Local self-similarity based registration of human ROIs in pairs of stereo thermal-visible videos

Atousa Torabi[*,a], Guillaume-Alexandre Bilodeau[a]

[a] *LITIV, Department of Computer and Software Engineering, École Polytechnique de Montréal, P.O. Box 6079, Station Centre-ville, Montréal (Québec), Canada, H3C 3A7*

Abstract

For several years, mutual information (MI) has been the classic multimodal similarity measure. The robustness of MI is closely restricted by the choice of MI window sizes. For unsupervised human monitoring applications, obtaining appropriate MI window sizes for computing MI in videos with multiple people in different sizes and different levels of occlusion is problematic. In this work, we apply local self-similarity (LSS) as a dense multimodal similarity metric and show its adequacy and strengths compared to MI for a human ROIs registration. We also propose a LSS-based registration of thermal-visible stereo videos that addresses the problem of multiple people and occlusions in the scene. Our method improves the accuracy of the state-of-the-art disparity voting (DV) correspondence algorithm by proposing a motion segmentation step that approximates depth segments in an image and enables assigning disparity to each depth segment using larger matching

[*]Corresponding author

Email addresses: atousa.torabi@polymtl.ca (Atousa Torabi), guillaume-alexandre.bilodeau@polymtl.ca (Guillaume-Alexandre Bilodeau)

window while keeping registration accuracy. We demonstrate that our registration method outperforms the recent state-of-the-art MI-based stereo registration for several realistic close-range indoor thermal-visible stereo videos of multiple people.

Key words: Local self-similarity, Mutual information, Multimodal video registration, Dense stereo correspondence, Thermal camera, Visible camera, Visual surveillance

1. Introduction

In the recent years, there has been a growing interest in visual surveillance using thermal-visible imaging system for civilian applications, due to the reduction in the price of infrared sensors. The advantages of jointly using a thermal camera with a visible camera have been discussed comprehensively in [1, 2, 3]. For human monitoring applications in uncontrolled settings, the joint use of these two sensors improves the quality of input data. For example, in a scene where there are shadows on the ground, poor color information under low lighting conditions, or similarity of the human body/clothing with the background, the combined data enables better detection and tracking of people. Moreover, for human activity analysis, the joint use of thermal and visible data enables us to better detect and segment the regions related to the object that people may carry based on their temperature differences compared to the human body.

In the literature, several methods including data fusion algorithms, background subtraction, multi-pedestrian tracking, and classification have been proposed for long-range thermal-visible videos of multiple people [4, 5, 6].

However, a fundamental and preliminary task associated with the joint use of thermal-visible data is accurately matching features of a pair of images captured by two different sensors with high differences in imaging characteristics. This task is challenging, especially for close-range scene due to the large scale objects with different detailed patterns in a pair of thermal and visible images. For a pair of close-range videos, it is very difficult to find the correspondence for an entire scene. Therefore, registration is focused on image region of interest (ROI) where for human monitoring applications are human body regions. Matching corresponding regions belonging to a human body in a pair of visible and thermal images is still challenging, because corresponding pixels have different intensities and have different patterns and textures due to the differences in thermal and visible image characteristics.

In few previous related works, MI is the only similarity measure used in dense multimodal stereo matching [7, 8, 9]. Fookes *et al.* proposed a MI-based window matching method that incorporates prior probabilities of the joint probability histogram of all the intensities in the stereo pair in the MI formulation [9]. Thier matching method is less sensitive to MI window sizes. However, in their experiment, they only used negative and solarized images that have similar patterns within corresponding ROIs as opposed to thermal and visible images. Egnal has shown that mutual information (MI) is a viable similarity metric for matching disparate thermal and visible images [10]. Chen *et al.* proposed a MI-based registration method for pairs of thermal and visible images with the assumption that each window bounding a ROI represents a single human [8]. In their method, occluded people that are merged into one ROI may not be accurately registered since an ROI

may contain people within different depth planes. As a solution to improve registration of occluded people in a scene, Krotosky and Trivedi proposed a disparity voting (DV) matching approach [7]. DV is performed by horizontally (column by column) sliding small width windows on rectified thermal and visible images, computing MI for pairs of windows, and finally for each column, counting the number of votes associated to each disparity and assigning one disparity to each column based on a Winner Take All (WTA) approach. Their method can handle occlusion horizontally (two neighboring columns might be assigned to different disparities), but it cannot accurately register people with different height where a shorter person is in front of a taller one (vertical occlusion) since all pixels of a column inside an ROI are assigned to only one disparity.

In the abovementioned papers, the correctness and confidence of MI compared to other viable similarity metrics is not discussed. Based on our experiments, in videos where people have textured clothes, where human ROI segmentation is imperfect (i.e., partial misdetection or false detection), and where there are occlusions, MI is unreliable for matching small width windows like the one suggested in [7]. For MI-based stereo matching, choosing the appropriate image window size is not straightforward due to the aforementioned difficulties. Also, there is always a trade-off between choosing larger windows for matching evidence, and smaller windows for the precision and details needed for an accurate registration.

In this work, we apply local self-similarity (LSS) to the problem of thermal-visible stereo correspondence for close-range human monitoring applications. LSS has been proposed by Shechtman and Irani in [11] and has been pre-

viously applied to problems of object categorization, image classification, pedestrian detection, image retrieval by sketching, and object detection [12, 13, 14, 11]. To the best of our knowledge, nobody has previously applied LSS as a thermal-visible dense stereo correspondence measure. LSS, similarly to MI, computes statistical co-occurrence of pixel intensities. However LSS, unlike MI, is firstly computed and extracted from an individual image as a descriptor and then compared between images. The property of LSS, which makes this measure more interesting for our application, is that the basic unit for measuring internal joint pixel statistics is a small image patch that captures more meaningful image patterns than individual pixels as used in MI computation. This property is useful for matching thermal and visible human ROIs with different direct visual properties such as colors and patterns but similar layout/body shape which is an indirect image property. The algorithms presented in this manuscript are based on our previous work [15], but they are further developed with detailed analysis and new evaluations.

In section 2, we present a theoretical analysis of LSS and MI as dense multimodal correspondence measures. In section 3, we quantitatively assess the reliability and accuracy of MI and LSS as dense stereo similarity measures in various close-range challenging human monitoring scenarios. In section 4, we propose our LSS-based registration that accurately performs for multiple people and occlusions. Finally, in section 5, we qualitatively and quantitatively compare our LSS-based stereo registration method and a recent state-of-the-art MI-based stereo registration method for human monitoring applications.

2. Theoretical analyses of MI and LSS for thermal-visible human ROI correspondence

2.1. Mutual information

Mutual information (MI) is the classic dense similarity measure for multimodal stereo registration. The MI between two image windows L and R is defined as

$$MI(L,R) = \sum_{l}\sum_{r} P(l,r) log \frac{P(l,r)}{P(l)P(r)}, \quad (1)$$

where $P(l,r)$, is the joint probability mass function and $P(l)$ and $P(r)$ are the marginal probability mass functions. $P(l,r)$ is a two-dimensional histogram $g(l,r)$ normalized by the total sum of the histogram. $g(l,r)$ is computed as for each point, the quantized intensity levels l and r from the left and right matching windows (L and R) increment $g(l,r)$ by one. The marginal probabilities $P(l)$ and $P(r)$ are obtained by summing $P(l,r)$ over the grayscale or thermal intensities.

The unit of measure for MI is the pixel, which forces existing underlying visual properties (i.e., pixel colors, intensities, edges, or gradients) in thermal and visible images to be identical for a contribution in MI computation. In our application, MI computes the statistical co-occurrence of pixel-wise measures, such as patterns related to human body regions on pairs of thermal and visible images. Based on our experiments, MI is unreliable for matching differently textured corresponding human body ROIs and partially misdetected or falsely detected human body ROIs caused by erroneous foreground segmentation in thermal and visible images. MI only performs well when the joint probability histogram is sufficiently populated inside MI windows.

