Local self-similarity as a dense stereo correspondence measure for themal-visible video registration

Atousa Torabi and Guillaume-Alexandre Bilodeau
LITIV Lab., Department of computer engineering and software engineering
École Polytechnique de Montréal
P.O. Box 6079, Station Centre-ville, Montréal, (Québec), Canada, H3C 3A7
{atousa.torabi, guillaume-alexandre.bilodeau@polymtl.ca}

Abstract

The robustness of Mutual Information (MI), the most used multimodal dense stereo correspondence measure, is restricted by the size of the matching windows. However, obtaining the appropriately sized MI windows for matching thermal-visible pair of images of multiple people with various poses, clothes, distances to cameras, and different levels of occlusions is quite challenging. In this paper, we propose local self-similarity (LSS) as a multimodal dense stereo correspondence measure. We integrated LSS as a similarity metric with a disparity voting registration method to demonstrate the suitability of LSS for a visible-thermal stereo registration method. We have analyzed comparatively LSS and MI as multimodal correspondence measures and discussed LSS advantages compared to MI. We have also tested our LSS-based registration method in several indoor videos of multiple people and shown that our registration method outperforms the most recent MI-based registration method in the state-of-the-art.

1. Introduction

In the recent years, the advantages of multimodal visual surveillance systems have been studied in several works [15, 2, 11]. Furthermore, several methods including data fusion algorithms, background subtraction, multi-pedestrian tracking, and classification for thermal-visible videos have been proposed [3, 8, 5]. However, the main difficulty associated with the joint use of thermal and visible videos, before any further analyses, is the matching and the registration of pairs of images captured by two different types of sensors, where one records thermal signatures and the other records the color information of the scene. Due to the numerous differences in imaging characteristics of thermal and visible cameras, most correspondence measures used for registering images of single imaging modality are not applicable. Also, it is very difficult to find correspondence for an entire scene. For people monitoring application, partial image region of interest (ROI) registration approach is the most common approach for thermal-visible video registration. In this approach, the problem is simplified to register the pixels associated with the objects' ROIs, such as moving people in a scene. However, the patterns and textures of corresponding thermal and visible ROIs related to a person are often not the same. People might have clothes with colorful textures which are visible in color images, but not in thermal images. Moreover, there might be some textures visible on thermal images caused by different clothing (e.g. light clothes/warm clothes) and the amount of emitted energy from different part of human body which are not visible in color images. In fact, for registering people in thermal and visible image pairs, the human body's spatial layout (shape) is the most used visual information for matching human ROIs.

MI is the most commonly used multimodal dense stereo correspondence measure in the literature [4, 7, 1]. MI measures the statistical co-occurrence of pixel-wise information such as local textures and patterns of matching regions. Egnal [4] has shown that mutual information (MI) is a viable similarity metric for matching disparate thermal and visible images recorded from a scene with a multimodal imagery system. He has also shown that MI, as multimodal correspondence measure, outperforms the traditional dense stereo correspondence box matching method, such as normalized cross-correlation-based methods. For people monitoring applications, Chen et al. proposed a MI-based registration method that matches boxes in the two images with the assumption that each box represents one single object [1]. In their method, occluded people extracted by a background subtraction that are merged in one ROI may not be accurately registered since an ROI may contain people within different depth planes. As a solution to accurately register occluded people in a scene, Krotosky and Trivedi

proposed a disparity voting (DV) matching method [7]. DV is performed by horizontally sliding (column by column) small width matching boxes on rectified thermal and visible images, computing MI for pairs of boxes, and finally counting the number of votes for each disparity based on a Winner Take All (WTA) approach. However, in both of these papers, authors have not discussed the limitations of MI. In uncontrolled settings, when people have clothes with different patterns, where there are partial ROI misdetections (some human body boundaries are missing), or occlusions, MI is unreliable for matching small width windows like the one proposed in [7]. MI-based matching fails to correctly match image boxes where the joint probability histogram is not sufficiently populated. Choosing the appropriate image box size is not straightforward due to the aforementioned difficulties. Also, there is always a tradeoff between choosing larger matching boxes for matching evidence, and smaller matching boxes for the precision and details needed for an accurate registration.

