ON REGISTRATION OF REGIONS OF INTEREST (ROI) IN VIDEO SEQUENCES

Hua-mei Chen
Department of Computer Science and Engineering
The University of Texas at Arlington, Arlington, TX 76019

Pramod K. Varshney
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY 13244

Mohamed-Adel Slamani
ITT Industries, 2560 Huntington Avenue, Alexandria, VA 22303

E-mail: \textsuperscript{1}hchen@cse.uta.edu, \textsuperscript{2}varshney@syr.edu, \textsuperscript{3}adel.slamani@itt.com

ABSTRACT

This paper addresses the problem of registering regions of interest in two video sequences. Potential applications include blob fusion and target tracking in blurry sequences. It is assumed that the moving target is tracked successfully in one of the two sequences and is represented by a bounding box in each frame of the first sequence. The goal is to find the corresponding bounding box for each frame of the second video sequence. The registration algorithm developed in this paper is based on mutual information. To facilitate the registration process, the two cameras are assumed to be calibrated such that the geometrical transformation required to register the corresponding bounding boxes is a 2D rigid body transformation without rotation. Visual and IR video sequences are used to test the proposed approach.

1. INTRODUCTION

Multisensor surveillance is an active research topic as described in [1]. The major objective of deploying multiple sensors is to (1) extend the area being monitored (2) track individual objects seamlessly over the extended area, and (3) combine and fuse video sequences from two or more types of sensors in order to obtain a visualization that is more suitable for the human operator. One example of the latter is given in Fig. 1. Fig. 1(a) shows a visual image and Fig. 1(b) shows a corresponding MMW image of three people one of which (the right most) is carrying a gun under his coat. Observing either one of these two images solely, it is hard to recognize the weapon. However, the fused image, shown in Fig. 1(c), gives a good indication to the human observer and more confidence in declaring that the bright portion in the right half of the MMW image is a gun.

In order to develop a surveillance system that integrates multiple types of video sequences, the corresponding regions of interest (ROIs) need to be registered accurately. We call this process “bounding box registration”, since the ROI (i.e., the tracked moving target) is often represented by a bounding box in each video frame. Notice that here we are interested in the fusion of the corresponding ROIs rather than the entire frame because, in general, it is not feasible to register entire corresponding frames precisely through simple geometrical transformations due to the cameras not being identically located and having different scene depths. In other words, when the backgrounds are registered, the foreground may not be aligned and vice versa. Fig. 2 illustrates this situation. Fig. 2 (a) and (b) show the first frame of each of the IR and visual sequences, respectively. Fig. 2 (c) is the transformed (b) image with registered foreground (person) using a 2D rigid body transformation without rotation. Fig. 2 (d) is another transformed (b) image but with registered background. Fig. 2 (e) and (f) are the result of superimposing images in (a) and (c) and in (a) and (d), respectively. Clearly, when the background is aligned, the foreground is misregistered and vice versa.

In addition to the fusion of the corresponding ROIs, “bounding box registration” has another potential application namely target tracking in a blurry video sequence. One such modality is millimeter wave (MMW) video sequence. Image frames from a MMW sensor are blurry by nature as seen in Fig. 3, so it is extremely hard to develop a robust tracking algorithm using information from MMW sensors alone. However, millimeter wave
sensors have been shown the capability to penetrate one’s clothing and therefore have great application value in concealed weapons detection (CWD) [2,3]. As long as our bounding box registration algorithm is modality-independent, target tracking in MMW video sequence is possible.

Our bounding box registration algorithm is based on the maximization of mutual information (MI). The advantage of this technique is that it is modality-independent, as long as MI is a valid similarity measure for the two imaging sensors involved. This paper is organized as follows. In section 2, a brief introduction of MI based registration is provided. Simplex method is used to maximize the MI measure in this paper and is described in Section 3. To facilitate our approach, some assumptions are made. They are identified in Section 4. The proposed bounding box registration algorithm is presented in Section 5. Video data used in this work is described in Section 6, along with experimental results. Finally, conclusions and future work are given in Section 7.

