Geospatial content summarization of UAV aerial imagery using mosaicking

GEOSPATIAL CONTENT SUMMARIZATION OF UAV AERIAL IMAGERY USING MOSAICKING

Rumana Aktar\textsuperscript{a}, Hadi Aliakbarpour\textsuperscript{a}, Filiz Bunyak\textsuperscript{a}, Toni Kazic\textsuperscript{a}, Guna Seetharaman\textsuperscript{b}, and Kannappan Palaniappan\textsuperscript{a}

\textsuperscript{a}Department of Electrical Engineering and Computer Science, University of Missouri-Columbia; rayy7@mail.missouri.edu, \{aliakbarpourh, bunyak, kazict, pal\}@missouri.edu Columbia, MO, USA
\textsuperscript{b}US Naval Research Laboratory, guna@ieee.org, Washington D.C., USA

ABSTRACT

In this study, we devise a method for summarizing geospatial content of aerial imagery using mosaicking. Mosaicking has been a popular method for combining large sequence of image information and providing a wide field of view in computer vision. We propose a feature based registration method which brings the complexity and homogeneity nature of maize data under consideration and formulates a two step registration method which reduces error accumulation acquired through traditional frame-to-frame registration. Experimental result shows good quality mosaic generation without any metadata information available for long sequence of aerial video. This geospatial summarization of crop field data can be helpful for phenotypic analysis, monitoring of field, checking movement and growth of plants, change in field dynamics, etc.

Keywords: Aerial imagery, mosaicking, registration, summarization, feature based registration.

1. INTRODUCTION

![Frame 0001](a) Frame 0001  (b) Frame 0167  (c) Frame 0240  (d) Frame 0261  (e) Frame 0308  (f) Frame 0321

Figure 1. Few random images from a sequence of 680 frames (each of 1000x1000 pixels) of maize field data.

Geospatial content summarization of dynamic video into mosaic is a popular technology in computer vision community. Application of mosaicking includes a variety of fields such as wide-field of view generation, video compression, medical imaging, surveillance of military field, monitoring of urban area or agricultural field, checking of restricted areas after catastrophic events, such as earthquakes, tsunamis, damage to nuclear power plants, etc.

In remote sensing specially for aerial imagery, mosaicking has been proved to be quite popular. One of the exhaustive attempts in aerial imagery exploitation is reported in. It seperates video into natural components corresponding to the scene and a framework for mosaicking, change detection and tracking moving object is demonstrated. Gonales at el. used robust feature detector, SIFT with information from image segmentation followed by a robust outlier removal method to register images. Tao at el. developed a graph based approach for scene stitching in large-scale aerial data. They exploited the property of frames being overlapped in time and space, and stitched the images by temporal grouping, retrieved spatial cross-group and finally generated a single mosaic from each cross-group. A novel temporal segmentation algorithm is proposed and one mini-mosaic is produced from each segment by feature based image registration in by Raphael at el.. Zhigang at el. proposed a geo-referenced video mosaicking for monitoring environment. In order to avoid
frame-to-frame error accumulation problem which results from frame-to-frame based registration, they exploit external geo-location data from GPS, laser profiler and INS system, and generate mosaic from airborne data. In, the error accumulation problem is overcome with the help of an external reference which can be a map image. A frame is matched both to the previous frame and external map. The transformation from these two different mapping is combined in order to get a robust and frame-to-frame error accumulation free mosaic. The work by Molina et al. report registering and mosaicking of long aerial video sequences which cycle over large area in few passes. In, no metadata information is used, instead the registration is solely image based. In such case, the mosaic suffers from error accumulation as it often results into large drifts with the increasing number of images in the sequence. In order to overcome this drifting problem, Molina et al. proposed to split the video into individual cycles of scene, and performed registration with drift correction. They exploited the property of cyclic coverage of data and calculated the drift correction for the first cycle only. Later, the drift correction result from the first cycle is used to correct the error in subsequent cycles. Hoang et al. proposed a feature based registration algorithm which detects and tracks moving objects in UAV videos. Once the frames are aligned by registration, forward/backward frame differencing and morphological operation is performed in order to estimate blob. Finally, by the use of spatial temporal blob detection, object tracking is accomplished. David et al. propose a system of mosaicking multicamera images where harris corner is chosen as feature detector due to its high repeatability under radiometric and geometric changes as well as availability of abundant corners points in urban scene images. In order to match key features in overlapping images, minimum spanning tree based coherence is implemented.

