MOSAICING OF TEXT CONTENTS FROM ADJACENT VIDEO FRAMES

NAGABHUSHAN P, 1 VIMUKTHA EVANGELEEN JATHANNA2*
1Department of Studies in Computer Science, University of Mysore, Mysore
2Department of Studies in Computer Science, University of Mysore, Mysore
*Corresponding Author: Email vimuktha_ej@yahoo.com

Received: November 06, 2011; Accepted: December 09, 2011

Abstract - Incorporation of video facility in mobile devices has opened up various challenging research problems in the field of image processing. One such challenge is to mosaic video containing the document source. In this paper we propose an approach to mosaic the fragmented video frames containing text for the purpose of text localization and recognition by decomposing the fragmented frames into vertical strips and then matching the strips of consecutive frames to stitch the content. Similarity between strips are found using Scale Invariant Feature Transform - SIFT match algorithm as it is invariant to scale, rotation and geometric distortions like blurring / resampling of local image orientation planes. The process of blending the content of the matched strips is done with the help of a translation function on the best matched points estimated by homography using RANSAC were wrapped using translation function.

Key words - Vertical strips, Track movement, SIFT, RANSAC, Mosaicing, Homography

Introduction
The video facility in mobile phones has almost been utilized as video cameras. Hence there are wide varieties of images captured through mobile devices. The traditional scanner which was used to scan documents, manuscripts, textbooks is now replaced by the digital cameras. The advantage of capturing text in the form of videos is that, longer text regions can be captured and it is also possible to improve the quality of text by combining multiple frames.

Generally, if we visualize the structure of a video it is composed of many still images called frame that forms the complete moving picture. However, any video file by itself possesses the property of spatial, temporal redundancy due to the video coding standards. The standard frame rate is 24 to 25 frames per second. If we assume that we have a clip of one minute clip then

Number of frames = 24 x 60 sec = 1440 frames / min. (1)

of which many of the consecutive frames are content wise similar having very little or no difference in them with reference to time t. Hence localizing and recognizing the text in these video frames becomes computationally expensive.

It is also true that the document source having wider visible area cannot be captured in one shot using cameras. Hence in practice, we capture the text source in parts with respect to the field of view that a camera is capable of capturing resulting in large number of frames containing the combination of the target. Also, the text source in general is captured using horizontal - left to right scan called Track Movement or vertical – top to bottom scan called Boom Movement. This results in a very small variation in the spatial content of each frame i.e., according to observation information is added at the end of few columns in the next frame in case of horizontal scan or at the end of few rows in case of vertical scan with respect to the reference frame. At the same time the information is subtracted in the first few columns and rows of the reference frame in case of horizontal and vertical image capturing.

Hence an efficient method to track the spatial content transition of consecutive frames and merging the additional content of the successive frame with the present frame content will be useful. In this paper we propose a method to blend consecutive frame information using decomposition of frames into vertical image strips and matching them using SIFT descriptors. We exploited the available RANSAC algorithm to estimate the homography, extract best match for points that coincide. Then the homography was used to transform the image for wrapping using simple translation function

Literature Survey
There are several ongoing researches on camera based document mosaicing and video mosaicing. Few of the methods depicted in the literature are discussed below.

Lowe et al’s Scale Invariant Feature Transform – SIFT was used in [1] to develop automated panorama reconstruction of images. The SIFT features were extracted from video frames and were matched using k-nearest neighbourhood. The homographies were estimated between the matched pairs using RANSAC and was verified by probabilistic model. Then each of the
connected components derived from the graph search method was subjected to bundle adjustment with joint camera parameters and was subjected to multi band blending to provide panoramic view which was invariant of scaling, rotation and geometric distortions

Liang et al [2] proposed a mosaicing technique for camera captured images. Overlapping regions within a small area with perspective distortions were used for image registration. The seamless blending was obtained by sharpness based text component selection. Mosaicing was done based on feature based alignment.

Zappala et al [3] and Peleg & Gee [4] proposed methods on document image mosaicing by estimating the motion through point matching. They considered the features based on the domain through exhaustive search procedure to extract best matches for mosaicing. The method works for 50% overlapping regions.

Hemanth Kumar et.al [5] proposed a novel approach for mosaicing split images based on simple pixel correspondence and Euclidian distance.

Junichi Hoshino in [7] proposed a fast method for panoramic image mosaicing using one-dimensional flow estimation. The approach first estimated the 1D flow vectors at each pixel in x and y directions and then using iterative least squares methods the global affine parameters were estimated from 1D flow. The accurate parameters were obtained by iteratively repeating optimization in x and y directions. The experimental results showed that the image mosaicing is possible in near-real time.

Hemanth Kumar et.al in [8] proposed a technique to mosaic the two split images of a large document based on matching the sum of values of pixels of window in the split images. The method compared the sum of values of pixels of window in split images to identify overlapping region in the split images.

