A LSS-based Registration Of Stereo Thermal-Visible Videos Of Multiple People Using Belief Propagation

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

In this paper, we propose a novel stereo method for registering foreground objects in a pair of thermal and visible videos of close-range scenes. In our stereo matching, we use Local Self Similarity (LSS) as similarity metric between thermal and visible images. In order to accurately assign disparities to depth discontinuities and occluded Region Of Interest (ROI), we have integrated color and motion cues as soft constraints in an energy minimization framework. The optimal disparity map is approximated for image ROIs using a Belief Propagation (BP) algorithm. We tested our registration method on several challenging close-range indoor video frames of multiple people at different depths, with different clothing, and different poses. We show that our global optimization algorithm significantly outperforms the existing state-of-the art method, especially for disparity assignment of occluded people at different depth in close-range surveillance scenes and for relatively large camera baseline.

Key words: IR camera, Visible camera, Dense stereo matching, Belief

*Corresponding author

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

propagation, Local self-similarity, Multimodal video surveillance

1. Introduction

In the recent years, by reduction in the price of infrared sensors, there has been a growing interest in visual surveillance using thermal-visible imaging system for civilian applications. The advantages of jointly using a thermal camera with a visible camera have been discussed comprehensively in [1, 2, 3, 4]. Combining visible and infrared information allows to better handling shadow, reflection, noise, misdetection, and missing information. 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.

A fundamental issue associated to data fusion of close-range thermal-visible imaging is accurately registering corresponding information and features of images with dramatic visual differences. For a close-range scene, matching corresponding features in a pair of visible and thermal videos is much more difficult than for a long-range scene. People might be in very different sizes due to their distances to the camera, in different poses, and at different levels of occlusion. They might have colorful/textured clothes that are visible in color images, but not in thermal images. On the other hand, there might be some textures observable in thermal images caused by the amount of emitted energy from different parts of the human body that are not visible in a color image. Due to the high differences between thermal and visible image characteristics, finding correspondence for entire scene is very challenging. Instead registration is focused on the foreground

ROIs.

The dense two-frame stereo correspondence is the only viable approach for registering possibly occluded objects at mutiple depths in the scene.Stereo matching is a well-studied subject for unimodal imaging system. An extensive taxonomy of two-frame stereo correspondence algorithms is described in [5]. However, this subject is new for multimodal visual surveillance applications. We summarize the problems associated to multimodal dense stereo as follows:

- **Dissimilar patterns**.This problem is specific to multimodal dense stereo. It is caused by the different types of image modalities. The corresponding regions in two images might be differently textured or one textured while the corresponding one is homogenous.

- **Depth discontinuities**. This difficulty is caused by segmentation results that contain two or more merged objects at different depths in the scene. In this case, correct disparities might be significantly different between neighboring pixels located on the depth boundaries.

- **Occlusions**. Some pixels in one view might be occluded in the other view. Therefore they should not be matched with pixels in the other view.

The main motivation of our proposed algorithm is the limitation of current approaches for registering occluded people ROIs. In this paper we present a global optimization algorithm for partial image ROI registration. we formulate a multimodal stereo matching in a Markov Random Fields (MRFs) framework using color and motion information as smoothness assumptions in order to elegantly handle depth discontinuities, occlusions, and non-informative pixels caused by

dissimilar patterns (corresponding pixels that do not contain similar visual information). Applying global optimization to multimodal stereo problem is challenging since most similarity measures, which are used for color images, are not viable for multimodal images. We integrate LSS as similarity measure in our global optimization algorithm.

The rest of the paper is organized as follows: The overview of the current multimodal registration approaches that gives insight about the limitations of exisiting methods is presented in section 2. In section 3, we describe the strengths of LSS as a viable image feature for matching thermal and visible images. In section 4, the overview of our registration system is presented, and, in section 5 each step of our algorithm is described in details. Our experiment is presented in section 6 and demonstrate that our method is efficient for video surveillance applications and outperforms the current state-of-the-art method. Finally, in section 7, we conclude this paper by describing the advantages and limitations of our algorithms.

2. Related Works

In the thermal-visible video surveillance research context, the majority of the image registration approaches are related to global image registration that globally transform a reference image on the second image. Krotosky and Trivedi give a comparative survey of multimodal registration approaches [6]. Global transformation approaches, either extract low-level image features such as edge features [7], or temporal-spatial features such as object trajectories [8, 9] to estimate a transformation matrix that transforms one image on another with the assumption that all the objects in the scene approximately lie in one depth plane. A few works in literature cover a video registration method appropriate for close-range people

monitoring. These methods have been categorized as partial image ROI registration [6].

In previous partial image registration approaches excluding ours [10, 11, 4], MI is the only similarity measure used in local dense correspondence algorithm for human monitoring applications [6, 12, 13]. The accuracy of MI as a similarity metric is directly affected by the MI window sizes. For unsupervised human monitoring applications, obtaining appropriate MI window sizes for the registration of multimodal pairs of images containing multiple people with various sizes, poses, distances to cameras, and different levels of occlusion is quite challenging. In the video surveillance context, Chen *et al.* proposed a MI-based registration method for pairs of thermal and visible images that matches windows on foreground regions in the two images with the assumption that each window contains one single depth plane [12]. In their method, the problem of depth discontinuity inside an ROI was not addressed. Later, Krotosky and Trivedi proposed a MI-based disparity voting (DV) matching approach [6]. Their method, for each ROI column, computes the number of votes related to each disparity and assigns a disparity with maximum votes. Their method theoretically considers depth discontinuities that may occur between neighboring columns, but it ignores vertical depth discontinuity where the pixels on a column belong to multiple depths. For example, two people with different heights, where the shorter person is in front of the taller one. To the best of our knowledge, in our context of visual surveillance, all the existing methods for multimodal stereo matching are local correspondence approach.

