. In addition, a qualitative evaluation is carried out to show that our method can be applied in selecting the most frontal characters from a set of images captured in various viewpoints.**

I. Introduction

Consider a crumpled receipt or a folded document which one would like to capture and save using a mobile device. It is often impossible to find the precise location and pose to capture the entire source with perfect quality. This is because some portions of the documents would not be directly facing the image plane while other portions may be out of focus, or experience inconsistent lighting. (Figure 1)

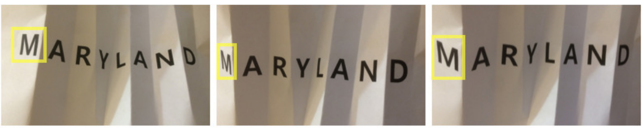


Fig. 1. Set of frames showing a folded document in different poses representing the case of crumpled document.

One possible solution is to capture and model the 3-D structure to "flatten" the document using dewarping algorithms to reconstruct the original planar surface. However, these methods either require external sensors such as structured light [1], [2] or light grid projectors [3] which makes them inconvenient or even impossible for typical users or cannot handle complex distortion. It also may not be desirable in outdoor environments. Instead of seeking to recover the whole document at once, an alternative approach may be to attempt to recover locally "optimal" portions of the image, from a collection of possible poses.

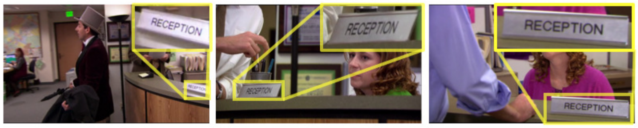


Fig. 2. Set of frames extracted from a video which shows different poses of an object of interest.

In another task, consider having an interest in a planar object, such as a book cover or business logo, in a movie or a long video. If one wants to find a frame which best depicts that object with respect to its pose, one may have to manually browse through the entire video. An example set of frames for such a case is shown in Figure 2.

As suggested in the case of crumpled documents, one may assert that this can be handled by applying a pose estimation solution for planar objects which seeks to estimate the relative pose of an object with respect to a reference (frontal) image. This has been addressed in a number of articles including [4]–[6] which were shown to have reliable and stable performance. However, these methods all share the same limitation in that they assume the reference model (frontal image) is provided a priori. This makes them unsuitable for handling this case because the assumption of having an ideal frontal image beforehand directly conflicts with the very purpose of our goal. Homography decomposition [7]–[9], on the other hand, does not require this assumption and can estimate the surface normal of a planar surface with respect to the optical axis of a camera when given a pair of images. However, it suffers from highly unstable performance and also provides results which are ambiguous.

We claim that these problems can be handled in a common framework which relies on analyzing the poses of the local planar targets and selecting the best one when given images or a video which span different viewpoints. Without loss of generality, the best shot of a planar target can be considered as the one capturing the pose closest to the frontal pose of the target.

In this paper we develop the concept of *evaluating the frontalness* of the image of a planar source by measuring how well the surface normal of a planar object aligns with the optical axis of a camera. We show that measuring the relative frontalness can be analyzed by noting that if an image is assumed to be a true frontal image (as a reference), but is not, it shows limited ability to represent other non-frontal images. In other words, a less frontal image has less representability for different poses of an object than a more frontal image. Based on this observation, we estimate the relative frontalness by comparing the objective space errors for a given image pair, first setting one of the two images as the true frontal image (reference image), then setting the other. Objective space error values are acquired by applying a state-of-the-art pose estimation algorithm for planar objects [4].

II. Our Method

A. Overview

We assume that we have a short video or camera burst of a planar source, captured from different orientations, sufficient to adequately capture at least one instance that would be considered acceptably "frontal". Given a pair of candidates, our goal is to evaluate the relative frontalness of the images and select the one which is more frontal. Through multiple pairwise comparisons, we can ultimately find the best or most frontal candidate. Since our method does not use any temporal information, it can be applied to any unordered set of images in an equivalent manner.

