Thermal-Visible Registration of Human Silhouettes: a Similarity Measure Performance Evaluation

Guillaume-Alexandre Bilodeau[a,*], Atousa Torabi[b], Pierre-Luc St-Charles[a], Dorra Riahi[a]

[a]*LITIV Lab., Department of Computer and Software Engineering, ÉcolePolytechnique de Montréal, P.O. Box 6079, Station Centre-ville, Montréal Québec, Canada, H3C 3A7*
[b]*LISA, Dept. IRO, Université de Montréal, Montréal, Québec, Canada, H2C 3J7*

Abstract

When dealing with the registration of information from different image sources, the *de facto* similarity measure used is Mutual Information (MI). Although MI gives good performance in many image registration applications, recent works in thermal-visible registration have shown that other similarity measures can give results that are as accurate, if not more than MI. Furthermore, some of these measures also have the advantage of being calculated independently from each image to register, which allows them to be integrated more easily in energy minimization frameworks. In this article, we investigate the accuracy of similarity measures for thermal-visible image registration of human silhouettes, including MI, Sum of Squared Differences

*Corresponding author
Email addresses: gabilodeau@polymtl.ca (Guillaume-Alexandre Bilodeau), torabi@iro.umontreal.ca (Atousa Torabi), pierre-luc.st-charles@polymtl.ca (Pierre-Luc St-Charles), dorra.riahi@polymtl.ca (Dorra Riahi)

(SSD), Normalized Cross-Correlation (NCC), Histograms of Oriented Gradients (HOG), Local Self-Similarity (LSS), Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), Census, Fast Retina Keypoint (FREAK), and Binary Robust Independent Elementary Feature (BRIEF). We tested the various similarity measures in dense stereo matching tasks over 25 000 windows to have statistically significant results. To do so, we created a new dataset in which one to five humans are walking in a scene in various depth planes. Results show that even if MI is a very strong performer, particularly for large regions of interest (ROI), LSS gives better accuracies when ROI are small or segmented into small fragments because of its ability to capture shape. The other tested similarity measures did not give consistently accurate results.

Keywords: Multispectral imagery, similarity measures, dense stereo matching, thermal-visible registration, thermal camera, visible camera

1. Introduction

In the recent years, there has been a growing interest in visual surveillance using multimodal sensors in both civilian and military applications. One of the fundamental issues associated with thermal-visible imagery is the matching and registration of pairs of images captured by the two different types of sensors. Unlike visible sensors that capture reflected light, IR sensors capture thermal radiations reflected and emitted by an object in a scene. Although not very well documented, most similarity measures used for registering visible images are not applicable for thermal-visible image registration because of the differences in imaging characteristics of thermal and visible

cameras. Furthermore, scene content at room temperature cannot be registered because it does not convey any textural information. For this reason, most works in thermal-visible imagery focus on hot bodies, like for example, people.

For the registration of people, two families of methods exist. First, there are the sparse registration methods. In these methods, people are considered as planar objects that can be registered by just matching some points. Most of these methods are based on matching contour points. Using edges is one of the most popular method as their magnitudes and orientations may match between infrared and visible for some object boundaries [1, 2, 3]. Raw edges alone are not very reliable, so they may be considered as connected groups for correspondence [1, 3]. Other methods use polygonal approximation of people and matches them using their vertices [4, 5]. Although these methods allow fast registration of people, the resulting transformation does not allow to capture fine depth details, like the position of the arms or legs relative to the body.

For accounting for depth details, the second family is based on dense correspondences between the visible and the thermal human silhouettes. Typically, in that case, registration is performed at every pixel on the human body by comparing the pixels inside a window using a similarity measure. We define as a similarity measure any function that returns a value that indicates a level of similarity. For example, Mutual Information (MI) is one of the most popular similarity measure [6, 7, 8, 9], but recently, Local Self-Similarity (LSS), a local image descriptor, was also proposed for this purpose. A matching window is described as a dense collection of LSS descriptors, and

the similarity of two windows is the Euclidean distance between the respective collections of descriptors [10, 11].

The proposal of new local image descriptors (LID) is a very active field in computer vision. These LIDs are typically proposed to allow discriminative matches between regions in visible images. For each of these new LIDs, the question is: can they be used to successfully register people in visible and thermal images? This question is worth answering for two reasons: 1) MI does not provide accurate registration all the time [6, 7, 8], and 2) sometime, surprisingly, other LIDs like Local sSelf-Similarity (LSS) have been shown to outperform MI [10, 11]. In this article, our goal is to study the applicability of various LIDs or other similarity measures to the problem of registering people (or any other bodies not at room temperature) in visible and thermal imagery. To test the viability of various similarity measures, we use them in the context of typical windows-based matching, where the potential measures or LIDs are applied over all the pixels in a window to find correspondences.

