UNIVERSITÉ DE MONTRÉAL

VIDEO REGISTRATION FOR MULTIMODAL SURVEILLANCE SYSTEMS

ATOUSA TORABI
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE INFORMATIQUE)
AVRIL 2012

© Atousa Torabi, 2012.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

VIDEO REGISTRATION FOR MULTIMODAL SURVEILLANCE SYSTEMS

présentée par : TORABI Atousa
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment accepté par le jury d'examen constitué de :

M. LANGLOIS J.M. Pierre, Ph.D., président
M. BILODEAU Guillaume-Alexandre, Ph.D., membre et directeur de recherche
M. SAUNIER Nicolas, Ph.D., membre
M. LAURENDEAU Denis, Doctorat, membre

ACKNOWLEDGEMENT

My utmost gratitude and appreciation is extended to my supervisor Dr. Bilodeau in the Department of Computer engineering and software engineering at École Polytechnique de Montréal. I am obliged for his judicious advice and criticism, steady encouragement and availability, and support during my doctoral training.

I am also indebted to my colleagues at LITIV research Laboratory. Their ability, resourcefulness, and dedication to the experimental work were instrumental to my success. I will always be grateful to the program's faculty and students who, with exemplary dedication and talent, taught me about the marvels of computer vision. I sincerely thank LITIV analyst, Mr. St-Onge for his help in the implementation phase of my project and Dr. Pal for his valuable mentorship for my research proposal.

I could never express enough gratitude for the unconditional encouragement and enthusiasm from my family, and friends.

Finally, I would like to thank members of the jury for their valuable comments on my thesis.

RÉSUMÉ

Au cours de la dernière décennie, la conception et le déploiement de systèmes de surveillance par caméras thermiques et visibles pour l'analyse des activités humaines a retenu l'attention de la communauté de la vision par ordinateur. Les applications de l'imagerie thermique-visible pour l'analyse des activités humaines couvrent différents domaines, notamment la médecine, la sécurité à bord d'un véhicule et la sécurité des personnes. La motivation derrière un tel système est l'amélioration de la qualité des données dans le but ultime d'améliorer la performance du système de surveillance. Une difficulté fondamentale associée à un système d'imagerie thermique-visible est la mise en registre précise de caractéristiques et d'informations correspondantes à partir d'images avec des différences significatives dans les propriétés des signaux. Dans un cas, on capte des informations de couleur (lumière réfléchie) et dans l'autre cas, on capte la signature thermique (énergie émise). Ce problème est appelé mise en registre d'images et de séquences vidéo.

La vidéosurveillance est l'un des domaines d'application le plus étendu de l'imagerie multispectrale. La vidéosurveillance automatique dans un environnement réel, que ce soit à l'intérieur ou à l'extérieur, est difficile en raison d'un nombre élevé de facteurs environnementaux tels que les variations d'éclairage, le vent, le brouillard, et les ombres. L'utilisation conjointe de différentes modalités permet d'augmenter la fiabilité des données d'entrée, et de révéler certaines informations sur la scéne qui ne sont pas perceptibles par un système d'imagerie unimodal. Les premiers systèmes multimodaux de vidéosurveillance ont été conçus principalement pour des applications militaires. Mais de nos jours, en raison de la réduction du prix des caméras thermiques, ce sujet de recherche s'étend à des applications civiles ayant une variété d'objectifs.

Les approches pour la mise en registre d'images pour un système multimodal de vidéosurveillance automatique sont divisées en deux catégories fondées sur la dimension de la scène : les approches qui sont appropriées pour des grandes scènes où les objets sont lointains, et les approches qui conviennent à de petites scènes où les objets sont près des caméras. Dans la littérature, ce sujet de recherche n'est pas bien documenté, en particulier pour le cas de petites scènes avec objets proches. Notre recherche est axée sur la conception de nouvelles solutions de mise en registre pour les deux catégories de scènes dans lesquels il y a plusieurs humains. Les solutions proposées sont incluses dans les quatre articles qui composent cette thèse. Nos méthodes de mise en registre sont des prétraitements pour d'autres tâches d'analyse vidéo telles que le suivi, la localisation de l'humain, l'analyse de comportements, et la catégorisation d'objets.

Pour les scènes avec des objets lointains, nous proposons un système itératif qui fait de façon simultanée la mise en registre thermique-visible, la fusion des données et le suivi des personnes. Notre méthode de mise en registre est basée sur une mise en correspondance de trajectoires (en utilisant RANSAC) à partir desquelles on estime une matrice de transformation affine pour transformer globalement des objets d'avant-plan d'une image sur l'autre image. Notre système proposé de vidéosurveillance multimodale est basé sur un nouveau mécanisme de rétroaction entre la mise en registre et le module de suivi, ce qui augmente les performances des deux modules de manière itérative au fil du temps. Nos méthodes sont conçues pour des applications en ligne et aucune calibration des caméras ou de configurations particulières ne sont requises.

Pour les petites scènes avec des objets proches, nous introduisons le descripteur Local Self-Similarity (LSS), comme une mesure de similarité viable pour mettre en correspondance les régions du corps humain dans des images thermiques et visibles. Nous avons également démontré théoriquement et quantitativement que LSS, comme mesure de similarité thermique-visible, est plus robuste aux différences entre les textures des régions correspondantes que l'information mutuelle (IM), qui est la mesure de similarité classique pour les applications multimodales. D'autres descripteurs viables, y compris Histogram Of Gradient (HOG), Scale Invariant Feature Transform (SIFT), et Binary Robust Independent Elementary Feature (BRIEF) sont également surclassés par LSS.

En outre, nous proposons une approche de mise en registre utilisant LSS et un mécanisme de votes pour obtenir une carte de disparité stéréo dense pour chaque région d'avant-plan dans l'image. La carte de disparité qui en résulte peut alors être utilisée pour aligner l'image de référence sur la seconde image. Nous démontrons que notre méthode surpasse les méthodes dans l'état de l'art, notamment les méthodes basées sur l'information mutuelle. Nos expériences ont été réalisées en utilisant des scénarios réalistes de surveillance d'humains dans une scène de petite taille.

En raison des lacunes des approches locales de correspondance stéréo pour l'estimation de disparités précises dans des régions de discontinuité de profondeur, nous proposons une méthode de correspondance stéréo basée sur une approche d'optimisation globale. Nous introduisons un modéle stéréo approprié pour la mise en registre d'images thermique-visible en utilisant une méthode de minimisation de l'énergie en conjonction avec la méthode Belief Propagation (BP) comme méthode pour optimiser l'affectation des disparités par une fonction d'énergie. Dans cette méthode, nous avons intégré les informations de couleur et de mouvement comme contraintes douces pour améliorer la précision d'affectation des disparités dans les cas de discontinuités de profondeur. Bien que les approches de correspondance globale soient plus gourmandes au niveau des ressources de calculs par rapport aux approches

de correspondance locale basée sur la stratégie Winner Take All (WTA), l'algorithme efficace BP et la programmation parallèle (OpenMP) en C++ que nous avons utilisés dans notre implémentation, permettent d'accélérer le temps de traitement de manière significative et de rendre nos méthodes viables pour les applications de vidéosurveillance. Nos méthodes sont programmées en C++ et utilisent la bibliothèque OpenCV.

Nos méthodes sont conçues pour être facilement intégrées comme prétraitement pour toute application d'analyse vidéo. En d'autres termes, les données d'entrée de nos méthodes pourraient être un flux vidéo en ligne, et pour une analyse plus approfondie, un nouveau module pourrait être ajouté en aval à notre schéma algorithmique. Cette analyse plus approfondie pourrait être le suivi d'objets, la localisation d'êtres humains, et l'analyse de trajectoires pour les applications de surveillance multimodales de grandes scène. Aussi, Il pourrait être l'analyse de comportements, la catégorisation d'objets, et le suivi pour les applications sur des scènes de tailles réduites.

ABSTRACT

Recently, the design and deployment of thermal-visible surveillance systems for human analysis attracted a lot of attention in the computer vision community. Thermal-visible imagery applications for human analysis span different domains including medical, in-vehicle safety system, and surveillance. The motivation of applying such a system is improving the quality of data with the ultimate goal of improving the performance of targeted surveillance system. A fundamental issue associated with a thermal-visible imaging system is the accurate registration of corresponding features and information from images with high differences in imaging characteristics, where one reflects the color information (reflected energy) and another one reflects thermal signature (emitted energy). This problem is named Image/video registration.

Video surveillance is one of the most extensive application domains of multispectral imaging. Automatic video surveillance in a realistic environment, either indoor or outdoor, is difficult due to the unlimited number of environmental factors such as illumination variations, wind, fog, and shadows. In a multimodal surveillance system, the joint use of different modalities increases the reliability of input data and reveals some information of the scene that might be missed using a unimodal imaging system. The early multimodal video surveillance systems were designed mainly for military applications. But nowadays, because of the reduction in the price of thermal cameras, this subject of research is extending to civilian applications and has attracted more interests for a variety of the human monitoring objectives.

Image registration approaches for an automatic multimodal video surveillance system are divided into two general approaches based on the range of captured scene: the approaches that are appropriate for long-range scenes, and the approaches that are suitable for close-range scenes. In the literature, this subject of research is not well documented, especially for close-range surveillance application domains. Our research is focused on novel image registration solutions for both close-range and long-range scenes featuring multiple humans. The proposed solutions are presented in the four articles included in this thesis. Our registration methods are applicable for further video analysis such as tracking, human localization, behavioral pattern analysis, and object categorization.

For far-range video surveillance, we propose an iterative system that consists of simultaneous thermal-visible video registration, sensor fusion, and people tracking. Our video registration is based on a RANSAC object trajectory matching, which estimates an affine transformation matrix to globally transform foreground objects of one image on another one. Our proposed multimodal surveillance system is based on a novel feedback scheme between

registration and tracking modules that augments the performance of both modules iteratively over time. Our methods are designed for online applications and no camera calibration or special setup is required.

For close-range video surveillance applications, we introduce Local Self-Similarity (LSS) as a viable similarity measure for matching corresponding human body regions of thermal and visible images. We also demonstrate theoretically and quantitatively that LSS, as a thermal-visible similarity measure, is more robust to differences between corresponding regions' textures than the Mutual Information (MI), which is the classic multimodal similarity measure. Other viable local image descriptors including Histogram Of Gradient (HOG), Scale Invariant Feature Transform (SIFT), and Binary Robust Independent Elementary Feature (BRIEF) are also outperformed by LSS.

Moreover, we propose a LSS-based dense local stereo correspondence algorithm based on a voting approach, which estimates a dense disparity map for each foreground region in the image. The resulting disparity map can then be used to align the reference image on the second image. We demonstrate that our proposed LSS-based local registration method outperforms similar state-of-the-art MI-based local registration methods in the literature. Our experiments were carried out using realistic human monitoring scenarios in a close-range scene.

Due to the shortcomings of local stereo correspondence approaches for estimating accurate disparities in depth discontinuity regions, we propose a novel stereo correspondence method based on a global optimization approach. We introduce a stereo model appropriate for thermal-visible image registration using an energy minimization framework and Belief Propagation (BP) as a method to optimize the disparity assignment via an energy function. In this method, we integrated color and motion visual cues as a soft constraint into an energy function to improve disparity assignment accuracy in depth discontinuities. Although global correspondence approaches are computationally more expensive compared to Winner Take All (WTA) local correspondence approaches, the efficient BP algorithm and parallel processing programming (openMP) in C++ that we used in our implementation, speed up the processing time significantly and make our methods viable for video surveillance applications. Our methods are implemented in C++ using OpenCV library and object-oriented programming.

Our methods are designed to be integrated easily for further video analysis. In other words, the input data of our methods could come from two synchronized online video streams. For further analysis a new module could be added in our frame-by-frame algorithmic diagram. Further analysis might be object tracking, human localization, and trajectory pattern analysis for multimodal long-range monitoring applications, and behavior pattern analysis, object

categorization, and tracking for close-range applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BRIEF	Binary Robust Independent Elementary Feature
BP	Belief Propagation
DV	Disparity Voting
FOV	Field Of View
FPS	Frame Per Second
GC	Graph-Cuts
HCI	Human Computer Interaction
HOG	Histogram Of Gradient
IM	Information Mutuelle
LID	Local Image Descriptor
LSS	Local Self-Similarity
MAP	Maximization of A Posteriori
MRF	Markov Randon Field
MI	Mutual Information
NCC	Normalized Cross-Correlation
NP	Nondeterministic Polynomial
RANSAC	RANdom SAmple Consensus
ROI	Region Of Interest
SIFT	Scale Invariant Feature Transform
WTA	Winner Take All

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Multispectral Imaging Systems

Over recent years, we witnessed a rapid growing interest in research, design and deployment of multispectral sensing systems in a variety of human analysis applications such as medical and video surveillance system. Depending on the application, different modalities can be used such as video, audio, thermal vibrations, *etc*. In fact, the joint use of multimodal sensors is one mean for augmenting the quality of the input data with the ultimate goal of improving overall system performance. This is often the main motivation for development of multispectral systems. However, besides the advantages of such a system, its complexity also increases by the addition of a new sensor, not to mention its cost. In the book of Zhu et al. (Zhu and Huang (2007)), the authors address the issues associated to various aspects of multimodal sensing systems.

A multimodal sensing system usually consists of three main components that are multimodal sensing, multimodal data fusion, and finally automatic multimodal data analysis. The details of these components vary from one to another system. In our research, we are interested in imaging sensors.

Figure 1.1 shows the main components of a multimodal video surveillance system for human analysis. In the sensing component, sensors are either two or multiple imaging mo-

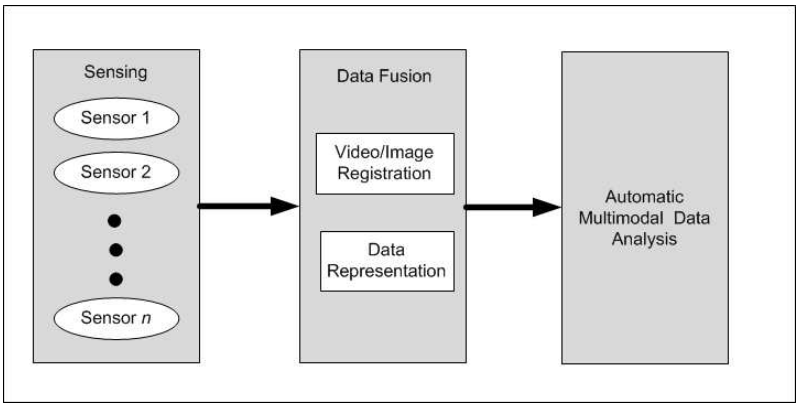


Figure 1.1 Multimodal video surveillance system components.

dalities. The cameras should record videos synchronously. Data fusion is the most crucial component in a multimodal video surveillance system. In this part, the synchronized video frames coming from multiple cameras should be aligned, and augmented (combined) data should be represented properly for targeted application. Finally, the higher-level data analysis such as object tracking, and human activity pattern analysis is done in the automatic multimodal data analysis module.

1.1.2 Thermal-Visible Sensing For Human Image ROI Analysis

The reduction in the price of thermal cameras resulted in a growing interest in human image Region Of Interest (ROI) analysis using thermal and visible cameras. The advantages of jointly using thermal and visible cameras as a multimodal imaging system have been studied and discussed in few works (Zhu and Huang (2007); Socolinsky (2007)). Thermal-visible imaging system for human analysis has been applied in both civilian and military applications. Figure 1.2 illustrates the application domains of thermal-visible imaging system in the literature.

For medical applications, temperature is important information that can be extracted from thermal images and used to detect and to diagnose diseases such as skin tumor and arthritis. For medical applications, the combination of thermal and visible human image ROIs allows the rich information provided by visible cameras to be used to assist the search of thermal patterns in regions of interest on the thermal images. Attempts have also been made to combine thermal image with stereo visible image of a face for inflammation diagnosis (Ju *et al.* (2010)).

For in-vehicle safety system, Krotosky *et al.* (Trivedi *et al.* (2004)) used a stereo visible camera and a single infrared camera for driver posture analysis. In their work, the thermal

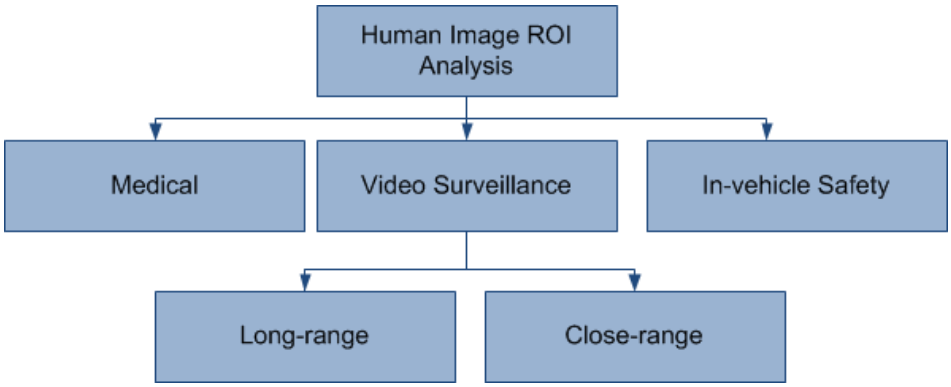


Figure 1.2 Human image ROI analysis application domains using a thermal-visible imaging system.

(a) (b)

Figure 1.3 Implicit object detection. Human ROIs are extracted using a background subtraction.(a) thermal (left) and visible (right) corresponding human ROIs of a person carrying a bag (b) thermal (left) and visible (right) corresponding human ROIs of a person carrying a hot pot.

and visible data are jointly used to detect the skin part of the occupant and especially its face position for making airbag-deployment decisions.

Automatic video surveillance in uncontrolled settings is a challenging task due to the infinite variety of environmental factors and challenging goals of human monitoring. Even the most advanced algorithms of object detection, tracking, and behavior pattern analysis might fail using a single imaging modality. Thermal sensors, in combination with visible sensors, open up new possibilities for performing system in challenging situation such as different illumination conditions and environmental variations. Two main advantages of the joint use of thermal and visible sensors are first the complementary nature of different modalities that provides the thermal and color information of the scene, and second, the redundancy of information captured by the different modalities, which increases the reliability of input data and consequently the robustness of the surveillance system. So far, thermal-visible surveillance systems are applied mostly in human localization and tracking in long-range scene. The multimodal close-range surveillance had the least attention. In the book entitled "Augmented Vision Perception in Infrared" (Hammoud (2009)), Hammoud gives a complete survey of state-of-the-art works for both close-range and long-range surveillance.

For long-range applications, the complementary nature of different modalities allows to better detect and keep track of monitoring targets (mostly people) in challenging environmental conditions such as fog, wind, and lack of illuminations.

For close-range human monitoring applications, the complementary features of the aligned visible and the thermal human ROIs in a pair of images, enable us to implicitly segment or detect different regions belonging to different objects in interaction with human body ROIs, based on the difference of object regions' temperatures and their visibility in each modality.

Such a property can be used to implicitly detect either hotter or colder objects compared to human body temperature. Figure 1.3 (a) shows an example of extracted human body ROIs in corresponding thermal and visible images using a background subtraction method. In this example, the bag has about the same temperature as the background ; therefore it is not detected in the thermal image. However, it is detected in the visible image based on its color difference with the background. Figure 1.3 (b) shows an example of a person carrying a hot pot. Since the pot has a higher temperature compared to the human body, it is implicitly detected in the thermal image. However, in the visible image, it is not possible to easily detect the hot pot. So the aligned thermal and visible ROIs enable us to take advantage of complementary features of these two modalities. Such advantages motivated computer vision community to continue studying and investigating algorithms for thermal-visible video surveillance systems for a close-range or nearly close-range scene.

1.1.3 Bi-modal Video Registration Approaches

The fundamental and preliminary task associated with the joint use of thermal-visible data is accurately matching features and aligning a pair of images captured by two different sensors. This problem is named video (or image) registration. In the literature, video registration problem is defined either as a low-level image processing problem or a high-level video processing problem. In the first case, the video registration is similar to low-level image registration ; the only required pre-processing is video synchronization that simplifies extracting a pair of corresponding thermal and visible video frames. In the second case, video registration problem defined as a high-level video processing problem that uses several pairs of video frames information rather than a pair of images information.

Figure 1.5 illustrates three state-of-the-art image (or video) registration approaches. Re-

Figure 1.4 Camera setup.

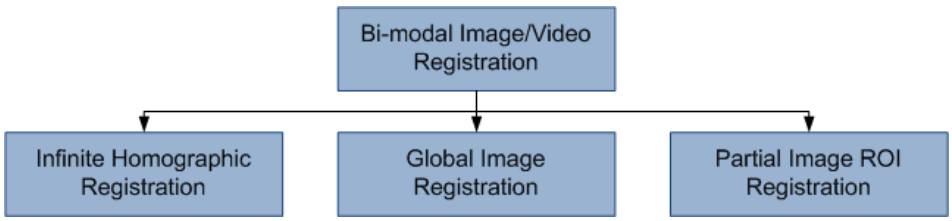


Figure 1.5 Bi-Modal Image Registration Approaches

gistration approaches vary based on the application domains (long-range or close-range) and factors such as camera positioning and desired accuracy of registered objects in the scene. To better understand the registration approaches, it is desirable to briefly outline the geometric framework behind the registration problem. The books (Hartley and Zisserman (2003a); Trucco and Verri (1998)) give an extensive mathematical definition of multiple views geometry.

Epipolar Geometry : Figure 1.6 illustrates epipolar geometry for a setup with two cameras. O_L (for left camera) and O_R (for right camera) are camera centers. The name epipolar geometry is used because the points at which the line through the centers O_L and O_R intersects the image planes are named epipoles. E_L is the image of the projection center of the right camera (visible camera) and E_R is the image of the projection center of the left camera (thermal camera). Given P_1 a 3-D point in the scene defined relative to each of the camera coordinate centers $P_1^L = (X, Y, Z)$ (left camera) and $P_1^R = (X^{'}, Y^{'}, Z^{'})$ (right camera), $p_1 = (x, y, 1)$ and $p_1^{'} = (x^{'}, y^{'}, 1)$ are 2-D projected points of P_1 on left camera image plane using K projection matrix and on the right camera image plane using $K^{'}$ projection matrix. Respectively, π is a plane in the scene defined by its surface normal of the plane and its distance from the camera center OL; the homography induced by π is $P_1^R = H_p P_1^L$ and the projection matrix is $p_1^{'} = Hp_1$ where $H = K^{'} H_p K^{-1}$. However, not all the points in the scene lie on the plane π, like point P_2 shown in figure 1.6. In this case, an additional parallax component needs to be added to take in account the projective depth of the other point (P_2) relative to plane π. So the transformation matrix that includes the parallax term is defined as,

$$p^{'} = Hp + \delta \tag{1.1}$$

where δ is the parallax term.

Based on these principles, in the following sections, we describe the aforementioned three bi-modal registration approaches.