In order to evaluate the relative frontalness of a target, we use a pose estimation error-based method. Typically, pose estimation is used to estimate the pose of an object with respect to a set of model points which are assumed to be known beforehand. However, in our case, the pose estimation algorithm is employed to measure the pose estimation error, or objective space error for an image with respect to another image. Thus, to compare the pose estimation errors for each image in a pair, the error is computed twice, once with the first image as the reference model and the second time with the other image as the reference model.

The intuition behind this process is that, when the true (or more) frontal image is used as a reference image, the pose estimation error is smaller than the case where non (or less) frontal image is used as the reference. This occurs because a true-frontal image can be used to reproduce non-frontal images by perspective projection, whereas the non-frontal image has a limited ability to reproduce other non-frontal images. Detailed explanation on our method is included in the following subsection.

B. Frontalness evaluation with known intrinsic camera parameters (K)

Let us first summarize the typical approach for a pose estimation procedure. Consider n coplanar model points $\mathbf{p}_i$ = $\begin{bmatrix} p_{ix} \; p_{iy} \; 0 \end{bmatrix}^T$ in reference coordinate system. These points can be transformed into the camera coordinates $\mathbf{v}_i$ by:

$$\mathbf{v}_i \propto R\mathbf{p}_i + \mathbf{t}, \tag{1}$$

where $\propto$ indicates that the left hand side is directly proportional to the right hand side, due to the fact that $\mathbf{v}_i$ can only be computed up to a scale. Note that R and $\mathbf{t}$ indicate the 3 dimensional rotation and translation vectors, respectively, which are also known together as extrinsic camera parameters. Under the assumption that the image coordinate system aligns with the reference coordinate system, the task of estimating the pose of a camera with respect to the reference coordinate system, is to estimate R and $\mathbf{t}$. So in principle, a pose estimation algorithm seeks to find the values for R and $\mathbf{t}$ that minimizes an error function. We use the object-space error, as used by [4], [6], [10], which can be written as:

$$E_{os}(\hat{R}, \hat{\mathbf{t}}) = \sum_{i=1}^{n} \parallel (I - \hat{V}_i)(\hat{R}\mathbf{p}_i + \hat{\mathbf{t}}) \parallel^2 \text{ with } \hat{V}_i = \frac{\hat{v}_i \hat{v}_i{}^t}{\hat{v}_i{}^t \hat{v}_i}. \tag{2}$$

For evaluating frontalness, we exploit the objective error itself which is being minimized in the pose estimation process instead of utilizing $\hat{R}$ or $\hat{\mathbf{t}}$. When given a pair of images, we first acquire a set of corresponding features from both images (in our case, SIFT [11] and RANSAC [12]). These feature coordinates are then normalized (i.e., transformed to camera coordinates) using the camera intrinsic parameters (represented by the matrix K) which are assumed to be known.

Using the transformed feature coordinates, we perform the pose estimation (Eq. 2) twice. In each case, one of the two images is chosen as the reference image. Lastly, we compare the two error values to decide which one better fits as the reference image or "which one is more frontal". Note that the smaller error value indicates that the reference image has been chosen well and this image serves better as a representative for the other image.

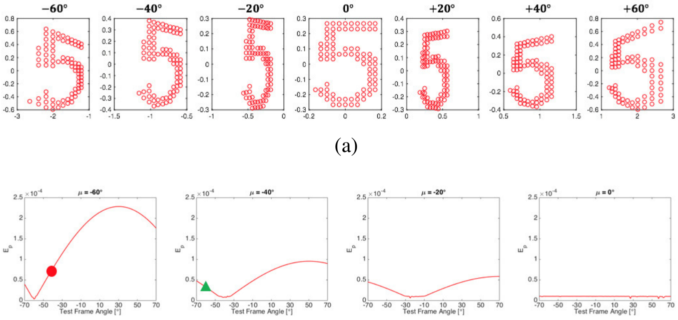
The overall process of frontalness evaluation given a set of corresponding feature coordinates extracted from a pair of images (image i and image j) is computed as shown below:

$$f^* = \begin{cases} i, & E_p(j|\mu = i)/E_p(i|\mu = j) \leq 1 \\ j, & \text{otherwise} \end{cases} \tag{3}$$

where f^* and μ indicate *image to be chosen (more frontal of the two)* and the *model frame*, respectively. Also note that $E_p(j|\mu = i)$ indicates the pose estimation error of image j when image i is set as the reference image.

To verify that our method of comparing E_p is a reasonable approach for frontalness evaluation, we have run a simulation using a synthetic dataset generated by a perspective camera model with known K. The images of a number "5" in various poses were captured by rotating the camera between -70° to 70° with respect to the y axis (Figure 3a). Each graph in Figure 3b is acquired by plotting the objective space error (E_p) for all the images in the dataset with respect to a reference model (μ). Observe that the E_p values generate a smoothly changing plot which is minimum when the reference model (μ) is used as the test model.

Now consider one example of comparing the E_p values which correspond to the two locations with the circle



(a)

(b)

Fig. 3. (a) Synthetic images of number “5” with various rotations captured by perspective camera model. (b) Objective space error plot for different reference images. X-axis: Test image angle ($-70°$ to $+70°$), Y-axis: E_p.

and the triangle marks in Figure 3b. It clearly shows that $E_p(-60°|\mu = -40°)$ is smaller than $E_p(-40°|\mu = -60°)$, and this verifies that the image with $-40°$ angle is indeed “closer to the true frontal” than the image with $-60°$. By comparing any two E_p values in two different plots, one can verify that the method can be applied in general.

C. K-Invariant projective space

In applying the method described in the previous subsection, we assume that the camera intrinsic parameters (K) are known. This means it remains a challenge for uncalibrated cameras where K is inconsistent or unknown [13]. There may be a case where K is constantly changing due to zooming even if a same camera is used. When the goal is to evaluate a set of randomly collected images from a web search, K is also unknown and most likely different for each image. In such cases, we need to transform the points from two images onto a space to make them invariant to the camera intrinsic parameters. This can be done by using a projective transformation as used in [14], [15].

Consider three non-collinear points in one image $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3)$ and their corresponding points in a second image $(\mathbf{p}_1^*, \mathbf{p}_2^*, \mathbf{p}_3^*)$, both in image coordinates. The image coordinates of these points are acquired by equations:

$$\mathbf{P} = \mathbf{K}\mathbf{V} \quad \text{and} \tag{4}$$

$$\mathbf{P}^* = \mathbf{K}^*\mathbf{V}^*. \tag{5}$$

where $\mathbf{P} = [\mathbf{p_1}\ \mathbf{p_2}\ \mathbf{p_3}]$, $\mathbf{V} = [\mathbf{v_1}\ \mathbf{v_2}\ \mathbf{v_3}]$, $\mathbf{P}^* = [\mathbf{p_1^*}\ \mathbf{p_2^*}\ \mathbf{p_3^*}]$, and $\mathbf{V}^* = [\mathbf{v_1^*}\ \mathbf{v_2^*}\ \mathbf{v_3^*}]$. Here, v_i is a point represented in camera coordinates as in Eq. 1. Since we assumed that these three points are not collinear, matrices $\mathbf{P}$ and $\mathbf{P}^*$ are non-singular which can define two different projective spaces, for example, γ and γ^*. Thus, we can transfer the points in the images onto the projective spaces as $\mathbf{w} = \mathbf{P}^{-1}\mathbf{p}$ and $\mathbf{w}^* = \mathbf{P}^{*-1}\mathbf{p}^*$, respectively. Considering these equations with Eq. 4 and Eq. 5, we can observe that $\mathbf{w}$ and $\mathbf{w}^*$ are invariant to K as shown below:

$$\mathbf{w} = \mathbf{P}^{-1}\mathbf{p} = \mathbf{V}^{-1}\mathbf{K}^{-1}\mathbf{K}\mathbf{v} = \mathbf{V}^{-1}\mathbf{v} \qquad \text{and} \tag{6}$$

$$\mathbf{w}^* = \mathbf{P}^{*-1}\mathbf{p}^* = \mathbf{V}^{*-1}\mathbf{K}^{*-1}\mathbf{K}^*\mathbf{v}^* = \mathbf{V}^{*-1}\mathbf{v}^*. \tag{7}$$

For generating the matrices $\mathbf{P}$ and $\mathbf{P}^*$, three non-collinear points from each images need to be chosen. These points are automatically chosen so as to maximize the spacing as recommended in [16]. We will show in the following section, that this approach indeed increases the accuracy of frontalness evaluation on images with unknown K.

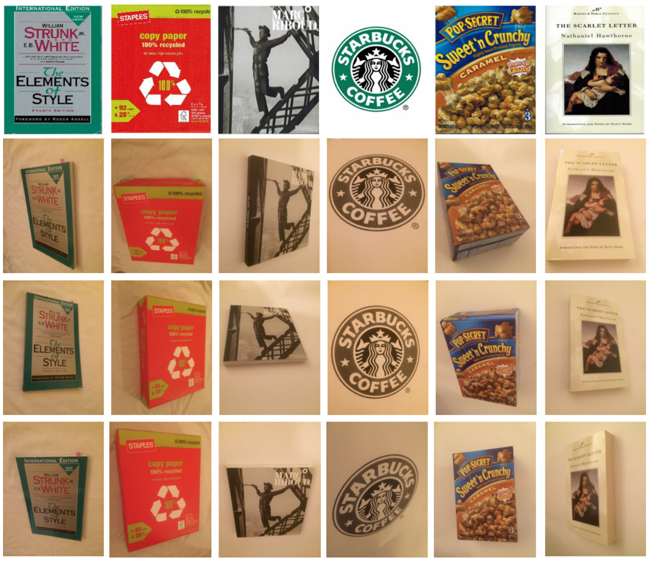


Fig. 4. Sample images from the dataset including scanned frontal images (top row) and corresponding non-frontal images (bottom two rows).

III. Experimental Evaluation

The experimental evaluation was carried out by targeting two real data scenarios based on the availability of camera intrinsic parameters (K). First, we evaluated our method assuming that the camera is calibrated (i.e., K is known). The camera intrinsic parameters were obtained beforehand using the the calibration method introduced in [17]. We compare the performance of our method with the homography decomposition-based method [8].

Second, we performed an evaluation on images under the assumption that K is unknown. In this scenario, we compare the performance of two different methods: 1) our method with a known or fixed K and 2) our method which uses a K-invariant space.

For both experiments, frontalness evaluation was performed on each possible pair of images deciding which of the two images is more frontal. The overall accuracy is computed as the percentage of the correct pairwise decisions over all possible pairs in the given dataset.

Lastly, we include two samples results which qualitatively verify that our method performs well in selecting the most frontal image from a set of images.

A. Experiment 1: Calibrated Camera, Known K

We have constructed a new dataset as there are no public dataset available targeting the evaluation of frontalness. The

TABLE I
FRONTALNESS EVALUATION ACCURACY

Homography-decomp	**68.35%**
Ours	**86.04%**

dataset consists of 1200 images which were captured using the camera on iPhone5s with the resolution of 3264 x 2448 (w x h). This includes 30 different planar objects (books, documents, boxes), with each object being captured in 40 different camera angles and distances. The images were captured so that the angle between the optical axis of the camera and the surface normal of the plane ranges between 0° to 50°, approximately, distributed in various random directions.