Recent global stereo algorithms have achieved impressive results by modeling disparity image as Markov Random Field (MRF) and determining disparities simultaneously by applying energy minimization method such as belief propaga-

tion [14, 15, 16], and graph cuts (GC) [17, 18]. Tappen and Freeman have shown that GC and BP produce comparable results using identical MRF parameters [19]. Sun *et al.* proposed a probabilistic framework to integrate into BP model, additional information (e.g., segmentation) as soft constraints [14]. Moreover, they have shown that the powerful message passing technique of BP deals elegantly with textureless regions and depth discontinuity problems. Later, Felzenszwalb and Huttenlocher proposed an efficient BP algorithm that dramatically reduced the computational time [15]. Their method is interesting for time sensitive applications like video surveillance. More recently, different extension of this efficient BP was proposed in several works [20, 21].

In our previous work, we have shown that local Self-Similarity (LSS), as a similarity measure, is viable for thermal-visible image matching and outperforms various local image descriptors and similarity measures including MI, especially for matching corresponding regions that are differently textured (high differences) in thermal and visible images [11]. Also we presented an extensive study of MI and LSS as similarity measure for human ROI registration in [4]. In [10, 4], we proposed a LSS-based local stereo correspondence using disparity voting approach for close-range multimodal video surveillance applications. In this work, we adopt LSS as similarity measure in an energy minimization stereo model using the efficient BP model [15].

3. MI And LSS For Multimodal Image Registration

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

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

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.

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 [22]. 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

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

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. LSS has two interesting characteristics for our application: 1) LSS is computed separately as a set of descriptors in one individual image and then it is compared between pair of images. In contrast, MI is computed directly between the two images. This characteristic makes LSS viable to be used in a global correspondence approach. 2) The measurement unit for LSS is a small image patch that contains more meaningful patterns compared to a pixel as used for MI computation. This property makes LSS describing layout accurately without being too sensitive to detailed texture variances. For multimodal human ROI matching, where human body have similar layouts in both modalities but they are not identical in textural appearance, LSS is a powerful feature.

In our application, before matching the LSS descriptors between pair of thermal and visible images, we discard the non-informative ones using a simple method. Non-informative descriptors are the ones that do not contain any self-similarities (e. g. the center of a small image patch is salient) and the ones that contain high self-similarities (a homogenous region with a uniform texture/color). A descriptor is salient if all its bin's values are smaller than a threshold. The homogeneity is detected using the sparseness measure of [23]. Discarding non-informative descriptors is like an implicit segmentation or edge detection, which 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. Figure 1 shows pixels having informative descriptors (white pixels) for a pair of thermal and visible images. The regions belonging to the human body boundaries and some image patterns are the informative regions.

(a) (b)

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

4. Overview Of Our Approach

Our registration algorithm is designed for video surveillance systems where the input data is a pair of synchronized thermal and visible videos. In our algorithmic design, it is feasible to add a new module for higher level processing, such as tracking. However, in this work, we only focus on the registration algorithm. The overall algorithm consists of several steps as shown in figure 2. At each time step t, the input data of our system is a rectified pair of thermal and visible frames at t and rectified visible frame at $t-1$. For the visible spectrum, two consecutive frames are needed to compute the optical flow in a later step of our algorithm. Due to the high differences in imaging characteristics of thermal and visible sensors, our registration is focused on the pixels that correspond to ROIs. As the first step of our algorithm, we extract image ROIs on pair of thermal and visible images using a background subtraction method [24]. Each image ROI is defined by its bounding box. The registration is applied on the pixels inside the box. In the thermal spectrum, a bounding box is surrounding a foreground region at time t. In the visible image, a bounding box is surrounding overlapping foreground

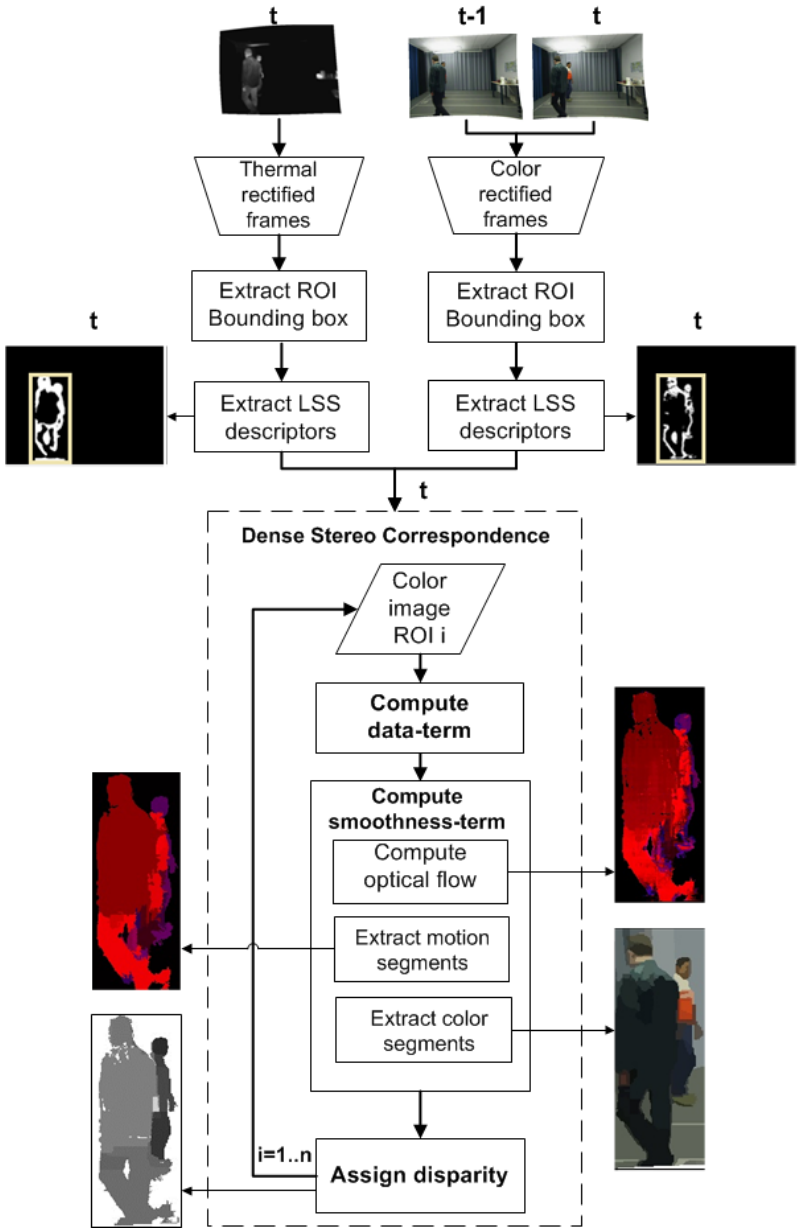


Figure 2: Block diagram of thermal-visible dense stereo matching algorithms augmented with input images, intermediate and disparity image results.

regions at time $t-1$ and t. In this way, for efficiency, the optical flow computations (later step) are performed only inside the visible image bounding box. The next step is extracting LSS descriptors for foreground pixels inside the bounding boxes at frame t. In figure 2, the image results of this step show pixels with informative LSS in white and non-informative ones in black (informative pixels are determined using the method described in section 3).