Infinite Homographic Registration

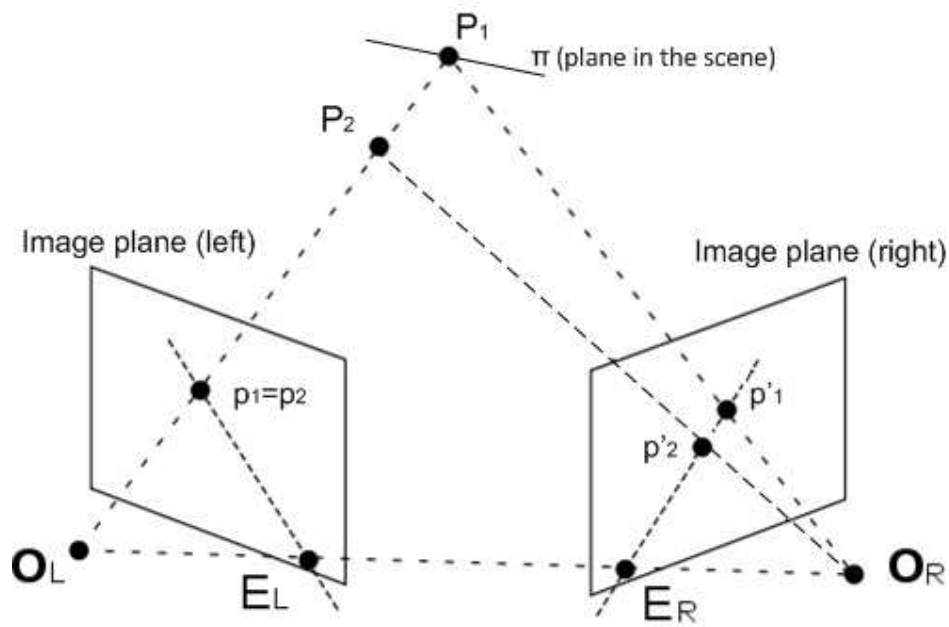


Figure 1.6 The epipolar geometry.

This approach can be used for long-range surveillance applications with the assumption that the captured scene is so far from the cameras (plane π is at infinity) and the depth differences of any two points in the scene is negligible compared to the distance of the imaged scene to the cameras (any two points P_1 and P_2 approximately lie on the infinite plane π). Under this assumption, for nearly collocated thermal and visible cameras, an infinite planar homography can be applied to the scene. The homography induced by π is $H_\infty = KRK^{'}$, where the homography between points is only the rotation (R) between the cameras and the projection matrices K and K'. Finally the transformation matrix is defined as,

$$p^{'} = H_\infty p. \tag{1.2}$$

The parallax term is negligible since the distance of the imaged scene to the camera tends to infinity.

Global Image Registration

This approach can be used for nearly long-range surveillance applications considering only foreground objects with the assumption that the depth differences of any two points on foreground objects in the scene is negligible compared to the distance of the imaged scene to the cameras (any two points P_1 and P_2 belonging to foreground objects lie approximately on

a one plane π in the scene). However, the plane π in the scene is not necessarily at infinity. Under this assumption the parallax term, δ, will be small for all objects in the scene and thus it is neglected. However in the scene where foreground objects are in different planes, only the objects lying in the plane π will be accurately registered and other objects will be misaligned.

Partial Image ROI Registration

This approach is for registering partial image ROIs (objects in the scene) with the assumption that objects are in multiple depth planes but a single object lie approximately in one plane π_i in the scene. Therefore the parallax effects are negligible between any two points belonging to one object, as each object is approximately lying in one single plane in the scene. This is the only registration approach that is applicable for close-range scenes. The accuracy of this approach is also limited to the accuracy of segmenting object region in the scene. Object segmentation is a challenging task, especially in uncontrolled scene where issues, such as illumination variations and occlusion, can cause imperfect segmentation results that contain two or more merged objects at different depths.

For global image registration approaches, either the whole left image or the foreground image of the left image are globally transformed on the right image using an approximated homography. For example, the homography may be approximated using a sparse two-image keypoint matching and computing a transformation matrix such as affine transformation matrix. But, for partial image ROI registration, there is no single global transformation for whole image since there are multiple objects at different depths in the scene. Therefore, registration is estimated by using a dense stereo correspondence algorithm.

1.1.4 Dense Stereo Correspondence Algorithms

Stereo vision refers to the impression of depth that is perceived from two or more disparate images of one scene captured from different viewpoints. In stereo vision, depth is inversely proportional to disparities (shifts) between pixels on two images.

Stereo matching is the process of taking two or more images of a scene and finding matching pixels or features between those images that later allows reconstructing the 3D geometry of the scene. Most of the recent stereo matching methods focus on dense correspondences (finding matches for every pixel in the whole image or image ROI). Before describing the dense stereo correspondence algorithms, we outline two basic definitions : 1) Image Rectification, an advantageous processing prior to matching, 2) Disparity map representation, the result of dense stereo two-frame matching.

Image Rectification : Given a pair of stereo images, rectification is a transformation of an image in such a way that pairs of conjugate epipolar lines on the left and right images

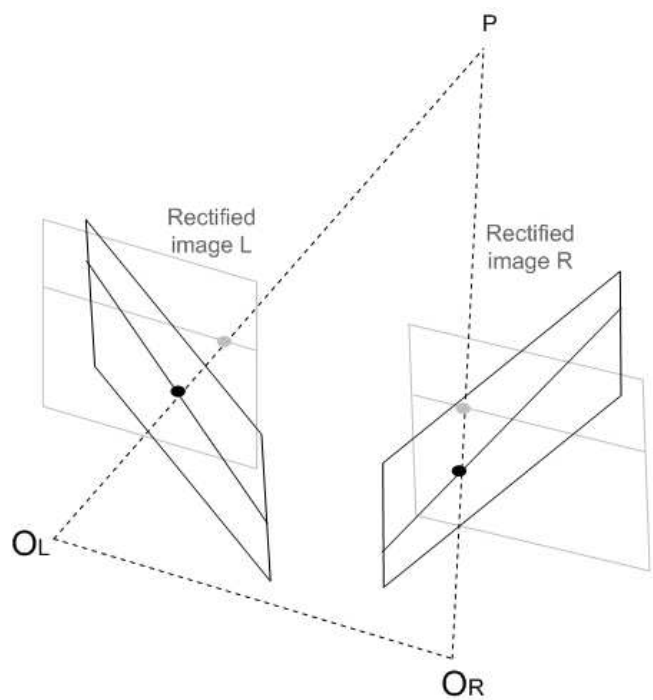


Figure 1.7 The stereo image rectification.

become collinear and parallel to the horizontal image axis. By knowing the intrinsic and extrinsic parameters of cameras computed by a stereo calibration method, such a transformation is feasible (Trucco and Verri (1998)). Figure 1.7 shows the image rectification for a pair of stereo images. The advantage of rectification is reducing 2-D search space in the image for correspondence algorithm to a 1-D scan-line search. In other words, to find the point $p = (x_l, y_l)$ on the left image, we just search along scan-line $y_r = y_l$ in the right image.

Disparity Map Representation : The term disparity describes the difference in location of corresponding points in the left and right images. Most stereo correspondence methods produce a univalued disparity map $d(x, y)$ as their result. The univalued disparity means that for each pixel, one disparity value either on horizontal or vertical direction is computed. In most works, disparity corresponds to horizontal disparity as synonymous with inverse depth (Scharstein and Szeliski (2002)). Given a reference image (left image) and matching image (right image) as the input of the correspondence algorithm, the correspondence between pixel $p = (x, y)$ in the reference image and pixel $p^{'} = (x^{'}, y^{'})$ in the matching image is computed as,

$$x^{'} = x + s \times d(x, y), y^{'} = y, \tag{1.3}$$

where $s = \pm 1$ is a sign chosen so that disparity is always positive.

Schrastein and Szeliski (Scharstein and Szeliski (2002)) give the taxonomy of stereo correspon-

dence algorithms. Considering the state-of-the-art bi-modal registration, we can categorize dense correspondence algorithms to two main categories of local and global stereo correspondence algorithms. In the following sections we describe these two categories.

Local Correspondence Algorithms

In local methods, the disparity map is produced based on a winner-take-all (WTA) matching method using local image regions usually bounded by windows on the reference images and performing scan-line search on the second image. This approach is named bloc matching. In this approach, the corresponding windows on reference and matching images are the ones with maximum similarity. In the literature, the classic similarity metric used in multimodal local correspondence algorithms is MI. The accuracy of local correspondence algorithms is usually limited to the matching window sizes and finding the best size is not trivial.

Global Correspondence Algorithms

Many global methods are defined in terms of energy function and goal is to find a disparity function d that globally minimizes energy over all the pixels of a complete image or image ROI. The energy equation is defined as,

$$E(d) = E_{data}(d) + E_{smooth}(d), \tag{1.4}$$

where $E_{data}(d)$ represents how well the disparity assignment, d, agrees with the input pair of images and $E_{smooth}(d)$ employs some assumptions usually between neighboring pixel disparities to make the minimization computationally tractable. This problem is naturally a discrete multi-labeling problem, where we would like to assign each pixel one of the L possible labels (disparities). The problem may be presented using a graphical model such as Markov Random Field (MRF) and labeling problem can be solved using an optimization method such as max-flow, graph-cut, and belief propagation.

In fact, global correspondence algorithms compute the disparities more accurately compared to the local methods, especially for partial image ROI that contains multiple merged objects at different depths. In global methods, information about the input images (e.g. edges) and restriction about disparity assignment may be formulated in the smoothness term.

In order to understand the background of our proposed global correspondence algorithm in this thesis, we briefly describe a general discrete multi-labeling problem and then its specifications for a global correspondence in a MRF framework.

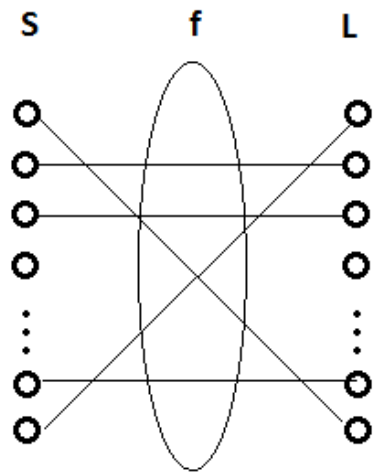


Figure 1.8 The discrete multi-labeling problem.

1.1.5 Discrete Multi-Labeling Problem

Several computer vision problems can be defined as a discrete multi-labeling problem. Labeling is also a natural representation for studying MRFs. Labeling is the problem of assigning a label from the label set L to each site in the set S. For example, for human detection, assigning label f_i from the set $L = \{human, non-human\}$ to site $i \in S$ where elements in S index the image pixels. Therefore, $f = \{f_1, f_2, ..., f_n\}$ is the labeling, which is a mapping from space S to L ($f : S \rightarrow L$). If L is a discrete set, like our example, and n is the number of sites, the solution space is $F = L^n$. Figure 1.8 illustrates a discrete multi-labeling problem.

1.1.6 Markov Random Fields In Stereo Correspondence

For the stereo matching problem, the MRF graphical model is an undirected graph, where each pixel is a vertex (site) and edges are represented by a neighborhood system, e.g. a four-connected neighborhood. The labeling problem assigns a label $f_p \in L$ (discrete set of disparities) to each pixel $p \in P$ (set of pixels/sites) in the image grid. For the MRF, the random field variables are $F = (F_p)_{p \in P}$. The probability that a random variable F_p takes the value f_p is $P(F_p = f_p)$ and the joint probability is denoted $P(F = f) = P(F_1 = f_1, ..., F_m = f_m)$. F is said to be a MRF on pixel set P with respect to a neighborhood system N, if and only if the Markov property is respected. The Markov property is defined as,

$$P(f_p | f_{P-p}) = P(f_p | f_{N_p}) \tag{1.5}$$

where $P-p$ is the pixel set excluding pixel p, f_{P-p} is a label assignment to pixel set excluding pixel p, and $f_{N_p} = \{f_{p'}|p' \in N_p\}$. N_p is the neighbor pixels of pixel p. Using a Bayesian labelling based on MRF, the best labelling can be approximated by estimating a Maximum of A Posteriori (MAP) (Li and Allinson (2008)). A simple posteriori probability can be defined as,

$$P(F|I) = \frac{P(I|F)P(F)}{P(I)}, \tag{1.6}$$

where I is a pair of stereo images and F is the labelling. As also mentioned in (Felzenszwalb and Huttenlocher (2006)), the MAP estimation for an appropriately defined MRF corresponds to finding a labelling with minimum energy. In (Boykov *et al.* (2001)), an energy function arising in the Bayesian labelling of first-order Markov Random Fields is defined as,

$$E(f) = \sum_{p_1,p2 \in N} V_{p,q}(f_p, f_q) + \sum_{p \in P} D_p(f_p), \tag{1.7}$$

where N is a set of neighboring pairs of pixels, the first term is E_{smooth} (the cost that two disparities f_p and f_q are jointly assigned to pixels p and q respectively) and the second term is E_{data} (the cost that a disparity f_p is assigned to pixel p) in equation 1.4.

1.2 Problematic Elements

For far-range surveillance, where the imaged scene is approximately planar, thermal and visible images may be aligned using a global transformation. For estimating such a transformation, a sparse keypoint matching is required. However, extracting low-level similar image features inside ROIs in thermal and visible images is difficult due the small size of objects. One interesting solution is using the spatio-temporal information of the scene, such as object trajectories and performing sequence-to-sequence matching rather than low-level image-to-image matching. In (Caspi *et al.* (2006)), a feature-based video sequence-to-sequence matching technique is proposed based on matching object trajectory points. However, trajectory-based matching involves another problem, which is computing trajectories of moving objects in the scene for a pair of video sequences. Since the matching features are trajectory points, the accuracy of the computed trajectories in both thermal and visible videos is improtant for image registration. Moreover, in unsupervised surveillance applications, the trajectories of people that newly entered in the scene might have an effect in transformation matrix estimation based trajectory-based matching. Therefore, these two problems are closely related to each other. In the literature, there is no research that addresses these two problems in an integrated framework. Such an integrated system is advantageous especially for online video surveillance applications.

For close-range scene, partial image ROI registration of a thermal and visible pair of videos is a challenging task. In order to accurately align moving objects in different depth planes in the scene, registration is estimated by using a dense stereo correspondence algorithm. Partial image ROI registration for multimodal video surveillance is a recent field of research in computer vision that is not well documented. Partial human body ROI registration for close range video surveillance has its own difficulties and requirements that motivate study of new similarity measures and stereo matching algorithms. Basically, for registration of pairs of images of a close-range scene, we need to deal with two main issues : 1) Selection of the similarity metric, and 2) Selection of the matching strategy. In the following, we describe each issue in details.

Selection Of Similarity Metric

The first issue is the selection of a viable similarity metric for matching thermal and visible human ROIs. Unlike visible sensors that capture reflected light, IR sensors capture thermal radiations reflected and emitted by an object in a scene. People might have colorful/textured clothes that are visible in color images but not in thermal images. On the other hand, there might be some textures observable in thermal images caused by different clothing characteristics (e.g. light clothes or warm clothes) and amount of emitted energy from different parts of the human body that are not visible in a color image. Due to the large differences between thermal and visible imaging characteristics, most similarity metrics used in single modal registration methods are not applicable. For image ROI matching, a similarity metric can be defined in two ways. First way is based on (either sparsely or densely) extracting local image features over the image ROI then defining a cost aggregation over two matching regions. Second way includes only one-step process by using inter-image similarity measures that directly compute the similarity between two image regions and skip the image feature extraction, such as MI.

In our context, color descriptors as matching image features are not applicable since the pixel intensities are totally different between thermal and visible images. However, some shape, pattern, and edge descriptors might be viable for thermal-visible human ROI matching. In recent years, Local Image Descriptors (LIDs) have gained popularity and dominance in computer vision tasks. The most popular LID category is the distribution-based descriptors. These descriptors use histograms or vectors to represent the appearance of edges or shape (Mikolajczyk and Schmid (2005)). They are computed either on a window centered on a keypoint, such as SIFT descriptor, or on an image patch such as LSS descriptor, and can be compared between images using simple $L1$ and $L2$ distances (cost function).

In the literature, MI is a classic multimodal similarity measure that has been widely used

in medical image registration (Pluim *et al.* (2003)). Egnal (Egnal (2000)) has shown that MI is a viable similarity metric for matching thermal and visible images. The robustness of MI as a similarity metric is restricted by the MI window sizes. For unsupervised human monitoring applications, obtaining appropriate MI window sizes for the registration of multimodal pairs of images containing multiple people with various sizes, poses, distances to cameras, and different levels of occlusion is quite challenging. Since the thermal-visible surveillance for human analysis is a recent subject in computer vision, there are no works in the literature that compare different similarity metrics for multimodal stereo matching.

Selection Of Matching Strategy

A multimodal ROI stereo correspondence algorithm should be robust to ROIs that contain multiple merged people at different depths (depth discontinuity), to textureless regions on the reference image with corresponding regions on the second image that may be either textured or textureless (large image characteristics differences), and to imperfect human ROI segmentation results. The image ROI segmentation methods such as background subtraction are not perfect and ROIs might be partially misdetected or some regions might be falsely detected.

Moreover, matching algorithms should be computationally efficient. For online video surveillance, computational time is an important factor. Usually, there is a trade-off between the matching accuracy and the computational time. Local stereo correspondence algorithms are usually faster than global stereo correspondence algorithms. For human image ROI correspondence, matching is focused only on the ROIs instead of the whole images, which reduces the processing time.

In the literature, most thermal-visible stereo correspondence algorithms are local stereo correspondence algorithms (described in 1.1.4). Since local approaches are based on the WTA bloc matching (matching two windows on the thermal and visible images with maximum similarity), they are not able to assign accurate disparities where there is depth discontinuity (people in different depth planes in the scene are merged in a single image ROI). Furthermore, in this approach, the selection of the size of the matching windows is manual. In the context of thermal-visible partial image ROI matching, there is one work that gives a comparative analysis of multimodal registration approaches (Krotosky and Trivedi (2007)). To the best of our knowledge, for video surveillance, no global correspondence algorithm has been proposed so far.

The problematic elements arising from the aforementioned issues can be summarized to the following questions that led the objectives of this research and the four articles that are included in this thesis.

Far-range surveillance :

- Is it possible to improve both tracking (a method of object trajectory computation) and registration by integrating a global image registration method and people tracking in a feedback framework for online multimodal video surveillance applications ?

Close-range surveillance :

- Is there an image descriptor or a similarity measure that is more robust for thermal-visible human ROIs matching than MI, which is the classic multimodal similarity measure ?
- So far, the existing state-of-the-art multimodal matching algorithms applied low-level image features and measures describing image texture. Since our input data is video, is there any high-level information of the scene, such as motion, that could be used for matching purposes in order to improve the existing state-of-the-art registration algorithms ?
- In our context, all existing state-of-the-art algorithms are local correspondence algorithms, which are not accurate for registration of occluded regions (depth discontinuity). Is it possible to improve the registration accuracy, especially for occluded regions, using an efficient global correspondence algorithm (considering the computational time limitation required for online applications) ?

1.3 Objectives Of Research

The main goal of this thesis is to propose solutions for human body ROI registration in a pair of thermal-visible videos for video surveillance applications.

For far-range surveillance applications, we aim to adapt a trajectory-based global image registration in an iterative framework with people tracking for an online video surveillance system. The main objective is to propose an automated system that requires no offline video processing. We aim to validate our system by extensive experiments using challenging indoor videos.

For close-range human monitoring applications, we are interested in accurate thermal-visible human ROI registration with assumption that people are in the different depth planes in the scene. Therefore, a stereo correspondence algorithm is required to compute a dense disparity map for the image ROIs in the scene. In our research, we address the problem of accurate disparity assignment for occluded people (depth discontinuity) in the scene.

For close-range surveillance applications, we summarized the detailed objectives of our research as the following items,

1. Comparing theoretically and quantitatively various viable LIDs and similarity metrics

for thermal-visible human ROI matching.

2. Integrating viable similarity measures and image descriptors with the state-of-the-art dense correspondence algorithms and evaluating their performance using realistic video surveillance scenarios.
3. Provide new solutions to improve the accuracy of the state-of-the-art dense local stereo correspondence algorithms to more accurately computing disparities for depth discontinuity regions caused by occluded people in the scene.
4. Providing new solutions for accurate disparity computation using a global correspondence approach for stereo registration of thermal-visible human ROIs considering time limitations.
5. Validating our methods by several experiments using challenging indoor videos and various scenarios and different number of moving people in the scene.

1.4 Contributions

In order to cope with the aforementioned difficulties, improve thermal-visible human ROI registration for far- and close-range surveillance applications, and satisfy the objectives of our research, we have documented our proposed solutions in four journal papers. The main contributions are summarized in the following items.

Far-range surveillance :

– For far-range surveillance, we proposed a novel integrated framework that iteratively improves both registration and tracking by feedbacks among system modules. Our proposed system has three main modules : 1) registration, 2) data fusion, and 3) tracking. Thermal-visible data fusion improves the input data for tracking in thermal and visible videos, which results in more accurate object trajectories compared to the trajectories computed using single modal videos. Using accurate trajectories as registration input data results in more accurate image registration. Moreover, the iterative estimation of global transformation based on "up to current frame" trajectory data prevents misalignment of newly entered people in the Field Of View (FOV) of the cameras. By considering the practical cases that not at every frames all the people are completely in one single plane in the scene (in one time step t), it is desirable to re-estimate the transformation matrix based on "up to current frame" trajectories.

Close-range surveillance :

– Performance evaluation has gained more and more importance in computer vision. This subject is well documented for similarity measures and image descriptors applied

in color image matching problems. However, for multimodal partial image ROI matching, to the best of our knowledge, there is no work in the literature that evaluates and compares local descriptors and similarity measures performance. Therefore, one of the contributions of this thesis is a performance evaluation of local descriptors and similarity measures for thermal-visible human ROI matching and then determining the characteristics that makes a descriptor or measure viable in this context.

- MI is a classic multimodal similarity metric that is widely used in local stereo correspondence algorithms in the literature. To the best of our knowledge, in the context of thermal-visible human ROI registration, there is no dense stereo correspondence algorithm in literature that uses other similarity measure than MI. One of the contributions of this research is integrating other viable similarity measures in state-of-the-art dense local stereo correspondence algorithms and comparing their performances with MI performance. We have integrated with success LSS and HOG (two LIDs) in a state-of-the-art dense correspondence algorithm named Disparity Voting (DV) (Krotosky and Trivedi (2007)).Our comparison of LSS-based, HOG-based, and MI-based dense correspondence algorithms shows that the LSS-based registration outperforms the similar MI-based registration in realistic close-range surveillance scenarios.
- Depth discontinuity, caused by merged people in a single image ROI (occluded people), is one of the difficulties that a registration algorithm should deal with it. In this research, a LSS-based local stereo correspondence algorithm is proposed that improves the state-of-the-art DV (Krotosky and Trivedi (2007)) algorithm to handle the depth discontinuity using the fact that the matching targets are moving people in the scene and motion segments in the image is a good estimate of the maximum number of existing depth segments in the scene. In fact the idea of using motion cue to improve the depth discontinuity region disparity assignment is one of our main contributions in this thesis that enables to automatically determine the suitable size of the matching window sizes for the different regions based on the size of the motion segments.
- Global correspondence algorithms are computationally more expensive than local methods. However, they more accurately estimate the disparity map compared to local correspondence methods, especially in depth discontinuity regions. In this research, a global correspondence algorithm for stereo registration of thermal-visible human ROIs is proposed. Our proposed algorithm uses a novel energy-minimization framework integrating LSS as similarity metric and motion and color cues as soft constraints in order to improve the accuracy of disparity assignment of the occluded people.

1.5 Thesis Structure

In chapter 2, a critical review of literatures is presented along with a summary of selected previous works on multimodal image registration. In chapter 3, the overview of proposed methods in this thesis is presented. Chapter 4 presents a trajectory-based global registration algorithm performing simultaneously with multiple people tracking for a nearly far-range surveillance applications in an article entitled *An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications* which is published in the journal of *Computer Vision And Image Understanding*. Chapter 5 presents the performance evaluation of new and famous LIDs and classic similarity measures for thermal-visible partial image ROI registration in an article entitled *A performance evaluation of local descriptors and similarity measures for thermal-visible human ROI registration* which is submitted in the journal of *Pattern Recognition Letters*, special issue on *Extracting Semantics From Multi-Spectrum Video*. In the chapter 6, we introduce LSS as a dense multimodal similarity metric for human ROI registration and propose a LSS-based registration using a local stereo correspondence algorithm in an article entitled *Local self-similarity based registration of human ROIs in pairs of stereo thermal-visible videos* that is submitted in the journal of *Pattern Recognition*. A global optimization based registration method using belief propagation is proposed in an article entitled *A LSS-based registration of stereo thermal and visible videos using belief propagation for human monitoring* that is submitted in the journal of *Computer Vision And Image Understanding*, special issue on *Advances In Machine Vision Beyond Visible Spectrum*, and presented in chapter 7. Chapter 8 presents a general discussion regarding to the different aspects of our research and the improvement of our methods compared to the state-of-the-art algorithms. Finally, chapter 9 concludes the thesis by summarizing our contributions and the future directions of this research.