We began some work in that regard in [12]. However, the experiments were more limited (around 300 test windows) and less similarity measures were compared. In this study, we experiment on more similarity measures and we test them on over more than 25 000 windows to have statistically significant results for human silhouette registration. We compare the similarity measures both for winner takes all (WTA) sliding window matching and disparity voting (DV).

In section 2, we present the similarity measures we tested. Section 3 describes the details of our camera setup, dataset, test procedure, and evaluation criteria. In section 4, we present and discuss our experimental results.

Finally, we conclude the paper with a general discussion in section 5.

2. Tested similarity measures

We tested three broad categories of similarity measures:

1. Similarity measures that are calculated across pixels of the two windows, namely Mutual Information (MI), Sum of Squared Differences (SSD), and Normalized Cross-Correlation (NCC);
2. Traditional LIDs that model data as distributions, namely Histograms of Oriented Gradients (HOG), Local Self-Similarity (LSS), Scale Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF). In this case, they are applied as dense collection of features compared with a distance to be used as similarity measure;
3. LIDs based on binary comparisons of pixels, namely Census, Fast REtinA Keypoint (FREAK), Binary Robust Independent Elementary Feature (BRIEF). In this case they are also applied as dense collections.

These similarity measures represent only a subset of possible measures, as any LIDs could be formulated as similarity measures. However, they represent a good sample of measures as they cover the main categories of measures and LIDs and they all show good discriminative power when comparing regions in visible images.

NCC is a classic similarity measure that consists in a pixel-wise cross-correlation of two image regions normalized by the overall intensity difference [13]. It is defined for two windows on a pair of images as

$$NCC(W_1, W_2) = \frac{\sum_{x,y}(W_1(x,y) - \bar{W_1}) * (W_2(x,y) - \bar{W_2})}{\sqrt{\sum_{x,y}(W_1(x,y) - \bar{W_1})^2 * \sum_{x,y}(W_2(x,y) - \bar{W_2})^2}}, \quad (1)$$

where W_1 and W_2 are two matching windows on a pair of thermal and visible images, and $\bar{W_1}$ and $\bar{W_2}$ are the mean pixel intensities in the windows. This measure relies basically on similar intensity patterns. This is similar for SSD, which is defined as

$$SSD(W_1, W_2) = \sum_{x,y}(W_1(x,y) - W_2(x,y))^2. \quad (2)$$

MI computes the statistical co-occurrence of pixel-wise image patterns inside a window on pair of images using

$$MI(W_1, W_2) = \sum_{X \in W_1} \sum_{Y \in W_2} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)}, \quad (3)$$

where $P(X,Y)$, is the joint probability mass function of intensities 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 W_1 and W_2. 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. MI relies on the co-occurrence of patterns that do not need to be similar. It can match a uniform region with a textured region. This is why it is successful with multimodal imagery.

To use LSS[14], SIFT[15], and SURF[16] as similarity measures, each of these feature descriptors are calculated densely for all the pixels in a window.

For a window of say 10×10, this gives a collection of 100 descriptors. The similarity measure SM of two windows W_1 and W_2 is given by

$$SM(W_1, W_2) = \sqrt{\sum_{x,y}(f_l(x,y) - f_r(x,y))^2}, \tag{4}$$

where f_l and f_r are feature descriptors in two matching windows on a pair of thermal and visible images. SIFT and SURF are gradient-based, while LSS is a local shape descriptor based on the comparison of the self-similarity of a central patch with neighboring patches. As for HOG[17], since it is window-based, we apply it directly on the whole windows. The HOG descriptors are then compared with the Euclidean distance with Eq. 4.

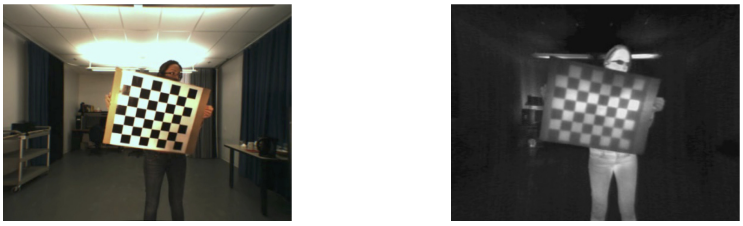
Census[18], BRIEF[19] and FREAK[20] are applied the same way as SIFT. They all consist of binary intensity comparisons within a window using different pre-determined patterns. In this case, the binary descriptors of two windows are compared using the Hamming distance to obtain a similarity measure.

3. Experimental method

We tested the 10 chosen similarity measures on more than 25 000 windows using both a WTA and DV procedure. In this section, we describe our test methodology.

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. The scene is relatively close range where different



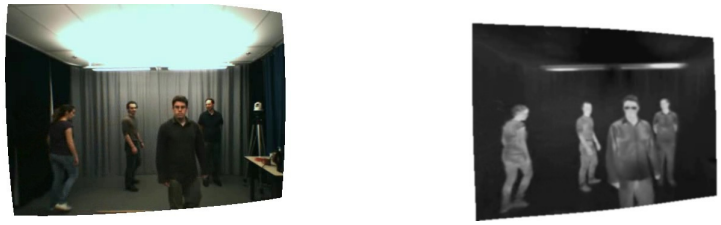
(a) (b)

Figure 1: Calibration images: (a) visible image (b) thermal image.

people in 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 [21]. As for calibration, we placed a checkerboard pattern in front of the cameras. Since in the thermal images, the checkerboard 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. 1 shows an example of our calibration images.