The main body of our registration algorithm begins after LSS feature extraction. Registration is done by matching visible ROIs on thermal ROIs. The reason for matching visible ROIs on thermal ROIs is that for color image, both color and motion cues are available to be used as complementary image cues in our registration model. However, for thermal image, the color cue is not defined. In our matching strategy, each bounding box on visible image is viewed as a smaller image. Registration is done separately for each bounding box. Disparities are assigned to all pixels inside a box using a global optimization that minimizes an energy function which is described in details in the following sections. Our energy function consists of a data term and a smoothness term as shown in the dotted block in figure 2. The data term is computed based on self-similarities matching between pixels that contain informative LSS descriptors. The smoothness term is computed using motion and color cues of pixels inside a bounding box in the visible image. To extract the motion cues, we compute the optical flow using a state-of-the-art method [25]. Then, we use mean-shift segmentation to cluster the motion vector fields extracted in the previous step [26]. To extract the color cues, we apply the same mean-shift segmentation on pixel intensities to compute the color segmentation. Figure 2 shows results of optical flow, motion segmentation, and color segmentation. Finally, the disparities are assigned to pixels inside the

bounding box using an efficient belief propagation method [15].

5. Detailed Description

We assume that a bounding box may contain one or more human body ROIs and background. In this section, we give a detailed description of our proposed multimodal dense stereo correspondence algorithm.

5.1. Thermal-Visible Stereo Model

We formulate the registration as a multi-labeling problem (we use the notation from [15]). We assume that P is the set of all pixels inside the image bounding box and that L is a set of labels, which are disparity quantities in our problem. A labeling f assigns a label $f_p \in L$ to each pixel $p \in P$. We model our stereo matching using a Markov Random Field (MRF) framework and estimate the quality of labeling using an energy function defined as,

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{(p,q) \in N} V(f_p, f_q). \tag{3}$$

where N are edges in the image graph and p represents a pixel. In our image graph, we use a four-connected neighborhood system. In this energy function, the first term is referred as the data term, which is the cost of assigning label f_p to pixel p. The second term is the smoothness term, which represents the cost of assigning labels f_p and f_q to two neighboring pixels.

5.2. Data Term

The data term only encodes the similarity distance of informative LSS descriptors on matching thermal and visible pixels for a preset disparity range. The

similarity distance is basically the $L1$ distance between two informative LSS descriptors on a pair of thermal and visible images.

$$D_p(f_p) = \begin{cases} L1(p_l, p_r) & \text{if } p_l, p_r \in \text{informative} \\ 1 & \text{otherwise} \end{cases}, \tag{4}$$

where p_l is the LSS descriptor of pixel p inside bounding box on the visible image and p_r is the LSS descriptor of possible matching pixel of p on the corresponding line of thermal image by disparity f_p. In our dataterm, if two matching pixels are containing informative LSS descriptors (more details section 3); we compute a normalized $L1$ distance as data term. Otherwise we, simply assign the maximum possible value for data term since matching is not defined if one of the pixels either on thermal or visible does not contain an informative descriptors. Then, we map the data term to values between $[0-255]$ as pixel intensity interval values.

5.3. Smoothness Term

In our stereo model for pair of thermal-visible videos, the smoothness term has a crucial role for passing the influence of messages from pixels with informative LSS far away to non-informative ones, while the influence in the depth discontinuous regions should fall off quickly. For this reason, we incorporated visual cues including motion and color segmentation in the stereo model as soft constraint to accurately determine disparities. The main advantage of this approach rather than a segment-based stereo algorithm such as [21], which assumes that depth discontinuity occurs on the boundary of segmented regions as a hard constraint, is that messages are still passed between segmented region; therefore it is more robust to incorrect segmentation results. In the following, we describe how we incorporate motion and color in our smoothness term.

Figure 3: (a) Image window (b) Foreground (c) Optical flow (d) Motion segments.

5.3.1. Motion

Since our data are videos of moving people at different depths in the scene, we incorporated the motion information in our smoothness term. Motion segmentation is a visual cue that provides a reasonable estimate of existing depth planes in the scene. We assume that each human ROI includes one or more motion segments, but each motion segment belongs to one and only one human ROI. In order to imply this assumption, we apply some level of over-segmentation. Thus, as a soft constraint, we consider that disparity discontinuities take place at some motion segment boundaries. However, not all the motion segment boundaries represent depth discontinuities.

We apply a simple two-frame motion segmentation using two consecutive color image frames $t-1$ and t. Firstly, we compute the motion vector field for all pixels (including foreground and background) inside the window of an ROI using an optical flow method based on block-matching [25]. Second, we apply the mean-shift segmentation method proposed in [26] (on foreground pixels) for segmenting the motion vector field computed in the previous step, and for assigning a mean velocity vector to each segment. We apply motion segmentation only on foreground regions inside the image window at frame t in order to extract also a segment associated to temporary stationary person for which its mean velocity

vector is zero. Mean-shift segmentation is applied on (2+2) feature point dimensions, where two dimensions are related to spatial dimensions and the two others are related to the two motion vector components in x and y directions. Figure 3 shows the motion segmentation result of three merged people in one ROI where two people are moving and the other one is temporary stationary. In order to visualize the motion segments, motion vectors are mapped to HSV color space. Our motion segmentation results in a set of regions $SM = \{sm_1, ..., sm_m\}$ inside the image window.

