Registration of Multimodal Stereo Images using Disparity Voting from Correspondence Windows

Stephen Krotosky and Mohan Trivedi
University of California, San Diego
Computer Vision and Robotics Research Laboratory
La Jolla, CA, USA
krotosky@ucsd.edu, mtrivedi@ucsd.edu

Abstract

This paper presents a method for registering multimodal imagery in short range surveillance situations when the differences in object depths preclude any global registration techniques. An analysis of multimodal registration approaches gives insight into the limitations of global assumptions and motivates the developed algorithm. Using calibrated stereo imagery, we use maximization of mutual information in sliding correspondence windows that inform a disparity voting scheme to demonstrate successful registration of color and thermal images. Extensive testing of scenes with multiple people at different depths and levels of occlusion shows high rates of successful registration and gives a reliable framework for further processing and analysis of the multimodal imagery.

1 Introduction

Surveillance applications are increasingly using multimodal imagery to obtain and process information about a scene. Specifically the disparate, yet complementary nature of visual and thermal imagery has been used in recent works to obtain additional information and robustness [1], [7]. The use of both types of imagery yields information about the scene that is rich in color, depth, motion and thermal detail. Such information can then be used to successfully detect, track and analyze people and objects in the scene.

In order to associate the information from each modality, corresponding data in each image must be successfully registered. In long range surveillance applications [7], the cameras are assumed to be oriented in such a way that a global alignment function will register all objects in the scene. However, this assumption means that the camera must be very far away from the imaged scene. When analysis of nearer scenes is desired or necessary, the global alignment will not hold.

A minimum camera solution for registering multimodal imagery in these short range surveillance situations would be to use a single camera from each modality, arranged in a stereo pair. Unlike colocating the cameras, arranging the cameras into a stereo pair allows objects at different depths to be registered. The stereo registration would occur on a local level, similar to the way unimodal stereo camera approaches give local registration for the left and right camera pairs. However, because of the disparate nature of the imagery, conventional stereo correspondence matching assumptions do not hold and care needs to be taken to ensure reliable registration of objects in the scene.

This paper demonstrates the successful registration of a color and thermal infrared multimodal stereo pair in short range surveillance applications. Local correspondences are used to provide accurate registration of multiple people that are at various depths in the scene with different levels of occlusion. Utilizing a technique that incorporates the maximization of mutual information for sliding correspondence windows, the approach generates a disparity voting matrix comprised of the best match at each window. This matrix allows for the estimation of the registration disparity image and a corresponding confidence image. Extended analysis shows successful registration for complex scenes with high levels of occlusion and numbers of people occupying the imaged space.

2 Related research in multimodal image registration

Many of the previous works in multimodal image registration have addressed the registration problem by assuming that a global transformation model exists that will register all the objects in the scene. Davis and Sharma [7], as well as Conaire *et al.* [6], use an infinite planar homography assumption to perform registration. This assumption

means that the imaged scene will be very far from the camera, so that an object's displacement from the registered ground plane will be negligible compared to the observation distance. While this type of assumption is appropriate for long distance and overhead surveillance scenes, it is not valid in situations where objects and people can be at various depths whose difference is significant relative to their distance from the camera.

Other global image registration methods assume that all registered objects will lie on a single plane in the image. It is impossible to accurately register objects at different observation depths under this assumption, as the displacement and scaling for each object will depend on the varying perspective effects of the camera. This means that accurate registration can only occur when there is only one observed object in the scene [8], or when all the observed objects are restricted to lie at approximately the same distance from the camera [10]. The global alignment algorithms proposed by Irani & Anandan [9] and Coiras, *et al.* [5] do not account or experiment with situations where there are objects at different depths or in different planes in the image. Both utilize the assumption that the colocation of the cameras and the observed distances are such that the parallax effects can be ignored.

Multiple stereo camera approaches have been proposed by Bertozzi *et al.* [1]. They used four cameras configured into two unimodal stereo pairs in each modality that yield two disparity estimates. Registration can then occur in the disparity domain. While this approach yields redundancy and registration success, the use of four cameras can be cumbersome both in physical creation, calibration and management, as well as in data storage and processing. A registration solution using the minimum (2) number of cameras is desired.