We captured four videos and manually annotated the disparity of people in 206 frames from extracted foregrounds using [22], which gives a total of 25819 ground-truth point pairs[1]. All videos include between 1 to 5 actors walking around and occluding each other (see Fig. 2). The specifications of each video are given in table 1. A second subset of ground-truth point pairs

[1]Our dataset is publicly available at http://www.polymtl.ca/litiv/en/vid/index.php



(a) (b)

Figure 2: Vid2cut1, rectified frames 526: (a) visible image (b) thermal image.

Table 1: Test video dataset. GTpp stands for ground-truth point pairs.

	Nb. Frames	Nb. Actors	GTpp for WTA	GTpp for DV
Vid1	4366	1 to 4	11166	1496
Vid2cut1	96	1 to 2	2217	706
Vid2cut2	451	2 to 4	6012	889
Vid3	477	1 to 5	6524	981
Total	5390	1 to 5	25819	3442

was selected for experiments with disparity voting that are more computationally expensive.

3.2. Test procedures

We tested each similarity measure under two matching procedures. The first consists in finding the best match for a point selected in the visible in the thermal image by scanning a row in the image column by column (see Fig. 3). This is the winner-takes-all (WTA) procedure and its algorithm is given in Algo. 1. For this test procedure, we found matches on all the annotated pixels in the four videos. The annotated pixel coordinate is used as the center

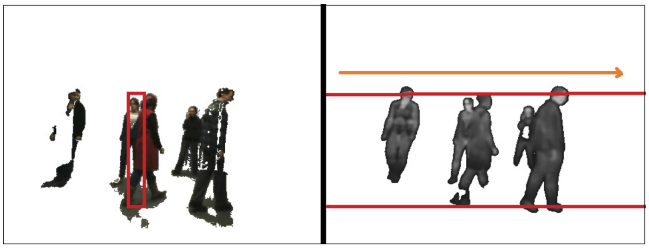


Figure 3: Visible-thermal 1-D sliding window matching.

of the matching window. The WTA test procedure 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 larger than their widths since our matching target is a human body ROI that has such proportions. Note that we tested the measures on foreground images that were obtained using background subtraction [22]. The background subtraction is not perfect and ROIs might be partially misdetected or some regions might be falsely detected. For each tested point, we record the column $best_j$ with maximum similarity value and its associated similarity value $bestSV_j$, in addition to the similarity value SV_j obtained for all the columns to evaluate the *power of discrimination* (see section 3.3.2). The similarity value is the value resulting from the application of a similarity measure, for example, the results obtained from Eq. 1, 2, 3, 4 and the Hamming distance.

The second test procedure is based on disparity voting (DV) as used previously by [6] in the context of thermal-visible registration. In this case, the disparity at a column is the result of a majority voting of disparities obtained from neighboring columns (see Algo. 2). The matching process was repeated using three rectangular window sizes of 10×130 (small), 20×130 (medium), and 40×130 (large) pixels. For our experiments, V in Algo. 2

Algorithm 1 Winner-takes-all testing procedure

procedure WTA TESTING
    NW: number of tested windows
    L: image width
    for each test window W_i centered at x,y in visible image, with i=1...NW **do**
        if testing LID **then**
            Extract LIDs for all pixels in W_i
        end if
        for each window W_j centered at j,y in thermal image, with j=1..L **do**
            if testing LID **then**
                Extract LIDs for all pixels in W_j
                Compute similarity value SV_j between LIDs in W_i and W_j
            else
                Compute similarity value SV_j between W_i and W_j
            end if
            Save $best_j$ and $bestSV_j$ if similarity of W_j is larger than for previous $best_j$
            Save SV_j
        end for
        Compare $best_j$ with ground-truth
        Save disparity error e_i
    end for
end procedure

was selected as the width of the rectangular window. This procedure was also performed on foreground images. As opposed to the WTA method, the DV method is more robust near occlusion boundaries because neighboring disparities are considered for taking a decision at a particular column.

3.3. Evaluation Criteria

To evaluate the similarity measures, we have used three metrics. The first two metrics are the *precision* and *recall* while the third metric is the power of discrimination.

3.3.1. Precision and Recall

We used a criterion based on the number of correct matches for all pairs of tested images as used in [23, 24]. *Precision* and *recall* are defined as follows:

$$precision = \frac{CM}{MR}, \tag{5}$$

$$recall = \frac{CM}{TM}, \tag{6}$$

where CM is the number of correct matches, MR is the number of matches retrieved, and TM is the total number of ground-truth matches. In a *precision* versus *recall* curve, a feature with high *recall* and low *precision* means that many correct matches as well as many false matches are retrieved. On the other hand, high *precision* and low *recall* means that most matches are correct but many others have been missed. In our experiment, TM is a fixed value that corresponds to the total number of tested windows.