There are three difficulties associated with motion segmentation. First, an image ROI belonging to objects closer to the camera might be too over-segmented and fragmented into several motion segments. Second, imperfect foreground segmentation causes some pixels inside an ROI not being assigned to any motion segments. Figure 4(a) and (b) show an example of over segmentation; (c) and (d) an example of imperfect background subtraction. Third, the occluded pixels at frame (t-1), which are visible at frame t, have no defined motion vectors. This last difficulty causes inaccurate motion segment boundaries that do not correspond to actual depth discontinuities in the image. Figure 5 shows an example of motion segmentation where the motion segment boundaries are inaccurate due to the existing occluded pixels. Applying motion segmentation on foreground regions eliminates those occluded pixels which are part of background. However, those which are inside an ROI containing two people like in our example, cause inaccurate motion segment boundaries. In order to avoid inaccurate disparity assignment caused by imperfect motion segmentation, we apply color segmentation as a complementary visual cue.

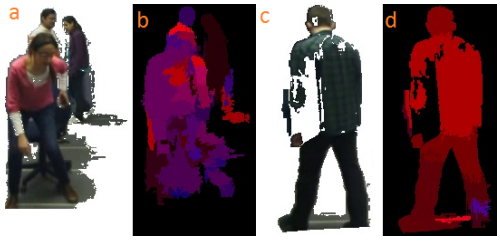


Figure 4: (a) Foregorund visible,(b) motion segmentation, example of oversegmentation, (c)Foreground visible, (d) Motion segmentation, example of misdetected regions

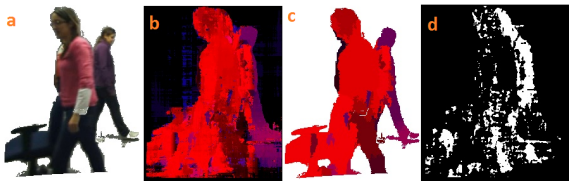


Figure 5: (a) Foreground visible (b) Optical flow (c) Motion segmentation (d) Occluded pixels (white pixels).

5.3.2. Color

We integrate the color visual cue as complementary information in our smoothness term to handle the three difficulties caused by motion segmentation. In fact color segmentation helps to more easily pass the influence of messages to neighboring pixels associated to previously aforementioned motion segmentation problems, while they are in a same color segment. We perform the color segmentation on all the pixels inside an image window to ensure that the pixels which were discarded from motion segments due to erroneous foreground regions are assigned to a color segment.

Color segmentation is done using the same mean-shift segmentation that we applied for motion segmentation [26]. In figure 2, the color segmentation block shows an example of our segmentation. We use an over segmentation to avoid

merging color regions belonging to more than one people.

5.3.3. Integrating Multiple Cues

The smoothness term encodes the prior information of the blob including motion segmentation and color segmentation as follows,

$$V(f_p, f_q) = \begin{cases} \alpha|f_p - f_q| & \text{if } p,q \in MS \wedge p,q \notin O \\ \beta|f_p - f_q| & \text{elseif } p,q \in CS \\ |f_p - f_q| & \text{otherwise} \end{cases} . \quad (5)$$

In our smoothness term, if two neighbor pixels p and q belong to same motion segment (MS) and they are not occluded pixels (O), the discontinuity cost is weighted by a constant α and increases with the distance between the two assigned disparities f_p and f_q. As a complementary cue, for the neighboring pixels which did not satisfied the previous condition, but that are in the same color segment, the discontinuity cost is defined in the same way, however weighted by another constant β. Finally, for the pixels which did not satisfy any of two previous conditions, the discontinuity cost is defined by the distance between the two assigned disparities. We define the constant value of β slightly higher than α to make the cost of assigning two different disparities to neighboring pixels inside one color segment slightly higher. The reason is that the confidence of color segment using over segmentation is higher than motion. In other words, pixels inside one color segment are more likely to belong to one and only one person in the scene than the motion segment.

5.4. Disparity Assignment

In our algorithm, an optimal labelling with minimum energy is approximated using the efficient loopy belief propagation proposed by Fezenswalb and Hut-

tenlocher [15]. Their method substantially reduce the complexity time of belief propagation approach from $O(nk^2T)$ to $O(nkT)$, where n is the number of pixels (nodes), k is number of possible disparities (labels), and T is the number of iteration. For stereo problem modeled in term of posteriori probabilities, BP algorithm is used for performing inference on MRFs by applying the max-product algorithm [14]. The equivalent computation used in [15] is negative-log probabilities, where the max-product becomes min-sum and the energy function definition (equation 3) can be used directly.

BP is based on a powerful iterative message passing on an image grid where each pixel represents a node and edges are connecting neighboring pixel using four-connection (up, down, right, and left). Messages are passed through the edges asymmetrically and adaptively to deal with textureless regions and depth discontinuities elegantly. A message between two nodes p and q at iteration i is defined as

$$m^i_{pq}(f_q) = Min_{f_p}\left(V(f_p,f_q) + D_p(f_p) + \sum_{r\in N(p)-q} m^{i-1}_{rp}(f_p)\right), \tag{6}$$

where $N(p)-q$ are the neighbors of node p other than q. And m^{i-1}_{rp} is the message send to pixel p from neighbor r (excluding q) in previous iteration $i-1$. After N iteration when the energy is minimized, in other words, when the disparity assignment has converged to optimal solution, a belief vector is computed for each node as,

$$b_p(f_p) = D_p(f_p) + \sum_{q\in N(p)} m^N_{qp}(f_p). \tag{7}$$

Finally, the disparity (label) which individually is assigned to each pixel p is the label with minimum value in final belief vector. In our implementation of efficient

(a) (b) (c)

Figure 6: (a) Camera setup. The halogen lights behind the cameras are used for calibration, (b) visible calibration image and (c) thermal calibration image.

BP [15], we used two of their techniques to speed up the processing time. First, by using their message updating that reduces the computational complexity from $O(k^2)$ to linear time $O(k)$. Second, by using their alternating message updating techniques for bipartite graph (like an image grid), which reduces the number of update message in each iteration to half. More details can be found in [15].

6. Experiments

6.1. Experimental setup

We tested our method using visible-thermal synchronized videos of a $5m \times 5m$ room at a fixed temperature of 24 $^{\circ}C$. The videos were recorded by stationary thermal and visible cameras with baselines of $10cm$ and $13cm$. The videos include up to five people moving throughout the scene. People have colorful, thick, or light clothes, which appear differently textured in thermal and visible images. Moreover, they may also carry objects, such as a bag that is only visible in one image modality. Figure 6 shows our camera setup and examples of calibration images in visible and thermal.