Choosing the appropriate window size is not straightforward due to the aforementioned difficulties. In fact, because of imperfect data, the appropriate size might not be available for the computation (e.g. region fragmentation, small objects).

2.2. Local self-similarity

Local self-similarity (LSS) is a descriptor that capture locally internal geometric layout of self-similarities (i.e., edges) within an image region (i.e., human body ROI) while accounting for small local affine deformation. Initially, this descriptor has been proposed by Sechtman and Irani [11]. LSS describes statistical co-occurrence of small image patch (e.g. 5×5 pixels) in a larger surrounding image region (e.g. 40×40 pixels). First, a correlation surface is computed by a sum of the square differences (SSD) between a small patch centered at pixel p and all possible patches in a larger surrounding image region. SSD is normalized by the maximum value of the small image patch intensity variance and noise (a constant that corresponds to acceptable photometric variations in color or illumination). It is defined as

$$S_p(x, y) = exp(-\frac{SSD_p(x, y)}{max(var_{noise}, var_{patch})}). \quad (2)$$

Then, the correlation surface is transformed into a log-polar representation partitioned into e.g. 80 bins (20 angles and 4 radial intervals). The LSS descriptor is defined by selecting the maximal value of each bin that results in a descriptor with 80 entries. A LSS descriptor is firstly computed for a ROI within an image then it can be compared with other LSS descriptors in a second image using a measure such as $L1$ distance.

Previously, Shechtman and Irani have shown that for matching image regions with similar shape/layout but with different direct visual properties such as color and edges, LSS is a more reliable similarity metric compared to other local image descriptors and match measures such as MI [11]. Shechtman and Irani have also shown that LSS is applicable for image retrieval by sketching [11]. In their work, the template image is a sketch of human body representing a body pose. The template is used to detect similar human body poses in several images with different color and textures. For this application, the advantage of LSS is that its unit measure, which is a small image patch, contains more meaningful patterns compared to pixel as used for MI computations. As it is described in Shechtman and Irani's work [11], this property makes LSS a suitable measure for matching image regions with different direct visual properties such as color, edges, or textures, as long as they have similar spatial layouts.

For matching thermal and visible human ROIs, we believe that the same LSS property as used for image retrieval by sketching in [11] is applicable. In fact, in thermal and visible images, the edges and pixel intensities within corresponding human body ROIs are not identical as it is required for matching using MI, but the human body layout is a common indirect visual property between thermal and visible corresponding human ROIs.

Registration accuracy is an important factor for human ROIs registration in a scene with multiple people and occlusions. In order to apply LSS as a similarity metric, it is required that this descriptor captures local image ROI layout within a small surrounding region while considering the required image details. Therefore, we experimentally found out that for a close-range

video, a patch of size 3×3 as a unit measurement within surrounding region of 20×20 pixels is sufficient to capture meaningful local image patterns of human body shape that are mostly belonging to edges. For an LSS-based window matching similar to MI-based matching, the overall geometric layout within a matching window is captured by a set of LSS descriptors respecting the relative geometric positions of the descriptors. Although the window sizes have a width smaller than the width of a complete human body and a height the same as the human body height; the set of descriptors within a window still captures partially body geometric layout. In fact, matching is performed by comparing two sets of descriptors belonging to two matching windows in thermal and visible images. A good match corresponds to two windows where the descriptors are similar both in values and their relative geometric positions.

In order to perform a better matching, we discard the non-informative descriptors from each set of descriptors. Non-informative descriptors are the ones that do not contain any self-similarities (i.e., the center of a small image patch is salient) and the ones that contain high self-similarities (i.e., a homogenous region with a uniform texture/color). A descriptor is salient, if all its bins' values are smaller than a threshold. The homogeneity (which also cause a non-informative descriptor) is detected using the sparseness measure of [16]. The sparseness measure is defined as

$$sparseness(X) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1} \tag{3}$$

where n is the dimensionality of descriptor x (in our method 80). This function evaluates to unity if and only if x contains only a single non-zero component, and takes a value of zero if and only if all components are equal. Dis-

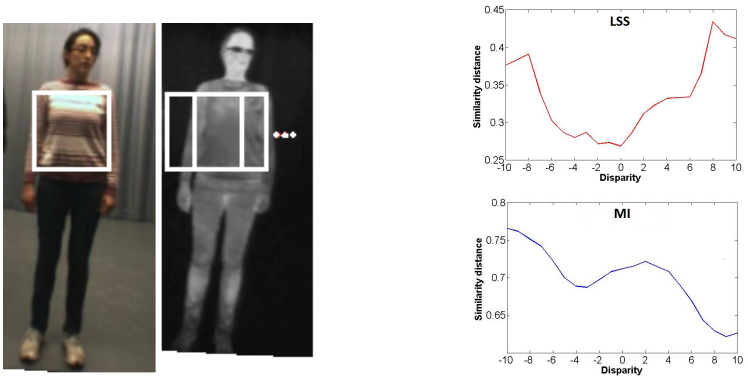
(a) (b)

Figure 1: Informative LSS descriptors. (a) Visible and informative LSS descriptors images (b) Thermal and informative LSS descriptors images.

carding non-informative descriptors is like an implicit segmentation or edge detection, which for window matching, increases the discriminative power of the LSS measure and avoids ambiguous matching. It is important to note that the remaining informative descriptors still form a denser collection compared to sparse interest points. Fig. 1 shows pixels having informative descriptors (white pixels) for a pair of thermal and visible images. Fig. 1 is an extreme example of discarding non-informative pixels to highlight by filtering out those pixels, the common visual property between thermal and visible human ROIs (i.e., human body shape) are obtainable without any explicit edge detection or segmentation.

2.3. Introductory examples of human ROI matching

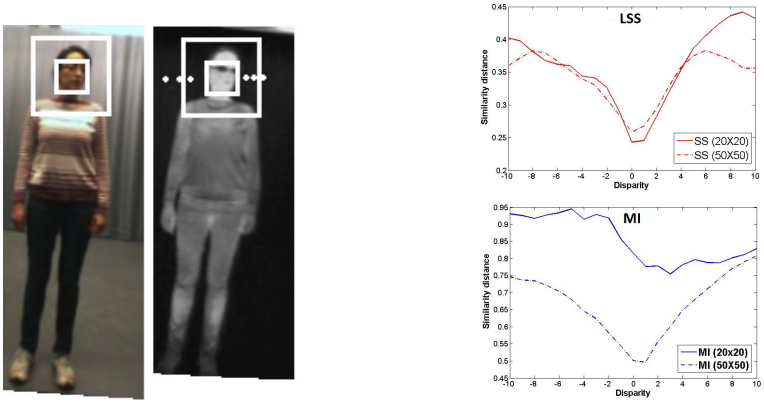
In order to illustrate the difficulties of thermal-visible human ROI matching and the advantages of LSS compared to MI, we present three introductory examples of human ROI matching using a simple sliding window matching approach and various window sizes. In these examples, matching is performed by computing the similarity distances of a fixed window on an image



(a) (b)

Figure 2: Matching corresponding textured and uniform regions in visible and thermal pair of images. (a) Aligned visible and thermal images and (b) Similarity distances of LSS and MI for disparity interval of [-10,10].