To evaluate the performance of each decision, the angle between a test image and the optical axis of the camera should be provided as groundtruth. Since it is difficult to directly measure and work with the optical axis of a camera, we computed the angle between each image in the dataset (non-frontal) with respect to its corresponding true frontal shot. The pose estimation method in [4], which is known to be one of the state-of-the-art in robustness and accuracy, was used to compute the angles and be saved as the groundtruth. The true frontal image of each planar object was acquired by scanning the frontal surface of the object using a flatbed scanner. Figure 4 shows some of the selected images of frontal (scanned) and non-frontal shots from the dataset.

Each decision is made in a pairwise manner. Thus, testing was performed on every possible image pair in the dataset, which sums up to 23.4k pairs. The frontalness evaluation accuracy of our method and the homography decomposition-based method (baseline) for the overall dataset is shown in Table 1. Our method clearly outperforms the baseline method.

To better analyze the capability of our method with different difficulty levels, we have defined the measure of difficulty, ν, which can be computed for each image pair. We use the cosine similarity as the measure which is shown below:

$$\nu = \frac{\mathbf{A} \cdot \mathbf{B}}{\| \mathbf{A} \| \| \mathbf{B} \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (8)$$

where $\mathbf{A}$ and $\mathbf{B}$ are the two surface normal vectors of the two given images which are provided by the groundtruth.

The plot in Figure 5 shows the performance of our method with respect to the 7 different difficulty levels along with two sample pairs with minimum and maximum difficulty. The accuracy goes up to 97% for the easiest pairs while it performs 71% for the most difficult ones. Note that, however, as the difficulty level goes up, the appearance of the image pairs begin to resemble with each other, thus having low risk even if the decision is incorrect.

The frontalness evaluation of each pair of images requires less than a second (0.54 seconds in average for the given dataset) with MATLAB implementation on Intel Core i5 PC (2.6GHz CPU, 4GB RAM) excluding the feature extraction time.

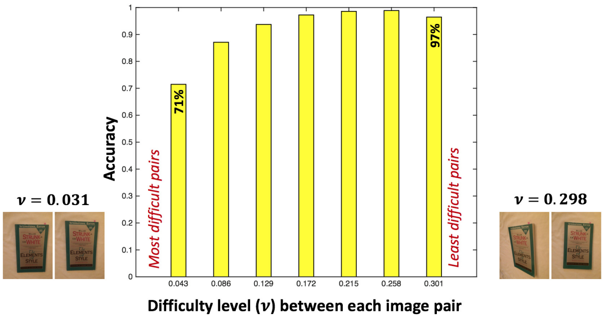


Fig. 5. Frontalness evaluation accuracy with respect to difficulty levels. Testing dataset size = 23.4k pairs.

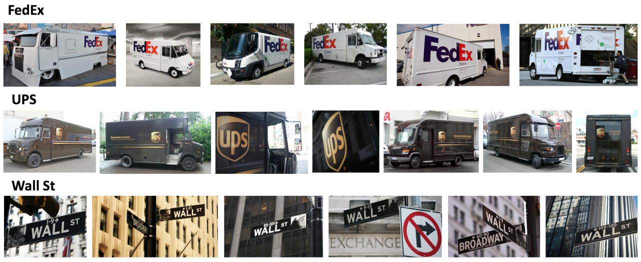


Fig. 6. Sample images from the dataset for cases with unknown K.

B. Experiment 2: Randomly Collected Images, Unknown K

Our method explained in Section III-A which assumes that K is given, is not suitable for handling images captured with cameras with unknown intrinsic parameters. To validate the effectiveness of using K-invariant space with a pose estimation-based method, we have collected images of 3 different planar objects (a FedEx logo, a UPS logo, and a Wall Street sign), each at various rotations. For each planar object, 20 non-planar images are included along with the one true frontal image for each object. Note that there are 190 possible pairs for each object for evaluation. The groundtruth for each pair was generated in an equivalent manner as described in Experiment 1. The images were downloaded from the internet and sample images are shown in Figure 6.

