A comparative evaluation of multimodal dense stereo correspondence measures

Atousa Torabi, Mahya Najafianrazavi, and Guillaume-Alexandre Bilodeau
LITIV Lab.
École Polytechnique de Montréal
Montréal, QC, Canada, P.O. Box 6079, H3C 3H7
Email: {atousa.torabi, mahya.najafianrazavi, guillaume-alexandre.bilodeau}@polymtl.ca

***Abstract*—In this paper, we compare the behavior of four viable dense stereo correspondence measures, which are Normalized Cross-Correlation (NCC), Histograms of Oriented Gradients (HOG), Mutual Information (MI), and Local Self-Similarity (LSS), for thermal-visible human monitoring. Our comparison is based on a Winner Take All (WTA) box matching stereo method. We evaluate the accuracy and the discriminative power of each correspondence measure using challenging thermal-visible pairs of video frames of different people with different poses, clothing, and distances to cameras for close-range human monitoring applications.**

I. Introduction

Multimodal imagery for human analysis has a variety of application domains, such as in-vehicle safety systems, medical monitoring, and video surveillance. For many applications, the joint use of two or more different imaging modalities improves the quality of the processed output. For example, for the extraction of a human body region of interest (ROI) in visible images, there are some difficult situations, such as color similarity of the human body/clothing with the background, or low contrast and reduced color information in night-time environment under poor lighting conditions. In these cases, only limited visual information can be captured by the visible imaging modality. On the other hand, a thermal sensor captures information of an object that cannot be seen in the visible spectrum, especially at night and in low light conditions. It provides enhanced contrast between the human body and its environment based on their temperatures. Thermal images allow human ROI to be extracted regardless of the lighting conditions and of the color similarities of the human clothing or skin with the background.

Even though thermal images provide rich information for hot objects, the human body (ROI) extraction may still be difficult when the human body or clothing are at the same or a temperature near the background or in a windy environment that change temperature. The joint use of thermal and visible imagery results in obtaining richer information from the scene that includes both the thermal signatures and the colors. Once a thermal and visible pair of images is registered, then it can be used to better detect, track, and analyze activities of human in a scene.

The main difficulty associated with the joint use of thermal and visible information of a scene is the matching and registration of pairs of images captured by two different types of sensors. Unlike visible sensors that record reflected light, IR sensors record thermal radiations reflected and emitted by an object in a scene. Due to the numerous differences in imaging characteristics of thermal and visible cameras, most correspondence measures used for registering images of single modality are not applicable. Moreover, it is impossible to find correspondences across an entire scene. Even for partial image ROI registration, matching corresponding regions of a person in a pair of visible and thermal images is problematic since the corresponding pixels have different intensities and regions may have different patterns and textures. People might have colorful/textured clothes that are visible in color images, but not in thermal images. Moreover, there might be some textures observable in thermal images caused by different clothing (e.g. light clothes/warm clothes) and the amount of emitted energy from different parts of human body.

In the literature, Mutual Information (MI) is the most commonly used multimodal dense stereo correspondence measure [1]–[3]. MI measures the statistical co-occurrence of pixel-wise information such as local textures and patterns of matching regions. Egnal [1] has shown that mutual information (MI) is a viable similarity metric for matching thermal and visible images. However, for monitoring people that may have textured clothes, MI is not necessarily a reliable correspondence measure in all conditions. MI-based multimodal matching may fail when people have textured clothes or are partially occluded.

For registering a pair of visible images, numerous works have studied different aspects, such as taxonomy and evaluation of dense stereo correspondence algorithms [4], evaluation of various area-based and feature-based matching approaches [5], evaluation of cost functions [6], and evaluation of different similarity measures and local descriptors [7]. However, due to several differences in image characteristics of visible and infrared modalities, most similarity measures and matching methods used for visible images are not applicable. More studies about viable similarity measures for multimodal image registration and for different applications are necessary. Krotosky *et al.* have given a general survey of different multimodal registration approaches from the algorithmic aspect [2]. To the best of our knowledge, there is no work that comparatively analyzes various multimodal correspondence measures for human body ROIs dense stereo registration in a pair of thermal

and visible images. Moreover, LSS and HOG (local image descriptors) that have been recently introduced have not yet been sufficiently studied for multimodal image registration applications. In this paper, we analyze the behavior and discriminative power of Normalized Cross-Correlation (NCC) [8], Mutual Information (MI) [1], [9], Local Self-Similarity (LSS) [10], and Histograms of Oriented gradients (HOG) [11], as dense stereo similarity measures. We use a Winner Take All (WTA) sliding box matching method for all the four measures. We also tested four different scenarios to investigate the effect of texture and distance to the camera with multiple people dressed differently and walking in the scene in several pairs of indoor and challenging thermal and visible videos.