In this work, we apply the local self-similarity (LSS) to the problem of multimodal stereo video registration. LSS has been proposed by Shechtman and Irani [9]. In the literatures, LSS has been applied for applications such as object categorization, image classification, pedestrian detection, and object detection [13, 14, 12]. To the best of our knowledge, LSS has never been used before as a dense stereo correspondence measure for multimodal video registration. LSS, similarly to MI, computes statistical co-occurrence of pixel intensities. However LSS, unlike MI, is firstly computed and extracted from an individual image as a descriptor and then compared between pair of images. The property of LSS, which makes this measure more interesting for our application, is that the basic unit for measuring internal joint pixel statistics is small image patches that capture more meaningful image pattern than individual pixels used in MI computation. In this work we give a comparative analysis of LSS and MI as multimodal correspondence measures.

Moreover, we use LSS as a similarity measure in a Disparity voting (DV) registration method [7]. We have tested qualitatively and quantitatively the MI-based registration [7] and our LSS-based registration in several close range indoor videos. We show that our LSS-based registration method outperforms the MI-based registration method [7] on challenging indoor videos with multiple occluded people in the scene, partial ROI misdetections, and people with different clothing.

2. LSS as a dense stereo correspondence measure

Mutual information (MI) is the most commonly used correspondence measure for multimodal stereo video registration. The MI between two image boxes is defined as

$$I(X,Y) = \sum_{X \in I_l} \sum_{Y \in I_r} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)}, \quad (1)$$

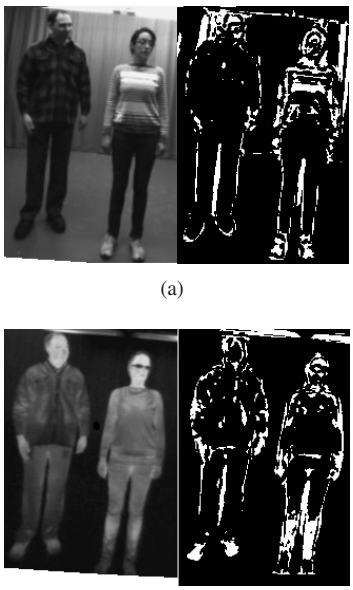
where $P(X,Y)$, the joint probability mass function, is the joint probability histogram of thermal and visible image boxes I_l and I_r normalized by the sum of the joint histogram entries. $P(X)$ and $P(Y)$ are the marginal probability functions. MI as similarity metric computes the statistical co-occurrence of pixel-wise measures such as image patterns inside thermal and visible image boxes. In our application, MI is not reliable for matching corresponding regions where the joint probability histogram is not sufficiently populated. This may happen when matching corresponding thermal-visible human ROIs with similar shapes (spatial layout) but different patterns, matching partially misdetected ROIs, or occluded ROIs (two or more people are merged in one single ROI).

We propose using local self-similarity (LSS) as a more robust and discriminative measure for thermal-visible stereo registration. LSS computes statistical co-occurrence of small image box (e.g. 4×4 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 image at the centered at pixel p and image boxes in a larger surrounding image region. SSD is normalized by the maximum value of the small image patch intensity variance and noise. It is defined as

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

Then, a LSS descriptor is defined as a partitioned log-polar representation of the correlation surface with e.g. 80 bins (20angles and 4 radial intervals).