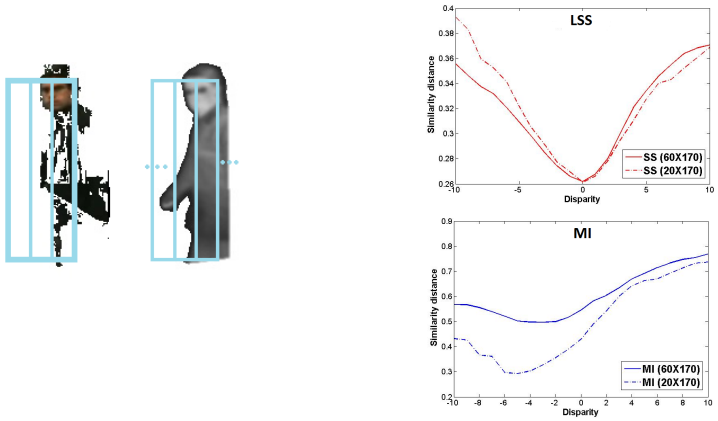
ROI of the visible image with a sliding window on the thermal image within a disparity range of [-10, 10], and then choosing the disparity that minimizes the similarity distance. In order to simplify the search to 1D, the two images were rectified, and then manually aligned so that a disparity of 0 corresponds to a ground-truth alignment (more details about multimodal camera calibration in section 3.1). We defined the LSS-based similarity distance between two windows L and R by the sum of the $L1$ distances of informative descriptors within those two windows, and the MI-based similarity distance as $1 - MI(L, R)$. Fig. 2 shows an example of matching a textured region in the visible image with a corresponding uniform region in the thermal image. Fig. 2 (b) shows the similarity distance results for both MI and LSS over a preset disparity range. For LSS, the similarity distance is correctly minimized at disparity 0. However for MI, the similarity distance is minimized incorrectly.



(a) (b)

Figure 3: Matching corresponding regions of visible and thermal within image windows of size 20×20 and 50×50 pixels. (a) Aligned visible and thermal images, (b) Similarity distances of LSS and MI for disparity interval of [-10,10].

This illustrate that MI is not a robust similarity metric for matching a textured region and a uniform region when there are not many similar patterns. Fig. 3 shows an example of matching windows of sizes 20×20 and 50×50 pixels on a head region. Fig. 3 (b) shows that MI is not a robust measure for matching 20×20 thermal-visible windows. However, using larger window of size 50×50 pixels containing more similar patterns and more similar spatial layout, MI-based similarity distance is correctly minimized at disparity 0. For this example, LSS-based similarity distance is correctly minimized at disparity 0 for both matching window sizes which illustrate the robustness of this measure for matching even with small window sizes. Fig. 4 shows an example of matching thermal-visible windows on regions with dramatic partial ROI misdetection using matching window sizes of 20×170 and 60×170 pixels. In the visible image, due to the color similarity of the



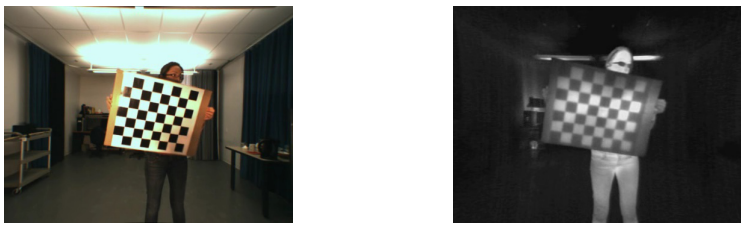
(a) (b)

Figure 4: Matching corresponding foreground pixels within 20×170 and 60×170 pixels windows in visible and thermal pair of images (a) Aligned visible and thermal images, (b) Similarity distances of LSS and MI for disparity interval of [-10,10].

ROI and the background, some parts of the body region are not detected. Fig. 4 (b) shows that MI fails to find the correct disparity offset with both window sizes. However, LSS find the correct disparity which illustrates the robustness of this measure for partial ROI misdetection.

3. Quantitative analyses of MI and LSS for thermal-visible human ROI correspondence

In our previous work, we have evaluated the performance of several local image descriptors and similarity measures for thermal-visible human ROI registration [17]. In this work, we only focus on the comparison between MI and LSS with new evaluation criteria and extensive experiments.



(a) (b)

Figure 5: Calibrating images: (a) Visible image and (b) Thermal image.

3.1. Video Acquisition and Calibration

We used synchronized visible-thermal videos of a $5m \times 5m$ room at a fixed temperature of 24 $^{\circ}C$ captured by stationary thermal and visible cameras with a 12 cm baseline. We used sets of video frames of a relatively close range scene where different people with different poses and clothing are walking at different depths (between 2-5 meters) from the camera baseline. In order to simplify the stereo matching to a 1D search, we first calibrated the thermal and visible cameras, and then rectified the images using the intrinsic and extrinsic calibration parameters. We used the standard technique available in the camera calibration toolbox of MATLAB ([18]). For calibration, we placed a checkboard pattern in front of the cameras. Since in the thermal images, the checkboard pattern is not visible at room temperature; we illuminated the scene using high intensity halogen bulbs placed behind the two cameras. This way, the dark squares absorb more energy and visually appear brighter than the white squares. Fig. 5 shows an example of our calibration images.

3.2. Experimental setup

Our experimental setup is designed to study the efficiency of MI and LSS as similarity measures for thermal-visible human ROI registration using realistic videos of multiple people in a close range scene. In our experiment, the ROIs within corresponding windows in a pair of thermal and visible images might be differently textured or one textured and the other uniform. Windows are centered at randomly picked points that are located inside visible human ROIs. We used a sliding window matching (see section 3.3) to find the corresponding image window on the thermal image. The matching process was repeated using three rectangular window sizes of 10×130 (small), 20×130 (medium), and 40×130 (large) pixels. The heights of the windows are chosen as the maximum possible height of a person in our experimental videos. The randomly picked points were located either on textured or textureless visible human ROI for relatively near targets (between 2 to 3 meters from the camera) or far targets (between 4 to 5 meters). Note that for close-range scene monitoring, the scale of targets considerably changes by walking one meter further away or toward the camera. Fig. 6 shows an example of randomly picked matching window. Our experiment is carried out using 300 matching windows (100 points using three window sizes).

3.3. Sliding window matching

For each thermal and visible pair of images, a window centered at a point on the human ROI at column j on the visible image is defined ($W_{l,j}$). Then, a 1D window matching search is done on the thermal image in order to find the corresponding window $W_{r,j+d}$ which minimizes a similarity distance SD. d is a disparity offset belonging to disparity interval set D. In our experiment,

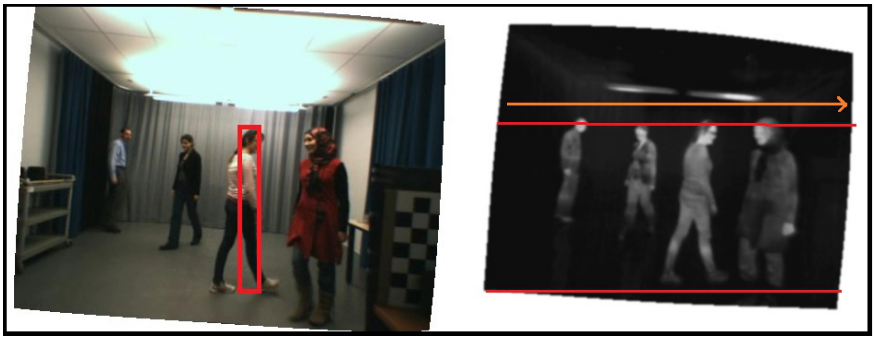


Figure 6: Thermal-visible 1-D sliding window matching.

the size of D is the same size as the image width. Fig. 6 illustrates the sliding window matching.