Chen, *et al.* [4] introduced the idea of registering partial image region-of-interests (ROI) instead of finding a global transformation. The main assumption of this approach is that each individual ROI corresponds to a person in the scene and is at a specific plane that can be individually registered with a separate homography. They propose that the imagery can be registered using a maximization of mutual information technique on bounding box ROIs that correspond to detected objects in one of the modalities. The matching bounding box is then searched over the other modality using a simplex method. Success gives registration for ROIs that correspond to objects at different depths.

However, a limiting assumption of this approach is that ROIs can always be properly segmented and tracked in one of the modalities so that the corresponding region can be identified using the maximization of mutual information technique. While [3] relaxes this somewhat by proposing an initial silhouette extraction for ROI construction, the assumption that ROIs will be properly segmented will often

not hold, especially in situations where occlusions can produce ROIs that contain two or more merged objects at different depths. Using ROIs that contain multiple objects will not register properly as the required assumption that an ROI is contained within a single plane will not hold. Additionally, Chen, *et al.*do not actually present any registration results where there are ROIs that are at significantly different depths in the scene or situations where occlusion or improperly formed ROIs are an issue.

3. Multimodal image registration using mutual information and disparity voting

Our registration algorithm [11] addresses the ROI and occlusion limitations of Chen, *et al.* [4]. We first reduced the search space needed in Chen's work by performing stereo calibration on the multimodal imagery. We eliminate the need for perfectly segmented ROIs by relying on reasonable initial foreground segmentation and using our disparity voting algorithm to resolve the registration for occluded or malformed ROIs. This approach gives robust registration disparity estimation with statistical confidence values. Figure 1 shows a flowchart outlining our algorithmic framework.

3.1 Multimodal image calibration and acquisition

Calibration can be performed using standard techniques, such as those available in the Camera Calibration Toolbox for Matlab [2]. Due to the large differences in visual and thermal imagery, some extra care needs to be taken to ensure the calibration board looks similar in each modality. A solution is to use a standard calibration board and illuminate the scene with high intensity halogen bulbs place behind the cameras.

The acquired and rectified image pairs are denoted as I_L, the left color image, and I_R, the right thermal image. Due to the high differences in imaging characteristics, it is very difficult find correspondences for the entire scene. Instead, foreground extraction is performed to give initial areas/objects to focus registration. The corresponding foreground images are F_L and F_R, respectively. Additionally, the color image is converted to grayscale for mutual information based matching.

3.2 Correspondence matching using maximization of mutual information

Once the foreground regions are obtained, the correspondence matching can begin. Matching occurs by fixing a correspondence window along one reference image in the pair and sliding the window along the second image that is the
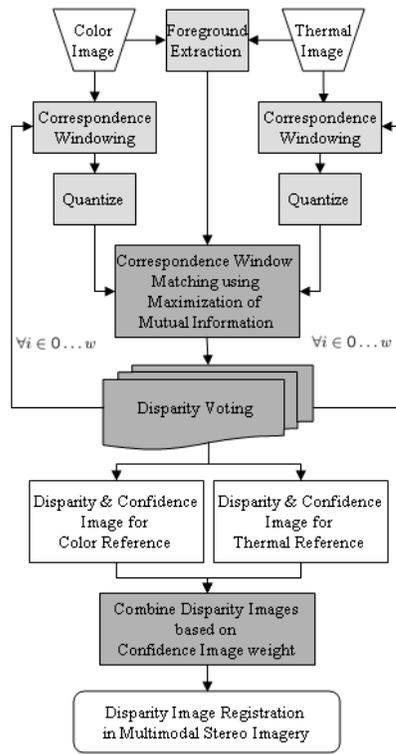
IEEE
COMPUTER
SOCIETY

Figure 1. Flowchart of multimodal image registration algorithm.

best match. Let h and w be the height and width of the image, respectively. For each column $i \in 0 \ldots w$, let $W_{L,i}$ be a correspondence window in the left image of height h and width M centered on column i. Define a correspondence window $W_{R,i,d}$ in the right image having height h and centered at a column $i+d$, where d is a disparity offset. For each column i, a correspondence value is found for all $d \in d_{\min} \ldots d_{\max}$.