Remember that for each tested window, we obtain a $bestSV_j$ value that corresponds to the maximum similarity value obtained when comparing two

Algorithm 2 Disparity voting testing procedure

procedure DV TESTING
    NW: number of tested windows
    L: image width
    V: Number of windows for voting
    for each test window W_i centered at x,y in visible image, with i=1...NW **do**
        for $k = -V/2...V/2$ **do**
            if testing LID **then**
                Extract LIDs for all pixels in W_{i+k}
            end if
            for each window W_j centered at j,y in thermal image, with j=1...L **do**
                if testing LID **then**
                    Extract LIDs for all pixels in W_j
                    Compute similarity value between LIDs in W_{i+k} and W_j
                else
                    Compute similarity value between W_{i+k} and W_j
                end if
                Save j if similarity value of W_j is larger than for previous j
            end for
            Add a vote for disparity $d = j - i + k$
        end for
        Select d with more votes
        Compare d with ground-truth
        Save disparity error e_i
    end for
end procedure

windows with a given approach. Depending on the tested window, the maximum similarity value can be large or small. Thus, in order to obtain different recall-precision points to draw our curves, we sort our results using their $bestSV_j$ values and define MR as the n first matches of the sorted results vector ($n \in [1, TM]$). We then define CM as a subset of MR in which all $best_j$ values have a disparity error smaller than 3 pixels (with respect to the ground truth). Hence, the first points ($n \ll TM$) should present high *precision* (albeit low *recall*) and the latter points ($n \gg 1$) the opposite.

3.3.2. power of discrimination

To assess the reliability of matches, correct ground-truth overlap is not the only important factor: the matches must also be discriminative. The *power of discrimination* metric describes the level of distinctiveness of a match compared to its neighboring matches on the SV_j versus column j curve. In order to evaluate the *power of discrimination* of similarity measures, we study the shape of SV_j over an interval of columns D for all the matches of all pairs of ground-truth points. Note that for some similarity measures, a good match is a minimum, while for others a good match is a maximum. To illustrate the *power of discrimination*, we will consider the former. In that case, a reliable match is located on an isolated minimum on the SV_j versus j curve and has a SV_j value much smaller than its neighboring points. In order to evaluate the isolation of the minimum, the SV_j values computed by the WTA test procedure on the interval of columns D around the minimum are first normalized and sorted increasingly and are transformed to the interval $[0, 1]$ which results in SV'_j. Then, N is computed by counting the number of values SV'_j that are less than a pre-computed small threshold α, ignoring the

minimum. α has the same value for evaluating all similarity measures. Third, a quality measure s (the s value) is computed by dividing N by the size of the interval D. So $s = 0$ corresponds to the most isolated minimum (best performance), and $s = 1$ corresponds to the least isolated minimum. Finally, for each similarity measure, a graph of accumulated frequencies (AF) of the s values of all matches is computed. Therefore, the similarity measure for which AF reaches a higher value at a smaller s value is the most discriminative. Fig. 4 (a) and (b) show an example where the minimum of SV_j is relatively isolated and N is 2, and fig. 4 (c) and (d) show an example where the minimum is not well isolated and N is 8, which results in higher value of s compared to the previous example.

4. Experimental Results

We will first present and discuss the results for the WTA test procedure, and we will then present the results for the DV test procedure.

4.1. Results with WTA test procedure

Fig. 5, 6, 7 present the *precision-recall* curves obtained for the three tested window sizes. Globally, as expected, a better precision is achieved with a larger window size, as more information is available to measure similarity. The two best performing similarity measures are LSS and MI. LSS tends to be better for smaller windows, while MI tends to be better for larger ones. The key of MI success is that it can consider similar different looking patterns if they co-occur in the matching windows. In the case of LSS, it works well because it is a local shape descriptor that evaluates the similarity of surrounding patches with respect to a center patch. Two shapes can be

