Panoramic Scanned Page Using Mobile Phone Camera

Naveed I Rao,Huijun Di
Tsinghua University
Pervasive Computing Lab
Beijing, 100084, China,
naveed03,dhj98@mails.tsinghua.edu.cn

GuangYou Xu
Tsinghua University
Pervasive Computing Lab
Beijing, 100084, China,
xgy-dcs@mails.tsinghua.edu.cn

Abstract

Today, almost all mobile phones introduced in market are equipped with built in cameras and in this work, we have proposed a new application by using this camera. A scanned image of a large document is achieved by completely traversing this camera on it. All successive frames are joined seamlessly to generate a panoramic image of the page. It is a difficult task due to, lateral movements of camera along with hand shaking, change in illumination due to hand shadow etc. Matter is further complexed by the fact that text alphabets does not possess distinct features which is a key requirement for panoramic image generation. A framework is designed for coarse to refine alignment of images based on the background modeling approach. Foreground segmentation and correspondence estimation are expressed as a unified labeling problem, and are solved efficiently via tree dynamic programming (TDP). Experiments proved our algorithm to be robust in performance.

1 INTRODUCTION

Integration of maximum devices in one piece of equipment is taking its realm in present times. A single phone can be used not only to make a telephone call but also can be enjoyed as audio/video player, capture still or movie clips, book reader, dictionary, schedule keeper etc. Recently there has been efforts to use its camera more than taking images i.e. to use it as a tool to bridge a gap between virtual space and physical space [12]. For example, visual codes are introduced at pages with encoded information about the same image (virtual image) at web, later these markers are used to track mobile phone position relative to the hard copy by comparing virtual and real image [12]; mobile phones cameras can be used as a virtual mouse [5]; mobile phone cameras are used as data entry devices [11].

In this work we present a new application of mobile phone camera i.e. scanning of a large document by scrolling camera horizontally and vertically on it, which later is transferred into a panoramic page and can be sent as single image file. In this task, various complexities are experienced from the field of computer vision; camera is moving and is undergoing panning, tilting and slightly zooming (as individual's hand can not maintain fix distance from the page); camera possess jitter in the motion as it is difficult for a person to traverse it smoothly; change in illumination is experienced due to shadow of hand. For a panoramic page, correct alignment among consecutive frames is required. Reliable alignment can be achieved by choosing the distinct feature points between successive frames. Written text alphabets lack these feature points and above mentioned problems makes the whole scenario more complex. A solution to this problem can be by computing the pose of the phone and relate it with world coordinates, but it needs a virtual image and special markers [12]. This is an application to our research work [8],[9],[10], with changes suitable for text based matching. For the sake of completeness , they are briefly explained here for detail treatment, please refer to [8],[9].

Paper is organized as, section 2 briefly explains framework overview, section 3 describes joint correspondence background modeling, experimental results are covered in section 4 and conclusion with future work is narrated in next section.

2 Framework overview

In our work, a panoramic background model (PBG) is built over a wide area, with a free moving camera. Due to parallax effect (caused by movement of camera's optical center) and lens distortion, a global transformation between current frame and PBG doesn't exist. Furthermore, because of facts such as noise, sub-pixel effect computational error, and disturbances created by change in illumination, the estimation of transformation is also inaccurate. Therefore we only assume an approximate alignment between current frame and model. Our PBG along with associated optimiza-

tion algorithm compute a dense (pixel level) refined matching between current frame and PBG. This refined matching aid in achieving good foreground segmentation results by eliminating the misalignment error. A specially designed updating scheme of PBG, ensure a stable system over a long period of time. Our framework is shown in figure 1 and the steps are explained below. Step 1: Approximate alignment between current frame and PBG is achieved through projective transformation in initial step. Same as in [1], we always compare current frame with model to estimate the transformation rather than frame by frame comparison to avoid registration error accumulation. A panoramic image of currently explored scene is first generated from PBG by taking the mean value of the most likely model, to enable image based matching. LK [6] is used to find sparse correspondence between current frame and this panoramic image. Since directly searching over whole panoramic image is infeasible for LK (due to time consuming and easily entrapped into local minimum), previously estimated projective transformation is used to give an initialization for LK (formally, use the previously computed transformation to synthesize a virtual image, and then apply LK, as shown in figure 1). Given the estimated sparse correspondence, M-estimator [4] is used to calculate the projective transformation between current image and PBG robustly. Step 2: generate an auxiliary image for motion compensation which is used in next steps as an input. Based on the estimated projective transformation in step 1, current frame is transformed into the coordinates of panorama and an auxiliary image is generated by cutting the transformed image out, as shown in figure 1.