For LSS, the descriptor computation and the matching are done in two separate processes, for each pair of image windows $W_{l,j}$ and $W_{r,j+d}$ centered at column j on the visible image and column $j+d$ on the thermal image. A normalized similarity distance $SD_{j,d}$, which is the sum of $L1$ distance of the corresponding pixels $p_l \in W_{l,j}$ and $p_r \in W_{r,j+d}$ having informative descriptors, is computed as

$$SD_{j,d} = \frac{\sum_{p_l,p_r} L1_{l,r}(p_l, p_r)}{N}, \tag{4}$$

where N is the number of corresponding pixels p_l and p_r contributing in the similarity distance computation and d is the disparity offset. This number is also proportional to the number of informative pixels inside an image ROI. The typical value of N for window size of 40×130 varies in the range of 200 to 1000 pixels and it is maximum when edges and boundaries inside matching windows are correctly overlapped. $L1_{l,r}$ is computed as

$$L1_{l,r}(p_l, p_r) = \sum_{k=1}^{80} |d_{p_l}(k) - d_{p_r}(k)| \tag{5}$$

where 80 is the number of local self-similarity descriptor bins.

For MI, SD is defined as

$$SD_{j,d} = 1 - MI(W_{l,j}, W_{r,j+d}), \tag{6}$$

where MI is the mutual information defined in equation 1. And finally the best disparity associated to best matching windows is computed by

$$d_{min} = argmin_{d}(SD_{j,d}), d \in D. \tag{7}$$

3.4. Evaluation Criteria

In our evaluation, we assess the precision-recall and power of discrimination of MI and LSS as explained in the following sections.

3.4.1. Precision and recall

We used a criterion similar to the one used in [19]. Precision and recall are defined as follows:

$$precision = \frac{\#correctmatches}{\#matchesretrieved} \tag{8}$$

$$recall = \frac{\#correctmatches}{\#totalcorrespondences} \tag{9}$$

In our experiment, *correctmatches* is the number of matches with a disparity error smaller than 3 pixels with respect to ground-truth and with SD (equation 4 or 6) smaller than a threshold t (t varies between minimum possible values where *matchesretrieved* become one and maximum value where *matchesretrieved* become all the matched windows *totalcorrespondence*). *totalcorrespondence* is a fixed value that corresponds to the number of tested windows (i.e., 100 windows of each size). *matchesretrieved* is the number

of matches with a SD below threshold t. $matchesretrieved$ varies from 1 to $totalcorrespondences$. In a precision and recall curve, a feature with high recall value and low precision value means that many correct matches as well as many false matches are retrieved. On the other hand, high precision value and low recall value means that most matches are correct but many others have been missed.

3.4.2. Power of discrimination

To assess the reliability of a similarity metric, not only its precision is important but also how that similarity metric possesses isolation characteristic (power of discrimination) is important as well.

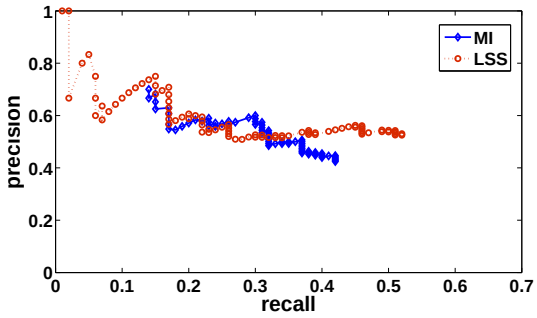
A similarity metric possesses a high power of discrimination, if its correct matches are located on isolated minimums over D (disparity range) and SD (equation 4) curve (that is, having SD value much smaller than its neighbors). In order to evaluate the isolation characteristic of MI and LSS, for their correct matches, we study the shape of SD computed along the disparity range $D = [q - 20 : q + 20]$, where q is the position of the global minimum (best match). We applied the same measure as in [20]. In order to evaluate the isolation of the global minimum, the SD values computed by the sliding window matching (section 3.3) are first sorted increasingly and are transformed to the interval $[0, 1]$ named SD'. Second, N is the number of values in SD' that are less than a pre-computed small threshold α, ignoring the global minimum. α has the same value for evaluating all descriptors and measures. Third, a quality measure s (the s value) is computed by dividing N by the size of the disparity range. So $s = 0$ corresponds to the most isolated minimum (best performance), and $s = 1$ corresponds to

the least isolated minimum (flat/constant SD versus d curve). Finally, for each correspondence measure, a graph of Accumulated Frequencies (AF) of the s values of all matches is computed (In fact AF is the distribution of s values belonging to correct matches). Therefore, the correspondence measure for which AF reaches a higher value at a smaller s value is the more discriminative.

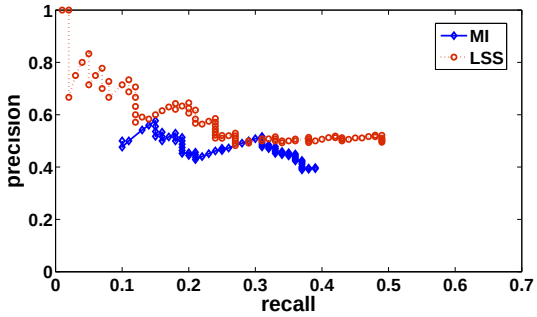
3.5. Results

First, we present the evaluation of MI and LSS using the precision and recall criterion as explained in section 3.4.1. Fig. 7 shows the precision-recall curves of MI and LSS for small, medium, and large window sizes as described in section 3.2. Overall, for all the three matching window sizes, LSS achieves higher values of recall and precision compared to MI. The largest size window achieves better precision than medium and small ones for both MI and LSS. However, MI is totally inefficient for small window sizes. This result shows the robustness of MI is closely related to the sizes of MI windows, which are required to be large enough to sufficiently populate the joint probability histogram. On the other hand, the precision of LSS is more consistent for three window sizes. For this experiment, we used a simple sliding window matching that ignores the occlusions. Using a more appropriate correspondence algorithm that we propose in section 4, will result in higher matching precisions.

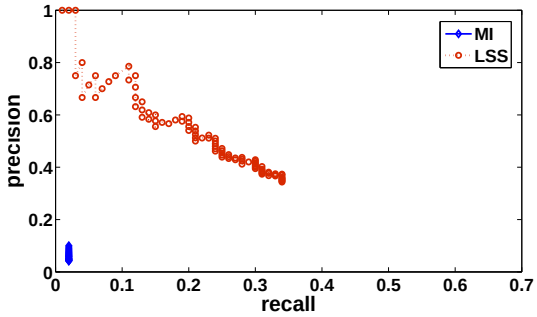
Fig. 8 shows the accumulated frequency distribution of s (details in section 3.4.2) obtained for MI and LSS using three window sizes as described in section 3.2. It can be seen, for LSS compared to MI, AF starts with a higher value and reaches to the higher value for a smaller S value. This shows



(a)

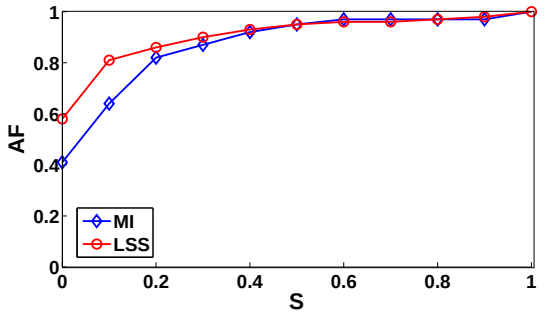


(b)

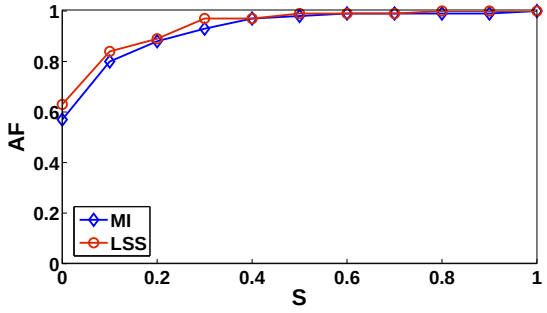


(c)

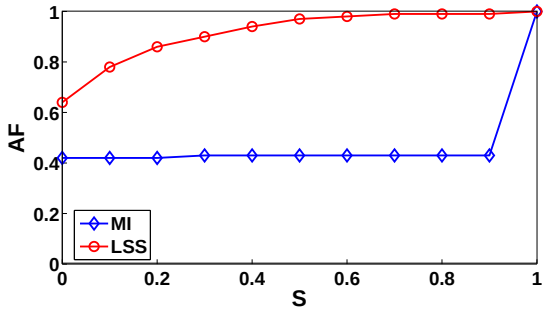
Figure 7: Precision- recall curves: (a) large window (40×130) (b) medium window (20×130) (c) small window (10×130).