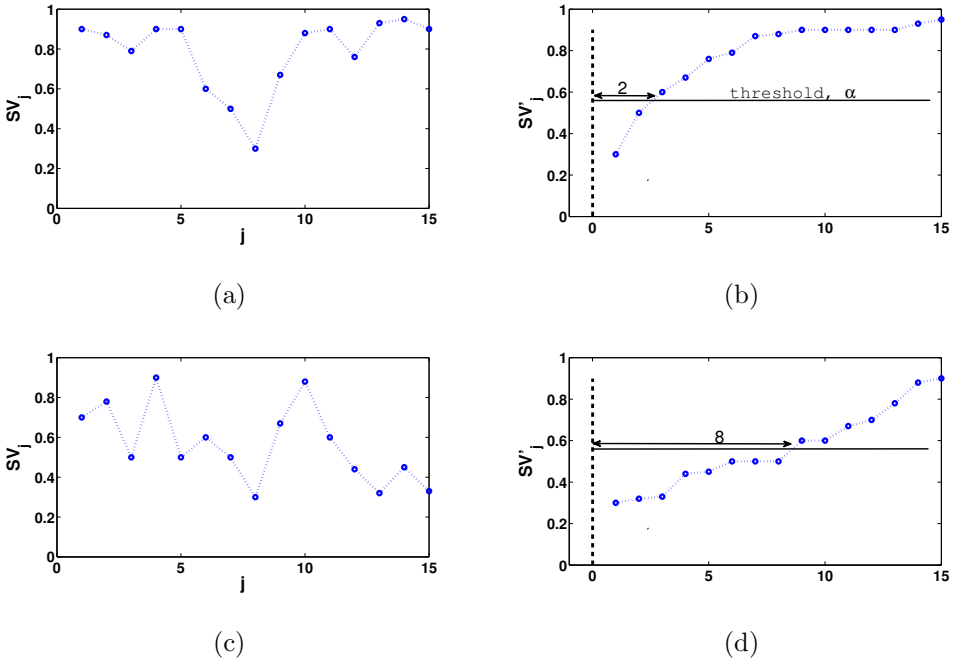


Figure 4: (a) and (c) Similarity distance SV_j versus column j curve. (b) and (d) Sorted SV_j' curve.

considered similar even if they are composed of globally different patterns if laid out the same way. In a sense, the principle behind MI and LSS are similar, which explains their similar performance. On the other hand, our experiments show that pixel by pixel comparisons as performed by SSD and NCC are not successful because the textural content between visible and infrared silhouette is too different. BRIEF and Census suffer from similar problems as they model fine small-scale textural patterns. Interestingly, even though FREAK is also based on binary comparisons, it ranks third or fourth depending on window size. This can be explained by the larger spatial extent of the comparisons done for constructing the FREAK descriptor, and by the fact that these comparisons are done on mean intensity values. As a result, it seems that FREAK is capable of capturing some shape information of the people's silhouette.

Also SURF gives performances that are close to FREAK, but superior to SIFT. These descriptors can sometimes give acceptable results, because they are gradient-based, and gradients were previously shown to be a common feature of visible and thermal human silhouette, particularly at the body's boundary [1, 2, 3]. The results with HOG are disappointing considering that this feature descriptor is quite similar to SIFT in essence and considering previous results obtained by Torabi *et al.* [12] that were reporting that HOG was performing relatively well. We believe that HOG may not perform as well as SIFT because of intensity normalization that may reduce the saliency of some edges in the thermal images. Furthermore, compared to previously reported results, our new results are coherent with them in the sense that HOG was clearly not as good as LSS and MI, but it seems that using a

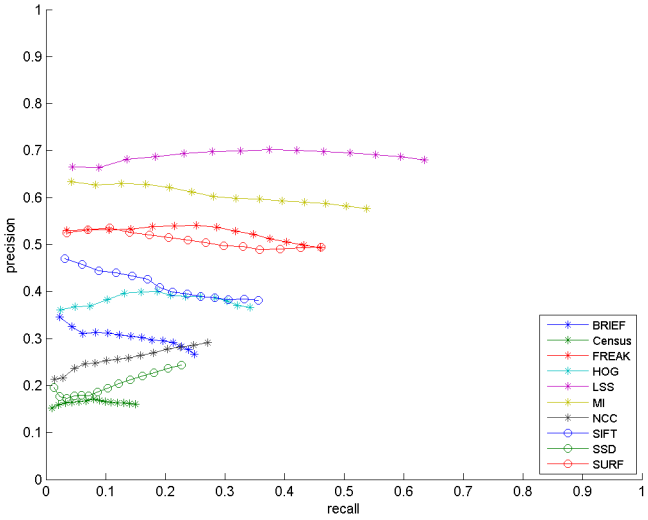


Figure 5: *Precision-recall* curves for small windows (10×130) for the WTA test procedure.

much larger number of test points resulted in a significant drop in its relative performance.

We would also like to point out, that the *recall* does not have a large impact on *precision*. It seems that the similarity measures are distinctive or not. That is, the absolute similarity value is not indicative of distinctiveness.