In the following section, we introduce the descriptors and measures that we have used in our experiments. In section III, we describe our systematic experiment scenarios, the matching techniques, and evaluation criteria that we have used in order to assess the accuracy and the discriminative power of the introduced descriptors and measures. In section IV, we present a comparative analysis and quantitative results related to each correspondence measure. Finally, in section V, we present the conclusions of our comparative analysis.

II. Tested descriptors and measures

In this section, we present four image descriptors and measures that are suitable or frequently used to extract common information of human ROIs in thermal and visible images.

A. Normalized Cross-Correlation

NCC is a classic similarity measure that has been widely used for single modality image template matching and image registration [8]. NCC consists in a pixel-wise cross-correlation of two image regions normalized by the overall intensity difference. NCC is defined for two bounding boxes on a pair of images as

$$C(X,Y) = \frac{\sum_{x,y}(I_l(x,y) - \bar{I}_l) * (I_r(x,y) - \bar{I}_r)}{\sqrt{\sum_{x,y}(I_l(x,y) - \bar{I}_l)^2 * \sum_{x,y}(I_r(x,y) - \bar{I}_r)^2}}, \quad (1)$$

where I_l and I_r are two matching bounding boxes on the pair of thermal and visible images.

B. Mutual Information

MI is a very popular similarity measure that has been widely used for multimodal image registration in different applications. MI computes the statistical co-occurrence of pixel-wise image patterns inside a bounding box on pair of images. MI is defined for two matching bounding boxes as

$$M(X,Y) = \sum_{X \in I_l} \sum_{Y \in I_r} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)}, \quad (2)$$

where $P(X,Y)$, is the joint probability mass function and$P(X)$ and $P(Y)$ are the marginal probability functions. $P(X,Y)$ is calculated by creating a two-dimensional histogram that records the number of co-occurrences of thermal and visible intensity values in I_l and I_r. The probabilities are then obtained by normalizing the histogram by the sum of the joint histogram entries. The marginal probabilities $P(X)$ and $P(Y)$ are then obtained by summing $P(X,Y)$ over the grayscale or thermal intensities.

C. Local Self-Similarity

LSS is a local image descriptor that has been previously applied in object detection and action detection in videos. In our previous work, we have shown that LSS is a viable local shape descriptor to be used for thermal-visible dense stereo matching [12]. While most image descriptors represent the photogrammetric properties of images (colors or gradients), LSS represents the indirect local image property which is the layout/shape of objects inside an image region. It can be used to match a textured region with other differently textured region as long as they have similar layouts. This property is interesting for human ROIs matching in thermal and visible images since the human body shape is similar in both types of images but they are differently textured. LSS represents the statistical co-occurrence of a small image patch in larger surrounding image region. LSS descriptor is a partitioned log-polar representation with 80 bins (20 angles and 4 radial intervals) of a correlation surface computed by sum of square differences (SSD) of small image patch centered at pixel p in a larger surrounding image region. SSD is normalized by the maximum value of the small image patch intensity variance and fixed a value for image noise. The correlation surface is defined as

$$S_p(x,y) = exp(\frac{SSD_p(x,y)}{max(var_{noise}, var_{patches})}). \quad (3)$$

Since the measurement unit of LSS is an image patch rather than a pixel, it can be customize to a suitable size for a given application. In our experiment, the size of the patch is 3×3 pixels and the size of surrounding image region is 20×20.