Given the two correspondence windows $W_{L,i}$ and $W_{R,i,d}$, we first linearly quantize the image to N levels such that

$$N \approx \sqrt{8Mh} \qquad (1)$$

where Mh is the area of the correspondence window. The result in (1) comes from Thevenaz and Unser's [12] suggestion that this equation is reasonable to determine the number of levels needed to give good results for maximizing the mutual information between image regions.

Now we can compute the quality of the match between the two correspondence windows by measuring the mutual information between them. The mutual information be-

tween two image patches is defined as

$$I(L, R) = \sum_{l,r} P_{L,R}(l,r) \log \frac{P_{L,R}(l,r)}{P_L(l)P_R(r)} \qquad (2)$$

where $P_{L,R}(l,r)$ is the joint probability mass function (pmf) and $P_L(l)$ and $P_R(r)$ are the marginal pmf's of the left and right image patches, respectively.

Now that we are able to determine the mutual information for two generic image patches, let's define the mutual information between two specific image patches as $I_{i,d}$ where again i is the center of the reference correspondence window and $i+d$ is the center of the second correspondence window. For each column i, we have a mutual information value $I_{i,d}$ for $d \in d_{\min} \ldots d_{\max}$. The disparity d_i^* that best matches the two windows is the one that maximizes the mutual information

$$d_i^* = \arg \max_d I_{i,d} \qquad (3)$$

The process of computing the mutual information for a specific correspondence window is illustrated in Figure 2. An example plot of the mutual information values over the range of disparities is also shown.
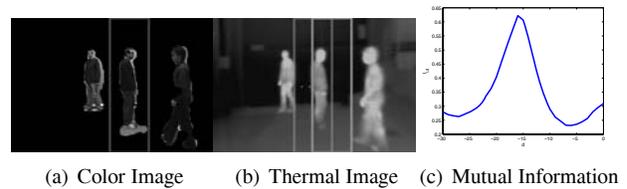

(a) Color Image   (b) Thermal Image   (c) Mutual Information

Figure 2. Mutual Information for Correspondence Windows.

3.3 Disparity voting with sliding correspondence windows

We wish to assign a vote for d_i^*, the disparity that maximizes the mutual information, to all foreground pixels in the reference correspondence window. Define a disparity voting matrix D_L of size $(h, w, d_{\max} - d_{\min} + 1)$, the range of disparities. Then given a column i, for each image pixel that is in the correspondence window and foreground map, $(u, v) \in (W_{L,i} \ \& \ F_L)$, we add to the disparity voting matrix at $D_L(u, v, d_i^*)$.

Since the correspondence windows are M pixels wide, pixels in each column in the image will have M votes for a correspondence matching disparity value. For each pixel (u, v) in the image, D_L can be thought of as a distribution of matching disparities from the sliding correspondence windows. Since it is assumed that all the pixels attributed to

a single person are at the same distance from the camera, a good match should have a large number of votes for a single disparity value. A poor match would be widely distributed across a number of different disparity values. Figure 3 shows the disparity voting matrix for a sample row in the color image. The x-axis of the image is the columns i of the input image. The y-axis of the image is the range of disparities $d = d_{\min} \ldots d_{\max}$. Entries in the matrix correspond to the number of votes given to a specific disparity at a specific column in the image. Brighter areas correspond to a higher vote tally.
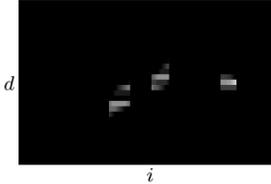

Figure 3. Disparity Voting Matrix for a sample row in the color image.

The complementary process of correspondence window matching is also performed by keeping the right thermal infrared image fixed. The algorithm is identical to the one described above, switching the left and right denotations. The corresponding disparity accumulation matrix is given as D_R.

Once the disparity voting matrices have been evaluated for the entire image, the final disparity registration values can be determined. For both the left and right images, we determine the best disparity value and its corresponding confidence measure as

$$D_L^*(u,v) = \arg\max_d D_L(u,v,d) \tag{4}$$

$$C_L^*(u,v) = \max_d D_L(u,v,d) \tag{5}$$

For a pixel (u,v) the values of $C_L^*(u,v)$ represent the number of times the best disparity value $D_L^*(u,v)$ was voted for. A higher confidence value indicates that the disparity maximized the mutual information for a large number of correspondence windows and in turn, the disparity value is more likely to be accurate than at a pixel with lower confidence. Values for D_R^* and C_R^* are similarly determined. The values of D_R^* and C_R^* are also shifted by their disparities so that they align to the left image:

$$D_S^*(u, v + D_R^*(u,v)) = D_R^*(u,v) \tag{6}$$

$$C_S^*(u, v + D_R^*(u,v)) = C_R^*(u,v) \tag{7}$$

Figure 4 shows examples of the disparity and confidence images obtained from (4) and (5), respectively. The disparities from the right thermal image have been used to shift the image pixels so that the corresponding pixels align. Notice how the disparity values in Figure 4a and Figure 4c are the same for corresponding people in the two images.
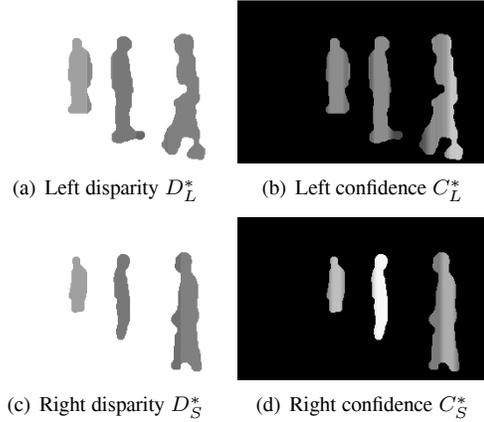

(a) Left disparity D_L^* (b) Left confidence C_L^*

(c) Right disparity D_S^* (d) Right confidence C_S^*

Figure 4. Disparity and confidence images.

Once the two disparity images are aligned, they can be combined. We have chosen to combine them using an OR operation. This tends to give the most complete results and can help to fill holes and errors in the foreground extraction of the two modalities.

$$D^*(u,v) = \begin{cases} D_L^*(u,v), & C_L^*(u,v) \geq C_S^*(u,v) \\ D_S^*(u,v), & C_L^*(u,v) < C_S^*(u,v) \end{cases} \tag{8}$$

The resulting image $D^*(u,v)$, shown in Figure 5, is the disparity image for all the foreground object pixels in the image. It can be used to register multiple objects in the image, even at very different depths from the camera.

Figure 5. The resulting disparity image D^* from combining the left and right disparity images D_L^* and D_S^* as defined in (8).

4 Results

The multimodal stereo image registration approach was tested using color and thermal data collected where the cam-

eras were oriented in the same direction with a baseline of 10 cm. The cameras were oriented so that the optical axis is approximately parallel to the ground. This position was used to satisfy the assumption that there would be approximately constant disparity across all pixels associated with a specific person in the frame. Placing the cameras in this sort of position is a reasonable thing to do, and such a position is appropriate for surveillance tasks. Video was captured as up to four people moved throughout an indoor environment. The goal was to obtain registration results for various configurations of people including different positions, distances from camera, and levels of occlusion.

(a) Initial Alignment   (b) Resulting Alignment

Figure 6. Registration results.

Figure 6 shows the result of registration for the example frame carried throughout the algorithmic derivation. Figure 6(a) shows the initial alignment of the color and thermal images, while Figure 6(b) shows the alignment after shifting the foreground pixels by the resulting disparity image D^* shown in Figure 5. The thermal foreground pixels are overlaid (in green) on the color foreground pixels (in pink).

The resulting registration in Figure 6 is successful in aligning the foreground areas associated with each of the three people in the scene. Each person in the scene lies at a different distance from the camera and yields a different disparity value that will align its corresponding image components. Registration algorithms that rely on global alignment cannot successfully handle this type of situation, yet the proposed algorithm can provide for successful multimodal registration on a per object level in cases when a global alignment is not obtainable.

Examples of successful registration for additional frames are shown in Figure 7. Columns (a) and (b) show the input color and thermal images, while column (c) illustrates the initial alignment of the people in the scene and column (d) shows the resulting alignment and overlay after the multimodal image registration has been performed. These additional examples show the success of the proposed registration technique during relatively dense scenes, where people are being significantly occluded and are at widely disparate depths from the camera.