In order to simplify the matching to a $1D$ search, the thermal and visible cam-

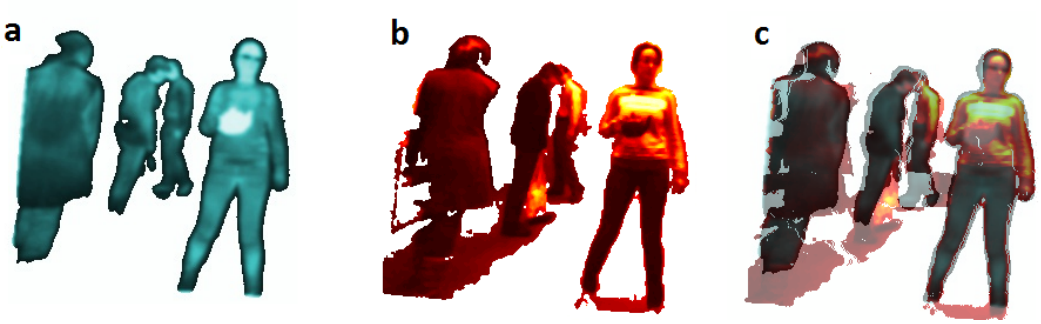


Figure 7: Detailed registration a person carrying a hot pot. (a) Foreground thermal image (green image), (b) Foreground visible image (red image), and (c) Registration of visible image on thermal image (overlayed red image on green image).

eras were calibrated using the standard method described in [27] and implemented in the camera calibration toolbox of MATLAB [28]. Since in the thermal images, the calibration checkboard pattern is not visible at room temperature, we illuminated the scene using high intensity halogen bulbs placed behind the two cameras. In this way, the dark squares of the checkboard absorb more energy and appear visually brighter than the while squares in the thermal images.

Figures 7 and 8 illustrate two examples of successful registration of visible image on thermal foreground images using our algorithm. Column (a) represents the foreground thermal image (green image), column (b) is foreground visible image (red image), and finally column (c) displays the registered images (i.e. correctly aligned and mis-aligned regions). Also, these two figures illustrate the benefit of combining thermal and visible information. People are at different depth levels and with different clothing (such as wearing scarf or jacket). Background subtraction is imperfect and includes false positive (shadows) and false negative (partial misdetections) errors. In figure 7, a person carries a hot pot that is clearly distinguishable in thermal image, but not as easy to detect in the visible image. In figure

Figure 8: Detailed registration of a person carrying a bag. (a) Foreground thermal image (green image), (b) Foreground visible image (red image), and (c) Registration of visible image on thermal image (overlayed red image on green image).

8, a person is carrying a bag at room temperature, and hence is not detected in the thermal image. Our global optimization approach has successfully estimated correct disparity for the bag region since it is connected to the person region in the image.

In order to assess our registration for video surveillance applications, we compared our proposed Local Self Similarity based Belief Propagation algorithm ($LSS + BP$) with the state-of-the-art Mutual Information based Disparity Voting algorithm ($MI + DV$) in [6] and with our previous work, Local Self Similarity based registration using DV matching ($LSS + DV$) in [10, 4]. We focus on two main aspects that demonstrate the efficiency of our method compared to previous works: 1) depth discontinuity handling of occluding/occluded people, and 2) the effect of different disparity ranges, whether small or large, on the registration performance. In the following sections, we present our comparative evaluation regarding these two aspects.

(a) (b)

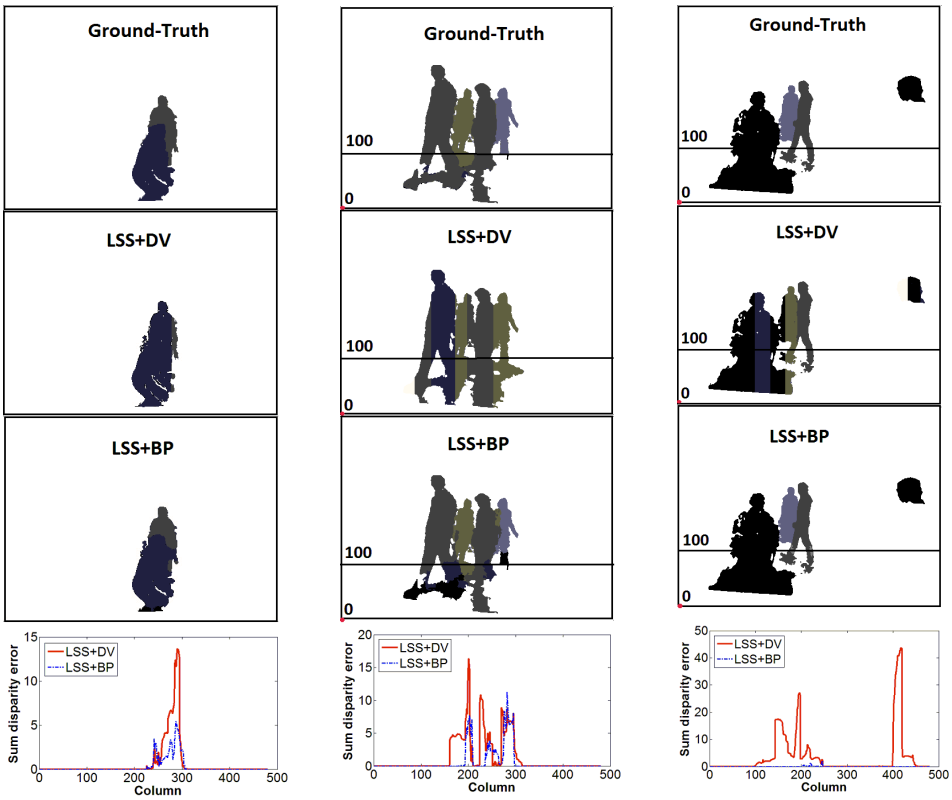
(c)

Figure 9: Three examples of our tested video frames.