(a)



(b)



(c)

Figure 8: Accumulated frequencies versus S value: (a)large window (40×130) (b) medium window (20×130) (C) small window (10×130).

LSS possesses a better isolation characteristics compared to MI.

Overall, the results show that comparing to MI, LSS is a more reliable similarity metric for matching differently textured human ROIs in thermal and visible images and it is less restricted by size of matching windows.

4. LSS-based multimodal ROI registration

In this section, we describe our novel multimodal ROI registration method using LSS. For a pair of thermal and visible video frames, our goal is to register the ROIs belonging to moving people in a scene in which they may be temporary stationary for a few frames. Our method addresses registration of multiple people merged into one ROI with different levels of occlusion and with partially erroneous foreground segmentation for realistic thermal-visible videos of a close range scene. We assume that each person at each instant lies approximately within one depth plane in the scene. Therefore, we propose that a natural way for estimating depth planes related to multiple moving people is by applying motion segmentation on foreground pixels with the assumption that each motion segment belongs to one person in the scene, but more than one motion segment may belong to a person.

We define the multimodal image registration as multiple labeling sub-problems. Then, we use the disparity voting matching approach to register each individual motion segment rather than a whole foreground blob. Let MS be the set of motion segments belonging to moving people in the scene, and D be a set of labels corresponding to disparities. Our registration method assigns a label $d_k \in D$ in the range between d_{min} to d_{max} to each pixel of a motion segment $ms_i \in MS$. Thus, our registration method has two

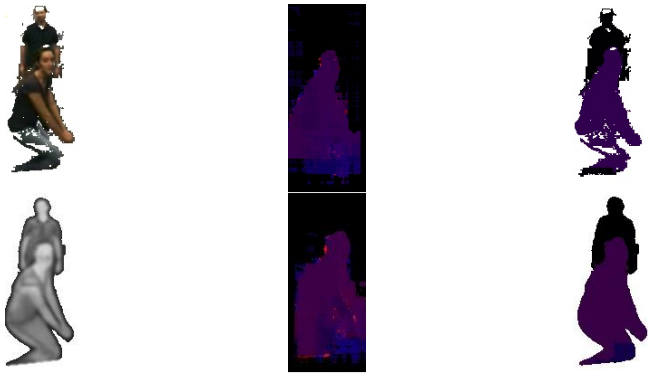
main parts: 1) motion segmentation that divides the registration problem as multiple labeling sub-problems and 2) disparity assignment which assigns disparity to each segment. The two parts of our method are described in the subsequent sections.

4.1. Motion segmentation

Our motion segmentation has three steps. Firstly, we extract foreground pixels using the background subtraction method proposed in [21]. Any background subtraction method with a reasonable amount of error is applicable. Secondly, we compute the motion vector field for foreground pixels using an optical flow method based on block-matching [22]. To speed up the process, the optical flow is only computed for regions inside the bounding boxes of the union of the foreground masks of two consecutive frames $t-1$ and t, instead of the whole image. Thirdly, we apply the mean-shift segmentation method proposed in [23] for segmenting the motion vector fields computed in the previous step and computing a mean velocity vector for the computed segments. Mean-shift segmentation is applied on (2+2) feature point dimensions, where two dimensions are related to spatial dimensions (horizontal and vertical directions) and the two others are related to the two motion vector components in x and y directions. Applying motion segmentation on ROIs results in a set of motion segments S defined as

$$SM = \{sm_1, ..., sm_m\}\,. \tag{10}$$

An average mean velocity vector $\hat{m}_i$ is associated to each sm_i using



(a) (b) (c)

Figure 9: First row: visible image and second row: thermal image. (a) Foreground image, (b) motion field vectors, and (c) motion segmentation (depth segments).

$$\hat{m}_i = \frac{\sum_{p \in sm_i} m(p)}{|sm_i|}, \tag{11}$$

where $m(p)$ is the motion vector of pixel p. Fig. 9 shows the motion segmentation results of two occluding people, where one of them is temporary stationary. Motion vectors are visualized by a mapping to HSV color space. Applying motion segmentation on foreground pixels enables us to determine also a depth segment associated to temporary stationary person for which its mean velocity vector is zero. Since in most indoor videos, the motion segmentation of thermal images are more accurate compared to visible images due to less partial ROI misdetection error, we perform motion segmentation for thermal images and we register the thermal motion segments on visible foreground images. However, it could also be done the opposite way.

4.2. Disparity assignment

At this step, we assign disparity to each motion segment individually. We use a disparity voting matching approach similar to the one that was previously proposed by Krotosky and Trivedi [7]. DV matching assigns one single disparity to all the pixels of a column of matching regions. However, different disparities can be assigned to two neighboring columns. Krotosky and Trivedi DV method uses MI as similarity metric and is performed on whole foreground blobs. Their method is able to resolve the horizontal part of an occlusion, but fails to assign correct disparity for the vertical part of an occlusion (in this case, the pixels of a column for a region associated to vertically occluded people should be assigned to a different disparity) (see fig. 11). To solve this problem, we propose performing DV on each motion segment separately. Moreover, based on our previous experiments, we use the informative LSS descriptors as similarity measure.

4.2.1. LSS-based DV algorithm

For each $sm_i \in S$, we build a disparity voting matrix of DV_i of size $(N, d_{max} - d_1 + 1)$ where N is the number of pixels of sm_i and $[d_1 - d_{max}]$ is a preset disparity range. This procedure is performed by shifting column by column $W_{l,j}$ on the reference image for all the columns $j \in s_i$, then doing window matching, the same as we previously described in section 3.3. Then, for each d_{min} computed by window matching, a vote is added to $DV_i(p_l, d_{min})$ for all $p_l \in (W_{l,j} \cap s_i)$. Since the width of windows are m pixels wide, we have m votes for each pixel belonging to s_i. Finally, the disparity map DM_i is computed as,

$$DM_i(p_l) = argmax_{_d}(D_i(p_l, d)), \tag{12}$$

5. Experimental validation and discussion

We have assessed our registration method with over 5000 video frames of up to 5 people with different clothing, various poses, distances to cameras, and with different level of occlusions. In these experiments, we used the same experimental setup as described previously in section 3.1. The first test video was captured during summer with people having lighter clothes (light clothes results in less heat patterns on the body in infrared) and with a fair amount of textures inside human ROIs in thermal and visible images. The background subtraction errors were mostly misdetection errors. Our other two test videos were captured during winter with people wearing winter clothes (thick clothes results in more heat patterns on the body in infrared), which causes patterns inside human body ROIs. The background subtraction results in our winter videos include both misdetection errors and falsely detected region as foreground. The disparity range was between 5 to 50 pixels.