CHAPTER 2

LITERATURE REVIEW

Thermal-visible video surveillance system for civilian applications is a new field of research that has not been yet well documented. The fundamental and preliminary task associated with the joint use of thermal-visible data is accurately matching features and aligning a pair of images captured by two different sensors. This problem is called multimodal image registration. In the literature, based on the range of the imaged scene that can be either long-range (cameras are far from the imaged scene) or close-range (cameras are close to the imaged scene), the registration methods are categorized into the three main approaches of infinite homography, global and partial image ROI registration.

The infinite homography registration is the most straight forward approach among the three methods. This approach is usually used as a simple pre-processing for higher-level analysis on multimodal data such as tracking. However, the other two approaches are more complex with their related literature focusing on the registration method. The detailed literature review on different aspects of global and partial image ROI approaches is presented in our four articles included in this thesis. In sections 4.1 and 4.2, we present the literature review of global image registration approach. In section 5.2, we present the literature review of viable similarity metric for partial image ROI registration. In section 6.1, we present the literature review of local stereo correspondence methods. Finally, in sections 7.1 and 7.2, we present related literature review of global stereo correspondence methods.

In this section, we will summarize the important state-of-the-art related to different registration approaches and add some missing details in our papers regarding those approaches.

2.1 Infinite Homography Registration

Infinite homography registration is the least difficult registration approach. It is applicable for long-range video surveillance where the imaged scene is very far. In such a case, the assumption that the whole scene is lying in plane at infinity is valid. In literature, using the infinite homography, several methods including data fusion algorithms (Han and Bhanu (2007)), background subtraction (Davis and Sharma (2007); O Conaire *et al.* (2005)), and multi-pedestrian tracking and classification (Leykin (2007)) for thermal-visible surveillance system have been proposed. In these works, it is assumed that visible and thermal cameras are nearly collocated and that the imaged scene is far, so that the deviation of people position

from the ground plane is negligible compared to the distance between the image scene and the cameras. Registration using infinite homography does not provide depth information of the scene.

Figure 2.1 shows corresponding visible and transformed thermal images from OTCBVS dataset (Davis and Sharma (2007)). The thermal image is transformed in the sense that the coordinates in visible and thermal images are transformed in a one coordinate system. Therefore, methods in literature using this dataset, like (Leykin (2007)), simply skip the registration required for data fusion. In this dataset (Davis and Sharma (2007)), videos are captured using cameras mounted adjacent to each other at location approximately 3 stories above ground. Visible and thermal images are aligned using infinite homography by matching manually-selected points.

2.2 Global Image Registration

Global image registration assumes that all registered objects will lie on a single depth plane in the scene. However, this approach does not perform well to accurately register objects at different depths (where single depth plane assumption is not valid) as the transformation for each object depends on varying perspective effects of the two cameras. In global approach only the foreground objects are considered.

In the literature, most works address the global image registration problem as a low-level image-to-image feature-based matching problem. In this approach, image features are first extracted and then a matching is done between the dense or sparse extracted features of a pair of images. Finally, based on corresponding features, a homography is estimated. For example, Coiras *et al.* proposed an affine transformation matrix that is estimated using the matching of triangles formed from edge features in thermal and visible images (Coiras *et al.* (2000)). In Han *et al.*, a hierarchical genetic algorithm based method is applied for matching the human

Figure 2.1 Transformed thermal image (left) pixels to visible image (right) pixel coordinate system (Davis and Sharma (2007)).

silhouettes in thermal and visible images using two pairs of corresponding points from two frames (Top-of-head and centroid) of a human walking on a straight line at a fixed distance from the camera (Han and Bhanu (2003)). In Bilodeau *et al.*, a set of viable keypoints on the boundary and on the skeleton of a region of interest (ROI) are proposed that may be applied for global registration (Bilodeau *et al.* (2011a)).

In these methods, the quality of image alignment is limited by the quality of low-level image feature extraction. In cases where cameras have in significantly different zoom or the imaged scene is far, due to the small size of the people in the images, extracting common features inside corresponding human regions in thermal and visible images is more difficult. Therefore, low-level feature extraction is sometimes problematic for matching purposes.

Some works have addressed global video registration as a higher-level problem using matching spatio-temporal features, such as object trajectories extracted from thermal and visible videos. In this approach, the transformation matrix is estimated based on trajectory point matching. In Caspi *et al.*, using two synchronized thermal and visible videos, a feature-based video sequence-to-sequence matching technique is proposed based on matching object trajectory points (Caspi *et al.* (2006)) where the matching criterion is the Euclidean distance between points. However in their method, the problem of object trajectory computation is not discussed. In ((Morin *et al.*, 2008; Bilodeau *et al.*, 2011b)), a similar trajectory-based registration is proposed. In this method, the object trajectories were computed separately for thermal and color video sequences using multiple object tracking in an offline process. Trajectory matching was improved over Caspi *et al.* by using a foreground pixel overlapping score as well as the number of matching trajectory points as registration criteria. However, since the trajectories were estimated separately using unimodal data, some trajectories were inaccurate and disconnected due to the imperfect object segmentation. For this reason, in this thesis, we proposed to tackle the problem of trajectory-based image registration and object tracking in a novel integrated, feedback framework with the final goal of improving both registration and tracking. Based on our analysis of previous works, to improve trajectory

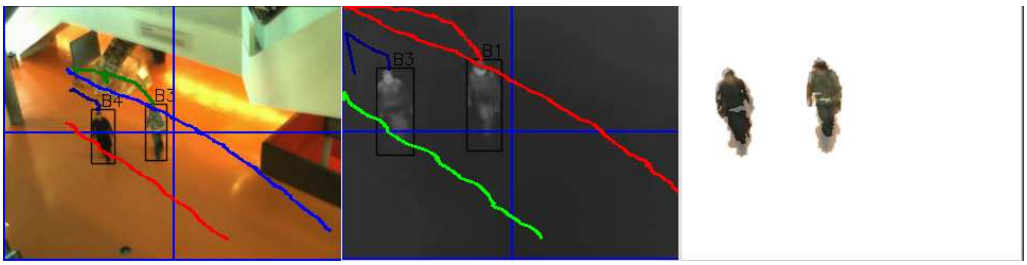


Figure 2.2 Tracking in visible video (left), tracking in thermal video (middle), and thermal-visible registration (right).

calculations, we propose to iteratively compute object trajectories using multimodal data as the input of tracking module then re-estimate a new registration (affine transformation matrix) using improved trajectories in each frame. Moreover, we aim to apply this approach for online applications Figure 2.2 illustrates the trajectory-based image registration approach (Torabi *et al.* (2010)).

2.3 Partial Image ROI Registration

Image registration for close-range videos of multiple people at different distances from the cameras is the most difficult problem. In the literature, partial image ROI registration is the only approach that considers that people in the scene may lie in different depth planes. For such scenes, there is no single global transformation, which accurately aligns all the people in the scene. Partial image ROI registration is the only viable approach for close-range multimodal surveillance. In the context of video surveillance, the main advantage of this approach is that it not only aligns the thermal and visible ROIs but also provides the depth information of people in the terms of disparities, which can be used as a feature for higher-level data analysis. The partial image ROI registration is based on a dense stereo correspondence that estimates a dense disparity map for each image ROI in the scene separately. The main problem associated with a partial image ROI registration is objects at different depths which are merged in a single image region. This problem is named depth discontinuity in stereo problems. For multimodal video surveillance, all existing partial image ROI registration methods in literature are formulated in a local dense stereo correspondence framework. However, there is none based on a global stereo correspondence. In fact, the global stereo correspondence is a well-studied subject for unimodal stereo problem and it has more accurate results, especially for the depth discontinuity regions, compared to the local correspondence approach. However, adopting global approach to multimodal stereo problem is not trivial due to the high differences in thermal and visible imaging characteristics. In fact, most global stereo approaches for unimodal images use pixel intensity as pixel-based image feature. The pixel intensity is not a viable feature for multimodal matching. In the following, we present a short overview of existing dense correspondence methods.

2.3.1 Local Dense Stereo Correspondence

Local dense stereo correspondence methods are based on WTA bloc or window matching on a pair of rectified thermal and visible images. In this approach, computing disparity is simply choosing the disparity with the minimum cost value over the matching windows. Fookes *et al.* proposed a MI-based window matching method that incorporates prior probabilities of

the joint probability histogram of all the intensities in the stereo pair in their MI formulation (Fookes *et al.* (2004)). Therefore, they detect textureless region and can adjust the size of MI window to improve matching. However, their experiment is only carried out on negative and solarized images that have similar patterns in their ROI. Egnal has shown that mutual information (MI) is a viable similarity metric for matching disparate thermal and visible images (Egnal (2000)). His work gives a comparison between MI and NCC, and theoretically describes the advantages of MI for thermal-visible image registration (Egnal (2000)). Chen *et al.* proposed a MI-based registration method for pairs of thermal and visible images that simply matches bounding boxes surrounding image ROIs in the two images with the assumption that each box represents one single human (Chen *et al.* (2003)). In their method, occluded people that are merged into one image ROI are not accurately registered since the image ROI contains people in the different depth planes in the scene. As a solution for improving registration of occluded people in a scene, Krotosky and Trivedi proposed a MI-based DV matching approach (Krotosky and Trivedi (2007)). DV is performed by horizontally (column by column) sliding small width windows on rectified thermal and visible foreground images, computing MI for pairs of windows, and finally counting the number of votes associated to each disparity and assigning one disparity to each column based on a Winner Take All (WTA) approach. Their method can handle horizontal occlusion, but it cannot accurately register people with different height where a shorter person is in front of a taller one (vertical occlusion) because in their method, all pixels of a column inside a ROI are assigned to only one disparity. Figure 2.3 (a) and (b) represents MI-based window matching on foreground visible and thermal images. Figure 2.3 (c) shows the corresponding disparity estimated by MI-based disparity voting method (Krotosky and Trivedi (2007)), and Figure 2.3 (d) shows registration results. It is shown that in depth discontinuity (occlusion) regions, window matching failed to accurately compute the disparities.

In uncontrolled settings, when people have clothes with different patterns, 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 (Krotosky and Trivedi (2007)). MI-based matching fails to correctly match image boxes where the joint probability histogram is not sufficiently populated. This shortcoming is the principal motivation for investigating other image features for thermal and visible image matching.

The most important limitation of local dense correspondence approach is determining appropriate matching window sizes. Choosing the appropriate window size is not straightforward due to the varying human ROI scales, poses, and imperfect object segmentation. Also, there is always a trade-off between choosing larger matching windows for matching evidence, and smaller matching windows for the precision and details required for an accurate

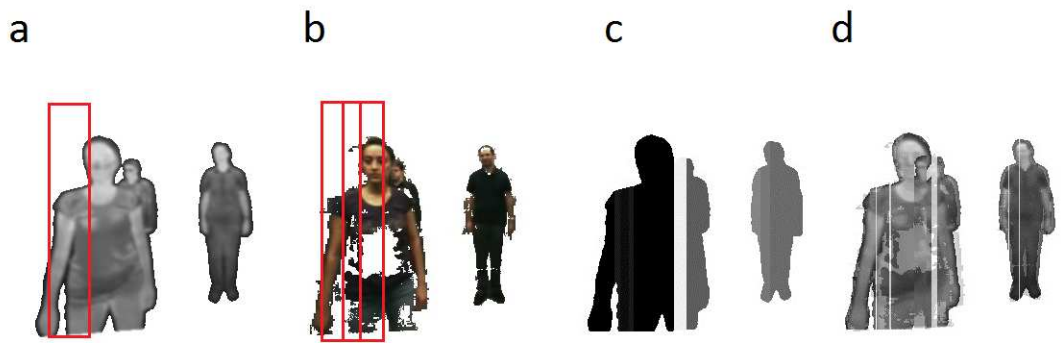


Figure 2.3 (a) and (b) MI-based window matching on pair of foreground thermal and visible images (c) dense disparity map for foreground regions of thermal image, and (D) registration results of thermal on visible image (Krotosky and Trivedi (2007)).

registration.

2.3.2 Global Dense Stereo Correspondence

Many global dense correspondence methods are formulated in an energy minimization framework. This approach produce more accurate disparities compared to local methods especially in depth discontinuity regions. Global dense stereo correspondence incorporates explicit smoothness assumptions and determines all disparities simultaneously by applying one of the energy minimization techniques such as dynamic programming, simulated annealing, belief propagation, and graph cuts. In an energy minimization framework, the similarity measure for matching is integrated in the data-term and some prior visual cues of images that can be used as information to handle depth discontinuity are integrated in the smoothness-terms. The most common visual cues are color segmentation and edge features. As an example, in (Sun *et al.* (2003)), color segments are used as cues that encourage the two neighboring pixels belonging to one segment is more likely to be assigned to one disparity than two neighboring pixels belonging to different color segments. Therefore, the cost of assigning disparity to two neighboring pixels in the same color segment is higher than two neighboring pixels belonging to different color segments. The same rule can be applied for the edges as the visual cue. Neighboring pixels which are not on the image edges are more likely to be assigned to one disparity level than the ones on the edges.

Over the past few years, there have been great advances in the development of algorithms for solving stereo problem using MRF models. While the MRF-based registration framework yields an optimization problem that is NP hard, good approximation techniques based on Graph Cuts (GC) (Boykov *et al.* (2001)) and on Belief Propagation (BP) (Weiss and Freeman (2001); Sun *et al.* (2003)) have been developed and demonstrated for stereo registration

problem. In the both methods, the computed local minima are the minima over large neighborhoods, which is good in the sense that it generates highly accurate results. In (Tappen and Freeman (2003)), authors present a comparison between the two different approaches for stereo matching. Several global stereo matching algorithms using GC (Deng *et al.* (2005); Bleyer and Gelautz (2007); Hong and Chen (2004)) and BP (Felzenszwalb and Huttenlocher (2006); Sun *et al.* (2005); Yang *et al.* (2009b)) have been developed for unimodal image registration. In practice, the quality of GC and BP are comparable. However, BP is more suitable for parallel execution for reducing the processing time (Yang *et al.* (2009b)).

In the recent years, BP became more popular compared to GC for stereo problems. The BP algorithm performs by passing messages around the graph defined by a four-connected image grid. BP algorithm is based on either max-product or sum-product rules. Originally, BP was computationally intensive for real-time applications. The BP computational time of original method is $O(TNL^2)$ (Sun *et al.* (2003)), where N is the image size, L is the number of disparity levels, and T is the number of the optimization iterations. Recently, in (Felzenszwalb and Huttenlocher (2006)), authors proposed an efficient sum-product belief propagation with a complexity reduced to $O(TNL)$ (linear time) using min convolution method and hierarchical estimate of messages. This method makes BP viable even for online applications such as video surveillance.

Figure 2.4 shows for a pair of visible images from Tuska data of Middlebury benchmark (Scharstein and Szeliski (2002)). A dense disparity map is computed for the whole image by using an efficient BP method (Felzenszwalb and Huttenlocher (2006)).

In the literature, there is no global dense correspondence method for multimodal surveillance applications. The main reason is that for a pair of color images, the similarity metric used for the data-term in the energy function is simply the pixel intensity differences. However, for pair of thermal-visible images, extracting viable common pixel-based features for the data-term to be used in an energy function is problematic. The last article that is included in this thesis discusses an efficient MRF-based registration method for close-range

Figure 2.4 Pair of visible images from two view-points (left and middle) and dense disparity map computed using efficient BP (right).

thermal-visible video surveillance and addresses this problem.

CHAPTER 3

OVERVIEW OF APPROACHES

Our proposed methods in this thesis address video registration for two applications domains : 1) Multimodal far-range surveillance 2) Multimodal close-range surveillance. Among the four articles included in the thesis, the first one discusses a registration method for far-range surveillance. The three others are about registration approaches for close-range surveillance as a major contribution of this thesis.

1. Our first article addresses the problem of image registration and object tracking in a novel integrated framework. We propose an iterative thermal-visible video registration, sensor fusion, and multimodal tracking for two synchronized streams of nearly long-range videos that are recorded by collocated visible and thermal cameras at different zoom settings. For our proposed methods, no camera calibration is needed. In this paper, we mainly focus on the system architecture, the feedback scheme, and the collaboration between the three modules of our system (image registration, sensor fusion, and tracking), but we also suggest a fusion score computed in our sensor fusion module as an improved registration criterion. This article covers our objective for long-range surveillance applications, in section 1.3.
2. In the second article, the viability of various LIDs and similarity measures for thermal-visible image registration of close-range scene is studied. Our evaluation uses a simple WTA block matching and assesses the viability of SURF, HOG, LSS, BRIEF, NCC, and MI by the precision-recall and the power of discrimination criteria. In this article, the performances of the three best metrics (LSS, MI, and HOG) are compared using a registration method based on local correspondence approach, named Disparity Voting (DV) (Krotosky and Trivedi (2007)). The comparison is carried out using realistic scenarios of human monitoring applications. This article covers our objectives 1, 2, and 5 for close-range surveillance applications, in section 1.3.
3. In third article, LSS is introduced as a dense multimodal similarity metric for images of a close-range scene. Its theoretical and quantitative adequacy and strengths are compared to MI in the context of visual surveillance systems using several examples. In the theoretical comparison, the properties of LSS (a local image descriptor) and MI (a similarity measure) for multimodal registration are studied. In the quantitative experiment, an evaluation by using a simple WTA window matching and comparing the

results with groundtruth data is carried out. Moreover in this part, a LSS-based registration of thermal-visible stereo videos based on a DV local correspondence algorithm is proposed. This registration consists of two steps : 1) motion segmentation, and 2) disparity assignment. It is shown that our proposed LSS-based registration method improves the accuracy of registration results compared to the state-of-the-art MI-based DV registration method (Krotosky and Trivedi (2007)). This third paper covers our objectives 2, 3, and 5 for close-range surveillance applications, in section 1.3.

4. In the last article, a global optimization-based registration for a thermal-visible human ROI registration is presented. In this method, the stereo matching is formulated in a novel energy-minimization framework integrating LSS as similarity metric. In this method, the disparity map is estimated using an efficient belief propagation (Felzenszwalb and Huttenlocher (2006)). This method handles depth discontinuities and homogenous regions by integrating Motion as principal visual cue and color as supplementary cue in smoothness term of an energy function. Extensive experiments are carried out for realistic surveillance scenarios and it is shown that our method outperforms the state-of-the-art local correspondence method (Krotosky and Trivedi (2007)). Our fourth article covers our objectives 4 and 5 for close-range surveillance applications, in section 1.3.

CHAPTER 4

AN ITERATIVE INTEGRATED FRAMEWORK FOR THERMAL-VISIBLE IMAGE REGISTRATION, SENSOR FUSION, AND PEOPLE TRACKING FOR VIDEO SURVEILLANCE APPLICATIONS

Abstract

In this work, we propose a new integrated framework that addresses the problems of thermal-visible video registration, sensor fusion, and people tracking for far-range videos. The video registration is based on a RANSAC trajectory-to-trajectory matching, which estimates an affine transformation matrix that maximizes the overlapping of thermal and visible foreground pixels. Sensor fusion uses the aligned images to compute sum-rule silhouettes (described in section 4.5), and then constructs thermal-visible object models. Finally, multiple object tracking uses blobs constructed in sensor fusion to output the trajectories. Results demonstrate the advantage of our proposed framework in obtaining better results for both image registration and tracking than separate image registration and tracking methods.

4.1 Introduction

In the recent years, there has been a growing interest in visual surveillance using multimodal sensors, such as thermal and visible cameras in both civilian and military applications. Zhu and Huang give a comprehensive introduction about multimodal surveillance systems in (Zhu and Huang (2007)). The advantages of jointly using a thermal camera and a visible camera have been studied and discussed extensively in some few works such as (Zhu and Huang (2007); Socolinsky (2007)). Two main benefits of the joint use of thermal and visible sensors are first the complementary nature of different modalities that provides the thermal and color information of the scene and second, the redundancy of information captured by the different modalities, which increases the reliability and robustness of a surveillance system. These advantages motivated the computer vision community to study and investigate algorithms for thermal-visible video surveillance systems.

For approximately planar far-range videos at different zoom settings, where extracting low level features inside ROIs are difficult due the small size of objects, using the spatio-temporal information of the scene, such as object trajectories and performing sequence-to-sequence

matching rather than low level image-to-image matching is an interesting solution. In Caspi *et al.*, a feature-based video sequence-to-sequence matching technique is proposed based on matching object trajectory points (Caspi *et al.* (2006)). However, trajectory-based matching involves another problem, which is computing trajectories of moving objects in the scene for a pair of video sequences. Since the features to match are trajectory points, the accuracy of computed trajectories in both thermal and visible video has a crucial effect on the image registration result.

In our previous work (Morin *et al.* (2008); Bilodeau *et al.* (2011b)), we proposed trajectory-based sequence-to-sequence video registration, where the object trajectories were computed separately offline for thermal and color video sequences using multiple object tracking, but with an improved trajectory matching that uses foreground pixel overlapping as well as trajectory point matching as registration criteria. In (Morin *et al.* (2008); Bilodeau *et al.* (2011b)), the image registration is similar to the one we used in this paper ; however, since the trajectories were estimated separately from tracking using data of a single modality, some trajectories (registration input data) were inaccurate and disconnected. Furthermore, the foreground pixel overlapping criterion could be misleading for some video frames due to the background subtraction errors. In this paper, we address the problem of image registration and object tracking in a novel integrated framework with the final goal of improving both registration and tracking. We propose an iterative, integrated, thermal-visible video registration, sensor fusion, and multimodal tracking for two synchronized streams of long-range videos recorded by collocated visible and thermal cameras at different zoom settings. For our proposed methods, no camera calibration is needed. The only assumption is the intersection of field of view between thermal and visible cameras. In this paper, we mainly focus on a feedback scheme and collaboration between the three modules of our system (image registration, sensor fusion, and tracking), but we also suggest a fusion score computed in the sensor fusion module of our system as an improved registration criterion.

Contribution. Our proposed integrated framework improves both registration and tracking by providing better quality for their input data. Thermal-visible sensor fusion improves the input data for tracking in thermal and visible videos, which results in more accurate object trajectories. Using accurate trajectories as registration input data results in more accurate image registration. In our experiments, we show that our proposed framework outperforms similar image registration methods previously proposed in the-state-of-the-art (Caspi *et al.* (2006); Bilodeau *et al.* (2011b)). Also, we propose a new transformation matrix selection method based on the fusion scores computed in our sensor fusion step. The algorithms presented in this manuscript are based on (Torabi *et al.* (2010)), but they are further developed with detailed analysis and new evaluations.

In the remainder of this paper, we present some background (section 4.2), then the architecture of the whole system (section 4.3), followed by a description of our image registration, sensor fusion, and tracking (sections 4.4, 4.5, and 4.6). Then, we discuss the performance of our proposed method (section 4.7). Finally, we conclude our paper (section 4.8).