Maize field imagery data is complex due to its homogeneous appearance of scenery and absence of distinctive feature points (Figure 1). Given no metadata information, we need to devise a method which compensates for registration error. This leads us to a two-step registration procedure explained in Figure 4. At first, the video sequence is divided among m groups and one mini-mosaic is generated from each group. Later, m mini-mosaics are registered again and transformed to produce the final mosaic. This novelty reduces image-to-image registration error and generate good quality mosaic in absence of any metadata.

2. METHOD

![Figure 2. Block diagram showing major steps our in feature based mosaicking.](https://example.com)

![Figure 3. Few random images from a sequence of 271 frames (each of 720x480 pixels) of VIRAT data.](https://example.com)

Image registration is a fundamental problem in computer vision and many studies have been dedicated for different domains and applications. One of the basic approach is to find overlap between images with distinctive key points using feature based registration. Once the matching function is computed, next step is to apply geometric transformation (for example- rotation, translation, scaling, shearing, etc.) so that all images align with same global coordinate. Normally the first frame of a sequence is chosen as the global/reference coordinate. In our feature based registration algorithm, there are four main steps as shown in Figure 2.
Figure 4. Two step (step1: row C and step2: row D) mapping model of frames for registering crop field imagery. Row A shows the sequence of frames in crop field imagery captured by drone. The video sequence is divided into \( m \) groups: Group\(_1\), Group\(_2\), ..., Group\(_m\) (each color represents a different group, row B) to avoid error accumulation in the process. First frame of each group is called the reference (dark colored frame, row C) while other frames in the group are shown in corresponding light color (row C). A single arrow from each frame to the reference in a group denotes direct mapping (row C) in first step registration. We get a mini-mosaic or set of frames with single coordinate system (represented by same color: dark blue for Group\(_1\) in row D) from each group. Next, we apply reference-to-reference frame mapping as second step registration (row D). This results into a single mosaic with base/first frame coordinate system (\( \text{mini-mosaic}_1 \)) from row D remains in original coordinate system while \( \text{mini-mosaic}_2 \) and \( \text{mini-mosaic}_m \) from row D have gone through some transformation in row E).
2.1 Feature Detection and Matching

The first step in feature based registration procedure is to find out distinctive feature points in images. Features can be corners, edges, contours, textures, intersection points, etc. If two images have some overlapping region, then there will be some common features, and we can use those to find the correspondence between images. Each of the widey used feature descriptors (such as Harris corner, PCA-SIFT, SURF, ASIFT, NCC, Structure Tensor, etc.) have their own strengths and weaknesses. For instance, Harris corner is an excellent corner detector but might not perform similarly for relatively homogeneous natured crop field data. SIFT shows promising results when comes to account for rotation, scaling and affine transformation while it is computationally expensive and poor at handling illumination change. On the contrary, SURF is usually faster with comparable performance to SIFT while it is reported to be unsatisfactory in case of rotation and illumination change. Unfortunately, both fail to detect enough feature points to map one image with another in our maize dataset.

An upgraded version of SIFT and more robust descriptor is ASIFT which covers the four parameters from SIFT (scaling, orientation, illumination change, and affine distortion), and can also use the latitudal and longitudinal angles of the orientation of the camera axis. ASIFT also introduced the notion of a transition tilt, which effectively measures the amount of distortion from one view to another using the camera axis parameters. The main drawback of ASIFT is that it is few times computationally expensive than SIFT or SURF.

Due to the complexity and homogeneity nature of crop field data, most of the feature descriptors except ASIFT fail to match good keypoints. Consequently, we use ASIFT as feature descriptor for maize data in our experiment. In contrast, to show that this method works with simple and less expensive feature detector and descriptor for other aerial imagery, we ran the algorithm for a sequence of VIRAT data. VIRAT data is full in distinctive and easily identifiable rich features (see Figure 3) and works smoothly with other feature descriptors. For instance, we used a fused descriptor including Structure Tensor, Normalized Cross Correlation and SURF.