LSS has two main advantages over MI as correspondence measure. Firstly, LSS is computed separately as set of descriptors in one individual image and then it is compared in a matching process across pair of images. This enables us to detect informative regions (region containing informative descriptors) inside human ROIs in the image and then using those regions for matching. Secondly, the measurement unit for LSS is a small image patch which contains more meaningful patterns compared to a pixel as used for MI computation. As it is described in Shechtman and Irani's work [9], this property makes LSS a suitable measure for matching textured region in one image with uniformly colored region or differently textured region in the other image, as long as they have similar spatial layout. Thus, for matching thermal and visible human ROIs of people wearing clothes with different patterns, LSS-based matching is more reliable than MI-based matching. Before matching two sets of descriptors in pairs of thermal and visible images, we discard the



(a)

(b)

Figure 1. Informative LSS descriptors for thermal and visible images in fig. 2. (a) Visible image and pixels with informative LSS descriptors and (b) thermal image and pixels with informative LSS descriptors.

non-informative descriptors. 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 (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 [6]. Discarding non-informative descriptors is like an implicit segmentation or edge detection, which for box matching, increases the discriminative power of LSS measure and avoids ambiguous matching. It is important to notice that the remaining informative descriptors still form a more dense collection compared to sparse interest points. Fig. 1 shows pixels having informative descriptors (white pixels) for a pair of thermal and visible images. The regions belonging to the human body's boundaries and image patterns are the informative regions. This is done without any explicit edge detection or segmentation.

3. Our LSS-based multimodal stereo registration method

Our method is applicable for pairs of rectified thermal and visible images. Our goal is registering the ROIs related to people in a scene. We integrated LSS, as similarity measure, in a WTA scanline-search box matching. For registration, we used the disparity voting method in [7]. This method resolves the problem of registering occluded people merged in an ROI. Therefore, it is a suitable choice for our application where multiple people are moving in a scene with various levels of occlusion. As preprocessing, we extract foreground pixels using the background subtraction method proposed in [10]. Any background subtraction method with a reasonable amount of error is applicable. For indoor videos, partial ROI misdetections and object fragmentation in thermal images are less frequent compared to visible images. In our method, we register each extracted ROI in thermal image on the visible foreground image.

3.1. LSS descriptors computation

Before the matching process, the LSS descriptors are extracted from individual images. For all the foreground pixels in the rectified thermal and visible images, LSS descriptors [9] are computed as it is described in section 2. LSS contains the co-occurrence information such the ROI pattern inside a patch, centered on a foreground pixel p, compared to a larger surrounding region. Local self-similarity is computed for foreground pixels using the original rectified thermal and visible images. This results in computed LSS descriptors that are less affected by partial ROI misdetection. This process associates a descriptor to each foreground pixel p. The LSS descriptors containing no self-similarity information (non-informative descriptors) are discarded, using the method described in section 2, and are not used in any further computation.

3.2. LSS-based image box matching

After extracting LSS descriptor, we perform matching between a pair of thermal and visible foreground images. For each ROI B_i in the thermal image, a box $W_{l,j}$ centered at jth column of B_i with the same height h as the height of the ROI and a width of m that is much smaller than the width of a person is defined. Also, a corresponding box $W_{r,j+d}$ is defined on the visible image with the same size as $W_{l,j}$, but centered at $j+d$ where d is a disparity offset of set of disparity D. For each pair of boxes $W_{l,j}$ and $W_{r,j+d}$, a normalized similarity distance $SD_{j,d}$, which is the sum of $L1$ distances of the corresponding pixels $p_l \in W_{l,j}$ and $p_r \in W_{r,j+d}$ having informative descriptors, is computed as

$$SD_{j,d} = \frac{\sum_{p_l,p_r} L1_{l,r}(p_l,p_r)}{N}, \tag{3}$$

where N is the number of corresponding pixels p_l and p_r contributing in the similarity distance computation. Then, $L1_{l,r}$ is computed as

$$L1_{l,r}(p_l,p_r) \sum_{k=1}^{80} |d_{p_l}(k) - d_{p_r}(k)| \tag{4}$$

where 80 is the number of LSS descriptor's bins. Finally, the best disparity $d_{min} \in D$ is computed as