Fig. 10 illustrates successful registrations with our method in one of our winter videos for three frames of people in different levels of occlusions.

5.1. Comparison of DV correspondence and our correspondence algorithm

In order to demonstrate the accuracy improvement of our method compared to a state-of-the-art disparity voting algorithm (DV) [7] in handling occlusions, we quantitatively compared our disparity results using motion

Figure 10: Registration results of foregrounds using imperfect background subtraction with false positive and false negative errors.

segmentation and the results of DV using for both LSS as similarity measure. We generated ground-truth disparities by manually segmenting and registering regions of foreground for each frame. Fig. 11 illustrates the comparison with ground-truth. Results in the first and second rows illustrate examples where two people in two different depths in the scene are in occlusion. LSS+DV method fails to assign correct different disparities to the columns containing pixels related to more than one individual since based on a WTA approach, a single disparity is assigned to all the pixels of each column. However, LSS+MS+DV succeed in assigning accurately different disparities to the two human body ROIs since the DV was applied to each motion segment individually. Accordingly, in fig. 11 (d), for the first and second rows, the sum of disparity errors of the columns corresponding to two occluded people is much higher for LSS+DV method compared to LSS+MS+DV method.

Indeed, to register merged objects in a single region, DV makes no assumptions about the assignment of pixels to individual objects and assigns

a single disparity to each column inside an ROI based on a maximization of the number of votes. In their matching approach [7], if a column of pixels belongs to different objects at different depth in the scene, the vote only goes for one of them based on WTA approach. However, in our registration method, motion segmentation gives a reasonable estimate of moving regions belonging to people in the scene, and applying the DV matching on each motion segment gives more accurate results since it is less probable that pixels in one column belongs to more than one object. Therefore, in the worst case, even with erroneous motion segmentation, our method will have at minimum the same accuracy as the DV algorithm.

Fig. 11, last row, illustrates the example of multiple occluding people. Although LSS+MS+DV registration results are not perfect because few small motion segments resulting from over segmentation were not matched correctly still the results are more accurate than for LSS+DV. Accordingly, in Fig. 11 (d), last row, the sums of disparity error for columns related to vertical occlusion is higher for LSS+DV compared to LSS+MS+DV. However, it is noticeable that in some columns, LSS+MS+DV has slightly higher errors caused by small motion segments misalignment.

Fig. 12 illustrates other registration results with LSS+MS+DV and LSS+DV. It is observable, that for LSS+DV method, the object misalignments happen where there are vertical occlusions while our method performs accurately in such a case.

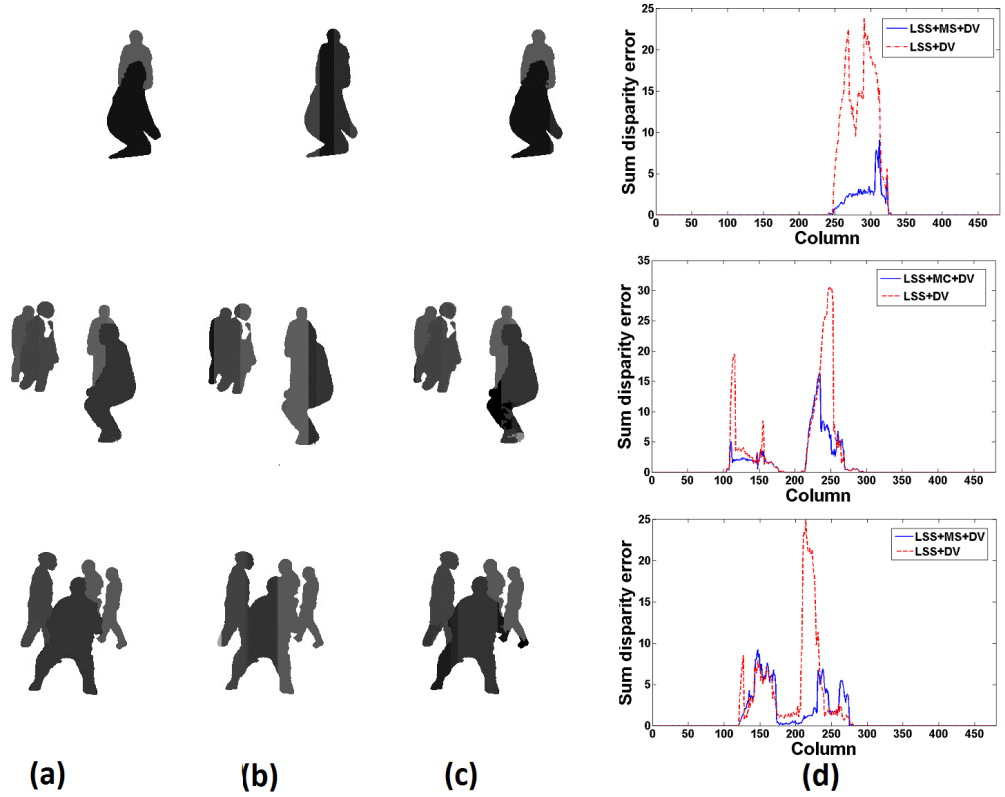
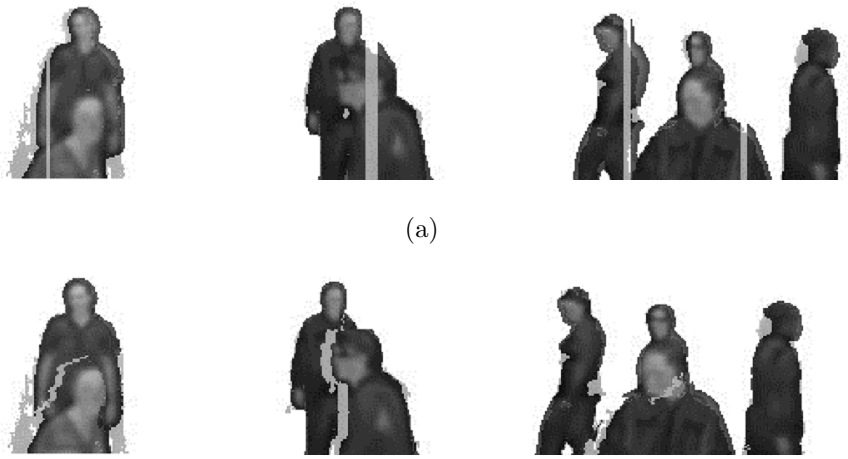


Figure 11: Comparison of LSS-based DV method and our proposed disparity assignment method(a) ground-truth disparity, (b) disparity estimation of DV matching using LSS as similarity measure (LSS+DV), (c) disparity estimation of our proposed method (LSS+MS+DV), and (d) Sum disparity errors over each column of pixels.



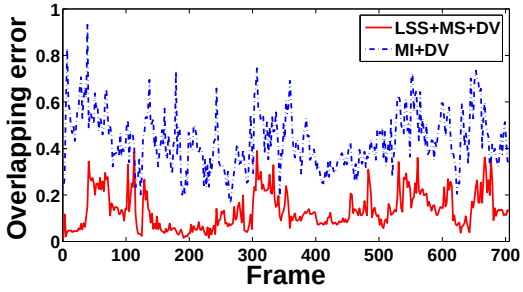
(a)

(b)

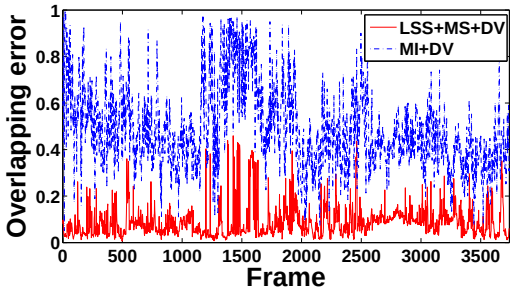
Figure 12: Comparison of LSS+DV and LSS+MS+DV detailed registration: (a) LSS+DV registration and (b) LSS+MS+DV registration.