D. Histograms of Oriented Gradients

HOG is an image shape descriptor that has been previously used for human detection [11]. HOG counts occurrences of gradient orientations in localized portions of an image. It characterizes object appearance and shape by local intensity gradients or edge directions. In practice, HOG is computed by dividing an image region, named a block, to small spatial image patches (cells) and, for each cell, accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. For each block, the combined histogram entries form a histogram with, for example, 36 bins (4 cells, 9 bins for each cell). This descriptor has not been used yet for multimodal dense stereo matching. In this paper, we assess the viability of this descriptor for use as a similarity feature in a multimodal dense stereo correspondence method. In our experiment, the size of the cells is 8×8 and the size of the blocks is 16×16.

III. Evaluation method

In order to evaluate the correspondence measures introduced in section II, we use series of video frames obtained from relatively close range scene where different people with different poses and clothing are walking close and far (between 2-5 meters) from the camera baseline. In our evaluation, we assess the accuracy and discriminatively power of each correspondence measure where matching human body ROIs in thermal and visible images are either differently textured or of different sizes. We used synchronized visible-thermal videos of $5m \times 5m$ room captured by thermal and visible cameras with a baseline of 12 cm. All videos were captured in an indoor environment using stationary thermal and visible cameras at a fixed room temperature (approximately 24 $^{\circ}C$). We defined four experimental scenarios based on where a manually picked point p was selected in the visible image. Each manually selected point p is inside a human body ROI. The scenarios are

- *TexturedFar*: Matching a bounding box located on a textured human body ROI for a target relatively far from the camera.
- *TexturelessFar*: Matching a bounding box located on a textureless human body ROI for a target relatively far from the camera.
- *TexturedNear*: Matching a bounding box located on a textured human body ROI for a target relatively close from the camera.
- *TexturelessNear*: Matching a bounding box located on a textureless human body ROI for a target relatively close from the camera.

The corresponding region on the thermal image can be either differently textured or homogenous. In our experiments, far is for a target moving at a distance between 4 to 5 meters from the camera and near is for a target moving at a distance between 2 to 3 meters from the camera. Note that for close-range scene monitoring, the size of target considerably changes by walking one meter further away or toward the camera. Fig. 1 and 2 show samples of videos frames for the four scenarios.

For each scenario, we have selected 5 video frames and within each frame, 10 points were manually selected. We have defined different sizes for matching boxes to examine the effect of bounding box size on each correspondence measure. For each selected point p on the visible image, we defined three rectangular bounding box sizes of 10×130, 20×130, and 40×130 pixels centered at pixel p. Then we performed bounding box matching between thermal and visible images. Note that the height of matching boxes is defined larger than its width since our matching target is a human body ROI which has such proportions

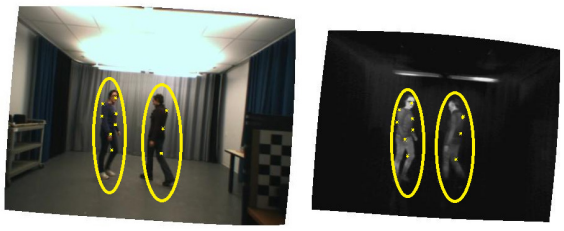
In order to simplify the correspondence search to 1D, the thermal and visible video frames were calibrated and rectified. For each thermal and visible pair of images and each correspondence measure, we first defined a bounding box centered at the manually picked point p on human ROI in the visible

(a)

(b)

Fig. 1. Examples of pairs of thermal and visible images for textured scenarios: (a) *TexturedNear*, (b) *TexturedFar.*

(a)



(b)

Fig. 2. Examples of pairs of thermal and visible images for textureless scenarios: (a) *TexturelessNear*, (b) *TexturelessFar.*

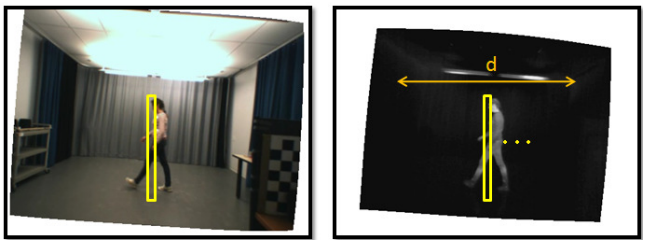


Fig. 3. Thermal-visible 1-D matching process.

image. Then, we performed a 1D box matching search on the thermal image in order to find the best correspondence on the thermal image based on a WTA approach. Figure 3 shows our matching process. The best match is the bounding boxes on the thermal and visible pair of images with the smallest Similarity Distance (SD) (section III-A). This procedure is repeated for various points on the different human body ROIs for all selected video frames.