$$d_{min} = argmin_{d}(SD_{l,j,d}), d \in D. \quad (5)$$

3.3. LSS-based DV disparity assignment

For each B_i, we build a disparity voting matrix DV_i of size $(N, d_{max} - d_1 + 1)$ where N is the number of pixels of B_i. This procedure is performed by shifting column by column $W_{l,j}$ in thermal image for all the columns $j \in B_i$, then doing image box matching. For each d_{min}, a vote is added to $DV_i(p_l, d_{min})$ for all $p_l \in (W_{l,j} \cap B_i)$. Since the width of image boxes are m pixels, we have m votes for each pixel belonging to B_i. Finally, the disparity map DM_i is computed as,

$$DM_i(p_l) = argmax_{d}(D_i(p_l, d)). \quad (6)$$

We align each blob B_i of thermal image on the visible image using the assigned disparity to each pixel $p \in B_i$. This process is repeated until registering all the ROIs in the thermal image on the visible foreground image.

4. Experimental validation and discussion

We used synchronized visible-thermal videos of a $5m \times 5m$ room captured by thermal and visible cameras with a baseline of 12 cm. We have chosen the camera baseline large enough to get distinguishable disparities at the pixel level for people in different depths. The cameras were oriented in the same direction with their optical axes parallel to each other and to the ground. Using this camera setup satisfies our assumption that each person lie approximately within one depth plane in the scene. We used videos of up to 5 people in the scene with various poses, clothes, distances to cameras, and different levels of occlusions. To compare MI and LSS as similarity measures, we prepared three examples which assess the accuracy of both measures for different real world cases. We also compare the registration accuracy of our LSS-based registration method (LSS-DV) with a MI-based registration method (MI-DV) [7].

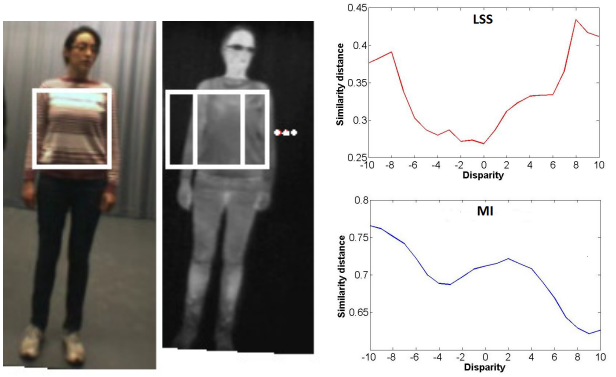
4.1. Comparing MI and LSS as multimodal similarity measures

For evaluating the suitability of MI and LSS as similarity measures for matching thermal and visible pair of images, we prepared three real world examples. We performed scanline-search box matching by computing LSS-based and MI-based similarity distances between a region bounded inside a box on the visible image and a second region on the thermal image bounded inside a similarly sized sliding box. We defined LSS-based similarity distance between two image boxes by the sum of $L1$ distances of informative descriptors within the thermal and visible image boxes. We defined the MI-based similarity distance as one minus the mutual information between the two image boxes $(1 - I(X, Y))$.