4.2 Related works

Despite the advantages of multimodal surveillance systems, jointly using two sensors of different modalities increases the complexity of a surveillance system and raises new problems such as image registration and multimodal data fusion. Several works are related to algorithms for thermal-visible data fusion. Conaire *et al.* compared the various fusion methods by evaluating the tracking performance of systems using different fusion methods for aligned pairs of images (Conaire *et al.* (2006)). Their image alignment is done by estimating the optimum planar homography using a manual process and then warping the thermal images. Also Sadjadi gave a comparative analysis of various fusion methods by proposing a set of measures to study directly their performance (Sadjadi (2005)). Furthermore, Conaire *et al.* proposed a framework that performs data fusion and tracking in one integrated system (Conaire *et al.* (2008)). In their framework, data fusion is based on fusing the output of multiple spatiogram trackers. In another work, Kumar *et al.* proposed a multimodal object detection based on fusion of blobs in thermal and visible foreground images (Kumar *et al.* (2010)). Their method addresses the problem of uncertainty in object detection for dynamic environment such as outdoor scenes. Their fusion method is based on a feedback scheme that performs a simple blob matching between fuse blobs in the previous frame and blobs detected individually in the current thermal and visible frames, followed by a belief fusion that determines the validity of foreground regions detected for each modality and a Kalman filter fusion method. However, in their method, they did not address the problem of object tracking (tracking is based on a simple blob matching) and image registration.

Moreover, a number of works have been published on computer vision methods appropriate for thermal-visible video surveillance applications including background subtraction, object detection (Davis and Sharma (2005, 2007)), multi-pedestrian tracking, and classification (Leykin and Hammoud (2006); Leykin (2007); Conaire *et al.* (2008); Hammoud (2009)). In the works mentioned above, especially the ones designed for approximately planar far-range scenes (Kumar *et al.* (2010); Conaire *et al.* (2008)), the problem of automatic video registration is not studied. However, in thermal-visible video surveillance applications, where the thermal and visible videos are captured by two synchronized cameras with different lenses or zooms and with different FOVs, the primary problem before data fusion or any further

analyses is automatic image registration. Due to the numerous differences in imaging characteristics of thermal and visible cameras, finding appropriate correspondence measure for matching multimodal images is challenging. Most methods used for registering images of single imaging modality are not applicable. It is also very difficult to find correspondence for an entire scene.

In the literature, some works have been proposed on multimodal image registration for various computer vision applications. Krotosky and Trivedi give a comparative analysis of multimodal image registration methods (Krotosky and Trivedi (2007)). Most of these works address the image registration problem as a low-level image-to-image feature-based matching problem. In this approach, image features are first extracted and then a matching is done between the dense or sparse extracted features of a pair of images. For example, Irani *et al.* proposed an image registration method by which local correlation values of the features extracted from a Gaussian pyramid of visible and thermal images are computed, and a global alignment using an iterative Newtonian method is performed (Irani and Anandan (1998)). In Coiras *et al.*, image registration is estimated from an affine transformation that maximizes the global edge-formed triangle matching (Coiras *et al.* (2000)). In Han *et al.*, a hierarchical genetic algorithm-based method is applied for matching the human silhouette in thermal and visible images using two pairs of corresponding points of a human walking on a straight line at a fixed distance from the camera (Han and Bhanu (2003)). In these methods, the quality of image alignment is limited to the quality of low-level image feature extraction. Especially for far-range scene people monitoring, extracting features inside blobs is more difficult because blobs are small. Therefore, low-level feature extraction is quite problematic. The other image-to-image matching approach for thermal-visible image registration is the dense stereo correspondence method which is basically a scanline- search box matching followed by a dense disparity map estimation based on the winner takes all (WTA) approach. For example, in Krotosky and Trivedi work, a mutual information (MI) based image registration method is proposed for calibrated pairs of thermal and visible images in a close range scene (Krotosky and Trivedi (2007)). The robustness of this method is limited by MI window sizes that are needed to be large enough to sufficiently populate the joint probability histogram of MI computation. For far-range people monitoring applications, this assumption is usually not satisfied due to the small size of blobs and lack of details of patterns inside blobs. Moreover, a simpler camera setup that does not need further pre-processing such as multimodal calibration is desirable.

4.3 Overview of methods

The input data of the system are synchronized video streams captured by a thermal and a visible camera that are collocated with intersecting fields of view (FOVs) at different zoom settings. We assume that the scene is planar, which means that difference of the distances of moving objects in the scene are much smaller than the distance of the scene from the camera (cameras are installed two levels upper than the imaged scene). Fig. 4.1 shows the camera setup. Cameras can rotate around the z-axis and move along the x-axis and y-axis relative to each other. The only requirement is the intersection of fields of view of the two cameras.

The input data of our system at each frame are pair of thermal and visible foreground images. We apply the background subtraction background method proposed by (Shoushtarian and Bez (2005)) to separate the foreground pixels from the background. Any reasonable background subtraction method with a fair number of false negative and false positive foreground pixels may be used. Fig. 4.2 shows the flowchart of our algorithm, which consists of two stages : 1) initialization ; and 2) the main loop for image registration, sensor fusion, and tracking. Initialization is performed at the beginning of the videos, where, for some frames, tracking is performed separately for the thermal and the visible video frames until we obtain enough object trajectory points in the scene to estimate a good transformation matrix. The second part of the algorithm consists of a loop on pairs of thermal and visible video frames, where image registration, sensor fusion, and thermal-visible tracking are performed respectively. The image registration estimates an affine transformation matrix, which is used to transform one image into the coordinates of the second one. The sensor fusion matches the color and thermal pixels of blobs using this transformation matrix, and combines thermal and color information. At this step, the matching quality of the computed blobs is also evaluated to decide whether a new transformation matrix should be estimated or if it should be skipped

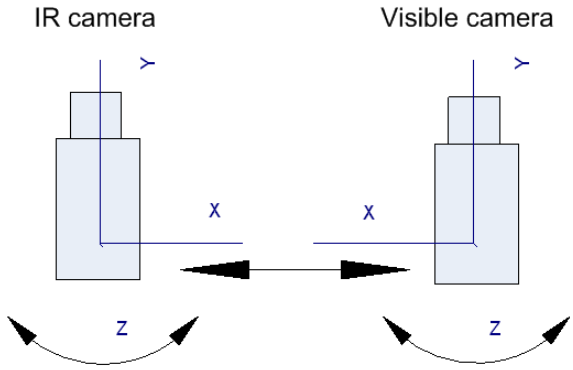


Figure 4.1 Camera setup

at the next frame. Finally, tracking is performed for thermal and visible videos using fused blobs obtained from the sensor fusion. These new trajectory points will be used for image registration computation at the next frame.

4.4 Thermal-visible image registration

At the beginning of the videos, a few trajectory points that are not collinear are required to compute a reasonable initial estimate of the transformation matrix that will be used for sensor fusion. For a fixed number of frames, tracking is performed separately in thermal and visible videos. Then, videos are registered and the overlapping error (Eq. 4.3) is computed. The registration is repeated until reaching a frame for which the overlapping error is less than a fixed threshold, to ensure the acceptable quality of image alignment required for sensor fusion. The number of initialization frames is subject to change from one video sequence to another, based on the frame rate of the video, the trajectory pattern of the moving objects in the scene, and the number of people walking in the FOV of the cameras at the beginning of the video.

Image registration is performed by aligning the thermal and color images using an affine transformation matrix H (Hartley and Zisserman (2003b)) computed by matching object trajectory pairs and point pairs from thermal and visible videos. Points are matched using a RANSAC-based algorithm. Our RANSAC-based method is based on matching randomly selected points on the object trajectories of synchronized thermal and visible videos, and finding the best matching points. The affine transformation matrix H is estimated using the normalized Direct Linear Transform (DLT) method (Hartley and Zisserman (2003b)) to find the least squares solution.

A pair of trajectories is composed of a trajectory from the thermal video and another from the visible video. For example, at frame t, if there are three trajectories for thermal video (T^1_{left}, T^2_{left} and T^3_{left}) and if there are two trajectories for visible video (T^1_{right} and T^2_{right}), then we have six pairs of trajectories that are used as the data pool for the RANSAC algorithm. We used the top-most point position of the human silhouette during tracking to construct a trajectory, since it is less sensitive to shadows on the floor that are falsely detected as part of the human silhouette. Fig. 4.4 shows matching trajectory points of a pair of trajectories.

Since the videos are synchronized, a pair of corresponding trajectory points in a trajectory pair is a pair of points with the same time stamp. Matching a possible pair of points with the same time stamp, instead of all the points, reduces the combinatorial complexity of the matching problem considerably.

Our RANSAC algorithm is a non deterministic iterative algorithm that estimates the

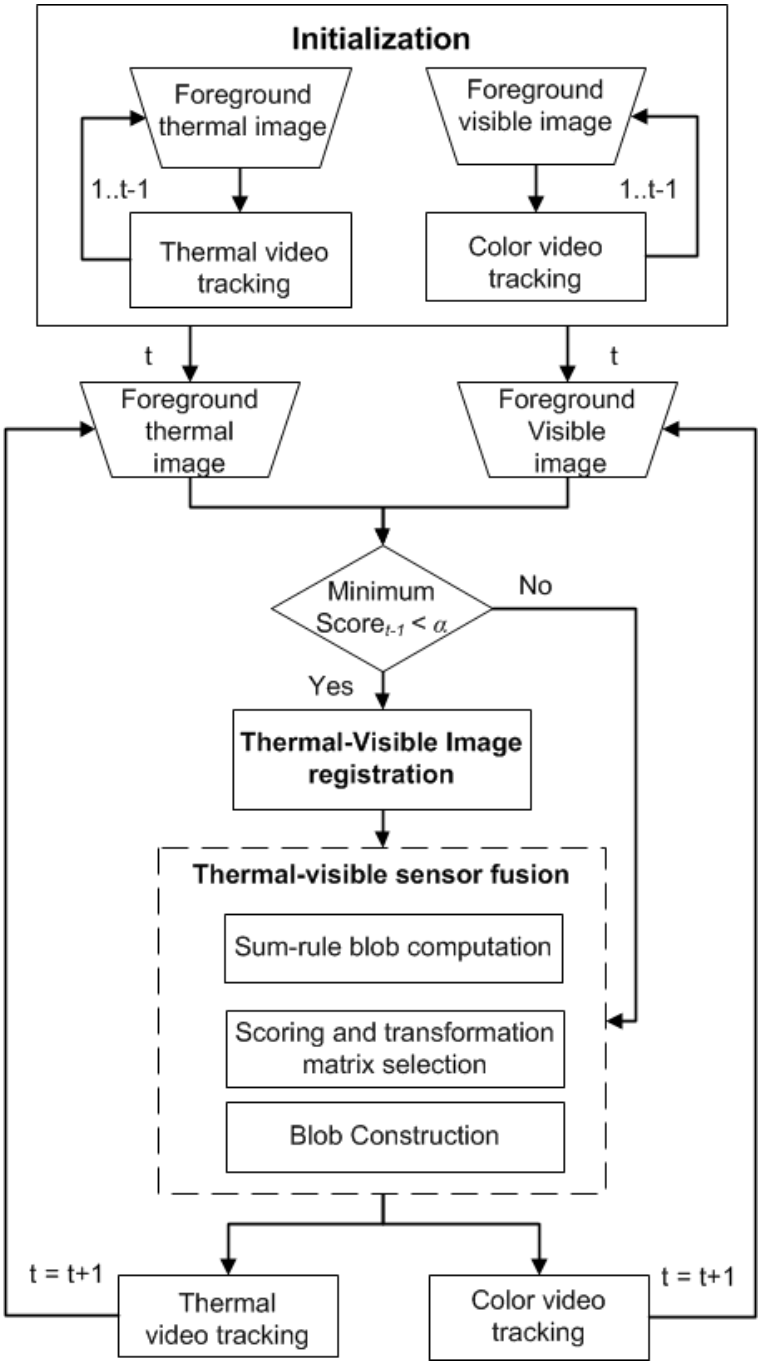


Figure 4.2 Flowchart of our system

Repeat N_1 times

1. Pick a pair of trajectories at random
2. Estimate the preliminary affine matrix H:
   Repeat N_2 times
   - Pick three pairs of points at random
   - Calculate H
   - Add inliers pairs of points:
     a. Evaluate Euclidean distance error
     b. Re-calculate H
3. Evaluate overlapping error
4. Add inliers pairs of trajectories:
   Repeat for all possible pairs
   - Add inliers pairs of points
   - Evaluate overlapping error

Figure 4.3 RANSAC-based algorithm for trajectory point matching

transformation matrix based on the matching of object trajectory points from a pair of thermal and visible videos. Fig. 4.3 shows the steps of our object trajectory point matching. It is composed of two RANSAC loops, one for the pairs of trajectories with N_1 iterations, and one for the pairs of points in a selected pair of trajectories with N_2 iterations. The number of iterations N is computed with

$$N = \frac{log(1-p)}{log(1-(1-\epsilon)^s)}, \tag{4.1}$$

where p is the confidence (in our experiments p is 0.99) and s is the minimum number of points required for the homography (e.g. $s = 3$ for affine transformation). ϵ, the probability of outliers, is computed by

$$\epsilon = 1 - \frac{N_p}{N_t}, \tag{4.2}$$

where N_p is the number of inlier pairs of points/trajectories and N_t is the total number of pairs of points/trajectories. In fact, the number of iterations depends on the number of inlier pairs of points/trajectories. The larger the number of inlier pairs, the fewer iterations are required. In our algorithm (Fig. 4.3), N_1 and N_2 are determined by Eq. 4.1 and 4.2.

H is calculated using three pairs of points selected at random. After that, all the points of the trajectory of the thermal video frame are transformed using the estimated H. Then, the Euclidean distance between these transformed points, and their corresponding points in the visible video are computed. Pairs of points for which the Euclidean distance is smaller than

a threshold T (typically, $T = 5$ pixels) are considered as inlier pairs. The best estimation of H is that computed with the largest number of inlier pairs of points. H is re-estimated using all the inliers pairs of points. Fig. 4.4 illustrates the matching of selected pairs of trajectory points.

After the first estimation of the transformation matrix H, its quality is evaluated using an overlapping error function OE defined for the foreground pixels of the pairs of thermal and visible video frames.

$$OE = 1 - \frac{N_{c \cap t}}{N_{c \cup t}}, \tag{4.3}$$

where $N_{c \cap t}$ is the number of overlapping foreground color and thermal image pixels, and $N_{c \cup t}$ is the number of foreground pixels from the union of the color and thermal images. The overlapping error as a second matching criterion enables our method to perform, even when there are a few trajectories in a pair of videos (i.e. overlapping pairs of trajectory points are a matching criterion).

For each possible pair of trajectories, the thermal image trajectory points are transformed into visible image coordinates, and then the inlier pairs of points are selected using Eq. 4.3. Using all inlier points, the H matrix is recalculated. Then, the overlapping error is computed for the new estimated matrix H. If the overlapping error for the new estimated matrix is less than the overlapping error of the previous estimation of H, the pair of trajectories is added to the set of inlier pairs of trajectories. This procedure is continued until all the possible pairs of trajectories have been evaluated.

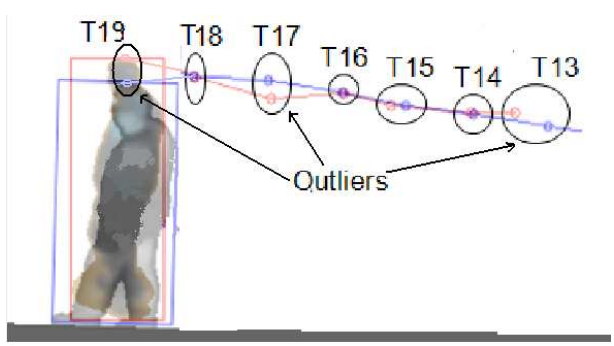


Figure 4.4 Matching trajectory points from thermal and visible video. $T14$, $T15$, $T16$, $T18$, and $T19$ are inliers.

4.5 Thermal-visible sensor fusion

Thermal-visible sensor fusion combines the information of the registered color and thermal foreground images. Fig. 4.5 shows our sensor fusion algorithm. M_n represents the transformation matrix estimated by image registration in the current frame, and M_b represents the current best matrix. If the image registration is not performed in the current frame, computations related to M_n shown in 4.5 are simply skipped.

In this work, a silhouette is defined as a binary object region, and a sum-rule silhouette is defined as a silhouette constructed using a sum of probabilities of foreground pixels in thermal and visible images. To compute a sum-rule silhouette, either foreground pixel coordinates of the thermal image should be transformed into visible image coordinates, or vice versa. Using either method, the computed sum-rule silhouette is the same. The sum-rule method was proposed by (Han and Bhanu (2007)), and is defined as

$$(X,Y) \in S : \textbf{IF} \;\; P(S \mid t(X,Y)) + P(S \mid c(X,Y)) > \alpha_{sum}, \tag{4.4}$$

where $t(X,Y)$ represents the thermal value at image coordinates (X,Y), $c(X,Y)$ represents the color value at image coordinates (X,Y) after transformation, S represents the sum-rule silhouette, and α_{sum} represents a threshold. The probabilities that a pixel belongs to the foreground in each sensor are computed as

$$P(S|t(X,Y)) = 1 - e^{-\|t(X,Y)-\mu_t(X,Y)\|^2} \tag{4.5}$$

where $\mu_t(X,Y)$ is the mean background value of the coordinates (X,Y) for the thermal image. $P(S|c(X,Y))$ is computed similarly for transformed visible image. The quality of a sum-rule silhouette is evaluated using a score function. A transformation matrix is selected, based on the scoring results of all the silhouettes inside one image. The score function for the thermal image is defined as follows :

$$SF_t(i) = \frac{sum\left(B^t_{j\in\{1,\ldots n\}} \cap S^t_i\right)}{sum\left(B^t_{j\in\{1,\ldots n\}}\right)}, i \in \{1,...,m\} \tag{4.6}$$

where m is the number of computed sum-rule silhouettes inside the intersecting FOVs of the two cameras, S^t_i represents the ith sum-rule silhouette computed in the thermal image, $SF_t(i)$ represents its score, and B^t_j are blobs in the original thermal foreground image that intersect with S^t_i. Since background subtraction is not perfect, object regions might be fragmented into smaller ones in the original foreground image. So, the blobs B^t_j that intersect S^t_i should all be fragments belonging to one object. If all blobs B^t_j are inside S^t_i, then S^t_i is

Repeat for thermal and visible images

1. Compute sum-rule silhouettes using matrix M_n
2. Compute sum-rule silhouettes using matrix M_b
3. Compute Blob scores:
   Repeat for all the blobs in the reference image
   - Calculate *Score* of silhouette computed using M_n
   - Calculate *Score* of silhouette computed using M_b
4. Compute overall score $Score_n$ of reference image
5. Compute overall score $Score_b$ of reference image
6. select transformation matrix:
   IF $Score_n > Score_b$ **THEN** $M_b = M_n$
7. Compute object model

Figure 4.5 Our sensor fusion algorithm

perfectly aligned and its score will be 1 (the maximum value). The same applies for visible images for computation of score function in visible $SF_c(i)$. The score of matrix M_n for one image is,

$$Score_n = \left\{ \frac{\sum_{i=1}^{m} (SF_c(i) + SF_t(i))}{2 \times m} \right\}_{M_n} \tag{4.7}$$

where m is the number of sum-rule silhouettes, $Score_n$ is the score of matrix M_n.The $Score_b$ (the score of matrix M_b for one image) is computed similarly using matrix M_b. Finally, if the score $Score_n$ of the new estimated matrix is higher than the score $Score_b$ of the best matrix, M_n replaces M_b.

Blobs are also constructed. In our work, a blob is defined as all the pixels (either connected or disconnected) with their visual features that belong to one object in an image. Blobs are the input data of tracking step. The sensor fusion improves the quality of input data by computing a sum-rule silhouette that handles the shortcomings of the background subtraction using a single sensor, such as blob fragmentation. Furthermore, sensor fusion provides the color and thermal information of the blob pixels that are used as features for tracking. For blob construction, if the score of a sum-rule silhouette (Eq. 4.6) is maximum which is 1, the sum-rule silhouette will be considered as a detected blob in the reference image. Otherwise, the original blob's fragments computed by background subtraction that intersect with the computed sum-rule silhouette will be clustered as one blob. In this way, the fragmentation problem is solved.

4.6 Multiple people tracking method

The object model used in our tracking method is the color-thermal histogram of the input blobs. This histogram has 54 bins for the HSV colors and 16 bins for the thermal intensities. For tracking, any method that computes and updates the trajectory of the objects frame by frame is applicable. Here, we use an online Multiple Hypothesis Tracking (MHT) method, which we proposed in previous work (Torabi and Bilodeau (2009)). Our tracking method identifies objects at each frame and estimates the best trajectories computed up to the current frame. In our previous work (Torabi and Bilodeau (2009)), the tracking was performed only for videos captured by a single visible camera. Therefore, we presented a method for handling blob fragmentation that used the spatial and temporal characteristics of blobs for a few frames, in order to reattach the blob fragments belonging to one object. In this work, instead of this fragmentation handling method, we applied data fusion, which combines the information from the thermal and color videos and improves the quality of the input data for tracking, and, consequently, improves the tracking results considerably. Tracking is performed separately for thermal and visible videos using constructed blobs with thermal-visible histogram as tracking feature.

Our tracking algorithm has three main steps that are described in the following sections. We use two graphs for tracking : an event graph to record all blob's events and store their appearance information while they are being tracked, and a hypothesis graph to generate hypotheses for handling data association of split objects.

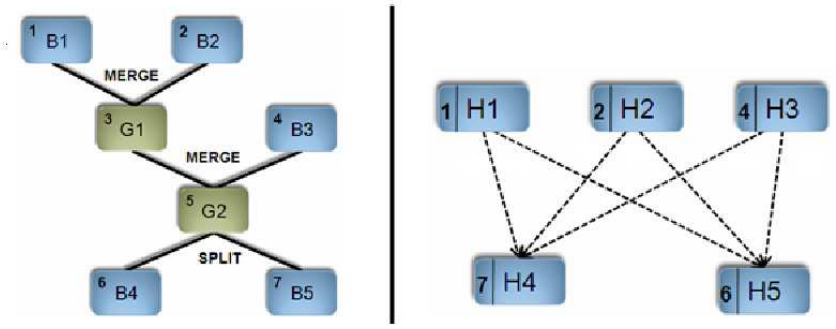


Figure 4.6 Event graph (left) and hypothesis graph (right). In the hypothesis graph, the number on the left of each hypothesis node corresponds to a track node in the event graph, with the corresponding number in the upper left corner.

4.6.1 Definition of event graph and hypothesis graph

Fig. 4.6 shows an event graph with its corresponding hypothesis graph. The event graph represents all blobs with their merging and splitting events during tracking. Each vertex of this graph (track node) stores a blob's appearance, including top-most point coordinates, its adaptive thermal-color histogram, blob events such as correspondence, merging, and splitting, and the frame number of the last update in the node. Edges represent merging and splitting events among the blobs. The hypothesis graph is a directed, weighted graph. The vertices of this graph (hypothesis nodes) simply correspond to the track nodes of the event graph that belong to entering blobs (blobs that appear in the scene) and split blobs (blobs that break away from a group, or a single blob). A group blob does not have hypothesis nodes. This is because these nodes are used to solve the data association problem before and after object interactions. The weight of each edge $n_i n_j$ that represents a hypothesis is defined as,

$$\omega\left(n_i n_j\right) = \left|AH\left(n_i\right) - AH\left(n_j\right)\right|, \tag{4.8}$$

where $\omega\left(n_i n_j\right)$ is the Euclidean distance between two adaptive color-thermal histograms of the two blobs belonging to the hypothesis nodes n_i and n_j. In practice, the edge information is stored in the nodes. Thus, for each hypothesis node n_i, three sets of nodes, called S (Source), E (End), and BH (Best Hypotheses), are defined as,

$$S\left(n_i\right) = \left\{n_j | \exists n_j n_i\right\}, \tag{4.9}$$

$$E\left(n_i\right) = \left\{n_k | \exists n_i n_k\right\}, \tag{4.10}$$

$$BH\left(n_i\right) = \left\{n_j \in S\left(n_i\right) | E_1\left(n_j\right) = n_i\right\}. \tag{4.11}$$

The sets defined by Eq. 4.9 and Eq. 4.10 are ordered based on the weights of their common edges with n_i. In Eq. 4.11, BH can be empty or contain one or more elements. E_1 is the first element of E. The sets S, E, and BH are used for object labelling and for finding trajectories. It is important to note that the event graph and the hypothesis graph may be composed of more than one component (subgraph), since the connections between nodes represent the interactions that have occurred between the blobs during tracking (two blobs that do not interact are not connected).