3 Joint Foreground Segmentation and Correspondence Estimations

As mentioned in previous section, the correspondences between auxiliary image and PBG are posed as our model parameters. These estimated correspondences provide dense matching, and eliminate the misalignment effects thus improving foreground segmentation. In this section, we will discuss the associated algorithm for foreground segmentation and correspondence estimation. There exists an inherent ambiguity in dense correspondence , a reasonable smoothness constraint is required to enable a matching procedure work accurately. In order to regularize the resulting correspondence, piecewise smoothness assumption of a scene is normally used [9][14], we also apply such a assumption in our problem domain. The rest formulation is similar to [8], except we discuss the panoramic case. For the sake of completeness and with the courtesy of authors, we briefly overview it here, but with our work i.e. PBG point of view. For each pixel in auxiliary image J_t, our goal is to find its correspondence in PBG, as well as classify it as background pixel or foreground pixel. It can be viewed as a labeling problem: assign an optimal label $l = (f, \Delta x, \Delta y)$ to each pixel in J_t while f gives segmentation result (0 for background, 1 for foreground) and displacement vector $(\Delta x, \Delta y)$ award its correspondence. The whole labeling space is

$$L = \{(0, \Delta x, \Delta y) | (\Delta x, \Delta y) \in D\} \bigcup \{(1, *, *)\} \quad (1)$$

where D is the domain of displacement vector $(x + \Delta x, y + \Delta y)$, $D = \{(i, j) | -a \leq i \leq a, -b \leq j \leq b\}$. Since correspondence is only modeled in-between background pixels, so when a pixel is labeled as foreground (f=1), we do not consider the correspondence for it.

To enforce scene smoothness, a minimal span tree (MST) [3]is generated from an undirected graph which is defined on the auxiliary image J_t: pixels as vertices; edges as piecewise connection between neighbors; and absolute intensity difference between neighboring pixels as a weight of edge to approximate the geometric smoothness. This MST contains most important edges in the graph that smoothness enforcing should be imposed. The Energy function for assigning optimal label is defined as

$$E(l) = \left(\sum_{v \in V} m(l_v) + \sum_{(u,v) \in E} S(l_v, l_u) \right) \quad (2)$$

where V and E are the set of all nodes and edges in MST respectively, l is the conjunction of labels for all nodes, u and v are nodes (also pixels, coordinates in image J_t), l_v and l_u are labels. In eqn. 2,$m(l_v)$ is the data measurement penalizing any disagreement of the label (foreground segmentation along with correspondence) with the observed data (image J_t), and is defined as negative logarithmic likelihood of the label, i.e.

$$m(l_v) = \begin{cases} -\log P_B \left(J(x, y), \Delta x, \Delta y \right), \text{for} f(x, y) = 0 \\ -\log\big((J(x, y), \Delta x, \Delta y) \big) \quad , \text{for} f(x, y) = 1 \end{cases} \quad (3)$$

$S(l_v, l_u)$ in eqn. 2 is the smoothness term defined in the form of Potts model [7]:

$$S(l_v, l_u) = \begin{cases} w_1(I_t(v), I_t(u)), if f_v \neq f_u \\ 0, if f_v = f_u = 1 \\ 0, if (f_v = f_u = 0) \& (\Delta x_v = \Delta x_u) \& \\ (\Delta y_v = \Delta y_u) \\ w_2(I_t(v), I_t(u)), otherwise \end{cases} \quad (4)$$

where w_1 and w_2 are two weights to penalize the discontinuity of labels between parent node and child node in MST.