6.2. Evaluation Of Disparity And Registration Accuracy In Occlusions

In order to demonstrate the disparity accuracy improvement of our matching approach compared to state-of-the-art DV matching approaches [6, 10, 4] for occlusion handling, we quantitatively compared the disparity results of our proposed *BP* and of *DV*. In order to perform a fair comparison, we use LSS as similarity measure in the two approaches. We generated ground-truth disparities for visible image by manually aligning visible foreground ROIs on corresponding thermal image.

Figure 9 shows three example frames of our tested video. Unlike previous work [6], our tested videos are realistic videos containing people in different poses (e.g. sitting, bending, and walking) rather than only walking people. Accordingly, figure 10 illustrates the comparison of disparity map for visible foreground pixels computed by $LSS + BP$ and $LSS + DV$ for three examples displayed in figure 9. In figure 10, for better visualization, the disparity levels in disparity maps are



(a) (b) (c)

Figure 10: Comparison of the disparity accuracy of $LSS + DV$ and $LSS + BP$ methods. Respectively columns(a), (b), and (c) are ground-truth (first row), $LSS + DV$ disparity map (second row), $LSS + BP$ disparity map (thrid row), and sum disparity error (forth row) for visible foreground pixels of examples in figure 9.

mapped to colors. Figure 9 column (a) illustrates an example where two people at two different depths in the scene appear in a single region. The columns (b) and (c) show the examples where multiple people are occluded. Figure 10, second row, shows $LSS + DV$ method fails to assign correct different disparities to the columns containing pixels related to more than one disparity level in occluded regions. In order to register people merged in a single region, DV method makes no assumptions about the assignment of pixels to individual person and assigns a single disparity to each column inside a ROI, based on a voting approach. If pixels on a column of image belong to different objects at different depths in the scene, the vote only goes for one of them based on WTA approach. However, $LSS + BP$ (third row) succeeds in assigning different disparities to the human ROIs using a global optimization. In $LSS + BP$, color and motion cues, integrated as soft constraints in an energy function, gives a reasonable estimate of moving regions belonging to people in the scene.

Accordingly, figure 10 last row displays the sum of disparity errors of the columns corresponding to visible foreground pixels of the three examples. In general, disparity error is higher for $LSS + DV$ method compared to LSS+BP method. However, in a few number of columns of plots (a) and (b), $LSS + BP$ has a slightly higher sum of disparity error. As it is shown in figure 10 (b) and (c), the sum disparity error is computed for the upper part of the image starting from row 100 in order to discard the disparity error caused by falsely detected regions belonging to shadows on the ground. Note that eliminating the lower part of image belonging to people's legs and the ground helps to better visualize the comparison between the disparity accuracy of two methods for occlusion handling.

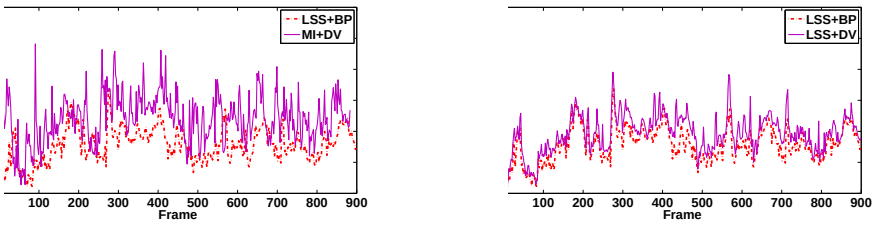
Figure 11 illustrates detailed registration of three video frames of people at

Figure 11: Comparison of $LSS+DV$ and $LSS+BP$ methods registration accuracy (large disparity range of $[5-50]$ pixels):(a) $LSS+BP$ detailed registration, (c) $LSS+DV$ detailed registration.

different levels of occlusion using $LSS+BP$ and $LSS+DV$ methods for a relatively large disparity range between $[5-50]$ pixels. In these examples, $LSS+DV$ fails to accurately register pixels related to depth discontinuity regions. In the following, we discuss the effect of a wide disparity range for WTA local matching approach such as DV compared to our proposed algorithm.

6.3. Evaluation Of Registration Accuracy Using Different Disparity Ranges

In this part of our experiments, we compared the registration results of $MI+DV$ [6], $LSS+DV$ [10, 4], and our proposed $LSS+BP$ for two videos using disparity ranges of $[2-20]$ pixels and $[5-50]$ pixels where in both videos, up to five people are walking throughout the scene. In order to perform a fair comparison, both videos are recorded in the same room with similar environmental factors but for one video, the camera baseline is $10cm$ and for the other one it is $13cm$. In



(a) (b)

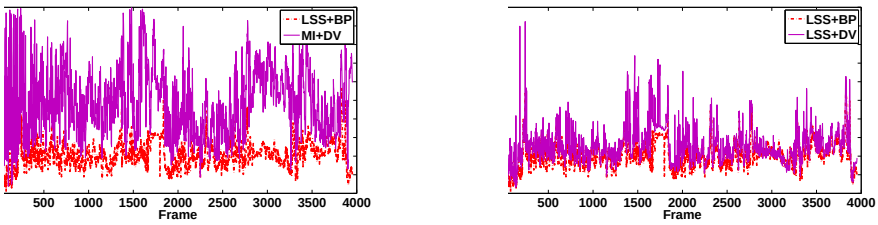
Figure 12: Overlapping error using disparity range $[2-20]$: (a) LSS+BP vs. MI+DV and (b) LSS+BP vs. LSS+DV.

order to perform a quantitative evaluation of the registration performance of the algorithms, we defined an overlapping error that gives an estimate of the registration errors. The overlapping error is defined as,

$$E = 1 - \frac{N_{v \cap t}}{N_t}, \tag{8}$$

where $N_{v \cap t}$ is the number of overlapped thermal and visible foreground pixels and N_t is the number of visible foreground pixels. The best performance with zero overlapping error is when all the visible pixels on the reference image have corresponding thermal pixels on the second image (we register the visible on the thermal image). This evaluation measure includes the background subtraction errors and also ignores misaligned visible pixels inside foreground regions of thermal image. However, since for the three methods, the background subtraction errors are included in the overlapping error, the differences between the overlapping errors are still good indicators for comparing overall registration accuracies for a large numbers of frames.

