**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 com- pletely 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 cam- era 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 im- ages based on the background modeling approach. Fore- ground segmentation and correspondence estimation are expressed as a uniﬁed labeling problem, and are solved efﬁ- ciently via tree dynamic programming (TDP). Experiments proved our algorithm to be robust in performance.*

**1 INTRODUCTION**

Integration of maximum devices in one piece of equip- ment 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 in- troduced 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 cam- eras can be used as a virtual mouse [5]; mobile phone cam- eras 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 trans- ferred into a panoramic page and can be sent as single im- age 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 indi- vidual’s hand can not maintainﬁxdistance from the page); camera possess jitter in the motion as it is difﬁcult for a per- son to traverse it smoothly; change in illumination is experi- enced 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 im- age 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 frame- work 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 er- ror, 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-

352

tion algorithm compute a dense (pixel level) reﬁned match- ing 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 pro- jective 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 toﬁndsparse cor- respondence between current frame and this panoramic im- age. Since directly searching over whole panoramic image is infeasible for LK (due to time consuming and easily en- trapped into local minimum), previously estimated projec- tive 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 transfor- mation 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 trans- formed 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 Cor- respondence 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 seg- mentation and correspondence estimation. There exists an inherent ambiguity in dense correspondence , a reasonable smoothness constraint is required to enable a matching pro- cedure work accurately. In order to regularize the result-

background pixel or foreground pixel. It can be viewed as a labeling problem: assign an optimal label l = (f,Δx,Δy) to each pixel in Jt while f gives segmentation result (0 for background, 1 for foreground) and displacement vector (Δx,Δy) award its correspondence. The whole labeling space is

L = {(0,Δx,Δy)|(Δx,Δy)∈D} {(1,∗,∗)} (1) where D is the domain of displacement vector (x +

Δx, y +Δy), D = {(i, j)|−a≤i≤a,−b≤j≤b}. Since correspondence is only modeled in-between back- ground 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 Jt: pixels as vertices; edges as piecewise connection between neighbors; and ab- solute intensity difference between neighboring pixels as a weight of edge to approximate the geometric smooth- ness. 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) = m(lv) + S(lv, lu) (2)

v∈V

(u,v)∈E

where V and E are the set of all nodes and edges in MST re- spectively, l is the conjunction of labels for all nodes, u and v are nodes (also pixels, coordinates in image Jt), lv and lu are labels. In eqn. 2,m(lv) is the data measurement penaliz- ing any disagreement of the label (foreground segmentation along with correspondence) with the observed data (image Jt), and is deﬁned as negative logarithmic likelihood of the label, i.e.

−log PB (J(x, y),Δx,Δy) , forf(x, y) = 0 m(lv) =

−log (J(x, y),Δx,Δy) , forf(x, y) = 1

(3)

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

⎧

⎪⎪⎪⎪⎪⎪w1(It(v), It(u)), if fv = fu⎪⎪⎪⎨0, if fv = fu = 1

ing correspondence, piecewise smoothness assumption of a scene is normally used [9][14], we also apply such a as- sumption in our problem domain. The rest formulation is

S(lv, lu) =

0, if (fv = fu = 0)&(Δxv =Δxu)&

⎪

⎪⎪(Δyv =Δyu)

⎪

⎪

⎪

⎩w2(It(v), It(u)), otherwise

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 Jt, our goal is toﬁndits correspondence in PBG, as well as classify it as

(4)

where w1 and w2 are two weights to penalize the dis- continuity of labels between parent node and child node in MST.

353

6WHS 6WHS

&XUUHQW ,PDJH

WH[W

WH[W

,W

/.

3DQRUDPLF ,PDJH

5REVW 0 (VW

$X[ ,PDJH

WH

WH[W

[W

WH[W

,W

3UHY 9LU ,PDJH

+

WH[W

**WH[W**

,W

3DQRUDPLF ,PDJH

+RPRW

5RXJK \*XHVV IRU 1H[W 6WDJH

3DQRUDPLF %\* 0RGHO

067

0HDQ %\* ,PDJH

3DQRUDPLF ,PDJH

'3

6HJPHQWHG 2XWSXW

+ ,VHJ

)\*

6HJPHQW

8S 'DWHG ,QGH[

5HILQH &RUUHV

8S 'D 0RGHO

WH

6WHS

6WHS

**Figure 1. Schematic Diagram of FrameWork**

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

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

m(lv) +

Ev(lr) , (5)

Emin =min

lr∈L

v∈Cr

where r is the root node of MST, Cr is the set of r s children,Ev(lr) is deﬁned recursively as

Ev(lp) = lminv∈L m(lv) + S(lv, lp) + Eu(lv) (6)

(u∈Cv)

The minimum energy in equ. 6 can be calculated ef-ﬁciently via tree dynamic programming [2], and associ- ated 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 pix- els are used to update the model, allowing objects to be part

of background with the passage of time. As the correspon- dence 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 in- put image, suppose the location of the model to be updated is (x + dx, y + dy). After the optimal label f∗, δx∗, δy∗

obtained for this pixel in section 3, (dx, dy) is calculated as

(Δx,Δy), if f∗= 0

(dx, dy) = (0, 0), if f∗= 1 (7)

Since correspondence is meaningless when the pixel is segmented as foreground, so we update the model at same

models at few locations may miss their updating, conse- quentially 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 char- acters in chinese language( same as ﬁgures). This combi- nation 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 mod- els. 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 suc- cessive 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

354

**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 model- ing for wide area surveillance and image synthesis. In *CVPR*, 2000.

[2] Monnet Antoine, Mittal Anurag, Paragios Nikos, and Ramesh Visvanathan. Background modeling and sub- traction 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 regis- tration 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. Sta- ble panoramic background model with a free moving camera. In *submitted in Advanced Concepts for Intel- ligent Vision Systems (Acivs 2007)*, 2007.

[11] Tapan S. Parikh. Cam: A mobile paper-based informa- tion services architecture for remote rural areas in the developing world. In *IEEE Symposium on Visual Lan- guages 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 Com- puting*, pages 265–271, Vienna, Austria, April 2004. Austrian Computer Society (OCG).

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

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

355