The results were then evaluated using two metrics as in [13]: 1) accuracy using the compatibility of correspondence result with ground-truth (section III-B and 2) discriminative power using the shape of the similarity distance (section III-C).

A. Matching method

For LSS and HOG, the descriptor computation and the matching are done in two separate processes, for each pair of image boxes $W_{l,j}$ and $W_{r,j+d}$ centered at column j on the visible image and column $j+d$ column on the thermal image. For LSS, 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. Then $L1_{l,r}$ 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's bins.

For HOG, for each pair of image boxes $W_{l,j}$ and $W_{r,j+d}$, the similarity distance is the Euclidean distance of the two descriptors of the thermal and visible image boxes and it is computed as

$$SD_{j,d} = \sqrt{(h_l(1) - h_r(1))^2 + ... + (h_l(M) - h_r(M))^2}, \tag{6}$$

where M is the number of descriptor bins of h_l and h_r for the pair of thermal and visible matching boxes.

For MI, SD is defined as

$$SD_{j,d} = 1 - MI(W_{l,j}, W_{r,j+d}), \tag{7}$$

where M is the mutual information defined in equation 2. Finally for NCC, SD is defined as

$$SD_{j,d} = 1 - C(W_{l,j}, W_{r,j+d}), \tag{8}$$

where C is the normalized cross-correlation defined in equation 1.

B. Compatibility of correspondence result with ground-truth

For each point p selected manually on the human body ROI in the visible image, the corresponding point p' on the thermal image is selected manually and used as a ground-truth matching point. The disparity error for pixel p is simply the L1 distance between p' and q, for which q is the center of the best corresponding bounding box computed by our matching process. The disparity error is computed for all the tested points. Then, the number of points which have disparity errors of more than 3 pixels (> 3) is counted and considered as the number of bad matches.

C. Evaluating the shape of similarity distance

To assess the reliability of a correspondence measure the number of bad matches is important, but the reliability of good matches (disparity error $<= 3$) is also important. We assessed the discriminative power of each correspondence measure by evaluating the shape of SD along the 1D search line for the good matches using the s value [13]. In fact, a good match is discriminative if it is located on an isolated minimum on the distance curve SD. It is unreliable if it is located on a minimum that is not well defined (with close SD values for its neighbors on the curve). In order to evaluate such minimum isolation, first the SD values computed by the matching process are sorted starting from minimum to maximum value and are transformed to the interval $[0, 1]$ named SD'. Second, a N value is computed by counting the number of values in SD' that are less than a pre-computed small threshold α. α can be estimated experimentally (more details in [13]). Note that α has the same value for evaluating all the correspondence measures. Third, a quality measure s (the s value) is computed by dividing N by the total number of SD values along the 1D search line. So $s = 0$ corresponds to the minimum possible distance or most isolated minimum, and $s = 1$ corresponds the maximum value or the least isolated minimum. Finally, for each correspondence measure, a graph of Accumulated Frequencies (AF) of s values is computed using a set of values S where $s \in \{0.1, 0.2, ..., 0.9, 1\}$. The AF for a s is computed by counting the number of points with s value between $[0, s]$ and then dividing by the total number of tested points to get a normalized value. Therefore, the correspondence measure for which AF reach one at the smallest s value has the most isolated minimum, thus the better discriminative power.

IV. Results

A. Quantitative matching results

In this section, we assess the compatibility of correspondence result with ground-truth data as described in section III-B. Table I shows the percentage of numbers of bad matches (disparity error > 3 pixels) for each correspondence measure and each scenario with three different box sizes. Based on our experiments, in general, the order of the accuracy of the correspondence measures from best to worst is LSS, MI, HOG, and NCC. Our results show that the matching error of NCC is large for all four scenarios. This means that this similarity measure is not viable for multimodal registration since it is unable to measure similarity in complex appearance relationships. Concerning the matching box size, Table I shows that a small box size of 10×130 results in a relatively large number of bad matches for all four similarity measures. For MI and LSS, a larger box size of 40×130 gives smaller errors

compared to the two other sizes, because a large bounding box covers the human body ROI's boundaries, which is the main similar information between thermal and visible human body ROIs.