2. MI BASED IMAGE REGISTRATION

Mutual information has its roots in information theory [4]. The mutual information (MI) of two random variables A and B is defined by

$$I(A,B) = \sum_{a,b} P_{a,b}(a,b) \log \frac{P_{a,b}(a,b)}{P_a(a) \cdot P_b(b)}$$

(1)

where \(P_a(a)\) and \(P_b(b)\) are the marginal probability mass functions, and \(P_{a,b}(a,b)\) is the joint probability mass function. MI measures the degree of dependence of A and B by measuring the distance between the joint distribution \(P_{a,b}(a,b)\) and the distribution associated with the case of complete independence \(P_a(a) \cdot P_b(b)\), by means of the relative entropy or the Kullback-Leibler measure [4]. MI is related to entropies by the equations

$$I(A,B) = H(A) + H(B) - H(A,B)$$

(2)

$$H(A) = -\sum_a P_a(a) \cdot \log P_a(a)$$

(5)

$$H(A | B) = -\sum_{a,b} P_{a,b}(a,b) \cdot \log P_{a,b}(a | b)$$

(10)

Our bounding box registration algorithm is based on the maximization of mutual information (MI). The advantage of this technique is that it is modality-independent, as long as MI is a valid similarity measure for the two imaging sensors involved. This paper is organized as follows. In section 2, a brief introduction of MI based registration is provided. Simplex method is used to maximize the MI measure in this paper and is described in Section 3. To facilitate our approach, some assumptions are made. They are identified in Section 4. The proposed bounding box registration algorithm is presented in Section 5. Video data used in this work is described in Section 6, along with experimental results. Finally, conclusions and future work are given in Section 7.

2. MI BASED IMAGE REGISTRATION

Mutual information has its roots in information theory [4]. The mutual information (MI) of two random variables A and B is defined by

$$I(A,B) = \sum_{a,b} P_{a,b}(a,b) \log \frac{P_{a,b}(a,b)}{P_a(a) \cdot P_b(b)}$$

(1)

where \(P_a(a)\) and \(P_b(b)\) are the marginal probability mass functions, and \(P_{a,b}(a,b)\) is the joint probability mass function. MI measures the degree of dependence of A and B by measuring the distance between the joint distribution \(P_{a,b}(a,b)\) and the distribution associated with the case of complete independence \(P_a(a) \cdot P_b(b)\), by means of the relative entropy or the Kullback-Leibler measure [4]. MI is related to entropies by the equations

$$I(A,B) = H(A) + H(B) - H(A,B)$$

(2)

$$H(A) = -\sum_a P_a(a) \cdot \log P_a(a)$$

(5)

$$H(A | B) = -\sum_{a,b} P_{a,b}(a,b) \cdot \log P_{a,b}(a | b)$$

(10)

To employ mutual information as a similarity measure, we need to utilize the concept of the two-dimensional histogram of an image pair, the joint histogram. The joint histogram \(h\) of an image pair can be defined as a function of two variables, A, the gray level intensity in the first image and B, the gray level intensity in the second image. Its value at the coordinate \((A,B)\) is the number of corresponding pairs having gray level A in the first image and gray level B in the second image. The joint probability mass function used in the calculation of MI of an image pair can then be obtained by normalizing the joint histogram of the image pair as

$$P_{a,b}(a,b) = \frac{h(a,b)}{\sum_{a,b} h(a,b)}$$

(8)

From the joint probability mass function, we may obtain the two marginal probability mass functions directly as

$$P_A(a) = \sum_b P_{a,b}(a,b)$$

(9)

$$P_B(b) = \sum_a P_{a,b}(a,b)$$

(10)

The MI registration criterion states that the image pair is geometrically aligned through a geometric transformation \(T\) when \(I(A(x),B(T(x)))\) is maximal. Notice that the marginal entropies in Equation (3) change with transformation \(T\) because the image overlap changes. The strength of the mutual information similarity measure lies in the fact that no assumptions are made regarding the nature of the relation between the image intensities in both modalities, except that such a relationship exists. Therefore, the MI criterion is very general and has been used in many different image registration problems [5-12], including registration of IR and MMW images [6].
3. SIMPLEX SEARCH METHOD