Isgrò et in [9] proposed a feature based image mosaicing method. In this method, the feature points were extracted from one of two images to be stitched, and then the corresponding points in the other images were calculated. An Euclidean transformation parameter were estimated from the corresponding points. The Euclidean transformation parameter was used to stitch the images. However, the method is unable to deal with scaling and perspective distortion because the Euclidean transformation includes only translation and rotation.

Lian et al. in [10] proposed a mosaicing method for camera captured document images. the method stitches document images captured from arbitrary angles using a digital camera. In this method, perspective distortions of document images are removed based on vanishing points estimated from text line direction and vertical character stroke direction. Then feature points of fronto-parallel document images are extracted and matched using PCA-SIFT. [11]

Tomohiro et.al. in [14] proposed a mosaicing method for camera-captured document images. by calculating the corresponding feature points using an image retrieval method called LLAH since it was invariant to perspective distortion. Feature points were matched without compensating perspective distortion. Document images were aligned using a perspective transformation parameter estimated from the correspondences.

To the best of our knowledge the problem of mosaicing text contents from adjacent video frames in document videos is not addressed in the literature and hence we have proposed a new approach to solve the problem.

Proposed Methodology
The proposed approach undergoes eight different stages for mosaicing / blending the text contents from adjacent frames. The video files used for experimentation are captured through mobile cameras placed at a fixed distance and camera motion is horizontal left to right - track to capture the text from the target location. The different stages involved in this approach is as given in the block diagram

**Fig(1).** Block Diagram of proposed method

**Frame Extraction**
A video is composed of many still images called frame that forms the complete moving picture. Hence as the first step towards solution, we segment the video into frames.

\[ V \rightarrow \{ f_1, f_2, f_3, \ldots, f_n \} \]  

where, \( f_1 \) to \( f_n \) are the frames having 3 dimensions. Two dimensions serve as spatial (horizontal and vertical) directions of the moving pictures, and one dimension represents the time domain.

**Decomposition of frames into vertical strips**
Before decomposing, an intermediate stage is involved for noise removal and quality enhancement. A 3x3 median filter is applied to remove noise. This is effective at preserving edges by removing very fine noise. Image sharpening based on intensity adjustment with Weiner filter to deblur motion blurred frames is used to enhance and improve the overall appearance of the frame.
Once the preprocessing stage is done, each video frame is subjected to decomposition. Here decomposition refers to dividing each frame into 3 vertical strips i.e. if M x N is the size of the frame then the first vertical strip

\[ B_1 = (M, NW) \text{ where } NW = n/3 \]  

(3)

Fig (2). Vertical strips of each frame where B1, B2 and B3 represents Vertical Strip 1, Vertical Strip 2 and Vertical Strip 3 respectively.

Let us assume that the camera is moved horizontally left to right (track motion) to capture the text from the target location. According to observations the transition in the information change of the successive frames can be visualized as follows:

1. Information entering into the current frame is present in case of right block i.e. vertical Strip 3
2. Information leaving out of the current frame is towards the left side of the frame i.e. vertical strip 1 gradually, depending on the camera track movement.

With respect to time the contents of vertical strip 1 in current frame is popped out and is added to previous frame content, vertical strip 2 contents is pushed into vertical strip 1, vertical strip 3 contents is pushed into vertical strip 2 and vertical strip 3 contains added information of vertical strip 3 of next frame resembling a queue data structure. Fig (3) shows the process of information transition in frames with respect to time t units

\[ \text{Frame}_\text{difference} = \text{abs(average(color of frame 1) - average(color of frame 1))} \]  

(4)

where, Frame_difference is a matrix that quantify the amount of displacement of each pixel with respect to motion.

SIFT based frame matching

Lowe in [6] described a method for mosaicing based on local descriptor called Scale Invariant Feature Transform – SIFT. It transforms the entire image into set of feature vectors. The key points are detected as a set of difference-of-Gaussian images which are local maxima and minima of gray level points in the given space. The set of each key point are invariant of image translation, scaling, and rotation, partially invariant to illumination changes and robust to local geometric distortion, the points that having low contrast and edge response are omitted. Orientations that are dominant are assigned to the key locations of the localized descriptors. The descriptors produce more key points for textured regions than uniform regions. The matching is done based on simple NN algorithm on the Euclidean distance of the located key points with a threshold value 0.8 to omit outliers.

Hence the next step of the approach is to extract and match SIFT [1] features between frame (1) and frame (2) to set the reference key frame. If the match function returns a value less than the threshold that is set then the index is incremented by 1 and compared with the next frame i+1 and it indicates that block contents of frame 1 is pushed out with block 2 contents and block3 has more contents added from the rest of the consecutive frames or the frame contains out of focus or scene change or blur. Once the reference frame is set we are sure that the minor spatial transition is in block3. Hence the reference frame’s block3 is compared with all the other frame’s block3’s till a value less than the threshold is found. If we have a value lesser than the threshold it indicates the need for changing the reference frame. The figure below shows the procedure

Fig (4) Original Frame Information in BTI video sequence
Fig (5) Information changes in the frames and decomposition of frame into vertical strips - Strip 3 of frame 1 and 2 has very minute changes but strip 3 of frame 15 has noticeable track moment.