Concerning the effect of texture, the differences between the correspondence measures are more apparent. For far scenarios, LSS and MI perform similarly. However for near scenarios where the textures are more noticeable, specifically *TextureNear*, LSS has much less error compared to MI for both 20×130 and 40×130 box sizes. This shows that LSS is a robust descriptor for matching human body ROIs that are differently textured. Moreover, for near scenarios, HOG performs better than MI. For HOG, the medium box size 20×130 has less errors in all scenarios. The reason is that a larger box size may contain dissimilar edges and textures in the visible and in the infrared which confuse the matching of similar edges. HOG is more sensitive to dissimilarities containing strong gradients compared to MI and LSS. For example, for *TexturedFar* and *TexturelessFar* because the background is textured with strong gradients (see Figure 1b and 2b), HOG has larger errors. This contrasts with the two other scenarios with larger targets where the majority of the bounding box is on the human body ROIs rather than the background. Therefore, HOG can be a good similarity measure choice as long as the bounding box is correctly located on human body ROI and that it contains strong distinctive gradients. It is important to notice that we used a simple box matching method that does not handle occlusion, depth discontinuity, and *etc.*. The errors could be reduced by applying more sophisticated matching method such as disparity voting or energy minimization based methods.

B. Comparative analysis of the discriminative power

In this section, we assess the discriminative power of the correspondence measures as described in section III-C. For each scenario, we have calculated the accumulated frequencies (AF) graphs of all the good matches of the correspondence measures (the points with $<=$ 3 pixel disparity error). To compute the s value of each testing point, we have selected a disparity interval $D = [q-10 : q+10]$, where q is the position of the minimum on the curve SD. Figures 4, 5, 6, and 7 show the accumulated frequencies (AF) for all the scenarios using a box size of 40×130. NCC in *TexturedNear* scenario has no good match, therefore we were not able to plot AF graph for it (Fig. 5). In general, the graphs show that for all four scenarios LSS begins with higher values of AF at smaller S values, which means that the proportion of good matches that are discriminative is higher compared to other correspondence measures. Second is MI, then HOG, and finally NCC. In the graph, the s value where AF reaches 1 means all the good matches have a s value between $[0, s]$. Recall that $s = 0$ is the maximum isolation (best discriminative power) and $s = 1$ is the minimum isolation (worst discriminative power). The levels of s where MI and LSS reach $AF = 1$ are almost similar. However, LSS starts with higher AF values, which means that the majority of the points in the good match population has a reasonable level of discrimination. For MI,

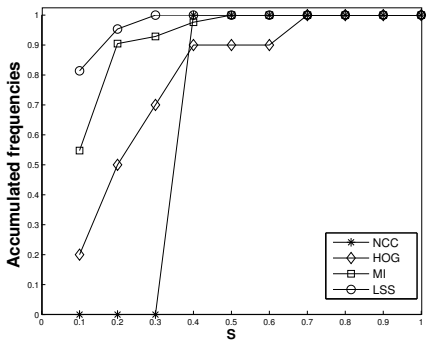


Fig. 4. Accumulated frequencies for s (*TexturedFar*)

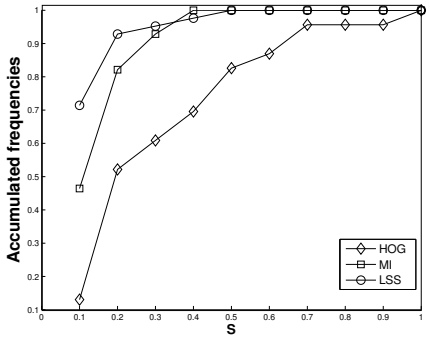


Fig. 5. Accumulated frequencies for s (*TexturedNear*)

some points have a high level of discrimination and some others have a low level of discrimination. Thus, LSS compared to MI is less sensitive to different textures inside human body ROIs as long as the two ROIs have similar layouts (similar shapes of body). LSS and MI are the best choices, as NCC and HOG are much less discriminative. LSS is superior to MI as its best matches are more discriminative and they are larger in number. Further, LSS is less sensitive to the matching box size.