Let us now consider the *power of discrimination* of our 10 similarity measures. Fig. 8, 9, 10 present the *power of discrimination* curves obtained for the three tested window sizes. Recall that these curves allow assessing the distinctiveness of the best match compared to the following best matches; that is, it describes how isolated the best match is, with a match being more isolated when $s = 0$. Thus, a more discriminative measure should have more correct matches with small s value. These curves show that MI has

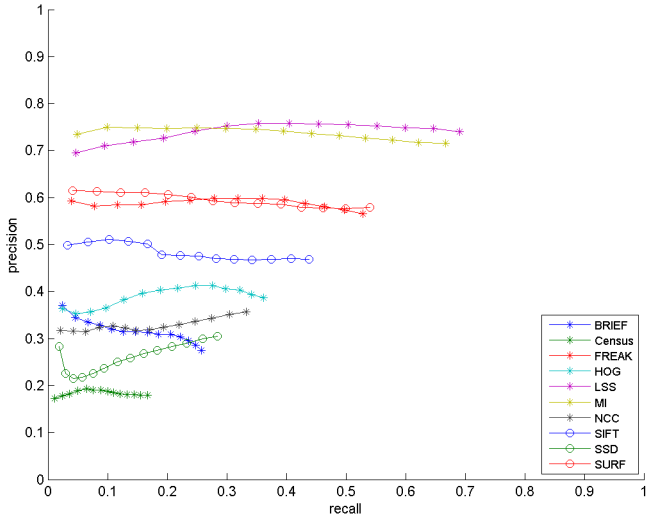


Figure 6: *Precision-recall* curves for medium windows (20×130) for the WTA test procedure.

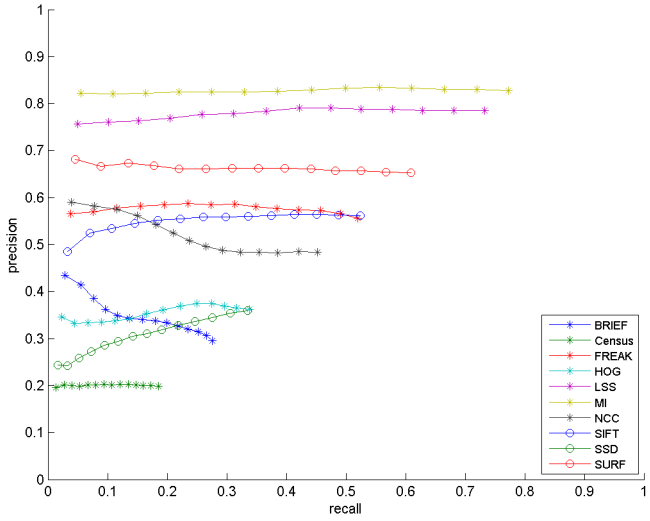


Figure 7: *Precision-recall* curves for large windows (40×130) for the WTA test procedure.

the most discriminative matches, followed by LSS and NCC. There is not much to add to the discussion about LSS and MI, except that MI seems to have a small advantage for distinctiveness, particularly with small windows. This contradicts to some extent the results of the *precision-recall* curves that show that LSS is better than MI for small windows. It should be interpreted as LSS gives plenty of matches for smaller window, but the good matches obtained by MI are from more isolated maximums (i.e. more certain). The results from NCC can be viewed a bit as a surprise, but what it tells is that NCC does not work in general for such an application, but when it gives a good match, this good match is very discriminative. Except for Census that is not at all discriminative, all other measures are more or less similar for their *power of discrimination*.

4.2. Results with DV test procedure

For the DV test procedure, we just report the *recall* because our experiments just give us a disparity for each point. Fig. 11, 12, 13 present the *recall* obtained for the three tested window sizes. Globally, using DV does not benefit the best similarity measures. The *recall* for MI, FREAK, and LSS does not improve much over the WTA procedure. This is because they are already quite discriminative. However, the voting mechanism helps most of the other methods. This is particularly true for BRIEF and SSD. BRIEF improves at the level of SURF and SIFT. This indicates that even if disparity voting will not systematically improve registration accuracy, it can improve robustness for less effective similarity measures. Concerning MI and LSS, we still get the same conclusions. LSS is a better for small windows, while MI is better for larger windows. Both methods are about on par for medium

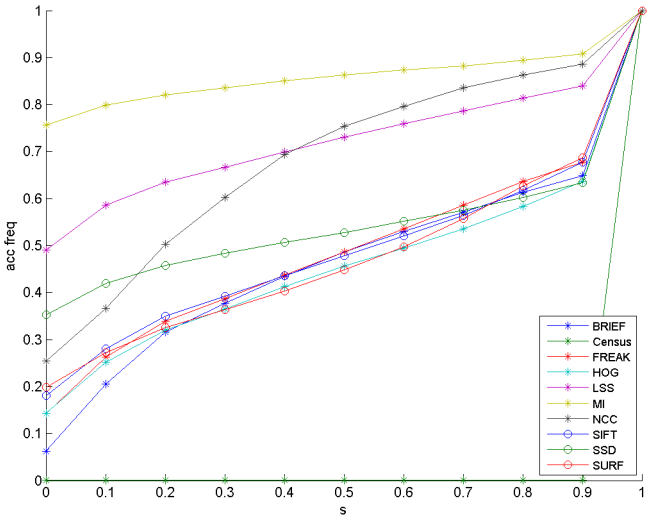


Figure 8: *Power of discrimination* curves for small windows (10×130) for the WTA test procedure.

