Homography-Based Orientation Estimation for Capsule Endoscope Tracking

Evaggelos Spyrou, and Dimitris K. Iakovidis
Department of Informatics and Computer Technology
Technological Educational Institute of Lamia
Lamia, Greece
vspyrou@teilam.gr, iakovidis@ctr.teilam.gr

Abstract—Wireless capsule endoscopy (WCE) is a developing imaging technology for screening the gastrointestinal system. Tracking of the swallowable capsule endoscope within the human body is typically performed by external sensor arrays enabling approximate localization and detection of its orientation. In this paper we propose a method for capsule orientation estimation based only on image features, without the requirement of any external equipment. The proposed method involves two steps: a) salient point detection and extraction of image features from these points, and b) detection of feature correspondences between consecutive frames and homography estimation. The capsule orientation is estimated by decomposition of the homography matrix into rotation and translation components. The results obtained by the application of this method on WCE video frames indicate its effectiveness and improved performance over the state of the art.

Keywords—wireless capsule endoscopy; orientation estimation; feature correspondence; homography

I. INTRODUCTION

The idea of a camera pill for imaging of the entire gastrointestinal system was conceived in the beginnings of the eighties, with a first implementation appearing in the mid-nineties [1]. Since then, this technology is developing and currently it is considered as a standard imaging technique for screening the gastrointestinal tract, especially the small intestine mucosa, which is not easily reached by the usual upper endoscopy and colonoscopy.

Wireless capsule endoscopy (WCE) is performed by a swallowable capsule with the size of a large vitamin that includes a miniature color video camera, a light, a battery and a video stream transmitter. The capsule is propelled by peristalsis through the gut and reaches the right colon typically in 5-8 hours. During its journey in the human body it captures thousands of video frames and transmits them wirelessly to a recorder. The tracking of the capsule within the body is typically performed by radio-frequency (RF) sensor arrays, mounted externally to the human body, receiving the signals transmitted by the capsule endoscope. The average position error reported for this technique is 37.7 mm, with a maximum error reaching 114 mm [2]. More recent tracking techniques are based on magnetic sensor arrays, which give a significantly lower average position error (3.3 mm) and an average orientation error of 3 degrees. However, the overall accuracy of the magnetic tracking depends highly on the number of external sensors used [2], in some cases rotation information is incomplete [3], and as the RF tracking approach it requires external equipment that can be discomforting for the patient.

In this paper we propose a method for the estimation of the orientation of the capsule based only on information extracted from WCE video frames. The proposed method involves image feature extraction, detection of feature correspondences between consecutive video frames, homography estimation, and homography decomposition into rotation and translation components. The feature correspondence method considered has been previously applied in the context of content-based image retrieval [4]. To the best of our knowledge its application for the problem under investigation has not been previously investigated. Such an approach does not require any external equipment for capsule orientation estimation. Furthermore, it could be used as a supplementary approach to improve the accuracy of the sensor-based localization techniques. A relevant method based on Lucas-Kanade optical flow computation (KLT) has been proposed in [3]; however, its experimental evaluation revealed an unacceptably large error for rotation angles greater than 30 degrees.

The rest of this paper consists of four sections. Section II describes the feature extraction algorithm considered, Section III describes the algorithm for the detection of feature correspondences and the orientation estimation, and the results of this study are presented in Section IV. Section V summarizes the conclusions that can be derived from this study.

II. IMAGE FEATURE EXTRACTION

As it has been discussed thoroughly in the bibliography, extraction of visual features, either global (from the whole image) or local (from image patches), presents serious limitations in certain problems. Global features may change notably due to e.g. small changes of viewpoint, zooming, changes of illumination, contrast etc. While the same stands for local features, an additional problem lies on the patch selection algorithms. Typically patches are extracted using segmentation or clustering techniques which appear to be very sensitive to image distortions such as the aforementioned. Thus, patches...
extracted from visually (to the human observer) images may be significantly different.