For DV methods, we used matching window size of 30 pixels wide that we experimentally found to have the minimum mean overlapping errors among the three

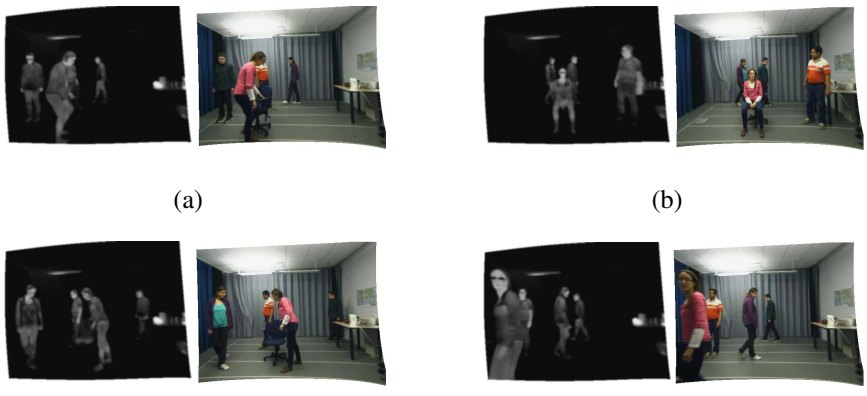


(a) (b)

Figure 13: Overlapping error using a disparity range of $[5-50]$: (a) LSS+BP vs. MI+DV and (b) LSS+BP vs. LSS+DV.

size of 10, 20, and 30 pixels. For DV, 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 especially in occlusion region. Therefore, we start our search by smaller windows sizes up to 30 pixels width. 30 pixels is the average of a human ROI width in our tested videos. Applying window size bigger than 30 pixels width is not efficient since the DV method will ne be able to compute accurate disparities for occluded regions (i.e. more than one person merged in a single image ROI in the image).

Figure 12 (a) and (b) illustrates overlapping error of $LSS+BP$ vs. $MI+DV$ and $LSS+BP$ vs. $LSS+DV$ over 900 video frames respectively. The mean overlapping error of $MI+DV$ is 0.24, $LSS+DV$ is 0.19, and $LSS+BP$ has the minimum error among the three methods which is 0.15. $LSS+DV$ has the second place and $MI+DV$ is the least accurate. However, the three methods have reasonable overlapping errors and are stable over 900 frames, considering the background subtraction errors as well. The standard deviation (std) value of $LSS+BP$ is 0.05, $LSS+DV$ is 0.06, and $MI+DV$ is 0.07. Again, $LSS+BP$ has the most stable



(a) (b)

(c) (d)

Figure 14: Example of Tested video frames of video with a disparity range of [2-20].

performance.

Figure 13 (a) and (b) illustrates overlapping error of $LSS+BP$ vs. $MI+DV$ and $LSS+BP$ vs. $LSS+DV$ over 4000 video frames respectively. For DV methods, we used matching window size of 30 pixels. The mean overlapping error of $MI+DV$ is 0.49, $LSS+DV$ is 0.25, and $LSS+BP$ is 0.20. Similarly to the previous experiment, $LSS+BP$ has the minimum error among three methods, $LSS+DV$ has the second place, and $MI+DV$ is the least accurate. The std value of $LSS+BP$ is 0.07, $LSS+DV$ is 0.25, and $MI+DV$ is 0.18. It is should be noted that for all three methods, overlapping errors have increased. However, compared to the other video, it is observable that the mean overlapping error of DV methods, especially $MI+DV$ significantly increased. Moreover, they have larger number of overlapping error outliers (large std) compared to the previous video, which shows some performance instabilities over the whole video. Furthermore, $LSS+DV$ performs better than $MI+DV$. This shows that LSS used as similarity metric is a more robust feature for multimodal matching compared MI in the case of visible

and infrared images. $BP + LSS$ was less influenced by the change of disparity range.

The main reason of the significant performance decrease of DV methods is that a larger disparity range used for horizontal matching increases the probability of false matching using a WTA approach, especially for scenes with imperfect foreground regions and corresponding regions that are differently textured in thermal and visible images. However, our proposed BP method that uses a BP global optimization approach is more robust, especially using larger disparity ranges. The overlapping error does not increase dramatically as the overlapping errors of DV methods.

Figure 14 shows four examples of tested video frames using a disparity range of $[2-20]$. For these video frames, figure 15 illustrates qualitatively the resulting disparity maps, and registrations of visible foreground image on thermal foreground image using $LSS + BP$, $LSS + DV$, and $MI + DV$. Figure 15, rows (d) and (e) show the disparity maps for the DV methods. In both methods, disparity assignments are inaccurate for depth discontinuity regions. However, $LSS + DV$ computes more accurate disparity map. Figure 15, rows (c) shows the disparity map of $LSS + BP$ method. It has more accurate results, especially for depth discontinuity regions. However, the last column shows some color and motion over-segmentation for the person close to the camera that results in less smooth disparity map inside the human body ROI compared to the farther objects.

7. Conclusions

In this paper, we proposed a stereo model for thermal-visible partial ROI registration using an efficient belief propagation algorithm that outperforms previous