4.6.2 Step1 : matching blobs

In the first step of our algorithm, a distance matrix is computed to find the blobs $B_i(t-1)$ and $B_j(t)$ that possibly correspond, along with their appearance dissimilarities in two consecutive frames. The appearance dissimilarity $D^t_{t-1}(i,j)$ is defined as

$$D^t_{t-1}(i,j) = \begin{cases} d(h_{B_i(t-1)}, h_{B_j(t)}) & \text{if overlapped} \\ -1 & \text{otherwise} \end{cases}, \tag{4.12}$$

where $d(h_{B_i(t-1)}, h_{B_j(t)})$ is the thermal-color histogram intersection between the ith blob in frame $t-1$ and the jth blob in frame t if the bounding boxes of the two blobs overlap (i.e. based on an assumption that corresponding blobs in two consecutive frame does not have a dramatic displacement ; therefore the bounding boxes surrounding two corresponding blobs are spatially overlapped). Otherwise, these two blobs cannot match each other and their corresponding element in the matrix is -1. The size of the distance matrix is $N \times M$, where N is the number of blobs in the frame $t-1$ and M is the number of blobs in the frame t. The thermal-color histogram intersection is defined as

$$d(h_{B_i(t-1)}, h_{B_j(t)}) = \frac{\sum_{k=1}^{K} min(h_{B_i(t-1)}(k), h_{B_j(t)}(k))}{\sum_{k=1}^{K} h_{B_i(t-1)}(k)}, \tag{4.13}$$

where $h_{B_i(t-1)}$ and $h_{B_j(t)}$ are the thermal-color histogram of the ith blob in frame $t-1$ and the jth blob in frame t, and K is the number of the thermal-color histogram bins.

A blob in frame $t-1$ matches a blob in frame t if the dissimilarity is not -1. Events such as entering, leaving, merging, and splitting are detected by finding the matching blobs in two consecutive frames using the distance matrix.

4.6.3 Step 2 : updating the graphs

The event graph and the hypothesis graph are updated based on the events detected in the matching process :

- If a blob in the current frame t is an appearing object, a track node in the event graph and a hypothesis node in the hypothesis graph are added.
- If correspondence is detected between two blobs in frames $t-1$ and t, the track node in the event graph belonging to the object is updated by adding its top-most point in the current frame t, adding the current frame number, and updating its adaptive thermal-color histogram using

$$AH_{B(t)} = \sum_{k=1}^{K} \alpha AH_{B(t-1)}(k) + (1-\alpha) h_{B(t)}(k). \tag{4.14}$$

In Eq. 4.14, $AH_{B(t-1)}$ is the adaptive thermal-color histogram of blob B at frame $t-1$, K is the number of thermal-color histogram bins, $h_{B(t)}$ is the thermal-color histogram of blob B at frame t, and α (varying between 0 and 1) is an adaptation parameter. The adaptive thermal-color histogram is used for generating a hypothesis (likelihood between two nodes) ; because it gives the global thermal-color information of the blob over several frames and helps reduce the effect of dramatic changes in the thermal-color distribution caused by short-time variations in lighting and temperature, as well as by shadows. Updating a track node for a correspondence event is equivalent to a sequential data association for blobs that are not in a situation of identification uncertainty. This is based on the fact that, if two blobs, one in each of two consecutive frames are found to be similar with a mutual matching, it is very likely that they are associated with the same object.

- If some blobs in frame $t-1$ are merged into a single blob in the current frame t, the tracking of the merging blobs is stopped and a new track node for the group blob is initiated in the event graph.
- If a blob in frame $t-1$ has disappeared from the FOV of the camera, its track node in the event graph is deactivated.
- If splitting is detected, for each split blob a track node in the event graph and a hypothesis node in the hypothesis graph are added and hypotheses are generated for the newly added nodes.

To generate the hypotheses for split blobs, hypothesis nodes are added. Then, the S, E, and BH sets of all the nodes that are in the same subgraph (i.e. part of graphs that their nodes are either direct or non-direct children of a root node) as the newly added nodes are updated. Generating a hypothesis only for the nodes in the corresponding subgraph and not for the other nodes in the hypothesis graph is part of our strategy to reduce the number of hypotheses.

To perform the update, newly initiated nodes are added to the E sets of the nodes from the previous frames in the subgraph, and the previous nodes in the subgraph are added to the S sets of the newly initiated nodes. Also, the BH sets of the newly added hypothesis nodes are created according to their S sets. In other words, all the nodes in the subgraph are connected, along with directed edges from the past hypothesis nodes to the new hypothesis nodes. The weight of each directed edge is the likelihood that the source node and the end node have the same appearance, and is calculated using Eq. 4.8.

If the first elements of the E sets are changed after updating (S sets and E sets are always ordered increasingly), the BH sets in the same subgraph are updated consecutively. This is based on the fact that the intersection of two BH sets for two different nodes should be

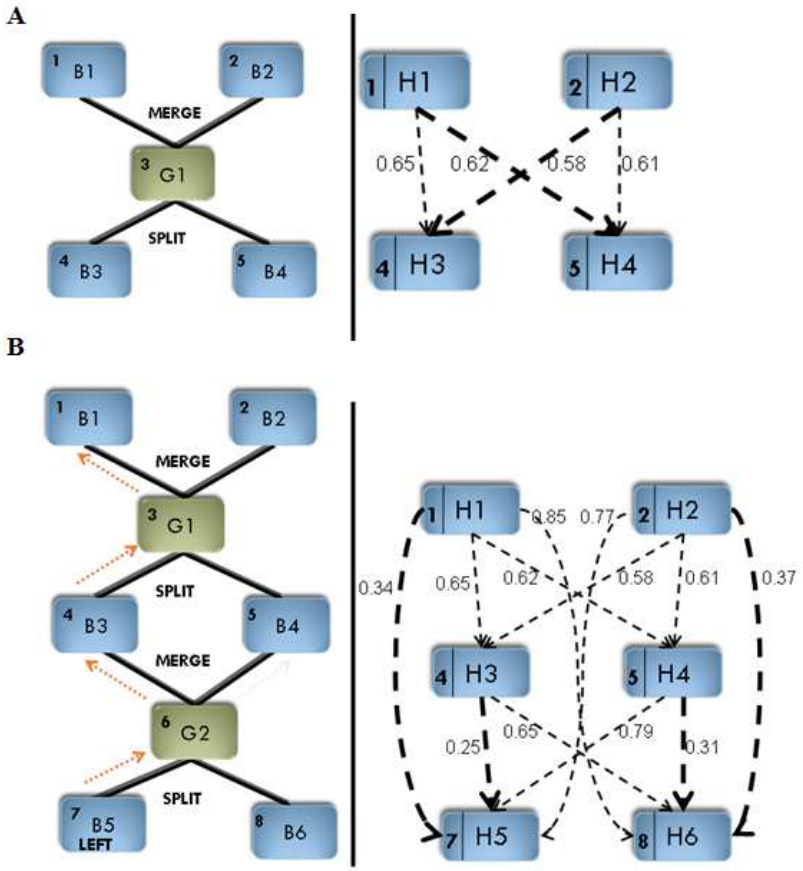


Figure 4.7 A) An event (left) and a hypothesis graph (right) after a merge/split. B) The same graph updated after a second merging and splitting. The number at the left of each hypothesis node corresponds to a track node in event graph with the same number in the upper left corner of the track node. The dashed arrows in the event graph show the history of one object.

empty. Figure 4.7 shows an example of graphs updating after a merging and splitting events.

4.6.4 Step 3 : object labeling and trajectory computation

The goal of object labeling is to assign a label to each tracked blob in the current frame. For a correspondence event, the blob's label in frame t is the same as it is in frame $t-1$. For merging, the merged blob's label in frame t is the label of all the merging blobs in frame $t-1$. For a blob entering frame t, the label is a new one.

For splitting, the label of a split blob in frame t is determined by processing the hypothesis graph. To do this, we traverse the hypothesis graph in bottom-up fashion, from the current frame, starting from the split blob's hypothesis node n_i. To do this, the TN (Traversing Node) set is initialized by,

$$TN_0(n_i) = \phi, \tag{4.15}$$

where ϕ represents an empty set of nodes. TN is updated by

$$TN_t(n_i) = (TN_{t-1}(n_i) \cup BH(n_{current})) - n_{next}. \tag{4.16}$$

In Eq. 4.16, $n_{current}$ is the current node during graph traversal (at first $n_{current}$ is n_i and $TN_{t-1}(n_i)$ is ϕ), $TN_t(n_i)$ is a set of possible next destination nodes in the current frame t, and n_{next} is the next node to traverse in the graph chosen with two criteria : 1) n_{next} exists in either $BH(n_{current})$ or $TN_{t-1}(n_i)$; and 2) n_{next} has the closest temporal relationship with $n_{current}$. It is important to note that, if there is more than one node in $BH(n_{current})$ or $TN_{t-1}(n_i)$ that obeys the n_{next} criteria, we traverse these nodes separately. Traversing the graph upward and updating the TN set are continued until we reach a node for which the TN set becomes empty (nowhere to go next). A split blob is given the label of the blob that we reach after traversal of the hypothesis graph. A hypothesis node belonging to a split blob that has an empty BH set before starting graph traversal is a new appearing object that is given a new label.

At each frame, object trajectories are computed by traversing the hypothesis graph in the same way as for labeling, to get its path into the hypothesis graph. However, in the hypothesis graph, some parts of the trajectory (when the object was tracked in a group) are missing, because group blobs have no nodes in the hypothesis graph. The missing parts of the path are recovered by completing it with the help of the event graph. Fig. 4.7 illustrates an example of trajectory construction for two objects that occlude each other twice. The represented values on hypothesis graphs are the weights belonging to this specific example. The thicker dashed arrow represents smallest weight that is generated from a source node.

4.7 Results and discussion

We have assessed the performance of our method using nine video sequences that we captured (LITIV dataset) and three video sequences of the OTCBVS dataset (Davis and Sharma (2005)). The LITIV dataset consists of videos of different tracking scenarios captured by a thermal and visible camera at 30 frames per second with different zoom settings and at different positions. The size of the images is 320 $\times$ 240. Fig. 4.12 gives qualitative results of our unified image registration, sensor fusion, and tracking. As columns (f) and (g) in the second row of Fig. 4.12 show, our system tracks objects solely at the intersection of the FOVs of the thermal and visible cameras, since sensor fusion requires the data from both sensors. In section 4.7.1, we quantitatively assess the performance of our image registration and show that our method outperforms state-of-the-art image registration methods (Caspi *et al.* (2006); Bilodeau *et al.* (2011b)). In section 4.7.2, we describe the quantitative results of our thermal-visible multiple people tracking and show the advantage of our integrated framework which performs multimodal tracking compared to separate tracking for thermal and visible videos.

4.7.1 Image registration evaluation

We have compared our image registration method with the image registration methods proposed by (Caspi *et al.* (2006)) and (Bilodeau *et al.* (2011b)), using the same background subtraction parameters for all methods. In (Caspi *et al.* (2006)) and (Bilodeau *et al.* (2011b)), the input data are trajectories generated from separate tracking for a thermal video and a visible video without sensor fusion. In contrast, in our method, the trajectories are generated by the tracking method described in section 4.6 performing iteratively with our image registration in an integrated framework. In (Caspi *et al.* (2006)), the registration criterion is the Euclidean point error of the object trajectory points in a pair of thermal and visible videos. In our proposed method and (Bilodeau *et al.* (2011b)), foreground pixel overlapping is used as a matching criterion (more details in section 4.4). However in (Bilodeau *et al.* (2011b)), image registration is based on a simple iterative scheme where the matrix selection is based on a simple foreground overlapping error rather than the blob fusion score used in this work.

To quantitatively compare the performance of image registration methods for each pair of videos, we constructed ground-truth (GT) foreground binary images using a manual image registration. For the manual image registration of each pair of videos, one pair of thermal and visible video frames was manually aligned, and, based on this alignment, the affine transformation matrix was computed and used as the GT transformation matrix. Then, two GT binary foreground images are constructed by manually selecting points forming

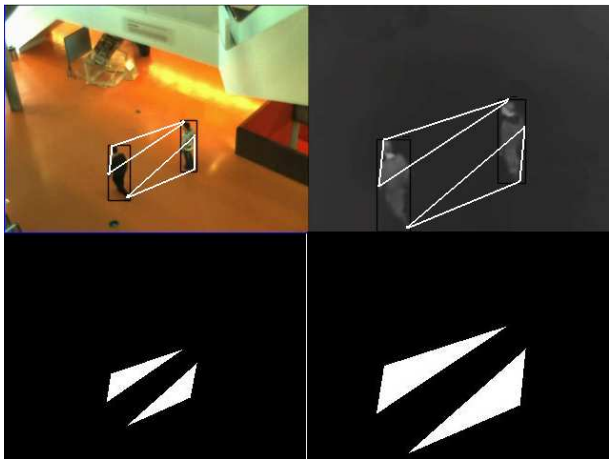


Figure 4.8 Top : manually selected polygons in IR and in visible images (Frame 90, Seq.1)) ; bottom : GT binary images

polygons on the thermal image and by transforming the polygon's pixel coordinates of the thermal image using the GT transformation matrix to obtain a GT foreground for the visible image. Fig. 4.8 shows the manually selected polygons and the GT thermal and visible binary foreground images. We used the GT foreground images for testing the overlapping error to ensure that the background subtraction error does not contribute to it. We used two metrics to validate our method : 1) the foreground pixel overlapping error (using an equation similar to Eq. 4.3) of the aligned GT foreground images using the matrices computed by our method and other two methods ; and 2) the average point error, which is the average pixel coordinate error in the x and y directions of the aligned polygons' corners after transformation of the GT foreground images.

For foreground pixel overlapping error comparison of our method and Caspi *et al.* (Caspi *et al.* (2006)), we have chosen video sequence 8 of the LITIV dataset. This pair of videos is challenging because there are several long term blob fragmentations due to background subtraction misdetection and partial occlusion caused by a stationary object that is part of the background in the scene. In addition, this pair of videos is captured with a thermal and a visible camera at different zoom settings with a small intersection of the FOVs, which makes image registration a challenging problem. Fig. 4.9 shows the blob fragmentations and the considerable object scale difference in a pair of thermal and visible image frames of video 8 (frame 300).

Fig. 4.10 shows the foreground pixel overlapping error (Eq. 4.3) for video pair 8 using our method, the method of (Caspi *et al.* (2006)), and manual image registration. Manual image registration also has a small overlapping error that is caused by rounding polygon coordinate values after transforming the points (our registration precision is at the pixel

Figure 4.9 Top : a thermal and a visible video frames (Frame 300, Seq.8), Bottom : corresponding thermal and visible foreground images

level). Around frames 350-400, due to several blob fragmentations occurring in the thermal video because of background subtraction misdetection, the overlapping error increases in the method of (Caspi *et al.* (2006)). Also, in several frames, this method cannot estimate an acceptable transformation matrix, since the trajectories in the thermal and visible videos are not similar in those frames. Therefore, the RANSAC algorithm did not succeed in estimating a transformation matrix based on matching the trajectories. In general, this plot shows : 1) our method estimates a good transformation matrix (error less than 30 percent) starting from around frames 110-120 ; 2) the transformation matrix estimated by our method is more stable over time compared to the method of (Caspi *et al.* (2006)), and 3) the overlapping error of our method is smaller than for the method of (Caspi *et al.* (2006)) over most video frames.

Our image registration, which performs iteratively with sensor fusion and tracking in an integrated system, has better image registration results than the method of (Caspi *et al.* (2006)), because : 1) the transformation matrices computed using more accurate trajectory points generated by tracking with sensor fusion are more precise than those computed using trajectories generated by separate tracking, because blob fragmentation is better handled ; this is especially true for videos where there are several long term blob fragmentations, such as video sequence 8 (Fig. 4.9) ; 2) using the foreground pixel overlapping criterion results in good estimates of the transformation matrix, even when there is a relatively small FOV intersection ; this makes trajectory matching a harder problem, since the trajectory patterns in the two videos are not similar, and 3) by using feedback, the matrix selection based on

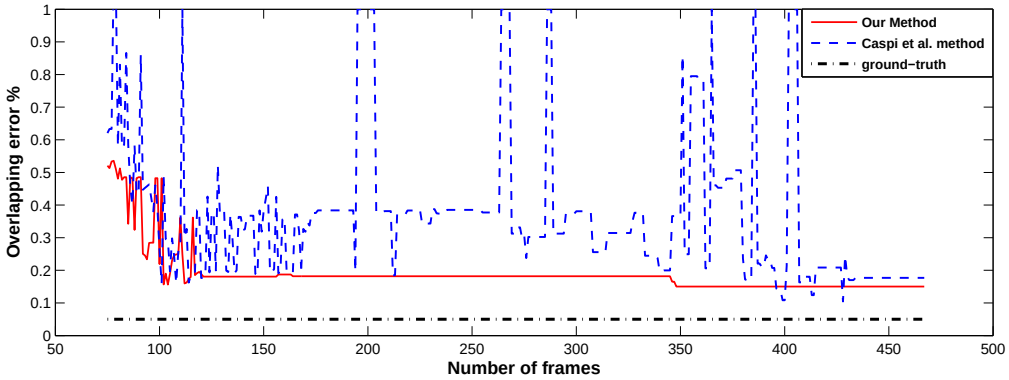


Figure 4.10 Overlapping error of our image registration method, of (Caspi *et al.* (2006)) image registration method, and of the manual image registration for video 8 frames 62-467.

the fusion score (section 4.5) replaces the previous transformation matrix by a new one only if it has better fusion score.

Fig. 4.11 shows the foreground pixel overlapping error (Eq. 4.3) for video pair 1 using our method, the method of (Bilodeau *et al.* (2011b)), and manual image registration. The reason why we have chosen video pair 1 is because it has a larger intersection of the FOVs (more similar trajectories), which enable us to show the performance of simple matrix selection and compare it with matrix selection based on fusion score that we used in this work. Plots in fig. 4.10 and fig. 4.11 show the transformation matrix selection in our method is more stable since there is less variation in the overlapping foreground errors compared to both state-of-the-art methods (Caspi *et al.* (2006); Bilodeau *et al.* (2011b)). Fig. 4.11 shows that even the simple matrix selection used in (Bilodeau *et al.* (2011b)) results in more stable registration results with less foreground overlapping error variations. However, because of the lack of accuracy of computed trajectories and the use of more sophisticated matrix selection such as the one used in our integrated framework, the overlapping errors vary more and even in some frames increase because of erroneous matrix selection compared to the errors of our proposed method.

Table 4.1 shows the average point errors of our image registration method and the (Caspi *et al.* (2006)) method for 12 video sequences. This table shows that, for video pairs 1, 3, 4, and 8, which are captured at considerably different zoom settings and a relatively small FOV intersection (less similar trajectory patterns) in both X and Y, the Euclidean distance errors of our system are less than with the (Caspi *et al.* (2006)) method. This shows that our method is more robust than the (Caspi *et al.* (2006)) method in challenging videos, where there are fewer similar trajectory patterns in the thermal and visible videos. This is basically because of two features of our method : 1) using the foreground pixel overlap criterion in

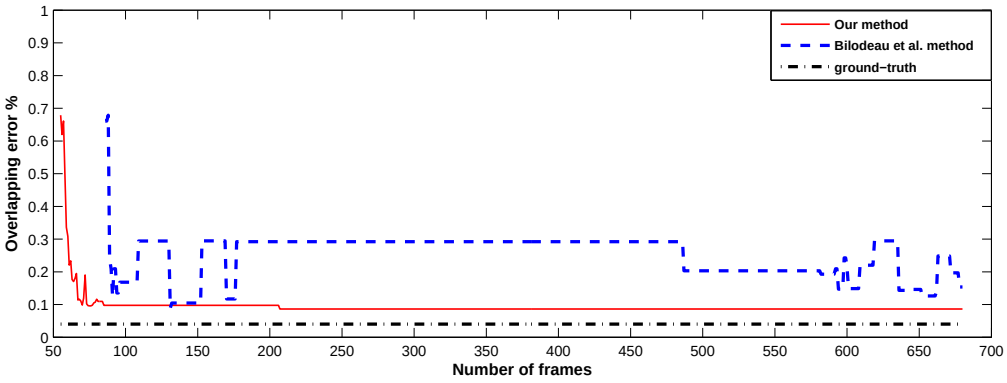


Figure 4.11 Overlapping error of our image registration method, of (Bilodeau *et al.* (2011b)) image registration method, and of the manual image registration for video 1 frames 55-680.

the RANSAC-based algorithm ; and 2) sensor fusion, which handles the fragmentation and gives more similar trajectories in both the thermal and visible videos. For the videos that are captured with the same zoom and with about the same FOV intersection (videos 2, 5, and 7) and in which there is a reasonable amount of short term blob fragmentation that does not significantly change the trajectories, our method and the (Caspi *et al.* (2006)) method give similar results. However, for video 6, where the FOVs of the two cameras are about the same, because of long term blob fragmentation that changes the trajectory patterns considerably, our method produces better results.

In our tests, videos from the OTCBVS dataset (videos 10, 11, and 12) are considered as unregistered sequences of images. In video 11, the average point errors are greater because there is only one person in this video and he is walking in a straight line. Thus, all the trajectory points are collinear, and so one of the assumptions required for estimating a precise affine matrix is not met.

4.7.2 Tracking evaluation

In this section, we quantitatively compare our tracking results using sensor fusion with separate tracking for the visible and thermal videos, but with the same data association method. In separate visible tracking, the color histogram is used as the tracking feature and

Table 4.1 Seqs. 1-9, videos from the LITIV dataset, and Seqs. 10-12, videos from the OTCBVS dataset (Davis and Sharma (2005)). Our image registration results and Caspi *et al.* (Caspi *et al.* (2006)) registration results. NF : number of video frames, SF : starting frame, which is the first frame after initialization in our method (section 4.4), NP : number of people in the scene, AE_X : Average Euclidean error in X of the polygons' corners for frames after initialization, AE_Y : Average Euclidean error in Y of the polygons' corners for frames after initialization.

Seq.	*Method*	NF	SF	NP	AE_X	AE_Y
1	our method	680	54	7	0.68	2.17
	Caspi *et al.*				4.75	14.79
2	our method	698	143	3	4.14	3.37
	Caspi *et al.*				6.30	3.96
3	our method	1238	200	5	2.84	2.74
	Caspi *et al.*				5.63	4.87
4	our method	329	60	2	3.89	2.84
	Caspi *et al.*				9.85	11.97
5	our method	563	100	3	2.85	3.08
	Caspi *et al.*				4.71	16.12
6	our method	1055	100	4	4.18	5.22
	Caspi *et al.*				9.86	14.07
7	our method	895	107	4	4.38	3.61
	Caspi *et al.*				4.34	2.67
8	our method	467	100	5	3.05	2.22
	Caspi *et al.*				8.89	11.21
9	our method	400	50	3	5.61	4.89
	Caspi *et al.*				7.29	7.79
10	our method	2031	180	2	1.29	1.57
	Caspi *et al.*				1.05	2.87
11	our method	650	123	1	5.92	9.03
	Caspi *et al.*				9.36	8.33
12	our method	1302	100	3	0.83	0.37
	Caspi *et al.*				6.93	2.83

in separate thermal tracking ; the pixel intensity histogram is used as the tracking feature. Table 4.2 shows the tracking results of our method and separate thermal and visible video tracking.