5.2. Comparison of our LSS-based registration with the state-of-the-art MI-based registration

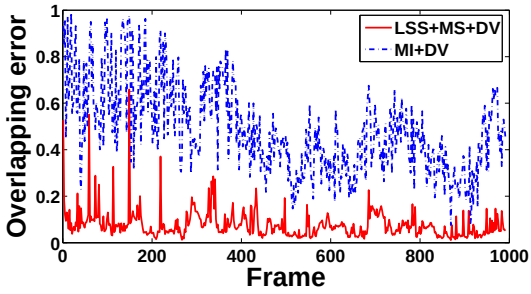
In order to demonstrate the improvement of our LSS-based registration method LSS+MS+DV compared to the state-of-the-art MI-based registration method, MI+DV, proposed by Krotosky and Trivedi [7], we qualitatively and quantitatively compared the two methods. Fig. 14 illustrates four examples of the disparity computation and the image registration results obtained using the two methods for our summer video. Note that our results are more accurate, especially for occlusions. Fig. 15 illustrates four examples for winter video 1. Note that MI+DV results are significantly poorer. These results demonstrate that for videos where there are falsely detected region as foreground and high differences of patterns inside human body ROIs, MI is not a reliable similarity measure. In contrast, LSS performs very well, except for



(a)



(b)



(c)

Figure 13: Overlapping error: (a) Summer video (702 frames), (b) Winter video 1 (3740 frames), and (c) Winter video 2 (992 frames)

few misalignments which occur for very small motion segments.

For a quantitative evaluation of the two registration methods, we defined an overlapping error that gives a quantitative estimate of the registration accuracy. The overlapping error is defined as,

$$E = 1 - \frac{N_{v \cap t}}{N_t}, \tag{13}$$

where $N_{v \cap t}$ is the number of overlapping aligned thermal foreground pixels on visible foreground pixels and N_t is the number of thermal foreground pixels. The best performance with zero overlapping error is when all the thermal pixels on the reference image have corresponding visible pixels on the second image. Note that our registration results are aligned thermal on visible images. This evaluation measure includes the background subtraction errors and also ignores misaligned thermal pixels which have falsely matched visible foreground pixels. However, since for both methods the background subtraction errors are included in the overlapping error, the differences between the two methods errors are still a good indicator for comparing overall registration accuracies for a large numbers of frames. Fig 13 illustrates the overlapping error using our LSS+MS+DV and MI+DV [7] methods for summer and winter videos. Based on table 1, the differences of mean overlapping error for the two methods over all frames (fig 13 (a)) are 0.30 for the summer video (fig 13 (a)), and 0.40 and 0.41 for the winter videos, (fig 13 (b) and (c), respectively). Also, table 1 shows the standard deviation of overlapping. The fluctuation of overlapping error for LSS+MS+DV method is much less than for MI+DV [7] method, especially for winter videos because of larger difference in textures on the objects. These results demonstrate

that our method performs more accurately and more consistently compared to MI+DV [7] method, especially for winter videos, in accordance with our qualitative results and previous discussions.

Table 1: Overlapping error (OE) for disparity voting (MI) and our proposed algorithm (LSS) with multiple people in the scene: frames with occlusion. SV: summer video, WV1 and WV2: winter videos, NO: number of objects, NF: number of frames, SM: similarity metric, and % OE (Ave - Std): average and standard deviation of overlapping error.

Video	NO	NF	SM	OE (Ave - Std)
SV	4	702	LSS	0.13 - 0.07
			MI	0.43 - 0.12
WV1	4	3740	LSS	0.09 - 0.06
			MI	0.49 - 0.17
WV2	5	992	LSS	0.07 - 0.07
			MI	0.48 - 0.19

6. Conclusion

In this paper, we applied LSS as a multimodal dense stereo correspondence measure and shown its advantages compared to MI, the most commonly used multimodal stereo correspondence measure in the state-of-the-art for human monitoring applications. We also proposed an LSS-based registration method, which addresses the accurate registration of regions associated to occluded people in different depths in the scene. In our results, we have shown the improvement of our registration method over the DV method proposed by [7]. Moreover, we have shown that our method significantly outperforms the state-of-the-art MI-based registration method in [7].

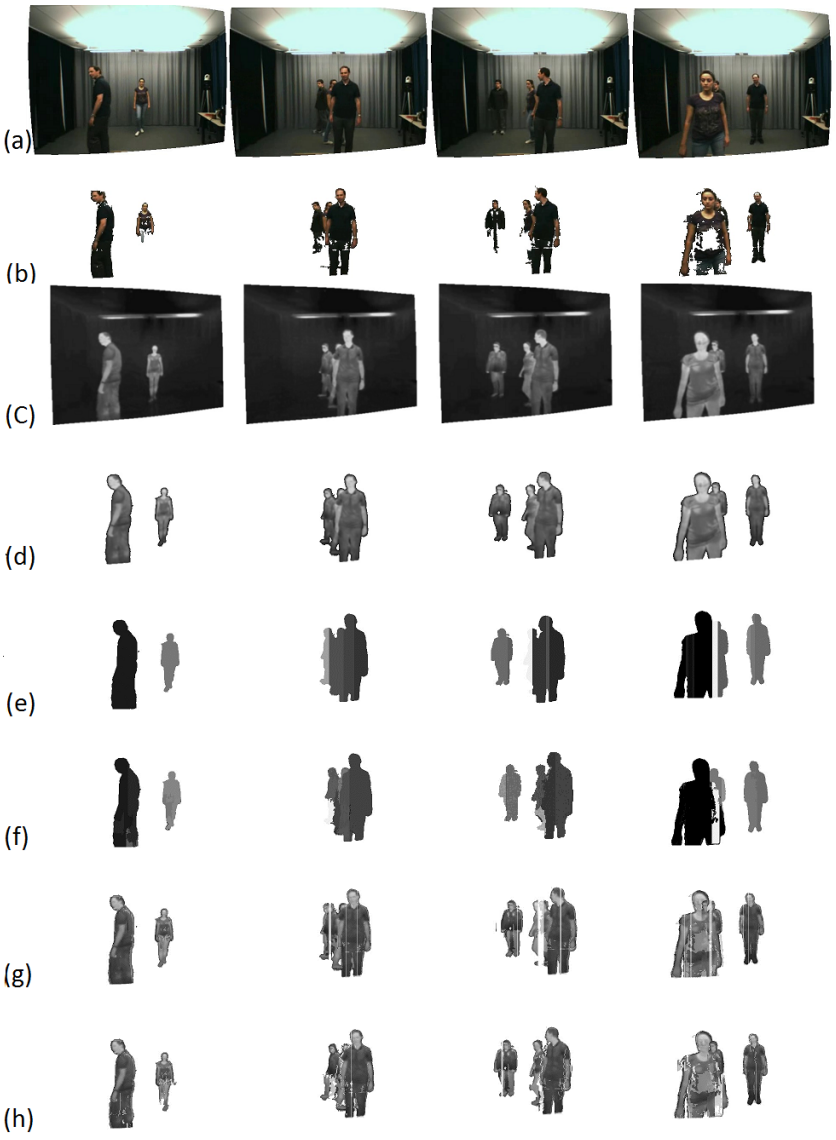


Figure 14: Comparison of MI+DV method in [7] and our proposed method LSS+MS+DV for our summer video using imperfect foreground segmentation (mainly misdetection). (a) visible image, (b) visible foreground segmentation, (c) thermal image, (d) thermal foreground segmentation, (e) MI+DV disparity image, (f) LSS+MS+DV disparity image, (g) MI+DV registration, and (h) LSS+MS+DV registration.