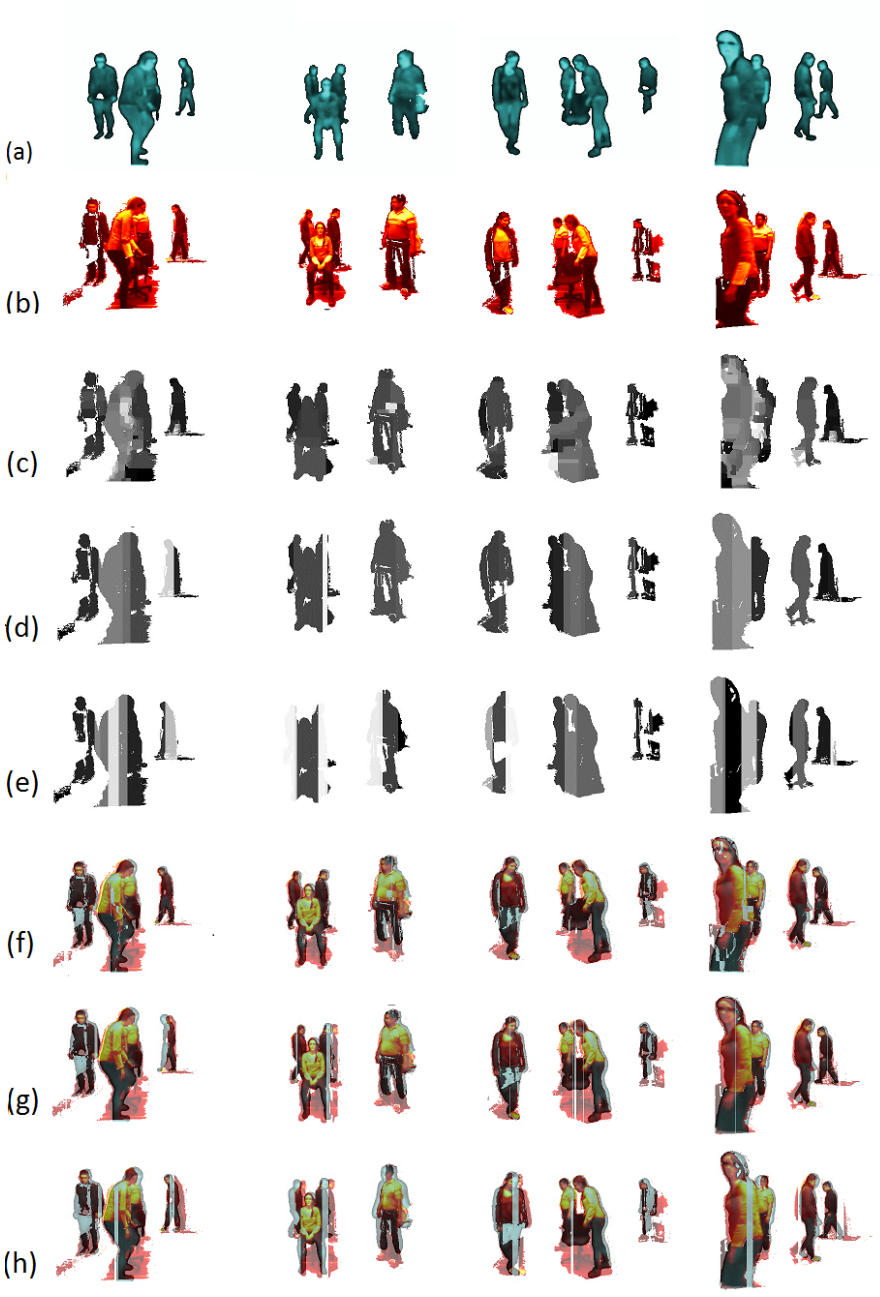


Figure 15: Qualitative Comparison: (a) thermal foreground image (green image), (b) visible foreground image (red image) (c) disparity map $LSS+BP$, (d) disparity map $LSS+DV$, (e) disparity map $MI+DV$, (f) registration of visible on thermal $LSS+BP$, (g) registration of visible on thermal $LSS+DV$, (h) registration of visible on thermal $MI+DV$.

state-of-the-art stereo registration designed for close range video surveillance applications. We have tested our methods on two indoor videos, and presented registration results over 4900 frames. Our results demonstrate that our method assigns more accurate disparity to pixels related to depth discontinuity regions and that it is more stable for large disparity range compared to previous works [6, 10, 4].

For video surveillance applications, processing time is an important factor. The processing time of our algorithm for each frame is approximately 2-6 seconds using a 3.40GHz multi-core desktop processor, while for *DV* method, it is between 1-3 seconds. For both methods, the processing time varies based on the number and size of foreground ROIs in the images and as more people are in the field of view of the cameras. Moreover, in our method, the number of iterations of belief propagation algorithm varies for different ROIs depending on the rate of for converging to the minimum energy (when between two consecutive iterations the energy over MRF nodes has not decreased). In our implementation we used lookup tables and parallel processing programming (openMP) in C++ to speed up the processing time significantly.

The registered thermal and visible images obtained using our algorithm can be used for further data analysis including tracking, behaviour pattern analysis, and object categorization based on the complementary information provided by data fusion.

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] A. Torabi, G.-A. Bilodeau, Local self-similarity-based registration of human rois in pairs of stereo thermal-visible videos, Pattern Recognition 2012.

[5] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision 47 (2002) 7–42.

[6] S. J. Krotosky, M. M. Trivedi, Mutual information based registration of multimodal stereo videos for person tracking, Comput. Vis. Image Underst. 106 (2-3) (2007) 270–287.

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

[8] A. Torabi, G.-A. Bilodeau, An iterative integrated framework for thermalvisible image registration, sensor fusion, and people tracking for video surveillance applications, Computer Vision and Image Understanding 116 (2) (2012) 210 – 221.

[9] A. Torabi, G. Masse, G.-A. Bilodeau, Feedback scheme for thermal-visible video registration, sensor fusion, and people tracking, in: Computer Vision

and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010, pp. 15 –22.

[10] 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.

[11] 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.

[12] 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.

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

[14] J. Sun, N.-N. Zheng, H.-Y. Shum, Stereo matching using belief propagation, Pattern Analysis and Machine Intelligence, IEEE Transactions on 25 (7) (2003) 787 – 800.

[15] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient belief propagation for early vision, Int. J. Comput. Vision 70 (2006) 41–54.

[16] Q. Yang, L. Wang, R. Yang, H. Stewenius, D. Nister, Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion

handling, Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (3) (2009) 492 –504.

[17] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, Pattern Analysis and Machine Intelligence, IEEE Transactions on 23 (11) (2001) 1222 –1239.

[18] M. Bleyer, M. Gelautz, Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions, Image Commun. 22 (2007) 127–143.

[19] M. Tappen, W. Freeman, Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, 2003, pp. 900 –906.

[20] Q. Yang, L. Wang, N. Ahuja, A constant-space belief propagation algorithm for stereo matching, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 1458 –1465.

[21] A. Klaus, M. Sormann, K. Karner, Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure, in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, Vol. 3, 2006, pp. 15 –18.

[22] 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.

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

[24] 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.

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

[26] 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.

[27] 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.

[28] J.-Y. Bouguet, Camera calibration toolbox for matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/.