**INTRODUCTION**

The objective of this research is to incorporate multiple sensors combining infrared (IR) camera, color (or RGB) camera and depth sensor to perceive the surgical environment. Features extracted from each modality can contribute to the cognition of complex surgical environment or procedures. Additionally, their combination can provide higher robustness and accuracy beyond what is obtained from one single vision modality. As a preliminary study, we implemented a fused tracking system for surgical instrument based on IR, RGB and depth sensor technique. Experiments showed that the computational speed and tracking accuracy of the hybrid tracking system can be improved.

**MATERIALS AND METHODS**

This paper considers the problem of improving the surgical instrument tracking accuracy by multi-sensor fusion technique in computer vision. We proposed a hierarchical fusion algorithm for integrating the tracking results from depth sensor, IR camera pair and RGB camera pair. Fig. 1 summarized the algorithm involved in this paper. It can be divided into the “low-level” and the “high-level” fusion.

The low-level fusion is to improve the speed and robustness of marker feature extraction before triangulating the tool tip position in IR and RGB camera pair. Specifically, the rough position of tracking tools from depth sensor is fused into the binocular IR tracker and the binocular RGB tracker. The depth data of the tool can be used as a priori for cross marker detection. Firstly, the working area of the tracking tool is supposed to be limited in a fixed volume \(v(x, y, z)\). The IR camera is modeled as a pin-hole camera,

\[
\begin{align*}
(u_{IR}, v_{IR}, 1)^T &= s_{IR} K_{IR} (X_{IR}, Y_{IR}, Z_{IR})^T \\
\text{where } K_{IR} \text{ the intrinsic parameter of IR camera in Kinect 1; } (X_{IR}, Y_{IR}, Z_{IR}) \text{ and } (u_{IR}, v_{IR}) \text{ the 3D and 2D coordinates of the same point in work space and its projection in image plane; } s_{IR} \text{ the ratio factor. With the readout } Z_{IR} \text{ at image point } (u_{IR}, v_{IR}) \text{ from depth sensor, } (u_{IR}, v_{IR}, Z_{IR})^T \text{ is used to solve out the other unknown variables } s_{IR}, X_{IR} \text{ and } Y_{IR}. \text{ Thereafter, we can establish the 2D pixels ~ 3D coordinates pairs } \sim (u_{IR}, v_{IR}) \sim (X_{IR}, Y_{IR}, Z_{IR}) \text{ in IR camera images. Comparing the 3D coordinates of each pixel with the required working volume } v(x, y, z), \text{ most of the } 2D \text{ pixels in the IR image will be filtered out to get a refined searching area } \alpha(u_{IR}, v_{IR}). \text{ Furthermore, the } \alpha(u_{IR}, v_{IR}) \text{ is transformed to a refined search area } \alpha(u_{IRGB}, v_{IRGB}) \text{ in RGB image by the homography transformation between IR image plane and RGB image plane.}
\end{align*}
\]

\[
\alpha(u_{IRGB}, v_{IRGB}) = H_{IRGB} \alpha(u_{IR}, v_{IR})
\]

Where \(\alpha(u_{IRGB}, v_{IRGB})\) and \(\alpha(u_{IR}, v_{IR})\) the homogeneous coordinates of \(\alpha(u_{IRGB}, v_{IRGB})\) and \(\alpha(u_{IR}, v_{IR})\); The 3x3 homography matrix \(H_{IRGB}\) correlates the pose of IR camera pair with RGB camera pair and can be got from so called “8-point algorithm” [1].

