ACCURATE ALIGNMENT OF PRESENTATION SLIDES WITH EDUCATIONAL VIDEO

Quanfu Fan (1,3), Kobus Barnard (1), Arnon Amir (2), Alon Efrat (1)

(1) Department of Computer Science, University of Arizona, Tucson AZ85721
(2) IBM Almaden Research Center, 650 Harry Rd., San Jose, CA95120
(3) IBM T. J. Watson Research Center, 19 Skyline Dr., Hawthorne, NY 10532

ABSTRACT

Spatio-temporal alignment of electronic slides with corresponding presentation video opens up a number of possibilities for making the instructional content more accessible and understandable, such as video quality improvement, better content analysis and novel compression approaches for low bandwidth access. However, these applications need finding accurate transformations between slides and video frames, which is quite challenging in capture settings using pan-tilt-zoom (PTZ) cameras. In this paper we present a nonlinear optimization approach for accurate registration of slide images to video frames. Instead of estimating the projective transformation (i.e., homography) between a single pair of slide and frame images, we solve a set of homographies jointly in a frame sequence that is associated with a given slide. Quantitative evaluation confirms that this substantively improves alignment accuracy.

Index Terms— distance learning, bundle adjustment

1. INTRODUCTION

Distance learning has become an important alternative way of learning for students in universities, and is widely used in large organizations for training purposes. Typically, instructional content is created by video capture of lectures and presentations, and is distributed by video streaming over the Internet. The video quality, however, varies greatly. A high quality production is still costly and labor intensive On the other hand, low-cost video production often suffers in quality with inadequate illumination compensation, significant color distortion, and unsharp images being common.

Spatio-temporal alignment of electronic slides with the corresponding presentation video opens up a number possibilities for improving the accessibility and understandability of the instructional content. This is demonstrated through the SLIC (**S**emantic **L**inked **I**nstructional **C**ontent) system [1] under development at the University of Arizona and IBM Almaden. The system makes extensive use of accurate alignment of slides to video, and offers a number of applications for improving the understandability of educational videos including (i) back projection of the high-resolution slide images into the video for improving the readability of the captured slides, (ii) a high-quality cross-media magnifying glass, (iii) slide color correction, (iv) slide text extraction, and (v) laser pointer gesture extraction. These applications, described in more detail elsewhere [2], are designed to improve and enrich distance learning, especially for visually impaired students.

Systems for analysis, indexing, search and browse of videos for education have been substantively studied, and several of them also integrate presentation slides (e.g., [3, 4, 5, 6, 7]. While each of these systems synchronizes slides with video (either manually or automatically) none of them take advantage of slides for improving video quality. Doing so requires accurate spatial registration of the slides with the video, and thus cannot be addressed simply by slide synchronization.

In this paper we describe an approach for accurate spatio-temporal alignment of slides with the corresponding presentation video. The initial alignment is based on our previous work [8], which automatically resolves the correspondences between slides and video frames, yielding their mappings (i.e., *homographies*) as a by-product. The focus of this work is compute more accurate and consistent homographies by integrating information from multiple images. In particular, we apply a non-linear optimization technique similar to *bundle adjustment* [9] to solve the homographies jointly in across frame sequences associated with each slide. Different from *bundle adjustment* which refines point correspondences in 2D images to produce jointly optimal 3D structure and camera parameters, our method solves the 2D slide structure and homographies simultaneously from a slide image and a set of sampled frames of the slide. The known slide image, analogous to the usually unknown 3D scene structure in *bundle adjustment*, allows us to obtain true homographies that are optimized for the frame sequence. For convenience, we refer our method as a *bundle adjustment* technique. The method links slide matching and frame matching through the *homography consistency* rule discussed in Section 2, and combines both of them in the optimization process. The frame matching ensures the consistency of homographies in the frame sequence, while the slide matching pushes the estimated homographies towards the true ones.

Quanfu Fan contributed to this work while he was at the University of Arizona.