In this work, simplex method is used to maximize the mutual information measure. The simplex method [13] is a method for minimizing a function of n variables. A simplex is a geometrical figure that is formed by N+1 points in an N dimensional space. An example of the simplex in three dimensions is given in Figure 4. The first step of this method uses a set of points to form an initial simplex. At each subsequent step in the algorithm, a new simplex is created according to three basic operations: reflection, contraction, and expansion. The definitions of these three operations are:

\[ P_{\text{ref}} = \overline{P} + \alpha(\overline{P} - P_h), \alpha > 0 \quad (11) \]
\[ P_{\text{con}} = \overline{P} - \beta(\overline{P} - P_h), 1 > \beta > 0 \quad (12) \]
\[ P_{\text{exp}} = \overline{P} + \gamma(\overline{P} - P_h), \gamma > \alpha \quad (13) \]

The constants \( \alpha, \beta, \) and \( \gamma \) are referred to as the reflection coefficient, contraction coefficient, and expansion coefficient, respectively. \( P_h \) is the vertex of interest and \( \overline{P} \) is the centroid of the vertices except for the vertex \( P_h \). With these three basic operations, the simplex method can successfully reach the local optimum without resorting to the derivatives.

Let us consider the minimization of a function of \( n \) variables. Let \( P_0, P_1, \ldots, P_n \) be the \( (n + 1) \) points in an \( n \)-dimensional space defining the initial simplex. Let \( y_i \) denote the function value at \( P_i \), and we define \( y_h = \max(y_i) \) and \( y_l = \min(y_i) \). So that \( P_h \) is the point resulting in the largest function value and \( P_l \) is the point resulting in the smallest function value within the \( (n + 1) \) points. Further \( \overline{P} \) is defined as the centroid of the points with \( i \neq h \). At each step in the process, \( P_h \) is replaced by a new point. If \( y_{\text{ref}} \) is between \( y_h \) and \( y_l \) then \( P_h \) is replaced by \( P_{\text{ref}} \). If \( y_{\text{ref}} < y_l \), that is reflection has produced a new minimum, then we check whether \( y_{\text{exp}} < y_l \). If \( y_{\text{exp}} < y_l \), we replace \( P_h \) by \( P_{\text{exp}} \), otherwise, we replace \( P_h \) by \( P_{\text{ref}} \). If \( y_{\text{ref}} > y_l \) for all \( i \neq h \), i.e. replacing \( P_h \) by \( P_{\text{ref}} \) leaves \( P_{\text{ref}} \) the maximum, then \( P_h \) is replaced by \( P_{\text{con}} \). Finally, if \( y_{\text{con}} > y_h \) then we replace all \( P_i \)’s by \( (P_i + \bar{P})/2 \) and restart the process. The whole process is terminated when \( (y_h - y_l) \) is less than a pre-determined threshold or the final simplex is smaller than a pre-determined size.

Let \( \gamma = \max(\gamma_1, \gamma_2, \ldots, \gamma_n) \).

![Fig.4 A simplex in 3D space](image)

4. ASSUMPTIONS

Some assumptions are made in this work. First, it is assumed that target tracking (i.e., human in our case) using the first video sequence can be done successfully and the tracked target is restricted to a sequence of bounding boxes. In other words, we are able to see the person being tracked in the bounding box in each frame of the first video sequence. Our goal is to find the corresponding bounding boxes in the second video (visual sequence in our case) that bound the person being monitored.

The second assumption we made was that the transformation required to perform bounding box registration is rigid. This assumption requires the lines of sight of both cameras to be parallel and close to each other. If, furthermore, the two cameras are close to each other, then the scale factor is approximately a constant throughout the whole sequence of image frames because the distances from the target to the two cameras are always approximately equal. To facilitate the registration process, we further assume that no rotation angle is involved in the rigid body transformation. In other words, we assume that rigid body transformations involving only vertical and horizontal displacements are sufficient for our purpose. This can be done by carefully calibrating the two cameras in advance.

The third assumption we made was that a coarse registration of the person in both sequences could be done by some prior knowledge and careful calibration. For example, if the lines of sight of both sensors are aimed at the same remote point and the field of view of each sensor and sampling rates are known, then the transformation that coarsely registers the two sequences will be \([\text{scale factor}, \text{vertical displacement}, \text{horizontal displacement}] = [s_0, 0, 0] \), where the scale \( s_0 \) depends on the individual field of view and the sampling rates. In our experiment, these parameters are obtained through manual registration because we were not able to calibrate the cameras involved and access information on the equipment used to acquire the video sequences.