Regarding the high-level fusion, the objective is to get a high accurate tracking result by fusing two measurements. We employ the covariance intersection (CI) algorithm [2] to estimate a new tracking result with less covariance. Given the measurement matrix \(P_{IR}\) and \(P_{RGB}\) which can be calculated offline in IR and RGB camera pairs, a probabilistic consistent estimate for the position can be described as,

\[
\begin{align*}
X_f &= P_f (\omega_{IR} P_{IR}^{-1} X_{IR} + \omega_{RGB} P_{RGB}^{-1} X_{RGB}) \notag \\
P_f &= (\omega_{IR} P_{IR}^{-1} + \omega_{RGB} P_{RGB}^{-1})^{-1}
\end{align*}
\]

where \(X_{IR}\) and \(X_{RGB}\) the measurement from IR and RGB camera pairs; \(P_f\) the covariance matrix of the estimation; \(\omega_{IR-P}\) and \(\omega_{RGB-P}\) the coefficients that can be obtained by,

\[
\begin{align*}
\omega_{IR} &= \frac{tr(P_{RGB})}{tr(P_{IR}) + tr(P_{RGB})} \notag \\
\omega_{RGB} &= \frac{tr(P_{IR})}{tr(P_{IR}) + tr(P_{RGB})}
\end{align*}
\]

where the symbol \(tr(\cdot)\) means the trace of a matrix.

To demonstrate the proposed algorithm, we designed a hybrid marker-based tracking tool that incorporates the cross-based feature in visible modality and retro-reflective marker based feature in infra-red modality to get a fused tracking of the customized tool tip. As shown in Fig. 2, \(P_m\) and \(P_r\) are the translation vector that represents the position of the tool tip relative to the coordinate frame of retro-reflective marker and cross-corner marker respectively.
Kinect is a motion sensing input device integrated with IR camera, RGB camera and depth sensor. We employ two Kinects to build the experimental setup. Four IR LED illuminators were employed to generate enough infrared light when tracking the target in IR camera mode. Depth data can be obtained directly from the Kinect SDK. However, interference between multiple structured light-based depth sensors should be considered. Berger et al. [3] and Schröder et al. [4] utilized the time-multiplexing based hardware to overcome the interference. Maimone et al. [5] applied a small vibration to reduce interference in frequency domain. In our project, we only use one IR laser emitter by covering a black sheet to block the other one.

RESULTS

To validate the proposed algorithm and developed dual-Kinect tracking system, we did a proof of concept experiment. We moved the customized tool in the front of the overlapped field of view. Harris corner detector to extract the cross feature in RGB images and region growing segmentation algorithm [6] to detect the retro-reflective marker in IR images were employed. Table I shows the processing speed per frame with and without depth-data acceleration in IR and RGB images, which indicated the speed of feature extraction in both IR and RGB images were improve greatly.

The box plots in Fig. 4 shows the sum of position errors in each direction (error = |Δx| + |Δy| + |Δz|) of three kind of tracking methods including the binocular IR camera pair, the binocular RGB camera pair and the CI-based fusion algorithm. The box plots indicate that the CI-based fusion approaches obviously tend to be better than the separate IR tracker or RGB tracker. The mean error and deviation of the fusion algorithm are all improved.

![Fig. 2 Hybrid marker-based tracking tool](image)

![Fig. 4 Box plot for sum of position errors in separate tracking method](image)

DISCUSSION

Due to the firmware limitation of Kinect, The RGB and IR data are only available as different settings for the same isochronous stream; therefore one can only obtain one or the other at a given time, which results in a relatively low refresh rate for position measurement. Even worse, the maximum frame rate of RGB camera with 1280 x 960 pixel resolution is at only 12 frames per second. Hence, the moving speed of the tracking target must keep lower than a given velocity, or the blurring RGB image may fail its tracking records. In addition, the issue of synchronization with multiple sensors in multiple Kinect will be further investigated.

TABLE I. SPEED OF IR AND RGB CAMERA PAIRS PER FRAME

<table>
<thead>
<tr>
<th>IR CAMERA PAIRS</th>
<th>RGB CAMERA PAIRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>without depth data</td>
<td>with depth data</td>
</tr>
<tr>
<td>0.0812s</td>
<td>0.0123s</td>
</tr>
</tbody>
</table>

REFERENCES