2.1.1 Transformation Model Estimation

Once the image features are accurately detected and matched across frames, next step is to transform the sensed or warped image to overlay with the reference image. Early registration techniques evolved through translation, similarity, affine transformation to most general projective or homography transformation. Given the scene is approximately planner, two images \( I(x, y, t) \) and \( I(u, v, t−k) \) can be related by homogeneous relationship:

\[
\begin{align*}
 u &= \frac{a_1 x + a_2 y + a_0}{c_1 x + c_2 y + 1} \\
 v &= \frac{b_1 x + b_2 y + b_0}{c_1 x + c_2 y + 1}
\end{align*}
\]

The first two parameters, \( x \) and \( y \), in \( I \) denote spatial coordinates and third one, \( t \), corresponds to time of image capturing. The homography matrix is shown in Eq. 3.

\[
\begin{bmatrix}
 u \\
 v \\
 w
\end{bmatrix} = \begin{bmatrix}
 a_1 & a_2 & a_0 \\
 b_1 & b_2 & b_0 \\
 c_1 & c_2 & 1
\end{bmatrix} \begin{bmatrix}
 x \\
 y \\
 1
\end{bmatrix}
\]

Here, \( w \) is a parameter in homogeneous coordinates. Given coordinates \((x, y)\) in the first image system, we need to calculate new coordinates \((u, v)\) in the second image system. There are 8 transformation parameters \((a_1, a_2, a_0, b_1, b_2, b_0, c_1 \text{ and } c_2)\) which are unknown and can be determined by four identical points in image \( I(x, y, k) \) and \( I(u, v, t−k) \).
Table 1. Unknown Parameters \(^ {23} \) of Homography Equation \(^3 \)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_1)</td>
<td>fixed scale factor in x direction with scale y unchanged</td>
</tr>
<tr>
<td>(a_2)</td>
<td>scale factor in x direction proportional to y distance from origin</td>
</tr>
<tr>
<td>(a_0)</td>
<td>origin translation in x direction</td>
</tr>
<tr>
<td>(b_1)</td>
<td>scale factor in y direction proportional to x distance from origin</td>
</tr>
<tr>
<td>(b_2)</td>
<td>fixed scale factor in y direction with scale x unchanged</td>
</tr>
<tr>
<td>(b_0)</td>
<td>origin translation in y direction</td>
</tr>
<tr>
<td>(c_1)</td>
<td>proportional scale factors x and y in function of x</td>
</tr>
<tr>
<td>(c_2)</td>
<td>proportional scale factors x and y in function of y</td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
  x_1 & y_2 & 1 & 0 & 0 & 0 & -x_1u_1 & -y_1u_1 \\
  x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2u_2 & -y_2u_2 \\
  x_3 & y_3 & 1 & 0 & 0 & 0 & -x_3u_3 & -y_3u_3 \\
  x_4 & y_4 & 1 & 0 & 0 & 0 & -x_4u_4 & -y_4u_4 \\
  0 & 0 & 0 & x_1 & y_2 & 1 & -x_1v_1 & -y_1v_1 \\
  0 & 0 & 0 & x_2 & y_2 & 1 & -x_2v_2 & -y_2v_2 \\
  0 & 0 & 0 & x_3 & y_3 & 1 & -x_3v_3 & -y_3v_3 \\
  0 & 0 & 0 & x_4 & y_4 & 1 & -x_4v_4 & -y_4v_4 \\
\end{bmatrix}
\begin{bmatrix}
  a_1 \\
  a_2 \\
  a_0 \\
  b_1 \\
  b_2 \\
  b_0 \\
  c_1 \\
  c_2 \\
\end{bmatrix} =
\begin{bmatrix}
  u_1 \\
  u_2 \\
  u_3 \\
  u_4 \\
  v_1 \\
  v_2 \\
  v_3 \\
  v_4 \\
\end{bmatrix}
\]

(4)

The unknowns in Table 1 can be calculated using the formulae in Equation 4 using Direct Linear Transformation \(^ {22} \) if points \((x_1, y_1), (x_2, y_2), (x_3, y_3)\) and \((x_4, y_4)\) from image, \(I(x, y, i)\) and \((u_1, v_1), (u_2, v_2), (u_3, v_3)\) and \((u_4, v_4)\) from image, \(I(u, v, t - k)\) are known. \(^ {23} \) Finally random sample consensus (RANSAC) \(^ {24} \) is used to remove outliers and establish a robust matching between images.