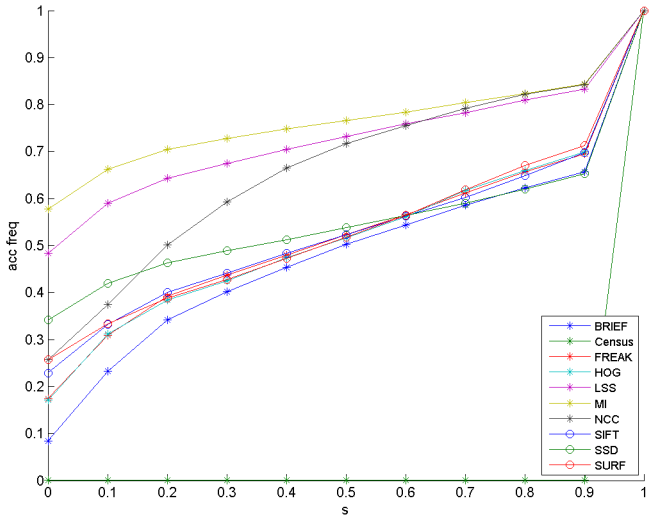


Figure 9: *Power of discrimination* curves for medium windows (20×130) for the WTA test procedure.

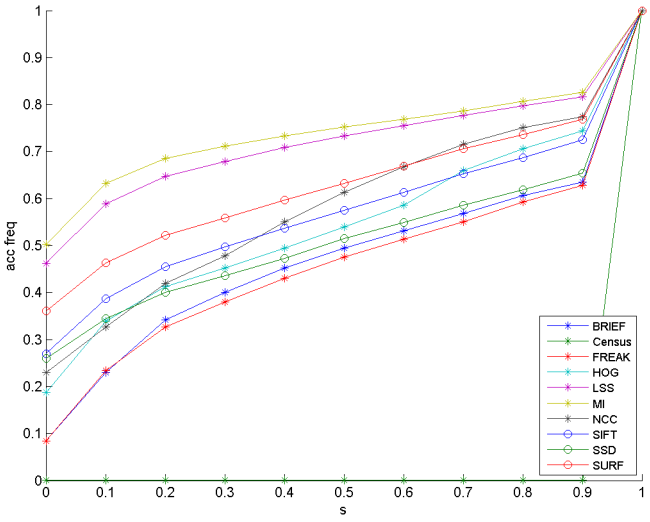


Figure 10: *Power of discrimination* curves for large windows (40×130) for the WTA test procedure.

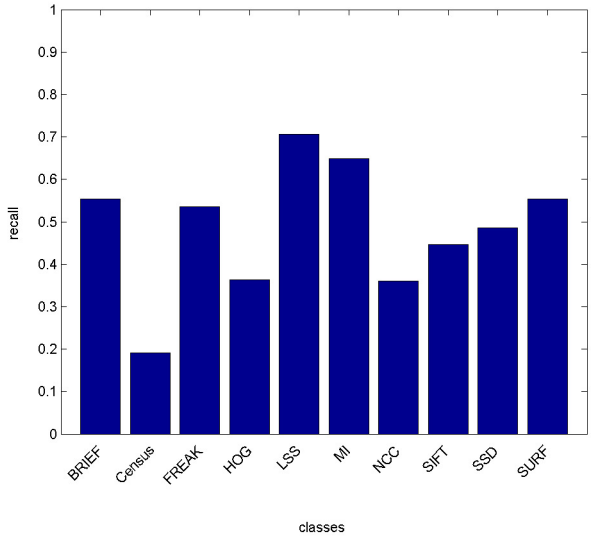


Figure 11: *Recall* for small windows (10×130) for the WTA test procedure.

windows. Again, globally, for larger window sizes, the results are better because more information is available. By more information, we mean that a larger window will necessarily include more boundary information that is very discriminative.

5. General discussion and conclusions

From our results, we observed that although MI is certainly a good choice for registering human silhouettes, it is not always the most accurate similarity measure for thermal-visible registration. As such, we should be attentive to new feature descriptors that may help us obtain future results that are even more accurate than what we have now. From our tests, we can conclude that:

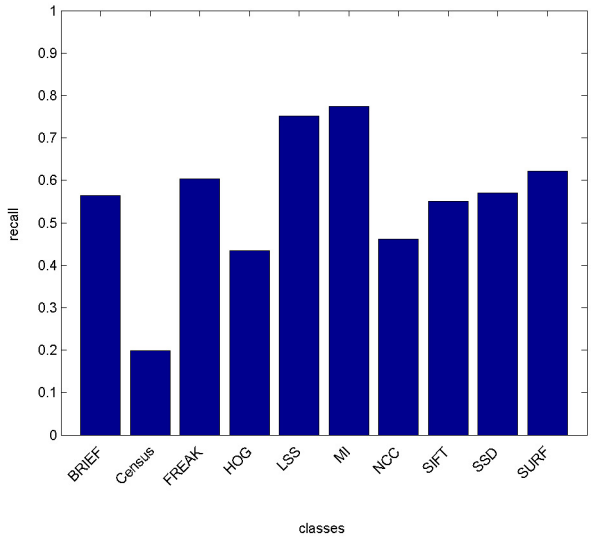


Figure 12: *Recall* for medium windows (20×130) for the WTA test procedure.