5. APPROACH

Based on the above assumptions, we developed the following algorithm to find the corresponding bounding boxes in the second video sequence.

Step 1: Obtain the transformation that registers the two video sequences coarsely. Denote it as \( P_0 = [s_0, \Delta x_0, \Delta y_0] \).

Step 2: For each pair of corresponding frames, reduce the number of intensity levels (usually 256) of both frames into \( N_i \) levels by linear binning, where \( i \) is the frame number and \( N_i \) is an integer that is determined by the following condition:

\[ \frac{n_i}{N_i^2} \approx 8 \quad (14) \]

where \( n_i \) is the number of pixels within the bounding box of the \( i \)th frame of the first video sequence. The reason for performing this intensity binning operation is that, to reasonably estimate the mutual information of two
images, the number of gray levels to represent each image must be related to the number of pixels within the overlap region. It is suggested in [14] that 8 is a reasonable ratio of the number of pixels within the overlap region and the number of joint histogram entries and is adopted in our work. In our implementation, this binning operation is carried out not on the entire image but on the regions restricted by the bounding boxes for the first sequence and the corresponding region for the second sequence.

Step 3: Register the corresponding frames by maximizing the mutual information. Each time the registration task is performed, the frame from the first sequence is served as the floating image and the frame from the second sequence is used as the reference image. Only those pixels that are within the bounding box of each frame of the first sequence are used to compute the MI measure. Nearest neighbor interpolation is used when estimating the joint histogram and simplex search algorithm [13] is used as the optimizer. \( P_0 \), found in Step 1, is always used as one of the vertices of the initial simplex. The purpose of keeping \( P_0 \) in the initial simplex is to ensure that the local maximum found by the simplex method is close to \( P_0 \), which is known to be a coarse solution. Except for the very first pair of frames, the transformation found in the previous pair of frames \( P_{i-1} \) is also adopted as one vertex of the initial simplex. The reason is that the transformation required to register the current pair of frames should be very close to the transformation required to register the previous pair of frames. Another vertex is then chosen arbitrarily.

Step 4. Use the transformation found for each pair of frames to transform the bounding box in the first sequence into the bounding box in the second sequence.

Steps 2-4 are repeated for the entire set of video frames to be registered.

6. EXPERIMENTAL RESULTS

6.1. Data description

The video data we used was originally collected by Trex Enterprises and was provided through ITT Industries. Three video cameras were used simultaneously to collect continuous images of size \([120\times160]\) at a rate of 30 frames/sec. They were visual, IR, and MMW video cameras. We have found, by observing the data, that the MMW camera was not well calibrated with either IR or visual cameras. On the other hand, the IR and visual sequences seem to better meet our assumptions. For this reason, we experimented with the IR and visual video sequences and the results are presented in the next section. In this experiment, the IR sequence was used as the first video sequence and the visual sequence was served as the second one. Therefore, we assumed that the target tracking has been performed on the IR sequence and the results are represented by a sequence of bounding boxes.

6.2. Experimental results

Fig. 5 shows the first frames of both video sequences along with the regions restricted by the corresponding bounding boxes. The bounding box in the first frame of the visual sequence was obtained by manual registration and the transformation parameter set found was \( P_0 = [\text{scale factor, vertical displacement, horizontal displacement}] = [0.439, -1.1541, -3.4275] \). The scale factor found here will not be changed during the subsequent process, and \( P_0 \) is always used as one of the vertices of the initial simplex. To present the result, we have manually registered the two video sequences and the results are used as the standard or ground truth for our experiment. We denote this standard as \( P'_i \), where \( i \) indicates the frame number. In Figs. 6-7, we show the registration deviation of each frame from the standard defined as:

\[
d(i) = \sqrt{((P_i(2) - P'_i(2))^2 + (P_i(3) - P'_i(3))^2) \quad (15)}
\]