Fig. 2 (a) shows an example of matching a textured region (inside the white box) in the color image with a corresponding uniform region in the thermal image. We rectified and manually registered the thermal-visible images such as a disparity of 0 corresponds to a perfect alignment. Fig. 2 (b) shows the similarity distance of MI and LSS between image boxes in the visible image and thermal images for disparity offsets varying between $[-10, 10]$. Fig. 2 (b) upper plot shows that for LSS, the similarity distance is correctly minimized at disparity 0. However for MI, the similarity distance minimum is not around 0. This would result in an inaccurate matching in a winner take all approach (Fig. 2 (b) lower plot). This illustrates that MI is not a robust similarity metric for matching a textured region and a uniform region where there is not many similar patterns and edges within thermal and visible corresponding image boxes. Fig. 3 shows an example of matching 20×20 and 50×50 pixels image boxes for the human's head region. Fig. 3 (b) lower plot shows that MI is not a discriminative measure for matching 20×20 thermal-visible image boxes. However, using larger image boxes of size 50×50 pixels containing more similar patterns and more similar ROI's spatial layout, MI similarity distance is correctly minimized at disparity 0. Fig. 3 (b) upper plot shows that LSS similarity distance is correctly minimized at disparity 0 for both image box sizes. This shows that LSS is more robust and discriminative for smaller size image matching boxes compared to MI. Fig. 4 (a) shows an example of matching thermal-visible image boxes using foreground images with box sizes of 20×170 and 60×170 pixels. In the visible image, due to the color similarity of the ROI and the background, some parts of the body region are not detected. Fig. 4 (b), lower plot shows that MI fails to find correct disparity offset using both sizes of image matching boxes. However, LSS find the correct disparity (Fig. 4 (b) upper plot). LSS descriptors are computed for ROI's pixels using the original images before background subtraction. This is feasible for LSS since descriptor computation and matching are done in two separate processes. However for MI, this is not possible. Therefore, the bad impact of partial ROI misdetection (human body region with unclear boundaries) on MI is inevitable. Moreover for LSS, the non informative descriptors in both thermal and visible ROIs are discarded. Therefore LSS is more discriminative measure compared to MI.

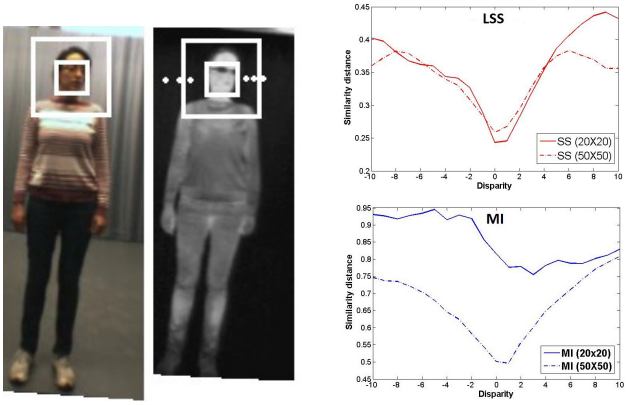
4.2. Comparison of MI-based and LSS-based multimodal registration

Table 1 shows the percentage of frames with at least one person that is visually misaligned (inaccurate registration) for two videos with different level of occlusion, people with different clothing, and partial ROI misdetection errors. The



(a) (b)

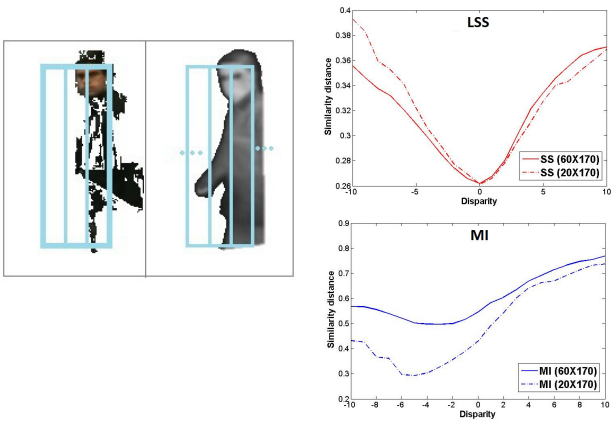
Figure 2. Matching corresponding textured and uniform regions in visible and thermal pair of images. (a) Aligned visible and thermal images and (b) Similarity distances of LSS and MI for disparity interval of [-10,10].