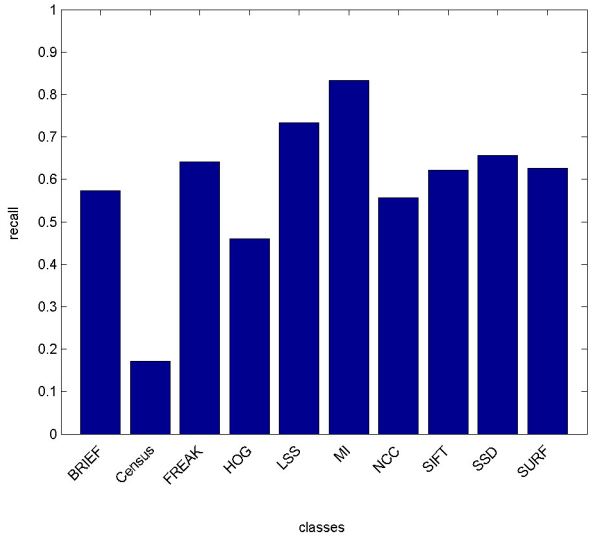


Figure 13: *Recall* for large windows (40×130) for the WTA test procedure.

- to obtain an accurate registration of thermal and visible silhouettes, the similarity measure chosen should be designed to account principally for the layout of patterns. This is the case for MI and LSS;
- similarity measures that model fine small-scale texture cannot work in this context;
- if the objects of interest are small or segmented into small fragments, or if there are many occlusions between objects which require small windows for treatment, LSS is the best choice to obtain accurate results;
- if the objects are large, MI is still the best similarity measure;
- *precision* and *recall* metrics are more useful metrics than the *power of discrimination* to evaluate a new similarity measure. This was illustrated by the results for NCC;
- disparity voting improves the robustness of the similarity measures, but the effect is more obvious for weaker similarity measures.

6. Acknowledgements

This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant No. 311869-2010.

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

[2] M. I. Elbakary, M. K. Sundareshan, Multi-modal image registration using local frequency representation and computer-aided design (cad) models, Image Vision Computing 25 (5) (2007) 663–670.

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

[4] G. A. Bilodeau, P. St-Onge, R. Garnier, Silhouette-based features for visible-infrared registration, in: Computer Vision and Pattern Recognition Workshops (CVPRW): Object Tracking and Classification Beyond the Visible Spectrum, 2011 IEEE Computer Society Conference on, 2011, pp. 68–73.

[5] S. Sonn, G.-A. Bilodeau, P. Galinier, Fast and accurate registration of visible and infrared videos, in: Computer Vision and Pattern Recognition Workshops (CVPRW): Perception Beyond the Visible Spectrum, 2013 IEEE Conference on, 2013, pp. 308–313.

[6] 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.

[7] 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.

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

[9] S. K. Kyoung, H. L. Jae, B. R. Jong, Robust multi-sensor image registration by enhancing statistical correlation, in: Information Fusion, 2005 8th International Conference on, Vol. 1, 2005, p. 7.

[10] A. Torabi, G.-A. Bilodeau, Local self-similarity-based registration of human {ROIs} in pairs of stereo thermal-visible videos, Pattern Recognition 46 (2) (2013) 578 – 589.

[11] A. Torabi, G.-A. Bilodeau, A lss-based registration of stereo thermalvisible videos of multiple people using belief propagation, Computer Vision and Image Understanding 117 (12) (2013) 1736 – 1747.

[12] 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.

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

[14] 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.

[15] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[16] H. Bay, T. Tuytelaars, L. Gool, Surf: Speeded up robust features, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision ECCV 2006, Vol. 3951 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, pp. 404–417.

[17] N. Dalal, 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, CVPR '05, 2005, pp. 886–893.

[18] R. Zabih, J. Woodfill, Non-parametric local transforms for computing visual correspondence, in: J.-O. Eklundh (Ed.), Computer Vision ECCV '94, Vol. 801 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1994, pp. 151–158.

[19] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: K. Daniilidis, P. Maragos, N. Paragios (Eds.), Computer Vision ECCV 2010, Vol. 6314 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 778–792.

[20] A. Alahi, R. Ortiz, P. Vandergheynst, Freak: Fast retina keypoint, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 510–517.

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

[22] B. Shoushtarian, H. E. Bez, A practical adaptive approach for dynamic

background subtraction using an invariant colour model and object tracking, Pattern Recognition Letters 26 (1) (2005) 5–26.

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

[24] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (10) (2005) 1615 –1630.