where \( P_i \) is the transformation parameter found for the \( i \)th frame pair. Fig. 6 shows this deviation from \( i = 1 \sim 135 \) and Fig. 7 shows this deviation from \( i = 220 \sim 790 \). To compare the result with that when a fixed parameter set \( P_0 \) is used, the registration deviation obtained in this case is shown in Fig. 8 and Fig. 9. Observing Fig. 6 and Fig. 8, we can see that the registration deviation obtained using our approach is smaller in general. In fact, the average registration deviation for Fig. 6 is 0.90 pixels and the average registration deviation for Fig. 8 is 1.14 pixels. However, the registration deviations for Fig. 7 and Fig. 9 are 1.54 pixels and 1.43 pixels, respectively. This shows a degradation of registration accuracy using the proposed approach. This is due to three strong peaks located around \( i = 222, 286, \text{ and } 312 \). These represent the occasional large errors resulting from the employed MI-based registration approach. To solve this problem, we first define the relative difference between two consecutive frames to determine an outlier as:

\[
r(i) = |P_{i-1}(2) - P_i(2)| + |P_{i-1}(3) - P_i(3)| \quad (16)
\]

An outlier is identified if the relative difference \( r(i) \) is larger then a predefined threshold value \( T \). In this experiment, \( T = 4 \) seems to be adequate. When an outlier is identified, we keep the transformation parameter set unchanged. That is:

\[
\text{if } r(i) > T, \text{ then } P_i = P_{i-1} \quad (17)
\]

With this approach to handle outliers, the occasional larger registration errors can be removed. Fig. 10 shows the registration deviation with this approach to handle
outliers. Clearly, those strong peaks are eliminated with this approach to handle outliers. The average registration deviation in Fig. 10 is 1.33 pixels. It is smaller than the average value in Fig. 9 of 1.43 pixels.

6.3. Computational efficiency

Our experiment was performed on a Dell Precision 530 with 1.8 GHz Intel Xeon CPU. In our implementation, the code was written using both MATLAB and Microsoft Visual C++. Number of frames in each video sequence was 790, and that accounts for about 26.3 seconds of video. The execution time for the bounding box registration task is 13.76 seconds. It is about half of the video playing time. Therefore, our algorithm has the potential to be implemented in real time, depending on the complexity of the tracking algorithm used.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an approach to bounding box registration of two video sequences. The registration algorithm is based on mutual information, a very general and powerful similarity measure. Although this work was initiated to track a moving person in a MMW image sequence with the help of another type of image sequence, we have not experimented with the proposed algorithm on MMW video data because we were not able to access the equipments to calibrate the MMW video camera with either IR or visual camera. However, our proposed algorithm is modality-independent. That is, as long as the mutual information is a valid similarity measure for the images collected from the cameras...
employed, the proposed algorithm is valid. Our previous work [6] has shown that MI similarity measure indeed is a valid measure for IR and MMW still images; therefore, we expect that we can successfully apply the proposed approach to MMW sequences. This needs to be further investigated.

At this preliminary stage, we assumed that the moving person can be tracked successfully in image sequences collected from an IR camera, and, as a result, the tracked person is restricted in a bounding box in each frame. The pixels in each bounding box are then used to compute the MI similarity measure. However, some pixels in the bounding box represent the background and can be misregistered when the subject within the bounding box is registered. This may influence the registration accuracy. Fortunately, promising results were obtained as long as the subject is dominant within the bounding box in terms of number of pixels. Most of the tracking algorithms represent the moving targets in the form of blobs before a bounding box is generated. We can use only the pixels in the blobs to compute the MI measure to reduce the processing time and improve accuracy.

Since we were not involved in the data collection process, nor were we provided the sensor parameters (i.e., geometrical relation of the cameras involved), camera calibration was not done for this work. With the sensor parameters and careful camera calibrations, the geometrical transformations required to register each corresponding frame can be better approximated using the 2D rigid body transformation model as described in this paper. We expect the performance of the proposed approach to improve significantly.

Our future work includes: 1. Collecting image sequences from calibrated MMW and visible/IR cameras and applying the proposed approach. 2. Developing a human tracking algorithm or improve/modify existing human tracking algorithms to meet our needs.

ACKNOWLEDGEMENT

This work was supported under NIJ and AFRL funding through a contract to ITT Industries. Images in Figure 1 were provided by Thermotex Corp.

REFERENCES