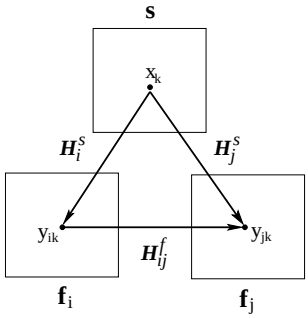


Figure 1. A slide **s** and two views (frames) f_i and f_j from it. The slide homographies of f_i and f_j are denoted as $\mathcal{H}_i^s$ and $\mathcal{H}_j^s$, respectively. The frame homography between f_i and f_j is $\mathcal{H}_{ij}^f$. A simple algebraic relationship exists among $\mathcal{H}_i^s$, $\mathcal{H}_j^s$ and $\mathcal{H}_{ij}^f$, i.e., $\mathcal{H}_j^s = \mathcal{H}_{ij}^f \mathcal{H}_i^s$.

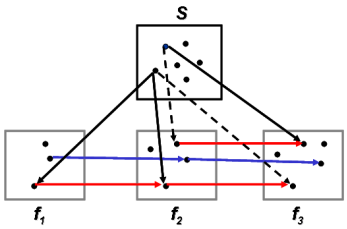


Fig. 2. Linking keypoint correspondences across a frame sequence.

2. REGISTRATION OF SLIDES WITH VIDEO

We have previously developed a method in [8] for robust matching of slides to video frames that combines image and temporal information. To achieve this, we extract scale invariant feature transformation (SIFT) keypoints [10] from both slide images and video frames, and match them subject to consistent projective transformation (homography) using random sample consensus (RANSAC) [11]. Further robustness is achieved by using a hidden Markov model (HMM) that integrates visual and temporal information.

While this approach achieves high accuracy in slide *recognition*, the mappings between slides and frames (i.e., homographies), available as a byproduct of the slide-to-video matching, are not sufficient for the applications mentioned in Section 1. Hence we have developed a bundle adjustment approach for computing more accurate and consistent homographies by integrating information from multiple images. Our method uses frame similarities to inform the slide homographies and push them towards better estimation. We begin by describing the geometric relationship between frames (*homography consistency*) and how we pair up frames for use in the homography refinement process.

Homography consistency. Views (frames) of the same slide are related by a simple geometric relationship. Let f_i and f_j be two frames that correspond to the same slide **s** (Fig. 1). We denote the slide homographies of f_i and f_j to **s** as $\mathcal{H}_i^s$ and $\mathcal{H}_j^s$, respectively and the frame homography between f_i and f_j as $\mathcal{H}_{ij}^f$. A slide keypoint x_k is mapped to a keypoint y_{ik} on f_i and a keypoint y_{jk} on f_j. Accordingly, y_{ik} should correspond to y_{jk} in the matching of f_i and f_j. So we have the following,

$$y_{ik} = \mathcal{H}_i^s x_k \ , \ y_{jk} = \mathcal{H}_j^s x_k \ , \ \textit{and} \ y_{jk} = \mathcal{H}_{ij}^f y_{ik} \qquad (1)$$

Merging the equations above yields the algebraic relationship between $\mathcal{H}_i^s$, $\mathcal{H}_j^s$ and $\mathcal{H}_{ij}^f$ as $\mathcal{H}_{ij}^f = \mathcal{H}_j^s (\mathcal{H}_i^s)^{-1}$. This relations defines the consistency constraints over slide homographies via frame homographies, which we call *homography consistency* and shows an indirect way to compute the slide homography for a frame.

Frame-frame matching. A key initial step is to establish homographies between all frames associated with a given slide. These frames often come from multiple video shots (i.e., an unbroken sequence of frames from one camera). By definition, frames within the same video shot (i.e., an unbroken sequence of frames from one camera) are visually similar, and matching two frames is typically accurate and efficient. Here we have found that considering the two neighboring images of the frame is sufficient. To match across shots, we find the best frame matches between two shots. All other homographies for the frame sequence can be determined as needed using the *homography consistency* rule.

Linking correspondences across frames. We make use of two types of keypoint correspondences in bundle adjustment, *frame-slide* correspondences, found through matching frames to slides (frame-slide matching), and *frame-frame* ones, from matching frames. Further, we can link them together to obtain a set of new correspondences.