(a) (b)

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

video seq. 1 is captured in summer when people have lighter clothes. The clothes have also very few textures. Our results show that LSS-DV outperforms MI-DV even in this simpler case. MI-based registration is inaccurate for most frames with dramatic background subtraction errors where there are not enough similar human body region boundaries existing inside matching image boxes. However, LSS is more robust to partial ROI misdetection errors compared to MI because descriptors are computed for foreground pixels using the original images (before background subtraction). Recall that this is feasible for LSS-based registration since descriptor computation and thermal-visible image matching are done in two separate processes. Furthermore, discarding non-informative descriptors before box matching increases



(a) (b)

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

LSS discriminative power. The video seq. 2 is captured in winter when people have warm clothes. The thick clothes cause some extra patterns appear in thermal images. Furthermore, some people have textured clothes (see Fig. 5). Our results show that for this video, MI-DV registration is unreliable and registration is inaccurate for almost all frames. However the LSS-based registration is considerably more robust to appearance differences between thermal and visible images, which shows its advantage compared to a similar MI-based registration method.

Seq.	**Method**	**NP**	**NF**	**NC**	**E**
1	MI-DV	4	726	620	15.61
	LSS-DV			673	7.31
2	MI-DV	5	546	27	95.05
	LSS-DV			533	2.38

Table 1. Registration errors for 726 frames. **NP**: number of people in the video, **NF**: total number of frames, **NC**: number of correctly registered frames and **E**: percentage of incorrectly registered video frames.

Fig. 5 shows qualitative registration results of our LSS-based and the MI-based registration methods for 5 frames of video seq. 2. Fig. 5 shows that people have thicker clothes. Moreover, there are considerable background subtraction errors in frames 248, 354, and 364. Fig. 5 rows (a) and (b) show the rectified thermal and visible pairs of images. Fig. 5 rows (c) and (d) show the thermal and visible foreground images. Fig. 5 rows (e) and (f) show respectively the registration results of MI-based and LSS-based methods. The considerable differences between the accuracy of registration results between two methods are easily recognizable. LSS is more robust to differences of appear-

ance and background detection errors.

5. Conclusion

In this paper, we applied LSS as a multimodal stereo correspondence measure, and shown its advantages compared to MI, the most commonly used multimodal stereo correspondence measure in the state-of-the-art. We also proposed an LSS-based registration method for people monitoring applications. Our results have shown that our method significantly outperforms the MI-based registration method in [7]. As future direction for this work, we will work on applying the LSS descriptor as a multimodal stereo correspondence metric for an energy minimization-based dense stereo correspondence method.

References

[1] 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)*, pages 313 – 318, 2003. 61

[2] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456 –1477, Oct. 2001. 61

[3] J. W. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Comput. Vis. Image Underst.*, 106, 2007. 61

[4] G. Egnal. Mutual information as a stereo correspondence measure. *Tech. Rep. MS-CIS-00-20, University of Pennsylvania*, 2000. 61

[5] J. Han and B. Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771 – 1784, 2007. 61

[6] P. O. Hoyer and P. Dayan. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004. 63

[7] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106(2-3):270 – 287, 2007. 61, 62, 63, 64, 66

[8] A. Leykin. Thermal-visible video fusion for moving target tracking and pedestrian classification. In *In Object Tracking and Classification in and Beyond the Visible Spectrum Workshop at the International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 61

[9] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1 –8, 2007. 62, 63

[10] B. Shoushtarian and H. E. Bez. A practical adaptive approach for dynamic background subtraction using an invariant colour model and object tracking. *Pattern Recogn. Lett.*, 26(1):5–26, 2005. 63

[11] D. Socolinsky. Design and deployment of visible-thermal biometric surveillance systems. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –2, 2007. 61

[12] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *IEEE 12th International Conference on Computer Vision (ICCV 2009)*, pages 606 – 613, 2009. 62

[13] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1030–1037, 2010. 62

[14] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. In *IEEE 12th International Conference on Computer Vision (ICCV 2009)*, pages 436 –443, 2009. 62

[15] Z. Zhu and T. Huang. Multimodal surveillance: an introduction. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –6, 2007. 61

Figure 5. (a) Visible images, (b) thermal images, (c) visible foreground images, (d) thermal foreground images, (e) MI-DV registration IR on visible, and (f) LSS-DV registration IR on visible.