2.1.2 Blending

Once the mapping function is generated, the final step is to transform the sensed or warped images. There are two ways to transform image pixels: forward and backward mapping. \(^2 \) The straightforward approach is forward transformation which directly maps sensed image pixels using mapping functions. On contrary, sensed image pixels are redefined and interpolated using coordinates of reference system with inverse of mapping functions. The latter one is used here as earlier one creates hole in output images. Instead of using any traditional blending methods, we overlay sensed or warped image on top of reference image. We name it pixel replacement as each pixel in reference image is replaced by pixels in sensed or warped image.

2.1.3 Global Transformation Estimation

For an ordered set of images to register, the traditional approaches with \(n\) frames follow Equation 3 where \(H_{i\leftarrow j}\) denotes the homography to transform image, \(I(x, y, i)\) to image, \(I(x, y, j)\). But we want to avoid it as this results into error accumulation due to chain multiplication of \(n\) homography matrices.

\[
H_{1\leftarrow n} = H_{1\leftarrow (n-1)} * H_{(n-1)\leftarrow n} = H_{1\leftarrow 1} * H_{1\leftarrow 2} * H_{2\leftarrow 3} * \ldots * H_{(n-2)\leftarrow (n-1)} * H_{(n-1)\leftarrow n}
\]

(5)

To overcome the drifting problem, we come up with a solution which avoids multiplying subsequent homography matrices. The mapping order of frames is explained in detail in Figure 4 where frames are first divided into \(m\) groups (each group having 40 or less images). The first frame of each group is called the reference with all frames directly mapped to it. Thus, each group results into a single mini-mosaic with reference coordinate system. In order to generate one global mosaic aligned with the first image (known as base), a next step transformation is calculated: \(HG_{x\leftarrow y}\) denote the homography to align frames from \(Group_y\) to the frames from \(Group_x\) with \(Ref_y\) and \(Ref_x\) as the reference frame respectively. Finally, Equation 6 is the general formulae for aligning any group (\(Group_m\)) of frames to base coordinate system. This cascading transformation is also shown in row \(D\) in Figure 4.
\[ HG_{1\rightarrow m} = HG_{1\rightarrow (m-1)} \cdot HG_{(m-1)\rightarrow m} \]
\[ = HG_{1\rightarrow 1} \cdot HG_{1\rightarrow 2} \cdot HG_{2\rightarrow 3} \cdot \ldots \cdot HG_{(m-1)\rightarrow m} \] (6)

3. RESULT

A maize field was imaged long after plant senescence using a DJI Phantom 4. No alterations to the vehicle or its camera were made. A 1.0 megapixel video of dimensions 1000 \times 1000 pixels was recorded at 15 frames/sec in 24-bit color depth. Each frame is downsampled to one-fourth of its size before feeding it to the mosaicking pipeline. Figure 5 (a) shows the mosaic generated from 680 images. The mosaic looks clear with easily identifiable maize plant rows (center of the mosaic), other vegetation field surrounding plants, truck (red colored at top-left corner), etc. As ASIFT \(^{16}\) is used as feature detector and descriptor, mosaic generation is time consuming and each frame processing takes several minutes. On the other hand, mosaic comprised of VIRAT dataset is shown in 5 (b) where per frame processing consumes few seconds only due to its effectiveness with less expensive descriptor such as Structure Tensor \(^{17}\), NCC \(^{17}\) and SURF \(^{15}\).

4. CONCLUSION

A robust algorithm for mosaicking aerial field imagery without any metadata is presented in this study. The novelty of this algorithm is finding the temporal consistency within the video sequence and utilizing a grouping technique to handle multiple coordinate systems for image registration. Though good quality mosaic is generated, seem are visible as we did not use any standard blending method. In addition, the mosaic also suffers from radial distortion. In future, we plan to integrate an advanced blending technique (such as poisson editing) in order to get a smooth, seamless mosaic.
5. ACKNOWLEDGMENT

We gratefully acknowledge the support of the U.S. Air Force Research Laboratory grant FA8750-14-2-0072 and the Missouri Corn Growers’ Association. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Government or any agency thereof.

REFERENCES