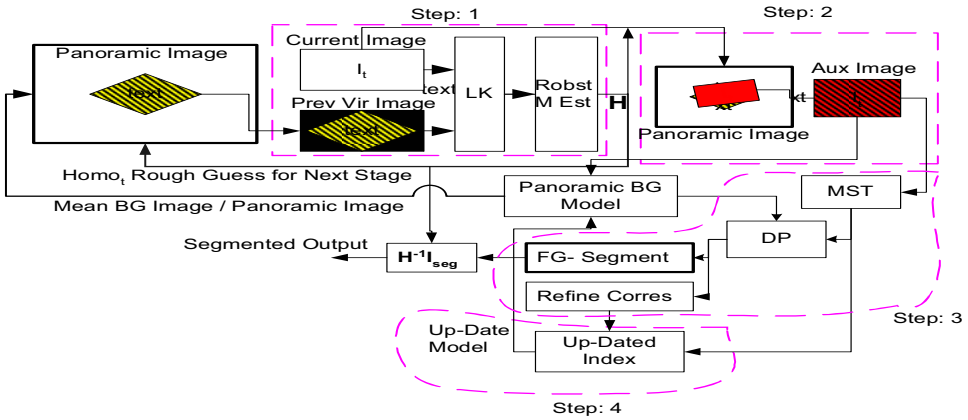

Figure 1. Schematic Diagram of FrameWork

Due to tree structure, the minimum energy can be written as

$$E_{min} = \min_{l_r \in L} \left(m(l_v) + \sum_{v \in C_r} E_v(l_r) \right), \tag{5}$$

where r is the root node of MST, C_r is the set of $r's$ children,$E_v(l_r)$ is defined recursively as

$$E_v(l_p) = \min_{l_v \in L} \left(m(l_v) + S(l_v, l_p) + \sum_{(u \in C_v)} E_u(l_v) \right) \tag{6}$$

The minimum energy in equ. 6 can be calculated efficiently via tree dynamic programming [2], and associated global optimal labels are then obtained, accordingly foreground segmentation and dense correspondence are achieved.

4 Updating Panorama

We use the same strategy as [13] to update background model, i.e. the on-line K-means approximation and all pixels are used to update the model, allowing objects to be part of background with the passage of time. As the correspondence is modeled explicitly, the difference is the decision, for which model (which location) should be updated based on the current pixel information. For the pixel (x, y) in input image, suppose the location of the model to be updated is $(x + dx, y + dy)$. After the optimal label $f*$, $\delta x*$, $\delta y*$ obtained for this pixel in section 3, (dx, dy) is calculated as

$$(dx, dy) = \begin{cases} (\Delta x, \Delta y), if f^* = 0 \\ (0, 0), if f^* = 1 \end{cases} \tag{7}$$

Since correspondence is meaningless when the pixel is segmented as foreground, so we update the model at same location (i.e. $dx = dy = 0$). It is worth noting that, since we match input image to the model, different pixels may be matched to the same model. At current frame, background models at few locations may miss their updating, consequentially their evolutions will be slow. Thus foreground object may exist for a longer time than [13] if it is presented during model initialization. However, this does not cause more false segmentation results for the related pixels, as these pixels are matched to other background models and are not detected as ghost.

5 Experimental Results

An A4 page is scanned using a mobile phone camera (figure 2a). This page contains some figures, text and characters in chinese language(same as figures). This combination will ensure the presence of distinct feature points in the near vicinity. When a camera is traversed on it, a panoramic page is filled step by step. In generated results , blurring effect is visible, which is due to gaussian models. In future, we intend to reduce this effect. Sequence is run on P-IV 1.67 GHz with 256MB RAM, and 2 frames per second is achieved. Detailed results can be downloaded at http://media.cs.tsinghua.edu.cn/ naveed/research.html