In our work, we adopt an approach based on a set of points that is extracted locally from images. More specifically, from each video frame we extract a set of interest (salient) points. Then, visual features are extracted from their surrounding area. This set of points should be robust to affine image transformations, since such are typical to video sequences. Moreover, their robustness to image variations ensures that the transformation between two images may always be calculated under certain visual changes, as in the case of the problem under investigation. An example of the interest points and respective SURF features extracted from two consecutive WCE video frames are illustrated in Fig. 1.

III. DETECTION OF FEATURE CORRESPONDENCES

After selecting a set of invariable interest points in order to capture the visual properties of a given video frame, the next step is to estimate a transform between two consecutive frames. We should expect that consecutive frames contain similar visual features, as the capsule’s camera viewpoint changes. The latter causes the appearance of similar features in a totally different spatial layout.

In order to detect the similar features and find the correspondences between video frames we adopted the well-known RANSAC (RAandom SAmple Consensus) algorithm [9]. Using RANSAC, we are able to determine the geometric transformation between two consecutive video frames given a set of tentative point-to-point correspondences. We exploit the basic advantage of RANSAC, i.e. its ability to determine a model in presence of many false such correspondences that are also called outliers. More specifically, RANSAC is used to estimate the transformation that maximizes the number of inliers that is the set of correspondences that support the model. It is obvious that RANSAC relies a lot in the correspondences among points, which will be provided at its initial step. In general, these correspondences are not available, thus need to be calculated each time.

Figure 2(a) illustrates an example of the correspondences...
detected among the initially extracted interest points using SURF features for the two consecutive WCE video frames of Fig. 1. One may observe the large number of tentative correspondences. The majority of these correspondences may be easily characterized as outliers. However, the selection of a valid model for the geometric transformation is not obvious to the human observer.

A homography [10] is a perspective transform that maps any given point \( x_i \) of a given image to a corresponding point \( x'_i \) of another. Given the set of correspondences of points of interest between two consecutive frames, i.e. the pairs \( x_i \leftrightarrow x'_i \), we can define the homography matrix \( H \) as:

\[
\begin{pmatrix}
    x'_1 \\
    x'_2 \\
    \vdots \\
    x'_n
\end{pmatrix} =
\begin{pmatrix}
    H_{11} & H_{12} & \cdots & H_{1n} \\
    H_{21} & H_{22} & \cdots & H_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    H_{n1} & H_{n2} & \cdots & H_{nn}
\end{pmatrix}
\begin{pmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_n
\end{pmatrix}
\tag{1}
\]

Then, in order to calculate this homography, we apply RANSAC. The mathematical model estimated by RANSAC in this case is a homography matrix, whereas the set of correspondences that do not fit the estimated homography are characterized as outliers.

The estimation of a homography using RANSAC has been applied very effectively in tasks such as stereoscopic camera calibration, a case where the images captured by two cameras differ only by means of a perspective transform. This simple idea is extended herein and instead of the two images taken by a stereoscopic camera, we consider the case of the camera moving slowly and capturing the same "scene" by different viewpoint. We should expect that in this case the homography will result to a rather small number of false correspondences since the variation of the viewpoint is not very high. However, many false correspondences may be introduced due to e.g. image noise (MPEG-1 block artifacts) significantly higher. This justifies our choice, as we adopt RANSAC, due to its ability to correctly estimate the homography even in the presence of a large number of outliers (false correspondences) present.

RANSAC begins with a random sample which should contain a number of points equal to the number of the model variables. In the case of homography, the model requires at least 4 initial points. Given a set of point correspondences between two consecutive video frames, the RANSAC algorithm that estimates the homography between them is summarized as follows:

1. Initially, 4 points are randomly selected. Under the assumption that all are inliers, they define a homography.
2. Using the model estimated under the assumption of the previous step, the number of inliers i.e. correspondences that fit the this model, is calculated. Their distance to the points estimated by the homography model should be smaller than a predefined threshold.
3. Random quadruplets of points are sampled. After every such iteration, both the maximum number of inliers that satisfy the model up to that point is kept, along with the corresponding model.
4. This process is repeated until a predefined number of iterations is reached, or the probability that new inliers may be determined using another model falls beneath a certain threshold.
5. When the aforementioned relation is satisfied, the homography model is redefined, using all inliers selected from the previous steps.