We compare the performance of two different methods: our method which assumes known/fixed K, our method which uses K-invariant space (KIS). When applying the method which assumes known/fixed K, we have used the K of our pre-calibrated camera to transform the points to camera coordinates in order to make a fair comparison. The performance comparison is shown in Figure 7 and it depicts the effectiveness of applying the K-invariant space. However, the overall performance does not quite reach the accuracy shown in the known/fixed K cases.

C. Qualitative Results

In addition, we show that our method can be used in selecting the best characters from a set of 40 images with

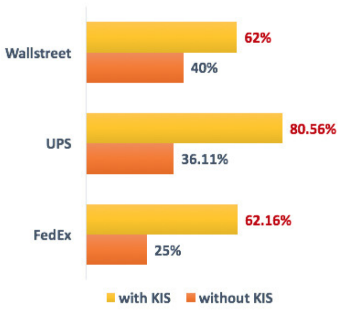
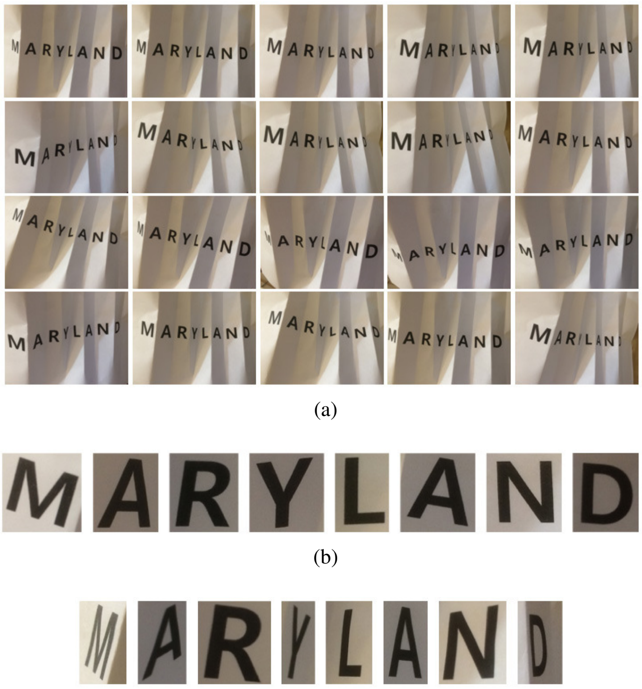


Fig. 7. Frontalness evaluation accuracy on dataset with unknown K. Using K-invariant space (KIS) shows its effectiveness.



(a)

(b)

(c)

Fig. 8. (a) Sample Images of a folded document captured in different viewpoints. (b) Characters with highest frontalness. (c) Characters with lowest frontalness.

various viewpoints. The sample images are shown in Figure 8a. Each character in different images are assumed to be residing on piecewise planar surfaces. Bounding boxes for the characters were manually assigned so that the evaluations are carried out within the same set of characters. Compare the best set of characters with the worst set of characters in Figure 8b and Figure 8c, respectively.

In addition, our method of performing the pairwise comparison of the E_p values can easily be used on a set of images to order them in terms of their frontalness. We have selected one of the objects from the dataset introduced in III-A and applied our method. The resulting ordered images are shown in Figure 9.