To give a extended numerical evaluation of the success of this algorithm, we have analyzed the registration results for over 2000 frames of captured video. Registration was considered successful if the color and infrared data corre-

(a) Color   (b) Infrared   (c) Unaligned   (d) Aligned

Figure 7. Example registration results.

sponding to each person in the scene were properly aligned. Table 1 shows the results of this evaluation. The data is broken down into groups based on the number of people visible in the frame.

Analysis showed that when there was no visible occlusion in the scene, registration was correct 100% of the time. This indicates that our approach can equal the "perfect segmentation" assumption of Chen *et al.* [4]. In the more challenging cases, where there are large amounts of occlusion in the scene, the success of our registration algorithm is still high, as shown in Table 2.

Registration errors often occur when there is a large amount of occlusion between two or more people in the scene. The errors occur because there is often very little foreground associated with the occluded person in one of the images and it is therefore difficult to accumulate a significant number of votes for the disparity corresponding to the occluded persons location. These errors are usually isolated for a single frame or two and quickly correct them-

COMPUTER
SOCIETY

Table 1. Multimodal Stereo Registration Results

# People in Frame	# Correct Registered	Total # Frames	% Correct
1	55	55	100.00 %
2	171	172	99.42 %
3	1087	1111	97.84 %
4	690	720	95.83 %
Total	2003	2058	97.33 %

Table 2. Multimodal Stereo Registration Results for frames with Occlusion

# People in Frame	# Correct Registered	Total # Frames	% Correct
1	0	0	n/a
2	51	52	98.08 %
3	653	677	96.45 %
4	581	611	95.09 %
Total	1285	1340	95.90 %

(a) Unaligned (b) Aligned

Figure 8. Examples of registration errors.

selves when the level of occlusion reduces. Figure 8 shows examples of registration errors. Notice how parts of the occluded person are misaligned or missing from the final alignment. Subsequent frames, with less occlusion, do not exhibit these errors.

5 Conclusions

In this paper we have presented and analyzed a method for registering multimodal stereo images in short range surveillance situations where parallax effects cause traditional global alignment assumptions to not hold. The method proves successful in dense scenes that include significant occlusion levels of people. The algorithm has given successful and reliable registration without relying on the limiting assumption that object regions be perfectly pre-segmented. An analysis of over 2000 frames yielded a registration success rate of over 97%. This level of registration quality gives the ability to segment occluded people using registration disparity and can be a valuable input for further detection, tracking and surveillance applications.

References

[1] M. Bertozzi, A. Broggi, M. Felias, G. Vezzoni, and M. D. Rose. Low-level pedestrian detection by means of visible and far infra-red tetra-vision. In *IEEE Conference on Intelligent Vehicles*, 2006.

[2] J.-Y. Bouguet. Camera calibration toolbox for matlab.

[3] H. Chen, S. Lee, R. Rao, M. Slamani, and P. Varshney. Imaging for concealed weapon detection. *IEEE Signal Processing Mag.*, pages 52–61, Mar. 2005.

[4] H. Chen, P. Varshney, and M. Slamani. On registration of regions of interest (ROI) in video sequences. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, 2003.

[5] E. Coiras, J. Santamaria, and C. Miravet. Segment-based registration technique for visual-infrared images. *Optical Engineering*, 39(1):282–289, Jan. 2000.

[6] C. Conaire et al. Background modeling in infrared and visible spectrum video for people tracking. In *IEEE CVPR Workshop on Object Tracking and Classification beyond the Visible Spectrum*, 2005.

[7] J. Davis and V. Sharma. Fusion-based background-subtraction using contour saliency. In *IEEE CVPR Workshop on Object Tracking and Classification beyond the Visible Spectrum*, 2005.

[8] J. Han and B. Bhanu. Detecting moving humans using color and infrared video. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2003.

[9] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *Computer Vision, 1998. Sixth International Conference on*, 1998.

[10] M. Itoh, M. Ozeki, Y. Nakamura, and Y. Ohta. Simple and robust tracking of hands and objects for video-based multimedia production. In *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2003.

[11] S. Krotosky and M. Trivedi. Multimodal stereo image registration for pedestrian detection. In *IEEE Conference on Intelligent Transportation Systems*, 2006.

[12] P. Thevenaz and M. Unser. Optimization of mutual information for multiresolution image registration. *IEEE Trans. Image Processing*, 9(12):2083–9, Dec. 2000.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 **IEEE**

IEEE
COMPUTER
SOCIETY