We first chain keypoint correspondences across consecutive frames in the sequence to form multiple *matching chains* where each pair of adjacent keypoints on a chain are matched up, as illustrated by the blue and red solid lines in Fig. 2. A chain can start from any frame and end at any frame. If any keypoint on a *matching chain* corresponds to a slide keypoint, then all the keypoints on the chain are linked to it by creating *slide-frame* correspondences. In doing so, we now match some frame keypoints that initially were not matched with a slide. (shown as dashed black lines in Fig. 2). In case a chain is independent of the slide, a *frame-frame* correspondence is generated from each keypoint to the first keypoint of the chain, which is called a *reference frame keypoint*. Keeping *frame-frame* correspondences helps in stabilizing frame matching across the sequence. From now on, unless specified, any keypoint correspondences refer to the newly constructed correspondences as described here.

Bundle adjustment of slide homographies. At this point we have a large number of correspondences that need to be explained by an inter-related set of homographies. To get a robust estimate of them, we jointly optimize the homographies using a bundle adjustment approach. Bundle adjustment is an iterative algorithm where a non-linear model is fitted to the observed data. Here the model refers to a set of slide-to-frame homographies of interest while the observations are the coordinates of frame keypoints corresponding to slide keypoints.

Bundle adjustment seeks to minimize the bulk errors of the image locations of observed and predicted points, which

in our case are expressed as the sum of squares of the re-projection errors of keypoint correspondences. One can visualize this as perturbing the observed points iteratively until they conform to a set of consistent transformations in the frame sequence while minimizing the total re-projection errors of all keypoint correspondences.

Let $(\mathbf{x_k}, \mathbf{y_{ki}})$ be a pair of slide-frame keypoint correspondences where $\mathbf{x_k}$ is a slide keypoint and $\mathbf{y_{ki}}$ is the corresponding keypoint from frame f_i. Then the re-projection error between them is defined by

$$\mathbf{e_k^s} = (\mathbf{x_k} - \hat{\mathbf{x}}_\mathbf{k})^\mathbf{2} + (\mathbf{y_{ki}} - \hat{\mathbf{y}}_\mathbf{ki})^\mathbf{2} = (\mathbf{x_k} - \hat{\mathbf{x}}_\mathbf{k})^\mathbf{2} + (\mathbf{y_{ki}} - \hat{\mathbf{H}}_\mathbf{i}^\mathbf{s}\hat{\mathbf{x}}_\mathbf{k})^\mathbf{2} \tag{2}$$

where $(\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_{ki})$ are the predicted locations of $(\mathbf{x_k}, \mathbf{y_{ki}})$ and $\hat{\mathbf{H}}_i^s$ is the predicted slide homography of f_i that is of our interest.

Similarly, the re-projection error of a pair of frame-frame keypoint correspondences $(\mathbf{y_{ki}}, \mathbf{y_{kj}})$ from f_i and f_j can be written as

$$\begin{aligned} \mathbf{e_k^f} &= (\mathbf{y_{ki}} - \hat{\mathbf{y}}_\mathbf{ki})^\mathbf{2} + (\mathbf{y_{kj}} - \hat{\mathbf{y}}_\mathbf{kj})^\mathbf{2} \\ &= (\mathbf{y_{ki}} - \hat{\mathbf{y}}_\mathbf{ki})^\mathbf{2} + (\mathbf{y_{kj}} - \hat{\mathbf{H}}_\mathbf{ij}^\mathbf{f}\hat{\mathbf{y}}_\mathbf{ki})^\mathbf{2} \\ &= (\mathbf{y_{ki}} - \hat{\mathbf{y}}_\mathbf{ki})^\mathbf{2} + (\mathbf{y_{kj}} - \hat{\mathbf{H}}_\mathbf{j}^\mathbf{s}\hat{\mathbf{H}}_\mathbf{i}^{\mathbf{s}\,-\mathbf{1}}\hat{\mathbf{y}}_\mathbf{ki})^\mathbf{2} \end{aligned} \tag{3}$$

where $\hat{\mathbf{H}}_{ij}^{f} = \hat{\mathbf{H}}_j^s \hat{\mathbf{H}}_i^{s\,-1}$ is the predicted homography between f_i and f_j.