V. Conclusions

In this paper, we comparatively evaluated the accuracy and the discriminative power of NCC, HOG, MI, and LSS,

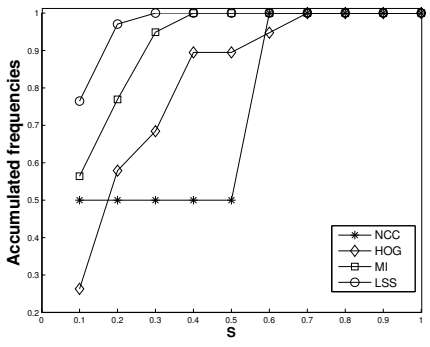


Fig. 6. Accumulated frequencies for s (*TexturelessFar*)

TABLE I
QUANTITATIVE MATCHING RESULTS OF THE FOUR SIMILARITY MEASURES. (ERR. %) IS THE PERCENTAGE OF THE NUMBER OF BAD MATCHES (DISPARITY ERROR $>$ 3 PIXELS).

Method	No. images	No. Points	Box size	TexturedFar (Err.%)	TexturelessFar (Err.%)	TexturedNear (Err.%)	TexturelessNear (Err.%)
NCC	5	50	10×130	100	98	100	100
MI				58	64	78	80
LSS				32	54	56	50
HOG				86	73	78	74
NCC	5	50	20×130	100	96	100	98
MI				20	46	70	74
LSS				22	46	20	54
HOG				69	52	36	30
NCC	5	50	40×130	98	96	100	95
MI				16	22	44	46
LSS				14	32	16	38
HOG				90	62	54	60

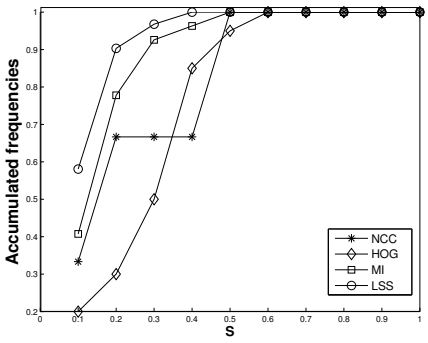


Fig. 7. Accumulated frequencies for s (*TexturelessNear*)

as dense thermal-visible stereo correspondence measures for human body ROI registration. Our experiments show that NCC fails in most scenarios as a similarity measure and is not suitable for this task. However, HOG can be a reasonable similarity measure in situation where there are sufficient similar strong edges and boundaries inside the matching bounding boxes. However, HOG is sensitive to dissimilarities containing strong gradients. MI is a reasonable and viable multimodal correspondence measure in the case where the joint probability is sufficiently populated inside the matching bounding boxes. However, MI is sensitive to the size of matching bounding boxes and fails when the human bodyROIs are differently textured. Finally, our experiments show that LSS is the most accurate and discriminative correspondence measure because it mostly describes the layout (human body shape) and it is less sensitive to differently textured regions inside human body ROIs in thermal and visible images.

VI. ACKNOWLEDGMENT

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by Canada Foundation for innovation (CFI).

REFERENCES

[1] G. Egnal, "Mutual information as a stereo correspondence measure," *Tech. Rep. MS-CIS-00-20, University of Pennsylvania*, 2000.

[2] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 270 – 287, 2007.

[3] H.-M. Chen, P. Varshney, and 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.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, pp. 7–42, 2002.

[5] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.

[6] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8.

[7] A. Gil, O. Mozos, M. Ballesta, and O. Reinoso, "A comparative evaluation of interest point detectors and local descriptors for visual slam," *Machine Vision and Applications*, vol. 21, pp. 905–920, 2010.

[8] J. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *Advances in Computing, Control, Telecommunication Technologies, 2009. ACT '09. International Conference on*, dec. 2009, pp. 819 –822.

[9] C. Fookes, A. Maeder, S. Sridharan, and 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*, sept. 2004, pp. 961 – 968.

[10] E. Shechtman and 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.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, ser. CVPR '05, 2005, pp. 886–893.

[12] A. Torabi and G. A. Bilodeau, "Local self-similarity as a dense stereo correspondence measure for themal-visible video registration," in *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW2011)*, 2010.

[13] R. Mayoral and M. Aurnhammer, "Evaluation of correspondence errors for stereo," *Pattern Recognition, International Conference on*, vol. 4, pp. 104–107, 2004.