False positive person identification, $+P$, mostly occurred during blob fragmentation, where a part of the human's body is detected as a new person. This can happen in the short term (1-2 frames) or the long term (several frames). As shown in Table 4.2, our sensor fusion succeeded in reducing the $+P$ error by handling blob fragmentation for both thermal and visible images in almost all the videos. The other error is the false negative person identification, $-P$. This error mostly occurs because of errors in people identification during a merge-split, or partial occlusion of a person by an object in the scene, where the person is falsely detected as a new object. Our system was able to reduce errors in people identification during a merge-split in our tested videos. The reason is that, in our method, a thermal-visible histogram is used as the tracking feature, which is more robust than separate color or thermal intensity histograms.

In Table 4.2, we also quantitatively compared the trajectories generated with our method and those generated by the separate video trackers using GT trajectories generated manually. The average Euclidean distance trajectory point error, AE_{ir-vi}, of our tracking method is significantly smaller than the separate visible/infrared trackers. This shows the effectiveness of sensor fusion for computing more accurate trajectories. In fact, our video registration and tracking results show that our sensor fusion plays a critical role in improving the quality of the whole system.

Table 4.2 Seq.1-9, videos from the LITIV dataset and Seq. 10-12 videos from the OTCBVS dataset (Davis and Sharma (2005)). Our thermal-visible tracking results and separate thermal-visible tracking results without sensor fusion. NF : number of frames, NP : number of tracked people, $+P_{ir-vi}$: false positive identified number of people in thermal and visible, $-P_{ir-vi}$: false negative identified number of people in thermal and visible, and AE_{ir-vi} : Average Euclidean distance trajectory point error compared with manually generated GT trajectories.

Seq.	*Method*	NF	NP	$-P_{ir-vi}$	$+P_{ir-vi}$	AE_{ir-vi}
1	Our method	680	7	0-0	0-0	3.57-2.12
	Separate			0-2	1-3	3.98-2.42
2	Our method	698	3	0-0	0-1	2.32-3.57
	Separate			4-4	2-1	2.74-2.47
3	Our method	1238	5	0-0	0-0	2.72-2.83
	Separate			0-4	5-0	3.27-2.74
4	Our method	329	2	0-0	0-0	5.02-3.12
	Separate			2-2	1-3	19.22-15.71
5	Our method	563	3	0-0	2-3	2.86-2.22
	Separate			2-2	3-3	2.83-3.17
6	Our method	1055	4	0-0	2-4	3.60-2.18
	Separate			0-0	4-6	10.48-7.54
7	Our method	895	4	2-2	0-3	2.27-2.46
	Separate			4-4	3-4	2.35-2.43
8	Our method	467	5	0-1	3-3	7.93-5.31
	Separate			2-1	11-8	14.56-5.26
9	Our method	400	3	0-0	2-2	3.06-4.70
	Separate			2-2	2-4	3.27-4.85
10	Our method	2031	2	0-0	1-0	2.51-1.38
	Separate			0-0	6-3	4.87-2.60
11	Our method	650	1	0-0	0-0	1.67-3.03
	Separate			0-0	4-0	1.22-1.92
12	Our method	1302	3	0-0	0-0	1.73-1.77
	Separate			0-0	3-0	0.81-0.75

4.8 Conclusions

In this paper, we have proposed an iterative integrated framework for thermal-visible video registration, sensor fusion, and multiple people tracking method with feedback designed for a pair of far-range, synchronized thermal and visible videos. Our video registration method is based on a RANSAC trajectory-to-trajectory matching that estimates an affine

(a) (b) (c) (d) (e) (f) (g)

Figure 4.12 Our results of video 1 at frames 99, 182, 300, and 652. (a) registration of the visible on the thermal image, (b) sum-rule silhouette aligned on the visible image, (c) sum-rule silhouette aligned on the thermal image, (d) and (f) tracking result for the visible image, and (e) and (g) tracking result for the thermal image

transformation matrix. Our sensor fusion method handles the object fragmentation caused by imperfect single sensor background subtraction using the aligned thermal and visible video frame pairs. Finally, our multiple people tracking methods inputs blobs constructed in sensor fusion and output the trajectories of moving people in the scene.

In our results, we have shown that sensor fusion improves tracking, and ultimately the accuracy of the object trajectories and registration. Our experiments show that our method outperforms similar methods previously developed, such as the methods in (Caspi *et al.* (2006); Bilodeau *et al.* (2011b)). Our proposed feedback scheme is flexible enough to use any other tracking method that generates trajectories online, and any other sensor fusion and object modeling that is needed for a specific video surveillance application.

CHAPTER 5

A PERFORMANCE EVALUATION OF LOCAL DESCRIPTORS AND SIMILARITY MEASURES FOR THERMAL-VISIBLE HUMAN ROI REGISTRATION

Abstract

In this paper, we compare the performance of some local image descriptors and similarity measures for multimodal dense stereo matching of image region of interest (ROIs). For thermal-visible image registration, the similarity metric should be distinctive and robust to the large differences in the thermal and visible image characteristics. At first, our evaluation uses simple Winner Take All (WTA) window matching and assesses the viability of SURF, HOG, LSS, BRIEF, NCC, and MI by precision-recall and power of discrimination criteria. We then compare the performance of the three best metrics (LSS, MI, and HOG based) in realistic scenarios of human monitoring applications using a more appropriate matching method robust to occlusions and depth discontinuities. We observe that the ranking of the metrics is independent of the matching method and that LSS-based matching performs best.

5.1 Introduction

In recent years, there has been a growing interest in visual surveillance using multimodal sensors in both civilian and military applications. The fundamental issue associated with thermal-visible imagery is the matching and registration of pairs of images captured by 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. Due to the numerous differences in imaging characteristics of thermal and visible cameras, most correspondence measures used for registering visible images are not applicable for thermal-visible image registration. Moreover, it is impossible to find correspondences across an entire scene, so often the registration is focused on a partial image region of interest (ROI). For human monitoring applications, matching corresponding human ROIs in a pair of visible and thermal images is still challenging due to people various sizes, poses, clothes, distance to cameras, and different levels of occlusions. In the scene, people might have colorful/textured clothes that are visible in color images but not in thermal images. On the other hand, there might be

some textures observable in thermal images caused by different clothing characteristics (e.g. light clothes/warm clothes) and amount of emitted energy from different parts of the human body that are not visible in color image.

In this paper, the feature/measure comparison is carried out between several distribution-based Local Image Descriptors (LIDs) and classic stereo correspondence measures using two different matching approaches, and different interest regions on gray-scale thermal and visible images for registration purposes. The first matching approach is a simple WTA sliding window matching tested on single human ROI with no occlusion. This experiment is carried out to investigate the possible viability of the tested descriptors and measures. The second matching approach takes into account occlusions and the existence of depth discontinuities caused by multiple people in the scene. This method is applied using only the viable descriptors and measures on realistic close range human monitoring videos. Compared to our previous work (Torabi *et al.* (2011)), this paper performs exhaustive evaluation by adding several LIDs to the comparison, using different matching approaches, and using a new evaluation criterion. The ranking of top measures is the same as in (Torabi *et al.* (2011)).

In section 5.2, we discuss related works. In section 5.3, we present our tested image descriptors and stereo correspondence measures. Section 5.4 describes the details of our camera setup, our dataset, our tested scenarios, our matching approaches, and our evaluation criteria. In section 5.5, we present and discuss our experimental results. Finally, we conclude the paper in section 5.6.

5.2 Related Work

Performance evaluation has become an important task in computer vision due to the increasing number of feature detectors, descriptors, and comparison methods for a variety of applications (Christensen and Philips (2002)). In the context of matching and recognition using visible images, Li and Allinson (Li and Allinson (2008)) give a comprehensive survey of current local descriptors. Moreover, Mikolajczyk and Schmid have evaluated the performance of local descriptors (Mikolajczyk and Schmid (2005)). In the context of thermal-visible partial ROI matching, there is one work that gives a comparative analysis of multimodal registration approaches (Krotosky and Trivedi (2007)), but there is no work for comparing performance of different image descriptors and similarity measures.

In previous works, Mutual Information (MI) is the only similarity measure used in stereo thermal-visible human ROI registration (Krotosky and Trivedi (2007); Chen *et al.* (2003); Fookes *et al.* (2004)). Authors did not discuss the accuracy of MI compared to other similarity metrics. For human ROI matching, MI is not necessarily a reliable correspondence measure,

especially for close range videos. MI-based matching may fail when there is imperfect ROI segmentation, differently textured corresponding thermal and visible ROIs, partial occlusions caused by stationary objects, and multiple occluded people in the scene. Moreover, it is limited by the choice of the size of the matching window and it is not easy to find the best size. In this paper, we aim to study other image descriptors to be used as a similarity metric, and we compare them with MI. We aim at finding a descriptor or measure that is better than MI, if possible.

Image descriptors can be classified into three categories which are gradient (or texture)-based, shape-based, and color-based. In our context, color descriptors are not applicable since the pixel intensities are totally different between thermal and visible images (thermal image reflects temperature information of the imaged scene while visible image reflects color information of the imaged scene). However, shape-based or gradient-based descriptors might possible be applicable for thermal-visible human ROI matching task (a pair of thermal and visible images contain similar patterns and human image ROI layout). In recent years, LIDs have gained popularity and dominance in computer vision tasks. The main advantage of LIDs is that they capture the geometric information of the scene by dividing an image region into smaller image cells and by computing different characteristics of appearance or shape for each cell individually. Therefore, they are more distinctive, robust to occlusion, and slight variations in viewpoint compared to global image descriptors that describe a whole image or a whole image ROI using one vector or histogram, such as color histograms, color moments, and edge histograms. In fact, an image ROI can be described by a set of LIDs, therefore for matching two image ROIs some of the descriptor might be so similar between two ROIs while some other descriptors related to unsimilar parts of ROIs might be totally different. The most popular LID category describing shape and gradient is the distribution-based category. The distribution-based LIDs use histograms/vectors to represent the appearance or shape (Mikolajczyk and Schmid (2005)). They are computed either on a keypoint, such as Scale Invariant Feature Transform (SIFT) descriptor, or on a small image patch such as Local Self-Similarity (LSS) descriptor (unit of measurement is a small image patch rather than a pixel), and can be compared using simple $L1$ and $L2$ distances.

Among shape and pattern descriptors, we have selected Local Self-Similarity (LSS) and Binary Robust Independent Elementary Feature (BRIEF). LSS was proposed initially by Shechtman and Irani in (Shechtman and Irani (2007)) and applied to the problems of object categorization, image classification, pedestrian detection, and object detection (Walk *et al.* (2010); Yang *et al.* (2009a); Vedaldi *et al.* (2009)). BRIEF is a computationally fast descriptor that was recently proposed by Calonder *et al.* (Calonder *et al.* (2010)) and it was shown to outperform SURF for recognition tasks. For our comparison, two gradient-based descriptors

were also selected, that is SURF (Speeded Up Robust Features) and HOG (Histogram of Oriented Gradients). SURF was initially proposed by Bay *et al.* (Bay *et al.* (2006)) and it is a speeded up version of SIFT. HOG was recently proposed by (Dalal and Triggs (2005)) and is recognized as an efficient descriptor for human detection. We have also tested two classic similarity measures, which are Normalized Cross Correlation (NCC) and Mutual information (MI). NCC has been widely used for single modality image template matching and image registration (Sarvaiya *et al.* (2009)) and MI is a classic multimodal similarity measure that has been widely used in medical image registration (Pluim *et al.* (2003)). Egnal (Egnal (2000)) has shown that mutual information (MI) is a viable similarity metric for matching thermal and visible images. In our experiment, we used the Open Computer Vision Library (OpenCV) implementation of the tested LIDs.

5.3 Tested Descriptors and Measures

5.3.1 Distribution-based Descriptors

Local Self-Similarity (LSS)

Unlike most local image descriptors that represent the photogrammetric properties of images (colors or gradients), LSS represents an 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 a differently textured region as long as both regions 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 describes statistical co-occurrence of small image patch (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 patch centered at pixel p and all possible patches in a larger surrounding image region. SSD is normalized by the maximum value of the small image patch intensity variance and noise (a constant that corresponds to acceptable photometric variations in color or illumination). It is defined as

$$S_p(x, y) = exp(-\frac{SSD_p(x, y)}{max(var_{noise}, var_{patch})}). \tag{5.1}$$

Then, the correlation surface is transformed into a log-polar representation partitioned into e.g. 80 bins (20 angles and 4 radial intervals). The LSS descriptor is defined by selecting the maximal value of each bin that results in a descriptor with 80 entries.

Since the measurement unit of LSS is an image patch rather than a pixel, it can be customized 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. These values were selected experimentally. They are small enough for a local descriptor that participates in 1-D window matching of two sets of LSS descriptors. We compute the LSS descriptor for all the pixels inside the matching windows. In our application of LSS for window matching, we discard the non-informative descriptors prior to matching. Non-informative descriptors are the ones that do not contain any self-similarities (e. g. the center of a small image patch is salient) and the ones that contain high self-similarities (a homogenous region with a uniform texture/color). A descriptor is salient if all its bin's values are smaller than a threshold. The homogeneity is detected using the sparseness measure in (Hoyer and Dayan (2004)). The sparseness measure is defined as

$$sparseness(X) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1} \tag{5.2}$$

where n is the dimensionality of descriptor x (in our method 80). This function evaluates to unity if and only if x contains only a single non-zero component, and takes a value of zero if and only if all components are equal. Discarding non-informative descriptors is like an implicit segmentation or edge detection, which for window matching, increases the discriminative power of the LSS measure and avoids ambiguous matching. It is important to note that the remaining informative descriptors still form a denser collection compared to sparse interest points. Fig. 5.1 shows pixels having informative descriptors (white pixels) for a pair of thermal and visible images. The regions belonging to the human body boundaries and image patterns are the informative regions. This is obtained without any explicit edge detection or segmentation.

(a) (b)

Figure 5.1 Informative LSS descriptors. (a) Visible image and informative LSS descriptors (b) Thermal image and informative LSS descriptors.

Binary Robust Independent Elementary Features (BRIEF)

BRIEF is a fast and relatively accurate local image descriptor that was presented recently by Calonder *et al.* (Calonder *et al.* (2010)). It was shown that in terms of speed and recognition performance, it outperforms other computationally fast descriptors such as SURF. BRIEF is defined as a bit vector out of test responses, which are computed on smoothed image patches. For BRIEF definition, we used the same notations as used in (Calonder *et al.* (2010)). A test τ is defined on a patch p of size $S \times S$ as

$$\tau(p; x, y) = \begin{cases} 1 & \text{if } (p(x) < p(y)) \\ 0 & \text{otherwise} \end{cases} . \tag{5.3}$$

where $p(x)$ is pixel intensity in a p at position $x = (u, v)^T$ on the smoothed image patch. BRIEF describes the local texture around a point of interest using a binary code. Choosing a set of $n_d(x, y)$-location pairs uniquely defines a set of binary tests. The BRIEF descriptor is defined as

$$f_{nd}(P) = \sum_{1 \leq i \leq n_d} 2^{i-1} \tau(p; x_i, y_i). \tag{5.4}$$

In our experiment, we used $n_d = 256$ (BRIEF-32) as it is suggested in the original paper (Calonder *et al.* (2010)). For stereo matching purpose, similarly to LSS, we compute the BRIEF descriptor for all the pixels inside the matching windows.

Speeded Up Robust Features (SURF)

The SURF descriptor is a type of local histogram of image gradient descriptor that was previously proposed by Bay *et al.* (Bay *et al.* (2006)). SURF describes SIFT-like features using integral images and it is a speeded up version of SIFT that was initially proposed by Lowe (Lowe (2004)) and widely applied in many computer vision applications, such as object recognition, video tracking. SURF computes a distribution of Haar wavelet responses within the interest point neighborhood. In our experiment, only 64 descriptor dimensions are used reducing the time for feature computation and matching. For stereo matching purpose, similarly to LSS and BRIEF, we compute the SURF descriptor for all the pixels inside the matching windows.

Histogram of oriented gradients (HOG)

HOG is an image gradient descriptor that has been previously used for human detection (Dalal and Triggs (2005)). 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). In HOG computation, histograms are normalized. Therefore, it captures ROI layout/shape, as boundary edges (stronger edges) have greater impact in the computation of the descriptor. In this paper, we assess the viability of HOG descriptor to be used as a similarity feature in a multimodal dense stereo correspondence algorithm. In our experiment, the size of the cells is 8×8 and the size of the blocks is 16×16 as suggested in the original work (Dalal and Triggs (2005)). We compute the descriptor inside a matching window as a grid where the distance between the centers of two descriptor blocks is 8 pixels.

5.3.2 Similarity Measures

Normalized Cross Correlation (NCC)

NCC is a classic similarity measure that has been widely used for single modality image template matching and image registration (Sarvaiya *et al.* (2009)). NCC consists in a pixel-wise cross-correlation of two image regions normalized by the overall intensity difference. NCC is defined for two windows on a pair of images as

$$C(L,R) = \frac{\sum_{x,y}(I_l(x,y) - \bar{I}_l) \times (I_r(x,y) - \bar{I}_r)}{\sqrt{\sum_{x,y}(I_l(x,y) - \bar{I}_l)^2 \times \sum_{x,y}(I_r(x,y) - \bar{I}_r)^2}}, \tag{5.5}$$

where L and R represent a pair of matching windows, I_l and I_r are the image ROI inside two matching windows on a pair of thermal and visible images. $I_l(x,y)$ represents the pixel intensity at position (x,y) belonging to corresponding image ROI. This measure relies basically on similar intensity patterns.

Mutual Information (MI)

MI is a very popular similarity measure that has been widely used in multimodal image registration for different applications, including medical and video surveillance systems. MI computes the statistical co-occurrence of pixel-wise image patterns inside a window on a pair of images. MI is defined for two matching windows as

$$M(L,R) = \sum_l \sum_r P(l,r) log \frac{P(l,r)}{P(l)P(r)}, \tag{5.6}$$

where $P(l,r)$, is the joint probability mass function and $P(l)$ and $P(r)$ are the marginal probability functions. $P(l,r)$ is a normalized two-dimensional histogram of $g(l,r)$ (an N by N matrix) so that for each point, the quantized intensity levels l and r from the left and right matching ROIs (L and R) increment $g(l,r)$ by one. The probabilities $P(l,r)$ are then obtained by normalizing the histogram $g(l,r)$ by the sum of the joint histogram entries. The marginal probabilities $P(l)$ and $P(r)$ are then obtained by summing $P(l,r)$ over the grayscale or thermal intensities. The unit of measure of MI as a similarity metric is pixel-based which urges that the common patterns in thermal and visible images to be exactly identical for a contribution in MI computation.

5.4 Experimental Setup

5.4.1 Video Acquisition and Calibration

We used synchronized visible-thermal videos of a $5m \times 5m$ room at a fixed temperature of 24 $°C$ captured by stationary thermal and visible cameras with a 12 cm baseline. We used series of video frames of a relatively close range scene where different people with different poses and clothing are walking at different depths (between 2-5 meters) from the camera baseline. In order to simplify the stereo matching to a 1D search, we first calibrated the thermal and visible cameras, and then rectified the images using the intrinsic and extrinsic calibration parameters. We used the standard technique available in the camera calibration toolbox of MATLAB (Heikkila and Silven (1997)). For calibration, we placed a checkboard pattern in front of the cameras. Since in the thermal images, the checkboard pattern is not visible at room temperature ; we illuminated the scene using high intensity halogen bulbs placed behind the two cameras. This way, the dark squares absorb more energy and visually appear brighter than the white squares. Fig. 5.2 shows an example of our calibration images.

5.4.2 Experimental Scenarios

Our performance evaluation is done using the two scenarios described in the following.

Scenario 1

The first scenario is designed to study the efficiency of different LIDs and similarity measures for thermal-visible image registration with respect to the differences in thermal and visible image characteristics. In this study, we focus on matching corresponding image windows on the thermal and visible ROIs, where the corresponding windows in each image

(a) (b)

Figure 5.2 Calibration images : (a) visible image (b) thermal image.

might be differently textured or one textured and the other uniform. Windows centered at manually picked points are located inside visible human ROI rather than on regions belonging to occluded people at different depths. We used the sliding window matching (see section 5.4.3) to find the corresponding image window on the thermal image. The matching process was repeated using three rectangular window sizes of 10×130 (small), 20×130 (medium), and 40×130 (large) pixels. The heights of the windows are chosen as a maximum possible height of a person in our experimental videos. The manually picked points are selected on textured or textureless visible human ROI for relatively near targets (between 2 to 3 meters from the camera) and far targets (between 4 to 5 meters). Note that for close-range scene monitoring, the scale of targets considerably changes by walking one meter further away or toward the camera. Figure 5.3 (visible image) shows an example of manually picked point with its surrounding window. Our experiment is carried out using 10 challenging video frames where within each frame 10 points on visible human ROIs were manually selected (total : 100 points).

Scenario 2

The second scenario is designed specifically for thermal-visible human monitoring application and for evaluating the performance of only viable LIDs and similarity measures as determined after applying the first scenario. In this scenario, we used an experimental setup similar to (Krotosky and Trivedi (2007)). First, we extract foreground pixels related to human body ROIs using the background subtraction method proposed in (Shoushtarian and Bez (2005)). Note that the background subtraction is not perfect and ROIs might be partially misdetected or some regions might be falsely detected. Our manually selected points on the foreground visible image are either located on the individual or on the boundary between occluded people in the scene. The human ROIs are either textured or textureless for far and

close targets. In order to compute the disparity map, we use disparity voting (DV) matching as described in section 5.4.3 . The matching process was repeated using three rectangular window sizes of 10×130 (small), 20×130 (medium), and 40×130 (large) pixels centered at 10 manually picked point p on 20 selected thermal and visible image pairs (total : 200 points).

5.4.3 Stereo Matching Approaches

We used two matching approaches as described in the following.

Sliding Window Matching

For each thermal and visible pair of images, a window centered at a manually picked point on the human ROI at column j on the visible image is defined ($W_{l,j}$). Then, a 1D window matching search is done on the thermal image in order to find the corresponding window $W_{r,j+d}$ which minimizes a similarity distance SD. d is a disparity offset belonging to disparity interval set D. In our experiment, the size of D is the same size as the image width. Figure 5.3 illustrates the sliding window matching approach.

For the image descriptors, a normalized similarity distance $SD_{j,d}$ of a pair of image windows $W_{l,j}$ and $W_{r,j+d}$, is computed as

$$SD_{j,d} = \frac{\sum_{p_l,p_r} L(p_l, p_r)}{N}, \tag{5.7}$$

N is the number of corresponding elements p_l and p_r that are participating in the similarity distance computation. For LSS, L is the $L1$ distance of the descriptors of the corresponding pixels $p_l \in W_{l,j}$ and $p_r \in W_{r,j+d}$ that are informative. For SURF, L is the $L2$ distance of the feature vectors. For HOG and BRIEF, since each image window has only one descriptor, SD

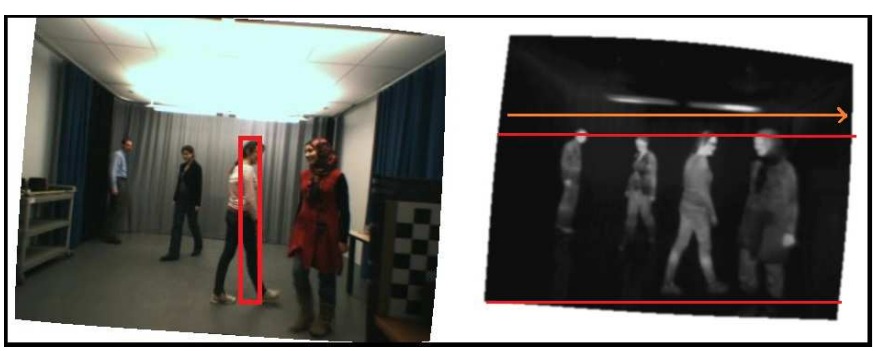


Figure 5.3 Thermal-visible 1-D sliding window matching.

is simply the $L2$ distance and the Hamming distance, respectively, of the two descriptors for the a pair of image windows $W_{l,j}$ and $W_{r,j+d}$. For MI and NCC, SD is defined as

$$SD_{j,d} = 1 - M(W_{l,j}, W_{r,j+d}), \tag{5.8}$$

where M is either MI defined in equation 5.6 or NCC defined in equation 5.5. And finally, the disparity associated to the matching windows that minimize SD is computed by

$$d_{min} = argmin_{d}(SD_{j,d}), d \in D. \tag{5.9}$$

Disparity Voting Matching

Disparity Voting (DV) was previously proposed by Krotosky and Trivedi in a MI-based registration framework (Krotosky and Trivedi (2007)). This algorithm is designed for registration of occluded or segmented ROIs that belong to moving people in a scene (Krotosky and Trivedi (2007)). It also handles the accurate registration of a merged region belonging to more than one people moving at different depth planes in the scene. In their method any ROI segmentation method with reasonable error is applicable.