Figure 2(b) illustrates an example of the inliers selected using RANSAC from the tentative correspondences of the two consecutive video frames of Fig. 1, depicted in Fig. 2(1).

Once the homography is calculated it is decomposed into rotation matrix \( R \) and a translation matrix \( T \) as in [11], under the assumption that the focal length of our camera is \( f=1 \). Thus, we may write that

\[
\begin{pmatrix}
    H_{11} & H_{12} & 0 \\
    H_{21} & H_{22} & 0 \\
    0 & 0 & 1
\end{pmatrix}
\tag{2}
\]

where \( H_{ij} \) are obtained from \( H \).

IV. EXPERIMENTAL RESULTS

The proposed method was evaluated on WCE videos acquired with Given Imaging PillCam capsules [12]. We adopted the same evaluation approach as in [3], where the correspondences between the original and rotated video frames are extracted. A total of eight different rotation angles were tested from 5 to 40 degrees with a rotation step of 5 degrees.
Figure 3. Inliers estimated using RANSAC, for a video frame and (a) its 10-degree rotated instance, (b) its 30-degree rotated instance.

Figure 4. Comparison between the proposed method and the method proposed in [3] in terms of error observed in rotation angle estimation.

Table I. Estimated Orientation Error per Rotation Angle

<table>
<thead>
<tr>
<th>Actual Rotation Angle</th>
<th>Estimated Rotation Angle</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4.9192</td>
<td>-0.0808</td>
</tr>
<tr>
<td>10</td>
<td>10.0936</td>
<td>+0.0936</td>
</tr>
<tr>
<td>15</td>
<td>15.1586</td>
<td>+0.1586</td>
</tr>
<tr>
<td>20</td>
<td>20.2015</td>
<td>+0.2015</td>
</tr>
<tr>
<td>25</td>
<td>25.2087</td>
<td>+0.2087</td>
</tr>
<tr>
<td>30</td>
<td>30.1456</td>
<td>+0.1456</td>
</tr>
<tr>
<td>35</td>
<td>33.0222</td>
<td>-1.0778</td>
</tr>
<tr>
<td>40</td>
<td>51.5815</td>
<td>+11.5815</td>
</tr>
</tbody>
</table>

The results obtained are apposed in Table I. This table shows that the error between the actual and the estimated rotation angle is rather low up to 35 degrees and increases for angles over 40 degrees. A further increase in the rotation angle causes larger errors.

Representative examples of the correspondences detected between rotated video frames by the proposed method is illustrated in Fig. 3. More specifically, Fig. 3(a) illustrates the correspondences detected for a small rotation angle, i.e. 10 degrees, and Fig. 3(b) illustrates the correspondences detected for a larger rotation angle, i.e. 30 degrees. In both cases the estimated rotation using our algorithm is accurate, as indicated in Table I.

Comparing the results with those presented in [3] (Fig. 4) it is evident that the proposed method outperforms the KLT-based method in the sense that it produces lower error rates for angles larger than 5 degrees. More importantly, that method, results in a very large error for angles greater than 30 degrees, whereas the error obtained by our method becomes large for angles greater than 40 degrees.

V. CONCLUSIONS

In this paper we proposed a novel imaging technique for the estimation of the orientation of capsule endoscopes. Our technique is based on the estimation of the homography between two consecutive video frames, since the homography matrix contains the necessary information for image rotation.

Our experimental results indicate that SURF features are appropriate for WCE sequences and tentative correspondences among them can be used efficiently along with RANSAC, in order to estimate a homography. It has been shown that the SURF-based method outperforms the state of the art method proposed in [3].

The proposed technique can used either autonomously or as a supplement to sensor-based approaches to capsule tracking. Within our future goals is to extend this method for full, image-based localization of the capsule in the gastrointestinal system.

REFERENCES