6 Future Work

In this paper, we have proposed and implemented a new application for mobile phone cameras to scan a large document by traversing it horizontally and vertically. A panoramic image is achieved by correctly aligned all successive images in the presence of hand shaking, shadow due to hand, etc. Generated results results experience blur and we intend to remove it in our future work. We also want to extend this work to scan curved page of book (By opening

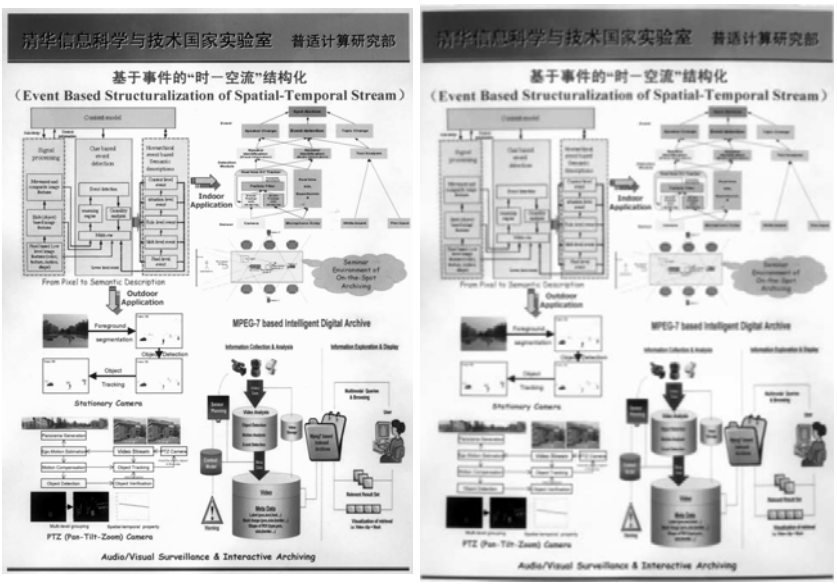

Figure 2. Original Image (a) and its panoramic scanned copy(b)

the book from center, one page is smooth and second faces curvature).

References

[1] Mittal Annurag and HuttenLocher Dan. Scene modeling for wide area surveillance and image synthesis. In *CVPR*, 2000.

[2] Monnet Antoine, Mittal Anurag, Paragios Nikos, and Ramesh Visvanathan. Background modeling and subtraction of dynamic scenes. In *ICCV*, 2003.

[3] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithm*. MIT Press, 1990.

[4] D. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. chap 15, 2003.

[5] Neema Moraveji Yuanchun Shi Hao Jiang, Eyal Ofek. Direct pointer: Direct manipulation for large-display interaction using handheld cameras. In *CHI*, 2006.

[6] B. D. Lucas and T. Kanade. An iterative image registration technique with an application in stereo vision. In *Seventh International Joint Conference on Artificial Intelligence (IJCAI-81)*, 1981.

[7] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *CVPR*, 2004.

[8] I Rao Naveed, Di Huijun, and Xu Guangyou. Joint correspondence and background modeling based on tree dynamic programming. In *ICPR*, 06.

[9] I Rao Naveed, Di Huijun, and Xu Guangyou. Stereo correspondence using bayesian network and dynamic programming on a color based minimal span tree. In *Advanced Concepts for Intelligent Vision Systems (Acivs 2006)*, 2006.

[10] I Rao Naveed, Di Huijun, and Xu Guangyou. Stable panoramic background model with a free moving camera. In *submitted in Advanced Concepts for Intelligent Vision Systems (Acivs 2007)*, 2007.

[11] Tapan S. Parikh. Cam: A mobile paper-based information services architecture for remote rural areas in the developing world. In *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'05)*, 2005.

[12] Michael Rohs and Beat Gfeller. Using camera-equipped mobile phones for interacting with real-world objects. In Alois Ferscha, Horst Hoertner, and Gabriele Kotsis, editors, *Advances in Pervasive Computing*, pages 265–271, Vienna, Austria, April 2004. Austrian Computer Society (OCG).

[13] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *TPAMI*, 2000.

[14] Olga Veksler. Stereo correspondence by dynamic programming on a tree. In *CVPR*, 2005.