Eq. (2) and (3) can be further combined as, $\tilde{\mathbf{e}}_k = \delta_k \mathbf{e_k^s} + (\mathbf{1} - \delta_\mathbf{k})\mathbf{e_k^f}$, where $\delta_k = 1$ for for slide-frame correspondences, and $\delta_k = 0$ for for frame-frame correspondences. Summing up, the re-projection errors of all keypoint correspondences gives the bulk errors that we would like to minimize $\mathbf{E} = \sum_{\mathbf{k=1}}^{\mathbf{n}} \tilde{\mathbf{e}}_\mathbf{k}$, where n is the total number of keypoint correspondences.

If only frame matching is considered, i.e., $\delta_k = 0$, then the solution of bundle adjustment is not the the true slide homographies, but only up to a common projective matrix from them. However, the optimized results can be used to compute a better transformation for any two frames in the sequence as the unknown common matrix is canceled off. The improved frame homographies in turn lead to better recovery of the slide homographies for the frame sequence, according to the *homography consistency* rule.

Note that if the sequence has only one single frame image, we will need to solve a homography between a frame and a slide. The algorithm degenerates to the well-known Gold Standard (GS) algorithm [12], which solves the homography by minimizing the re-projection error defined in (2).

We now look at how to minimize $\mathbf{E}$ given the initialization of the parameters

$$\mathbf{P} = [\hat{\mathbf{H}}_\mathbf{1}^\mathbf{s}, \hat{\mathbf{H}}_\mathbf{2}^\mathbf{s}, \cdots, \hat{\mathbf{H}}_\mathbf{m}^\mathbf{s}, \hat{\mathbf{x}}_\mathbf{1}, \hat{\mathbf{x}}_\mathbf{2}, \cdots, \hat{\mathbf{x}}_\mathbf{n}, \hat{\mathbf{y}}_\mathbf{1}, \hat{\mathbf{y}}_\mathbf{2}, \cdots, \hat{\mathbf{y}}_\mathbf{l}], \tag{4}$$

where $\hat{\mathbf{H}}^s$ represents slide homographies, $\hat{\mathbf{x}}$ is slide keypoints and $\hat{\mathbf{y}}$ is the *reference frame keypoints* defined in section 2. Denoting by $\mathbf{X}$ the measurement vector concatenated from all the observed points, we consider an approximate function f taking the parameter vector $\mathbf{P}$ to the predicted measurement vector $\hat{\mathbf{X}} = f(\mathbf{P})$. Thus the $\mathbf{E} = ||\mathbf{X} - \hat{\mathbf{X}}|| = ||\mathbf{X} - \mathbf{f}(\mathbf{P})||$. This is a typical non-linear optimization problem which we solve by the Levenberg-Marquardt (LM) algorithm [13]. As a standard technique for non-linear optimization problems, the LM algorithm can be seen as a combination of steepest descent and the Gauss-Newton method. It provides faster convergence and regularization for over-parametrized problems. At each step, the algorithm iteratively updates $\mathbf{P}$ by $\mathbf{\Delta P} = (\mathbf{J^T J})^{-1}\mathbf{J^T}\epsilon$ where J is the Jacobian $\partial\mathbf{X}/\partial\mathbf{P}$ evaluated at each step and $\epsilon = \mathbf{X} - \mathbf{f}(\mathbf{P})$ is the residual vector at the current step, i.e., the vector of differences between the observed and predicted points.

3. EXPERIMENTS

We evaluated the accuracy of homography estimation using the bundle adjustment approach and compare it to the previous frame-slide matching method (referred to as "RANSAC") using ground-truth data. Constructing ground truth for this problem is difficult, and to do so we combined human intervention with state-of-the-art image registration methodology to generate a *semi-ground-truth* data set for our evaluation purposes. The basic idea is to identify one sufficiently accurate homography for each video clip without slide change or camera switch. This was done either manually or via inspecting the homographies computed from RANSAC and bundle adjustment, which, if needed, were further improved by using an image registration algorithm. The identified homography was further used to compute other homographies. Using this method, we created the semi-ground-truth data for 6 videos in the CONF1 data set used in [8], which gives the most varied appearance of slides.