For image window $W_{l,j}$ on the visible image, a disparity voting matrix DV_j of size (F, D) is built, where F is the number of foreground pixels inside $W_{l,j}$. This procedure is performed by shifting column by column $W_{l,j}$ on the visible, then doing the sliding window matching described in the previous section and adding a vote in $DV_j(p_l, d_{min})$ for all p_l inside image window $W_{l,j}$. For a foreground pixel inside an image window, the sum of the votes for a preset disparity levels is the same as the width of the image window. Finally, the disparity map DM_j which assigns a disparity to each pixels inside the $W_{l,j}$ is computed as,

$$DM_j(p_l) = argmax_{d}(D_j(p_l, d)), \tag{5.10}$$

Fig. 5.4 shows an example of DV matching using foreground visible and thermal images (more details about DV method in (Krotosky and Trivedi (2007))).

5.4.4 Evaluation Criteria

Precision and Recall

We used a criterion based on the number of correct matches of all pairs of tested images similar to the one used in (Gil *et al.* (2010); Mikolajczyk and Schmid (2005)). Precision and recall are defined as follows :

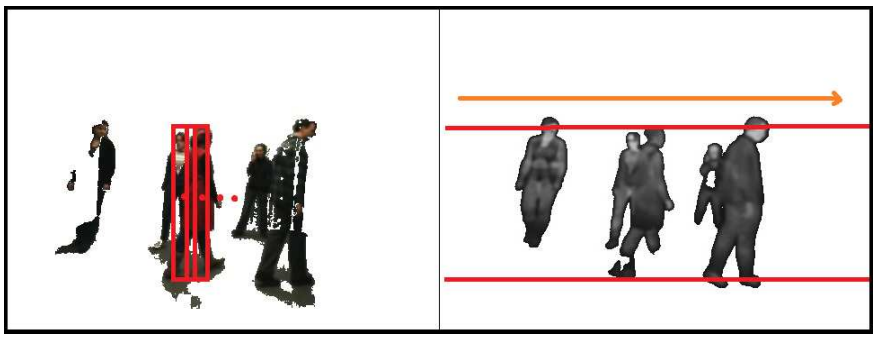


Figure 5.4 Thermal-visible DV matching on foreground pair of images.

$$precision = \frac{\#correctmatches}{\#matchesretrieved} \tag{5.11}$$

$$recall = \frac{\#correctmatches}{\#totalcorrespondences} \tag{5.12}$$

In our experiment, *correctmatches* is the number of matches with a disparity error smaller than 3 pixels with respect to ground-truth and with their SD value smaller than a threshold t (t varies between minimum possible values where *matchesretrieved* become one and maximum value where *matchesretrieved* become all the matched windows *totalcorrespondence*). *totalcorrespondence* is a fixed value that corresponds to the number of tested points (100 or 200). *matchesretrieved* is the number of matches with a SD value below threshold t. *matchesretrieved* varies from 1 to *totalcorrespondences*. In a *precision* versus *recall* curve, a feature with high *recall* value and low *precision* means that many correct matches as well as many false matches are retrieved. On the other hand, high *precision* value and low *recall* value means that most matches are correct but many others have been missed.

Power of Discrimination

To assess the reliability of matches, not only correct matches are important but also how discriminative are the matches. The power of discrimination verifies the distinctiveness of a match compared to its neighboring points on the SD versus disparity d curve. In order to evaluate the power of discrimination of LIDs, we used the similarity criterion from Section 5.4.3. We study the shape of SD versus a disparity range D for all the matches of all pair of images. A reliable match is located on an isolated minimum on the SD versus d curve and has a SD value much smaller than its neighboring points. In order to evaluate the isolation of the global minimum, the SD values computed by the sliding window matching (section 5.4.3) are first sorted increasingly and are transformed to the interval $[0, 1]$ named SD'.

Second, N is the number of values in SD' that are less than a pre-computed small threshold α, ignoring the global minimum. α has the same value for evaluating all descriptors and measures. Third, a quality measure s (the s value) is computed by dividing N by the size of the disparity range. So $s = 0$ corresponds to the most isolated minimum (best performance), and $s = 1$ corresponds to the least isolated minimum (flat/constant SD versus d curve). Finally, for each correspondence measure, a graph of Accumulated Frequencies (AF) of the s values of all matches is computed (In fact AF is the distribution of s values belonging to correct matches). Therefore, the correspondence measure for which AF reaches a higher value at a smaller s value is the more discriminative. Fig. 5.5 (a) and (b) show an example where the global minimum of SD is relatively isolated and N is 2 , and fig. 5.5 (c) and (d) show an example where minimum is not well isolated and N is 8, which results in higher value of s compared to the previous example. In our experiment, SD that is minimized with considerably smaller values compared to other points on the curve is considered accurate and distinctive for matching.

5.5 Experimental Results and discussion

5.5.1 Metric Viability Evaluation

In this section, the viability of LIDs including HOG, LSS, $SURF$, and $BRIEF$ and similarity measures NCC and MI, is evaluated as multimodal similarity metrics. We used the *scenario 1* described in section 5.4.2.

Figure 5.6 shows the precision-recall curves of the tested metrics for the three windows sizes described in 5.4.2. Overall for the three window sizes, the precision of LSS for different recall values is the highest and the last value of recall, which is equivalent to precision where $matchesretrieved$ is equal to $totalcorrespondences$ (this is obtained by varying the value of threshold t). For large and medium window sizes, MI is the second best. However for small window size, MI performance decreases dramatically, which shows its limitation to matching window sizes, which are required to be large enough to populate enough the joint probability histogram. The third best performance belongs to HOG. This metric has reasonably high values for large and medium window sizes, but it is not viable using small window size. SURF does not perform well for large and small windows since the recall and precision values are dramatically low. Similar results for BRIEF and NCC show that these metrics are not viable as thermal-visible similarity metric. These results correspond to what we may intuitively expect. Metrics that are more shape-based will perform better since the appearance in visible and infrared images is different. Thus, NCC, SURF, and BRIEF cannot perform well, as they rely heavily on intensity appearance. Although not shape-based, MI performs reasonably well

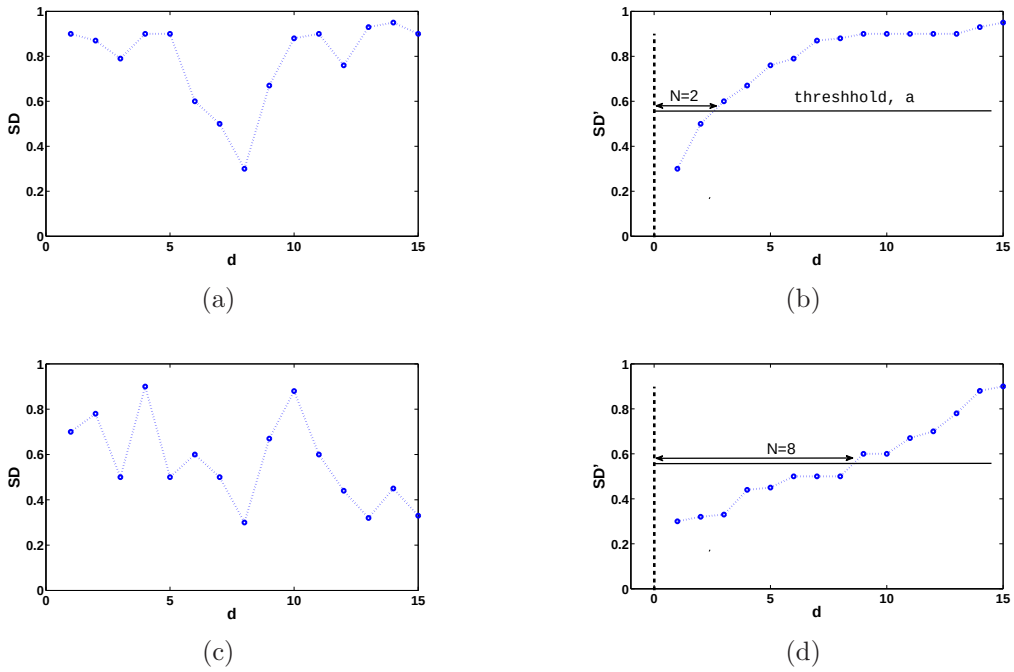
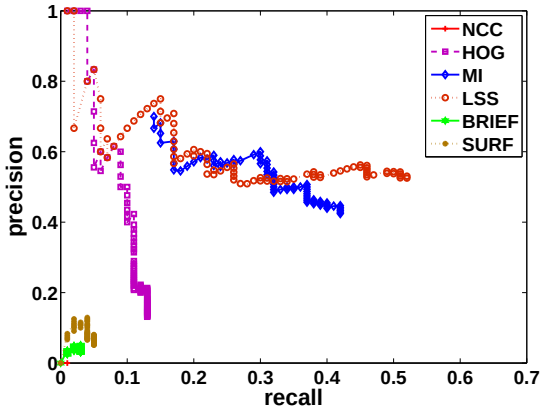


Figure 5.5 (a) and (c) Similarity distance SD versus disparity d curve. (b) and (d) Sorted SD curve.

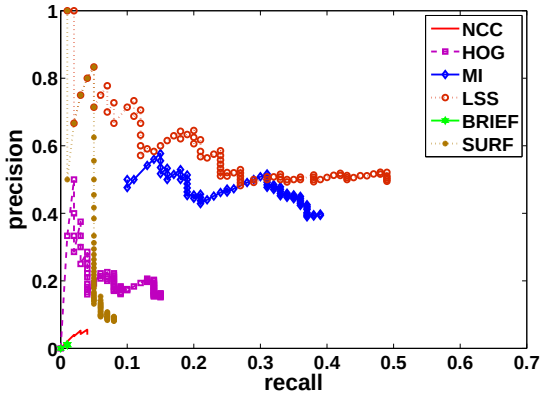
because it can match regions with different intensity appearance.

We also show the accumulated frequencies versus s values in Figure 5.7. For the three window sizes, LSS has the highest starting s values. BRIEF has the minimum discriminative power for all the three cases. For small window size, Figure 5.7 (c) shows that MI is not discriminative. For large and medium window sizes, all metric except BRIEF have reasonable discriminative power.

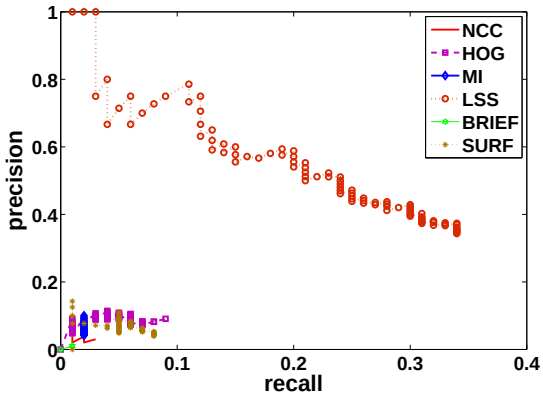
Table 5.1 shows the precision values (equation 5.11) in the case where $retrievedmatches$ is equal to $totalcorrespondences$ (maximum possible value), which in this experiment is 100. From best to worst, the metrics ranking is LSS, MI, HOG, SURF, BRIEF, and NCC. In order to be a viable thermal-visible metric, a good precision is a necessary condition. The power of discrimination is a second important complementary condition for consistent and stable performance. If a metric is not discriminant, the matches will not be reliable in the general case. Therefore based on our results, we picked the first three best metrics LSS, MI, and HOG as three viable multimodal similarity metrics for our purpose, which is multimodal human ROI registration for automatic human monitoring applications. Although MI and HOG are considered as viable, they are not viable for registering small objects, as the matching windows need to be relatively large. LSS performs well even with



(a)

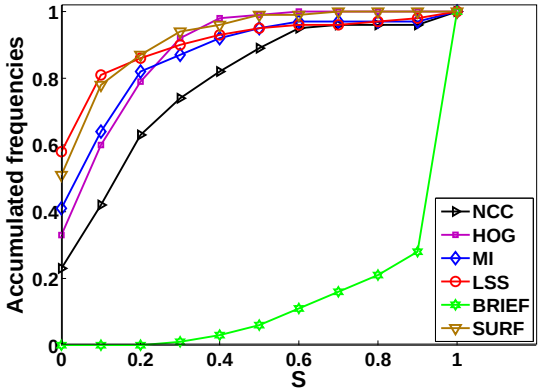


(b)

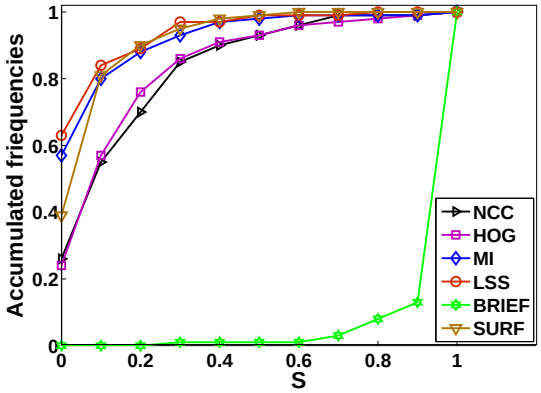


(c)

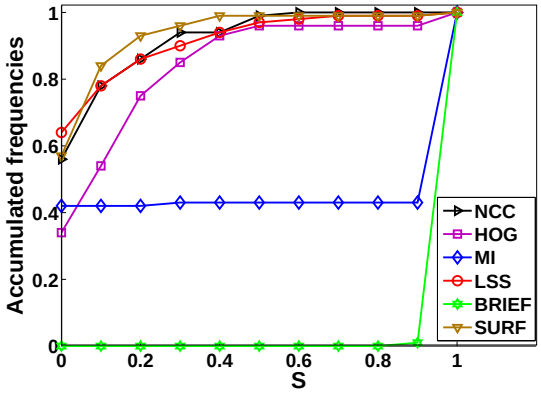
Figure 5.6 precision- recall curve : (a) large window (40×130) (b) medium window (20×130) (c) small window (10×130).



(a)



(b)



(c)

Figure 5.7 Accumulated frequencies vs. s value : (a)large window (40×130) (b) medium window (20×130) (C) small window (10×130).

smaller windows. Next, we will see if disparity voting, a more robust matching method, will change these conclusions.

Table 5.1 Matching precision of six tested LIDs/similarity measures for total 100 points on 10 pairs of selected thermal and visible images.

	precision(40×130)	precision (20×130)	precision (10×130)
NCC	0.01	0.04	0.03
HOG	0.13	0.15	0.09
MI	0.42	0.40	0.02
LSS	0.52	0.50	0.35
BRIEF	0.03	0.01	0.01
SURF	0.06	0.08	0.04

5.5.2 Comparison of Viable Metrics for Multi-modal Human ROI registration

In this section, we compare the performance of LSS, MI, and HOG-based stereo registration for automatic multimodal human monitoring applications. We used the *scenario 2* described in section 5.4.2. As it will be shown, using the segmented ROIs and the more robust registration method described in section 5.4.3 results in globally improved precision.

Table 5.2 shows the precision of LSS, MI, and HOG for *retrievedmatches* equal to *totalcorrespondences* (maximum possible value). Using a large window size for matching, LSS performs the best with 0.93 precision (MI is very close with 0.92). Using a small window size results in the lowest precisions for the three metrics. However, MI and HOG are more sensitive to window size compared to LSS. LSS has more consistent performance when varying the matching window sizes, which demonstrate the accuracy of this metric.

Based on our results for multimodal human ROI registration, overall LSS has the best performance, then, MI and HOG rank second and third, respectively. Using disparity voting increases the precisions for all three measures ; however the order of precisions remains the same. Thus, MI is not a bad choice for matching visible and infrared ROIs, but our results show that LSS is even a better choice. Indeed, although the photometric appearances of objects in visible and thermal image are different, their shapes tend to remain the same. Since LSS is designed to model shape, it is well suited for multimodal registration. Because MI is not based on the shape, it can fail when appearance changes unexpectedly in two matching windows, for example, in the case of heat-based textures that are not related and that do not co-occur with the visible modality local appearance. HOG has almost 75% reasonable precision using largest size window, however in general has low precision.

Table 5.2 Matching precision of three best LIDs/similarity measures for total 200 points on 20 pairs of selected thermal and visible images.

	precision(40×130)	precision (20×130)	precision (10×130)
HOG	0.74	0.33	0.14
MI	0.92	0.69	0.20
LSS	0.93	0.76	0.42

5.6 Conclusion

In this paper, we studied the performance of 6 local descriptors and measures for matching ROIs in visible and infrared images. Based on our evaluation metrics (precision-recall and power of discrimination), LSS and MI are viable similarity metrics for thermal-visible stereo registration. MI is a classic multimodal similarity measure and was known to be viable, but LSS was not previously considered for multimodal stereo matching. In fact, for the registration of human ROIs, we have shown that LSS is the most robust metric. It has reasonably good results for the three tested window sizes using realistic close range human monitoring scenarios, and outperforms MI.

CHAPTER 6

LOCAL SELF-SIMILARITY BASED REGISTRATION OF HUMAN ROIs IN PAIRS OF STEREO THERMAL-VISIBLE VIDEOS

Abstract

For several years, Mutual Information (MI) has been the classic similarity metric used in multimodal stereo matching approaches. The robustness of MI as a similarity metric is restricted by the MI window sizes. For unsupervised human monitoring applications, obtaining appropriate MI window sizes for the registration of multimodal pairs of images containing multiple people with various sizes, poses, distances to cameras, and different levels of occlusion is quite challenging. In this work, we apply local self-similarity (LSS) as a dense multimodal similarity metric and we evaluate theoretically and quantitatively its adequacy and strengths compared to MI in the context of visual surveillance systems. We also propose a LSS-based registration of thermal-visible stereo videos that consists of two steps of motion segment estimation and disparity assignment. We have assessed the performance of our method for realistic scenarios including several close range indoor thermal and visible video frames of a scene with multiple people at different depths and levels of occlusion. We demonstrate that our registration method outperforms a recent state-of-the-art MI-based stereo registration for human monitoring applications.

6.1 Introduction

In the recent years, there has been a growing interest of visual surveillance using multimodal sensors in both civilian and military applications. The combination of the thermal and visible modalities is one of the most used multimodal imagery system. The advantages of jointly using a thermal camera with a visible camera have been discussed comprehensively in (Zhu and Huang (2007); Collins *et al.* (2001); Socolinsky (2007)). For applications such as human monitoring and human behavior analysis, the joint use of two or more different imaging modalities provides richer information about the scene. For example, in challenging cases of visible modality, such as existing shadows on the ground, poor color information under low lighting conditions, or similarity of the human body/clothing with the background, once the images of the different modalities have been registered, better detection, tracking,

and analysis of human activities can be performed. The same applies for challenging thermal modality situations, such as where the human body or people clothing are at the same or at a temperature near the background or in a windy environment that changes temperature. Moreover, in high level of human activity analysis, the joint use of thermal and visible images enables us to more easily detect and segment the objects that one might hide in his clothes, or more easily segment the regions related to the object that people may carry.

In the literature, several methods including data fusion algorithms, background subtraction, multi-pedestrian tracking, and classification for thermal-visible surveillance videos have been proposed (Davis and Sharma (2007); Leykin (2007); Han and Bhanu (2007)). However, for close range videos, a fundamental and preliminary task associated with the joint use of thermal-visible data is accurately matching features of a pair of images captured by two different sensors. Due to the numerous differences in imaging characteristics of thermal and visible cameras, most methods used in single modality stereo matching are not applicable. Moreover, it is very difficult to find correspondence for an entire scene. For people monitoring applications, image region of interest (ROI) registration is one of the feasible approaches. In this approach, the problem of registration is simplified to aligning the pixels associated with the human body regions. However, matching corresponding regions belonging to a human body in a pair of visible and thermal images is still problematic. The corresponding pixels have different intensities and ROIs may have different patterns and textures due to the differences in imaging characteristics.

In previous works, MI is the only similarity measure used in dense multimodal stereo matching for human monitoring applications (Krotosky and Trivedi (2007); Chen *et al.* (2003); Fookes *et al.* (2004)). Fookes et al. proposed a MI-based window matching method that incorporates prior probabilities of the joint probability histogram of all the intensities in the stereo pair in the MI formulation (Fookes *et al.* (2004)). This matching method is less sensitive to MI window sizes. However, in their experiment, they only used negative and solarized images that have similar patterns in their ROI as opposed to thermal and visible images. Egnal has shown that mutual information (MI) is a viable similarity metric for matching disparate thermal and visible images (Egnal (2000)). Chen et al. proposed a MI-based registration method for pairs of thermal and visible images that matches boxes in the two images with the assumption that each box represents one single human (Chen *et al.* (2003)). In their method, occluded people that are merged into one ROI may not be accurately registered since a ROI may contain people within different depth planes. As a solution to improve registration of occluded people in a scene, Krotosky and Trivedi proposed a disparity voting (DV) matching approach (Krotosky and Trivedi (2007)). DV is performed by horizontally (column by column) sliding small width windows on rectified thermal and visible images,

computing MI for pairs of windows, and finally for each column, counting the number of votes associated to each disparity and assigning one disparity to each column based on a Winner Take All (WTA) approach. Their method can handle occlusion horizontally (two neighboring columns might be assigned to different disparities), but it cannot accurately register people with different height where a shorter person is in front of a taller one (vertical occlusion) since all pixels of a column inside a ROI are assigned to only one disparity.

In these papers, authors have not discussed the discriminative power and confidence of MI compared to other viable similarity metrics. Based on our experiments, in uncontrolled settings, where there are people with textured clothes, partial ROI misdetections, false detection, or occlusions, MI is unreliable for matching small width windows like the one proposed in (Krotosky and Trivedi (2007)). Moreover, MI-based matching fails often when the search range of window matching is relatively large. For MI matching, choosing the appropriate image window size is not straightforward due to the aforementioned difficulties. Also, there is always a trade-off between choosing larger windows for matching evidence, and smaller windows for the precision and details needed for an accurate registration.

In this work, we apply local self-similarity (LSS) to the problem of multimodal dense stereo matching for close range human monitoring applications. LSS has been proposed by Shechtman and Irani in (Shechtman and Irani (2007)) and has been previously applied to problems, such as object categorization, image classification, pedestrian detection, and object detection (Walk *et al.* (2010); Yang *et al.* (2009a); Vedaldi *et al.* (2009)). To the best of our knowledge, nobody has previously applied LSS as a thermal-visible dense stereo correspondence measure. LSS, similarly to MI, computes statistical co-occurrence of pixel intensities. However LSS, unlike MI, is firstly computed and extracted from an individual image as a descriptor and then compared between 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 a small image patch that captures more meaningful image patterns than individual pixels as used in MI computation. This property makes LSS a suitable measure for matching a textured region in one image with a uniformly colored region or differently textured region in another image as long as they have similar spatial layout (Shechtman and Irani (2007)). For thermal-visible human ROI registration, this property is advantageous since the human body might be differently textured, but the spatial layout (shape) is the most common visual information between thermal and visible corresponding ROIs. The algorithms presented in this manuscript are based on (Torabi and Bilodeau (2011)), but they are further developed with detailed analysis and new evaluations.