Fig (6) (a) SIFT descriptors on Frame1 and Frame2, (b) Frame 1 set as reference after SIFT matching (c and d) strip 3 of frame3 and strip 3 of frame 15

Fig (7) a) Reference frame b) Strip 3 of frame 15 c) SIFT descriptors on Frame 1 and strip 3 of frame 15

Fitting Homography using RANSAC in video strips
Homography maps points of two images with one to one correspondence. The SIFT key points sometime produces false match locations called outliers Hence it can be used as a matching function to find correspondence between two strips of successive frames. According to [1] a 2D homography is defined by 3x3 matrix represented as H which corresponds to pixel P of strip 1 of frame i to pixel P1 of strip 1 in frame i +1 can be estimated using 4 or more corresponding points obtained by SIFT descriptors using the equation

\[ HP = WP^1 \]  

where w is the scale parameter.

The matched points obtained by SIFT feature descriptor algorithm produces outliers. Hence RANSAC helps in fitting homographies between two strips of frames. RANSAC is an iterative algorithm that matches random 4 pairs of already matched points and uses it to compute homography. It then checks for the number of remaining key points that are consistent with the Homography H obtained. Hence the required number of iterations will increase depending on the amount of outliers.

3.5. Wrapping
The homography fitted image is wrapped using a simple translation function available by a solving a set of affine transformation using 1st order polynomial

\[ x_0 = a_0 + a_1x + a_2y \]
\[ y_0 = b_0 + b_1y + b_2y \]

Then the text contents of the adjacent frames are stitched horizontally. The mosaiced image after wrapping is as shown below

Fig (8) Homography estimation using RANSAC with actual number of inliers

Fig (9) Wrapped Image

Video Test Data
For our experimentation, data samples with different varieties of document videos were collected and tested for mosaicing. The video dataset exhibit the following attributes

a) The videos were captured using mobile camera with high 5.0 mega pixel, medium 3.5 mega pixel and 1.3 mega pixel for producing low quality as well as high quality videos of nearly 1500 frames each.
b) We considered documents with homogenous background (white) with black foreground and few colored samples.
c) The videos were captured with left to right track movement with fixed movement of 1 inch to avoid motion blur, out of focus and at a distance of 15 inches from the source.
d) The experiments were conducted to capture high speed videos. But as we are concentrating on the document images containing large text sources the quality of texts in the videos get deteriorated due to motion and text becomes blur or out of focus. The Fig(10b) depicts the effect of high speed video. Hence the mosaicing was not possible as the SIFT match points were less than the threshold value set in our experiments
The slow speed videos like keeping camera at a point for more than 50 seconds i.e. static recording results in the increase of similar content consecutive frames. Hence change in the reference frame and transition in the information becomes slow. The result of mosaicing such a video is illustrated in Fig 13 and Fig 17.

f) The videos considered in our experiment were captured in ordinary lighting conditions without using mobile camera flash. Illumination was handled to certain extent. The result for wrapping illuminated frames is as shown in Fig(11).

5. Complexity

The step of the SIFT match algorithm that takes most of the processing time is the descriptor matching, as it has a complexity of $O(N M D)$ comparisons, where $N$ is the number of features in the new test image i.e consecutive frame in our case, $M$ is the number of features in the reference frame and $D$ is the dimension given by $X \times Y$. Here in our experimentation we are reducing the dimension $Y$ by $Y/3$ columns hence, saving the time consumed for matching and fitting the data. The graphs for comparison of two frames for with and without strip decomposition for time complexity analysis is shown.

Results

The algorithm was implemented using Matlab 2009. Results obtained by our approach is as shown in figures Fig (13), Fig (14), Fig (15), Fig (16) and Fig(17).

The quality of mosaiced image obtained by the algorithm was analyzed with the Mean Square Error (MSE) metric as proposed in [13] by Azzari et.al. The reference frames were set as original image and rest other adjacent frames as distorted ones and were compared with each other. The average of all MSE’s obtained by comparing reference frame and the adjacent frame was compared with MSE of reference frame and mosaiced image at each instance of mosaicing. However the overall evaluation of the results can be done only after OCRing the text contents that are mosaiced to evaluate machine readability. The result of the approach is also evaluated.
based on the readability of mosaiced image by human perception.

7. Conclusion
The new strip based method for mosaicing the text contents in adjacent frames is proposed in this paper. The studies reveal that the mosaicing time is comparatively reduced in case of this approach. Also due to the use of SIFT descriptors for matching the mosaicing is invariant of scale, rotation and geometric distortions and works on several types of documents. We also created our own video test dataset containing document videos with large text source. Future work includes the removal of seam lines that are visible in the mosaiced image and also blending videos having lighting variations and distortions.

References