Quantitative Evaluation. Let $\hat{\mathbf{H}}$ be the ground-truth homography of a pair of slide and frame images and $\mathbf{H}$ be a test homography estimated from some algorithm. The estimation error of $\mathbf{H}$ is given as

$$\mathbf{e}(\hat{\mathbf{H}}, \mathbf{H}) = \sqrt{\frac{\sum_{\mathbf{i=1}}^{\mathbf{N}} (\hat{\mathbf{H}}\mathbf{x_i} - \mathbf{H}\mathbf{x_i})^\mathbf{2}}{\mathbf{N}}} \tag{5}$$

where $\mathbf{x_i}$ is a random point uniformly distributed in the slide image. The error $\mathbf{e}$ can be considered as the average projection error of a random image point from the test homography, relative to the ground-truth mapping. However, the geometric distortion of a projected slide is not merely determined by the estimation error. In general, larger slides (zoom-in slides) tend to tolerate more errors than smaller ones (zoom-out slides). For this reason, we normalize the estimation error by the scale $\mathbf{s}$ of the projected slide with respect to the original one, i.e., $\tilde{\mathbf{e}}(\hat{\mathbf{H}}, \mathbf{H}) = \mathbf{e}(\hat{\mathbf{H}}, \mathbf{H})/\mathbf{s}$. Here $\mathbf{s}$ is given by $\mathbf{s} = \sqrt{\mathbf{s_x^2} + \mathbf{s_y^2}}$ where s_x and s_y are the horizontal and vertical scalings, respectively. We approximate s_x by $\hat{h}_{11}/\hat{h}_{33}$ and s_y by $\hat{h}_{22}/\hat{h}_{33}$ where $\hat{h}_{11}$, $\hat{h}_{22}$, and $\hat{h}_{33}$ are the diagonal

Alg.	V_1	V_2	V_3	V_4	V_5	V_6
RANSAC	1.66 (0.09)	1.55 (0.07)	1.38 (0.13)	1.61 (0.15)	1.51 (0.09)	1.90 (0.07)
BUNDLE	**0.73 (0.07)**	**0.54 (0.04)**	**0.77 (0.10)**	**0.89 (0.11)**	**0.59 (0.05)**	**0.88 (0.04)**

Table 1. The means and standard errors (in the brackets) computed from the homography estimation errors of the sampled frames in the videos from CONF1.

Alg.	< 0.8	$0.8 - 1.2$	> 1.2
RANSAC	2.72 (0.09)	1.57(0.09)	1.47 (0.04)
BUNDLE	0.84 (0.07)	1.04 (0.05)	0.64 (0.03)

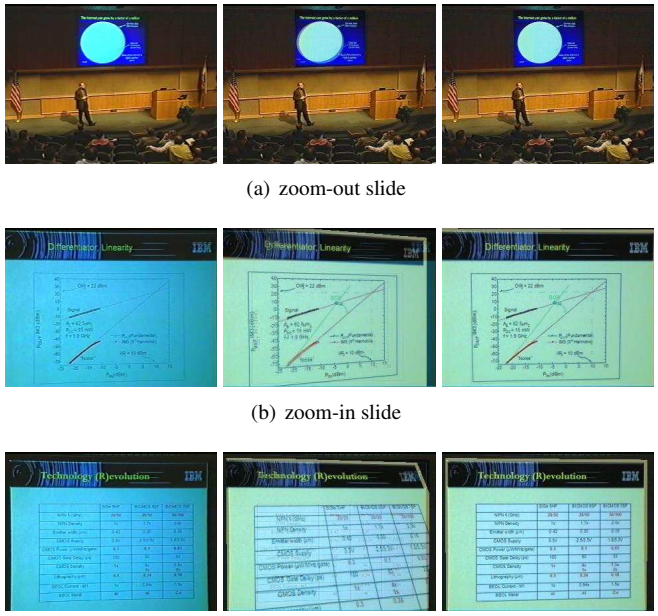
Table 2. The means and standard errors (in the brackets) under different scales of the projected slides in the video with respect to the original slides: small (< 0.8), normal ($0.8 - 1.2$) and large (> 1.2).

elements of $\hat{\mathbf{H}}$.