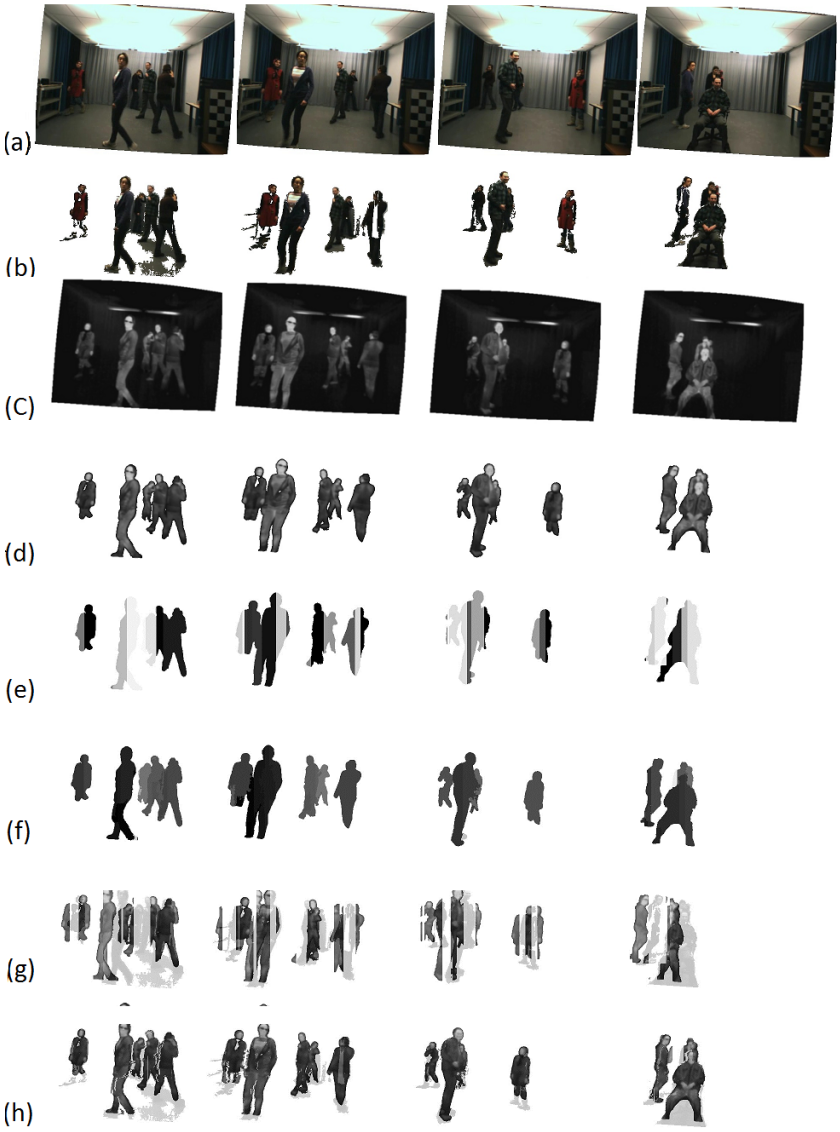


Figure 15: Comparison of MI+DV method in [7] and our proposed method LSS+MS+DV for our winter video using imperfect foreground segmentation (false detection and misdetection). (a) visible image, (b) visible foreground segmentation, (c) thermal image, (d) thermal foreground segmentation, (e) MI+DV disparity image, (f) LSS+MS+DV disparity image, (g) MI+DV registration, and (h) LSS+MS+DV registration.

As future direction for this work, we are working on improving the motion segmentation results to obtain more accurate segments and to avoid over segmentation.

References

[1] Z. Zhu, T. Huang, Multimodal surveillance: an introduction, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1 –6.

[2] R. Collins, A. Lipton, H. Fujiyoshi, T. Kanade, Algorithms for cooperative multisensor surveillance, Proceedings of the IEEE 89 (10) (2001) 1456 –1477.

[3] D. Socolinsky, Design and deployment of visible-thermal biometric surveillance systems, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1 –2.

[4] J. W. Davis, V. Sharma, Background-subtraction using contour-based fusion of thermal and visible imagery, Comput. Vis. Image Underst. 106.

[5] A. Leykin, Thermal-visible video fusion for moving target tracking and pedestrian classification, in: In Object Tracking and Classification in and Beyond the Visible Spectrum Workshop at the International Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[6] J. Han, B. Bhanu, Fusion of color and infrared video for moving human detection, Pattern Recognition 40 (6) (2007) 1771 – 1784.

[7] S. J. Krotosky, M. M. Trivedi, Mutual information based registration of multimodal stereo videos for person tracking, Computer Vision and Image Understanding 106 (2-3) (2007) 270 – 287.

[8] H.-M. Chen, P. Varshney, M.-A. Slamani, On registration of regions of interest (roi) in video sequences, in: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2003), 2003, pp. 313 – 318.

[9] C. Fookes, A. Maeder, S. Sridharan, J. Cook, Multi-spectral stereo image matching using mutual information, in: 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on, 2004, pp. 961 – 968.

[10] G. Egnal, Mutual information as a stereo correspondence measure, Tech. Rep. MS-CIS-00-20, University of Pennsylvania.

[11] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), 2007, pp. 1 –8.

[12] S. Walk, N. Majer, K. Schindler, B. Schiele, New features and insights for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), 2010, pp. 1030–1037.

[13] J. Yang, Y. Li, Y. Tian, L. Duan, W. Gao, Group-sensitive multiple kernel learning for object categorization, in: IEEE 12th International Conference on Computer Vision (ICCV 2009), 2009, pp. 436 –443.

[14] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, Multiple kernels for

object detection, in: IEEE 12th International Conference on Computer Vision (ICCV 2009), 2009, pp. 606 –613.

[15] A. Torabi, G.-A. Bilodeau, Local self-similarity as a dense stereo correspondence measure for themal-visible video registration, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, 2011, pp. 61 –67.

[16] P. O. Hoyer, P. Dayan, Non-negative matrix factorization with sparseness constraints, Journal of Machine Learning Research 5 (2004) 1457–1469.

[17] A. Torabi, M. Najafianrazavi, G. Bilodeau, A comparative evaluation of multimodal dense stereo correspondence measures, in: Robotic and Sensors Environments (ROSE), 2011 IEEE International Symposium on, 2011, pp. 143 –148.

[18] J. Heikkila, O. Silven, A four-step camera calibration procedure with implicit image correction, in: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, 1997, pp. 1106 –1112.

[19] A. Gil, O. M. Mozos, M. Ballesta, O. Reinoso, A comparative evaluation of interest point detectors and local descriptors for visual slam, Mach. Vision Appl. 21 (2010) 905–920.

[20] R. Mayoral, M. Aurnhammer, Evaluation of correspondence errors for stereo, Pattern Recognition, International Conference on 4 (2004) 104–107.

[21] B. Shoushtarian, H. E. Bez, A practical adaptive approach for dynamic background subtraction using an invariant colour model and object tracking, Pattern Recogn. Lett. 26 (1) (2005) 5–26.

[22] A. Ogale, Y. Aloimonos, A roadmap to the integration of early visual modules, International Journal of Computer Vision 72 (2007) 9–25.

[23] D. Comaniciu, P. Meer, Mean shift analysis and applications, in: The Proceedings of the Seventh IEEE International Conference on Computer Vision, (ICCV 1999), Vol. 2, 1999, pp. 1197 –1203 vol.2.