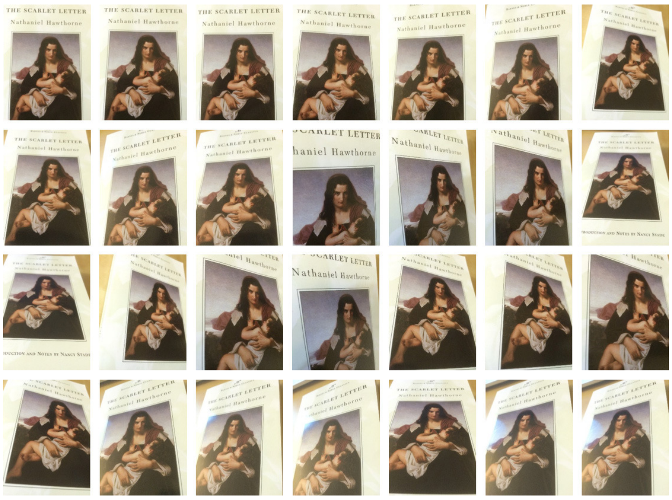


Fig. 9. Ordered images with respect to their frontalness, from high to low.

IV. Conclusion

In this paper, we have devised a novel method for evaluating the frontalness of planar objects. Our method takes a pair of images at a time to measure the relative frontalness between the two by exploiting the objective space error. Each run only requires a fraction of a second which makes it possible to be applied in real applications. Unlike the previous pose estimation methods that strictly require a true frontal image of the target object as a reference model, our method does not require any reference model. Moreover, by introducing K-invariant space, we show that the proposed method can be applied even when the camera intrinsic parameters are unknown. The approach can be applied to optimizing the reconstruction of severely crumpled documents from a short video scan, especially the cases where a character or any continuous content reside on two or more piecewise planar surfaces. In addition, bringing more efficiency in terms of computation time would trigger real time applications or auto capturing of planar objects using mobile devices.

Acknowledgment

The partial support of this research by the US Government through NSF Awards IIS-0812111 and IIS-1262122, the Xerox Foundation is gratefully acknowledged.

References

[1] M. S. Brown and W. B. Seales, “Document restoration using 3d shape: a general deskewing algorithm for arbitrarily warped documents,” in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, 2001, pp. 367–374.

[2] M. Pilu, “Undoing page curl distortion using applicable surfaces,” in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1, 2001, pp. 237–240 vol.1.

[3] A. Doncescu, A. Bouju, and V. Quillet, “Former books digital processing: image warping,” in *Document Image Analysis, 1997. (DIA ’97) Proceedings., Workshop on*, Jun 1997, pp. 5–9.

[4] G. Schweighofer and A. Pinz, “Robust pose estimation from a planar target,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2024–2030, Dec 2006.

[5] Z. Jia, A. Gallagher, and T. Chen, “Cameras and gravity: Estimating planar object orientation,” in *2013 IEEE International Conference on Image Processing*, Sept 2013, pp. 3642–3646.

[6] C. P. Lu, G. D. Hager, and E. Mjolsness, “Fast and globally convergent pose estimation from video images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 610–622, Jun 2000.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[8] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. SpringerVerlag, 2003.

[9] O. Faugeras and F. Lustman, “Motion and Structure From Motion in a Piecewise Planar Environment,” *Intern. J. of Pattern Recogn. and Artific. Intelige.*, no. 3, pp. 485–508, 1988.

[10] P. Wunsch and G. Hirzinger, “Registration of cad-models to images by iterative inverse perspective matching,” in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, vol. 1, Aug 1996, pp. 78–83 vol.1.

[11] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Computer Vision and Image Understanding*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[12] M. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun 1981.

[13] B. P. Wrobel, *Calibration and Orientation of Cameras in Computer Vision*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, ch. Minimum Solutions for Orientation, pp. 7–62.

[14] E. Malis, “Visual servoing invariant to changes in camera intrinsic parameters,” in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1, 2001, pp. 704–709 vol.1.

[15] E.Malis, “Visual servoing invariant to changes in camera-intrinsic parameters,” *IEEE Transactions on Robotics and Automation*, vol. 20, no. 1, pp. 72–81, Feb 2004.

[16] E. Malis and F. Chaumette, “2 1/2 d visual servoing with respect to unknown objects through a new estimation scheme of camera displacement,” *International Journal of Computer Vision*, vol. 37, no. 1, pp. 79–97, 2000.

[17] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.