We evaluated the RANSAC and bundle adjustment (BUNDLE) methods. The number of random points used for evaluation is fixed as $N = 100$ in the experiments. Table 1 shows the statistics (means and standard errors) of the estimation errors of the sampled frames from the 6 videos in CONF1. As expected, bundle adjustment significantly outperforms RANSAC on all test videos, achieving sub-pixel accuracy. The results clearly demonstrate that bundle adjustment can produce much more accurate and consistent homographies. Note that in our experiments, we excluded frames with an estimation error greater than 5.0 ($\mathbf{e} > \mathbf{5.0}$) in order to alleviate the effects from those completely failed estimations.

We further broke down the results based on the scale of the projected slides in the video with respect to the original slides. As shown in Table 2, bundle adjustment performs quite well on zoom-in or zoom-out slides, clearly indicating its powerful ability in addressing poor homography estimations in the case of camera zoom.

Fig. 3 shows the improvements of homography estimation by bundle adjustment over RANSAC under several challenging conditions: a) *zoom-in* slides; b) *zoom-out* slides; c) *blurry*. The slides are first back projected into the frames using the homographies computed from RANSAC and bundle adjustment, and then overlaid with the frame images.



(a) zoom-out slide

(b) zoom-in slide

(c) blurry slide

Fig. 3. Improvement of homography estimation by bundle adjustment (right) over RANSAC (middle). The original images are shown on the left. The back-projected images using the RANSAC homographies are clearly distorted, whereas the ones using BUNDLE are accurate. (Seeing this for the zoom-out slide requires viewing the image electronically and zooming in.)

4. CONCLUSIONS

We presented a nonlinear optimization approach for accurate spatial alignment of slides to the corresponding presentation video. Our approach demonstrates great capability in yielding more accurate and consistent projective mappings from slides to video frames. With the improved alignment of slides with video, we have implemented a number of tools to make instructional content more accessible in the SLIC system [1, 2].

5. REFERENCES

[1] SLIC: Semantically Linked Instructional Content, "http://vision.cs.arizona.edu/slic/," 2006.

[2] J. Torkkola, , A. Winslow, Q. Fan, R. A. Swaminathan, K. Barnard, A. Amir, A. Efrat, and C. Gniady, "Content enrichment in presentation video," Tech report, University of Arizona, *http://kobus.ca/SLIC/publication/SLICApplicationseTR.pdf*, 2009.

[3] L. A. Rowe and J. M. Gonzelez, "BMRC lecture browsers," in *http://bmrc.berkekey.edu/frame/projects/lb/index.html*, 1999.

[4] S. Mukhopadhyay and B. Smith, "Passive capture and structuring of lectures," in *ACM Multimedia (1)*, 1999, pp. 477–487.

[5] G. D. Abowd, C. A.and Feinstein A. Atkeson, C. E. Hmelo, R. Kooper, S. Long, N. Sawhney, and M. Tani, "Teaching and learning as multimedia authoring: The classroom 2000 project," in *ACM Multimedia*, 1996, pp. 187–198.

[6] A. Amir, G. Ashour, and S. Srinivasan, "Automatic generation of conf. video proceedings," in *Journal of Visual Communication and Image Representation*, 2004, pp. 467–488.

[7] ViaScribe, ," in *http://w3.ibm.com/able/solution_offerings/ViaScribe.html*.

[8] Q. Fan, K. Barnard, A. Amir, and A. Efrat, "Robust spatio-temporal matching electronic slides to presentation video," Technical report, *http://kobus.ca/slic/publication/spatiotemporal_matching_tr.pdf*, University of Arizona, 2008.

[9] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment – A modern synthesis," in *Vision Algorithms: Theory and Practice*.

[10] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 20, pp. 91–110, 2004.

[11] A. Fischler, M. and C. Bolles, R., "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, pp. 381–395, 1981.

[12] R Hartley and A. Zisserman, *Multiple view and geometry in computer vision*, chapter 3: Estimation–2D Projective Transformation, Cambridge University Press, 2002.

[13] D Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," in *SIAM J. Appl. Math.*, 1963, vol. 11, pp. 431–441.