In section 6.2, we give a theoretical analysis between LSS and MI as dense multimodal correspondence measures and explain the advantages of LSS compared to MI by showing

some problematic matching examples. In section 6.3, we assess quantitatively the reliability and accuracy of MI and LSS as dense stereo similarity measures in various nearly close range challenging human monitoring scenarios. In section 6.4, we propose our LSS-based registration which accurately registers occluded people in different depths. Finally, in section 6.5, we compared qualitatively and quantitatively our multimodal LSS-based stereo registration method and a recent state-of-the-art multimodal MI-based stereo registration method.

6.2 Theoretical analysis of MI and LSS as similarity metrics for dense stereo matching

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

$$MI(L,R) = \sum_{l}\sum_{r} P(l,r) log \frac{P(l,r)}{P(l)P(r)}, \tag{6.1}$$

where $P(l,r)$, is the joint probability mass function and $P(l)$ and $P(r)$ are the marginal probability functions. $P(l,r)$ is a normalized two-dimensional histogram of $g(l,r)$ (an N by N matrix) so that for each point, the quantized intensity levels l and r from the left and right matching windows (L and R) increment $g(l,r)$ by one. The probabilities $P(l,r)$ are then obtained by normalizing the histogram $g(l,r)$ by the sum of the joint histogram entries. The marginal probabilities $P(l)$ and $P(r)$ are then obtained by summing $P(l,r)$ over the grayscale or thermal intensities. The unit of measure of MI as a similarity metric is pixel-based which urges that the common patterns in thermal and visible images to be exactly identical for a contribution in MI computation. In our application, MI computes the statistical co-occurrence of pixel-wise measures, such as patterns inside human body regions on pairs of thermal and visible images. Based on our experiments, MI has the following shortcomings for multimodal ROIs stereo matching tasks : 1) MI-based matching may fail to match corresponding thermal-visible ROIs with similar layout, but with different textures, 2) MI-based matching fails using small size image windows where the joint probability histogram is not sufficiently populated. Choosing the appropriate window size is not straightforward due to difficulties, such as target size changes and occlusions where two or more people are merged into one single ROI, and 3) MI-based stereo matching may fail due to a partial ROI misdetection or a falsely detected region caused by erroneous background subtraction in thermal and visible images.

LSS describes statistical co-occurrence of small image patch (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 patch centered at pixel p and all possible

patches in a larger surrounding image region. SSD is normalized by the maximum value of the small image patch intensity variance and noise (a constant that corresponds to acceptable photometric variations in color or illumination). It is defined as

$$S_p(x,y) = exp(-\frac{SSD_p(x,y)}{max(var_{noise}, var_{patch})}). \tag{6.2}$$

Then, the correlation surface is transformed into a log-polar representation partitioned into e.g. 80 bins (20 angles and 4 radial intervals). The LSS descriptor is defined by selecting the maximal value of each bin that results in a descriptor with 80 entries. LSS has two main advantages over MI as a correspondence measure : 1) LSS is computed separately as set of descriptors in one individual image and then it is compared in a matching process across a pair of images. This enables the detection of informative regions (regions containing informative descriptors described in next paragraph) inside human ROIs in the image and then using those regions for matching, 2) the measurement unit for LSS is a small image patch that contains more meaningful patterns compared to a pixel as used for MI computation. As it is described in Shechtman and Irani's work (Shechtman and Irani (2007)), this property makes LSS a suitable measure for matching textured region in one image with uniformly colored region or differently textured region in another image, as long as they have similar spatial layouts. Thus, for matching thermal and visible ROIs of people wearing clothes with different patterns, LSS-based matching should be more reliable than MI-based matching. In our application of LSS for window matching, before matching the two sets of descriptors in the thermal and visible images, we discard the 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 (a homogenous region with a uniform texture/color). A descriptor is salient (non-informative) if all its bins' values are smaller than a threshold. The homogeneity (which also cause a non-informative descriptor) is detected using the sparseness measure of (Hoyer and Dayan (2004)). The sparseness measure is defined as

$$sparseness(X) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1} \tag{6.3}$$

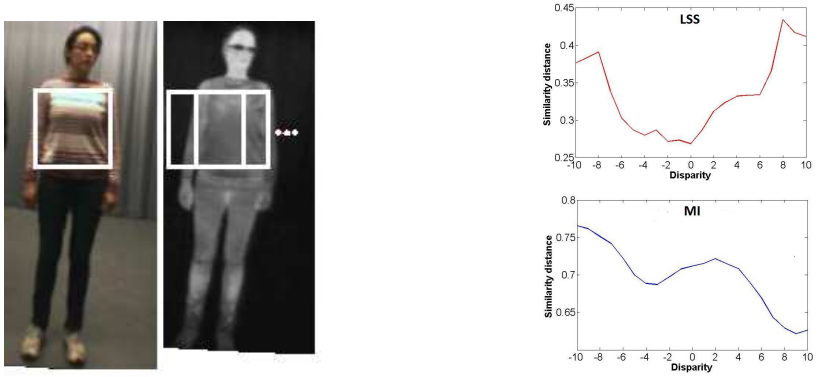
where n is the dimensionality of descriptor x (in our method 80). This function evaluates to unity if and only if x contains only a single non-zero component, and takes a value of zero if and only if all components are equal. Discarding non-informative descriptors is like an implicit segmentation or edge detection, which for window matching, increases the discriminative power of the LSS measure and avoids ambiguous matching. It is important to note that the remaining informative descriptors still form a denser collection compared to

sparse interest points. Fig. 6.1 shows pixels having informative descriptors (white pixels) for a pair of thermal and visible images. The regions belonging to the human body boundaries and image patterns are the informative regions. This is obtained without any explicit edge detection or segmentation.

We prepared three real world examples to illustrate the difficulties of multimodal human ROI matching and to show the advantages of LSS compared to MI. Matching is performed by computing the similarity distances of a fixed window on a region of the visible image with a sliding window on the thermal image within a disparity range of [-10, 10], and then choosing the disparity that minimizes the similarity distance. In order to simplify the search to 1D, the two images were rectified, and then manually aligned so that a disparity of 0 corresponds to a ground-truth alignment (more details about multimodal camera calibration in section 6.3.1). We defined the LSS-based similarity distance between two windows by the sum of the $L1$ distances of informative descriptors bounded in the thermal and visible windows, and the MI-based similarity distance as $1 - MI(L, R)$. Fig. 6.2 shows an example of matching a textured region in the visible image with a corresponding uniform region in the thermal image. Fig. 6.2 (b) shows the similarity distance results for both MI and LSS over the disparity range. For LSS, the similarity distance is correctly minimized at disparity 0. However for MI, the similarity distance is minimized incorrectly. This illustrate that MI is not a robust similarity metric for matching a textured region and a uniform region when there are not many similar patterns. Fig. 6.3 shows an example of matching windows of sizes 20×20 and 50×50 pixels on a head region. Fig. 6.3 (b) shows that MI is not a robust measure for matching 20×20 thermal-visible windows. However, using larger window of size 50×50 pixels containing more similar patterns and more similar spatial layout, MI-based similarity distance is correctly minimized at disparity 0. For this example, LSS-based similarity distance is correctly minimized at disparity 0 for both matching window sizes which demonstrate the

(a) (b)

Figure 6.1 Informative LSS descriptors. (a) Visible and informative LSS descriptors images (b) Thermal and informative LSS descriptors images.



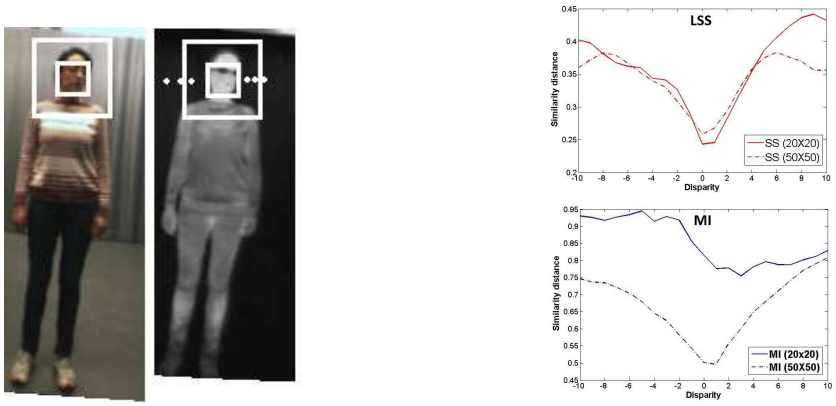
(a) (b)

Figure 6.2 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].

robustness of this measure for matching small window sizes. Fig. 6.4 shows an example of matching thermal-visible windows on regions with dramatic partial ROI misdetection using matching window sizes of 20×170 and 60×170 pixels. In the visible image, due to the color similarity of the ROI and the background, some parts of the body region are not detected. Fig. 6.4 (b) shows that MI fails to find the correct disparity offset with both window sizes. However, LSS find the correct disparity which illustrates the robustness of this measure for partial ROI misdetection.

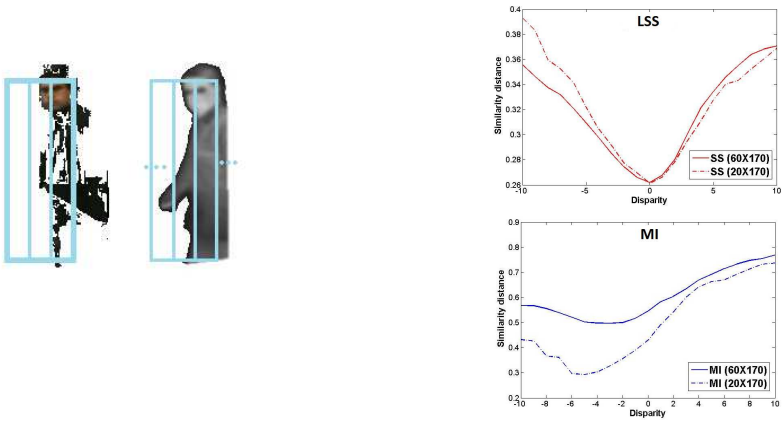
6.3 Evaluation of MI and LSS as similarity metrics for dense stereo matching

The goal of our evaluation is to assess the robustness and reliability of MI and LSS similarity measures in challenging scenarios, where, for instance, the human body ROIs contain different patterns in thermal and visible images. We also aim to examine the effect of matching window sizes on each similarity measure. The problems of erroneous foreground segmentation and occlusion are studied in section 6.4 using an appropriate matching approach. For our evaluation, we define a window centered around a manually picked point on a human ROI in the visible image and perform a simple 1D window search on the thermal image where the corresponding windows (best match) are computed based on a winner take all (WTA) approach. The simplified 1D search for correspondence matching is feasible using our multimodal image calibration described in the following subsection.



(a) (b)

Figure 6.3 Matching corresponding regions of visible and thermal within image windows of size 20×20 and 50×50 pixels. (a) Aligned visible and thermal images, (b) Similarity distances of LSS and MI for disparity interval of [-10,10].



(a) (b)

Figure 6.4 Matching corresponding foreground pixels within 20×170 and 60×170 pixels windows in visible and thermal pair of images (a) Aligned visible and thermal images, (b) Similarity distances of LSS and MI for disparity interval of [-10,10].

(a) (b)

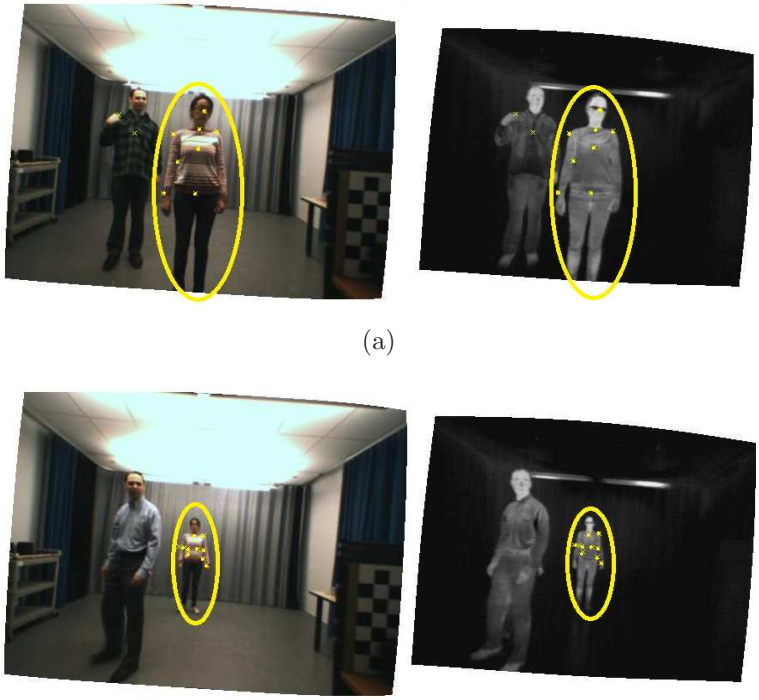
Figure 6.5 Calibrating images : (a) Visible image and (b) Thermal image.

6.3.1 Experimental setup

We used synchronized visible-thermal videos of $5m \times 5m$ room at a fixed temperature of 24 $°C$ captured by stationary thermal and visible cameras with a 12 cm baseline. In order to simplify the 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 (Heikkila and Silven (1997)). For calibration, we placed a checkboard pattern in front of cameras. Since in the thermal images, the checkboard pattern is not visible at the room temperature, we illuminated the scene using high intensity halogen bulbs placed behind the two cameras. In this way, the dark squares of the checkboard absorb more energy and checks visually appear brighter. Fig. 6.5 shows an example of our calibrating images. After calibration, to test MI and LSS, we used series of video frames of a close range scene where different people with different poses and clothing are walking at different depths (between 2-5 meters) from the camera baseline. We defined four experimental scenarios based on the position of manually selected window on the visible image. The windows for each scenario are selected manually by a human visual decision. The scenarios are

- *TexturedNear* : Window located on a textured human body ROI of a target relatively close to the camera.
- *TexturedFar* : Window located on a textured human body ROI of a target relatively far from the camera.
- *TexturelessNear* : Window located on a textureless human body ROI of a target relatively close to the camera.
- *TexturelessFar* : Window located on a textureless human body ROI of a target relatively far from the camera.

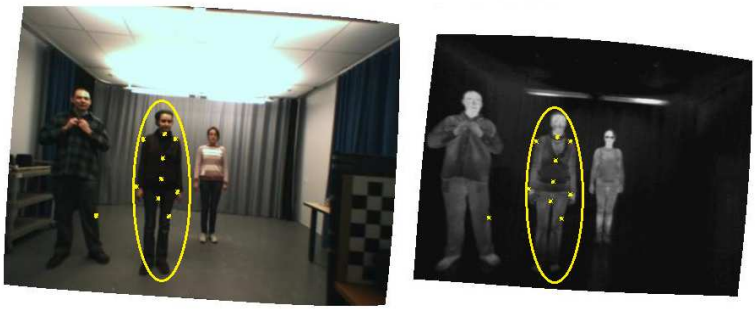
Figure 6.6 Thermal-visible 1-D matching process.



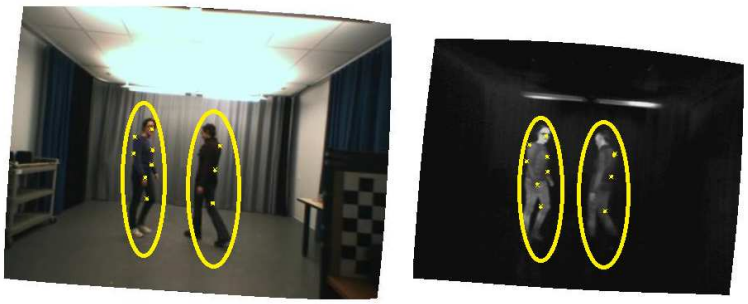
(a)

(b)

Figure 6.7 Examples of pairs of thermal and visible images for textured scenarios with selected points : (a) *TexturedNear*, (b) *TexturedFar*.



(a)



(b)

Figure 6.8 Examples of pairs of thermal and visible images for textureless scenarios with selected points : (a) *TexturelessNear*, (b) *TexturelessFar*.

Note that 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 targets considerably changes by walking one meter further away or toward the camera. Fig. 6.7 and 6.8 show samples of videos frames for the four scenarios. For each scenario, 5 challenging video frames were selected and within each frame, 10 points on human body ROIs were manually selected.

6.3.2 Dense correspondence matching

For each thermal and visible pair of images, a window centered at the manually picked point on human ROI column at j on the visible image is defined ($W_{r,j}$). Then, a 1D window matching search is done on the thermal image in order to find the best corresponding window $W_{r,j+d}$, where d is a disparity offset belonging to disparity interval set D. In our experiment, the size of D is the size of image width. Figure 6.6 shows our matching process. The best match on the thermal image is the one with the smallest Similarity Distance (SD), as explained in the following paragraph.

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

$$SD_{j,d} = \frac{\sum_{p_l,p_r} L1_{l,r}(p_l, p_r)}{N}, \tag{6.4}$$

where N is the number of corresponding pixels (N is smaller than number of foreground pixels ; however, it is still a large proportion of foreground pixels since the informative descriptors are dense) p_l and p_r contributing in the similarity distance computation and d is the disparity offset. Then $L1_{l,r}$ is computed as

$$L1_{l,r}(p_l, p_r) = \sum_{k=1}^{80} |d_{p_l}(k) - d_{p_r}(k)| \tag{6.5}$$

where 80 is the number of local self-similarity descriptor bins. d_{p_l} and d_{p_r} are LSS descriptors of p_l and p_r respectively. For MI, SD is defined as

$$SD_{j,d} = 1 - MI(W_{l,j}, W_{r,j+d}), \tag{6.6}$$

where MI is the mutual information defined in equation 6.1. And finally the best disparity

associated to best matching windows is computed by

$$d_{min} = argmin_{d}(SD_{j,d}), d \in D. \tag{6.7}$$

The matching process was repeated for each point p on the visible image, with three rectangular window sizes of 10×130 (small), 20×130 (medium), and 40×130 (large) pixels centered at pixel p. The heights of the windows are chosen as a maximum possible height of a person in our experimental videos.

6.3.3 Evaluation measures

In our evaluation, we compute the percentage of erroneous matches and the confidence of the good matches (discriminative power). Note that our matching method is based on a WTA approach, therefore the confidence and reliability of a good match is important information.

- *Matching error*

  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. The disparity error for pixel p is simply the Euclidean distance between p' and q, where q is the center of the best corresponding window computed by our matching process. The disparity error is computed for all the tested points. Then, the number of points that have disparity errors of more than 3 pixels (> 3) is counted and considered as the number of bad matches BM. We accept an error of up to 3 pixels to account for small errors in the manual ground-truth selection (Note that image size is 480×360 pixels).

- *Discriminative power*

  For all the good matches of each tested scenario (matching error $<= 3$), we assess the discriminative power of LSS and MI by studying the shape of the SD curve computed along the disparity range $D = [q - 20 : q + 20]$, where q is the position of the global minimum (best match). We applied the same measure as in (Mayoral and Aurnhammer (2004)). Recall that SD is the similarity distance as defined in section 6.3.2. A reliable good match is located on an isolated minimum on the SD curve and has a SD value much smaller than its neighboring points. In order to evaluate the isolation of the global minimum on the SD curve, the SD values computed by the matching process are first sorted increasingly and are transformed to the interval $[0, 1]$ named SD'. Second, N is computed by counting the number of values in SD' that are less than a pre-computed small threshold α, ignoring the global minimum (See more details in (Mayoral and Aurnhammer (2004))). α has the same value for evaluating both MI and LSS. Third, a

quality measure s (the s value) is computed by dividing N by the size of the disparity range. So $s = 0$ corresponds to the most isolated minimum (best performance), and $s = 1$ corresponds to the least isolated minimum. Finally, for each correspondence measure, a graph of Accumulated Frequencies (AF) of the s values of all good matches is computed. Therefore, the correspondence measure for which AF reaches a higher value at a smaller s value is the most discriminative.

6.3.4 Results

Table 6.1 shows the percentage of bad matches for MI and LSS. Results show that a window size of 10×130 results in a relatively poor matching performance for both LSS and MI. For all the scenarios and both measures, using a window size of 40×130 pixels, results in improved performance compared to matching using small and medium window sizes. The reason is that the large window size is about the same width as a human ROI in our experimental images, and includes more of the human body layout, which is the main similar information between thermal and visible human body ROIs. Table 6.1 also shows that for far scenarios, LSS and MI perform quite similarly as a similarity measure. However for near scenarios, specifically *TexturedNear* scenario, where the textures are more noticeable inside the human body ROIs , LSS has fewer matching errors compared to MI for both 20×130 and 40×130 matching window sizes. Also for *TexturelessNear* scenario, LSS performs better than MI since even if the human body ROI is textureless in visible, the corresponding region might be textured in thermal image. Fig.6.9 shows the AF graph of the *Textured* scenarios and Fig.6.10 shows the AF graph of the *Textureless* scenarios using a window size of 40×130. In the graphs, the s value where AF reaches 1 means all the good matches of a tested scenario have a s value between $[0, s]$. All four graphs show that LSS reaches a larger AF values at

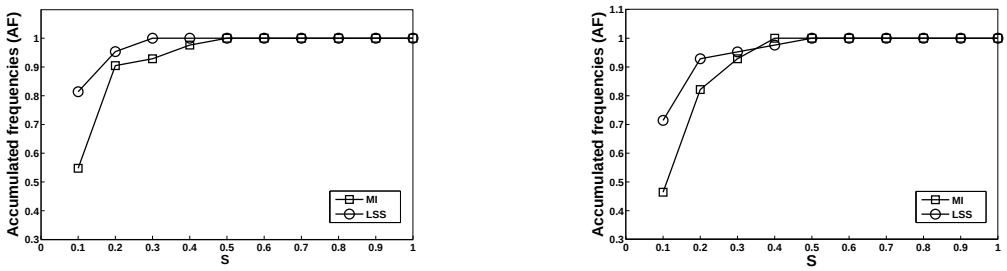


Figure 6.9 Accumulated frequencies (AF) using window size of 40×130 : (a) *TexturedFar*, (b) *TexturedNear*.

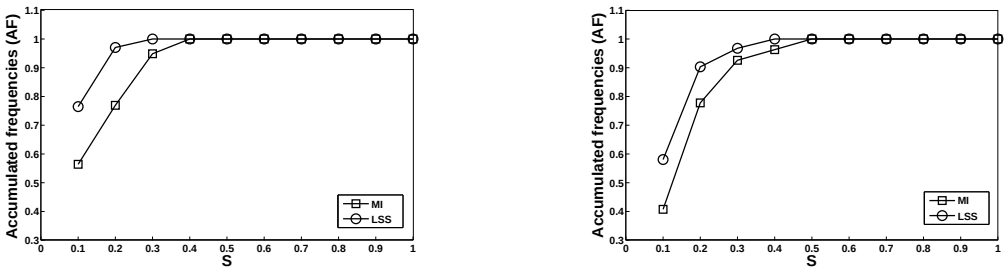
Table 6.1 Quantitative matching results of 50 points. (BM %) is the percentage of bad matches. M : Metric, WS : Window Size, TF : TexturedFar, TFL : TexturelessFar, TN : TexturedNear, and TLN : TexturelessNear

M	WS	TF (BM %)	TLF (BM %)	TN (BM %)	TLN (BM %)
MI	10×130	58	64	78	80
LSS		32	54	56	50
MI	20×130	20	46	70	74
LSS		22	46	20	54
MI	40×130	16	22	44	46
LSS		14	32	16	38

smaller s values compared to MI. This means that for LSS, the number of good matches with high confidence (high discriminative power) is larger compared to MI. Overall, from these results, we conclude that LSS is more robust as a multimodal similarity measure compared to MI for matching regions textured differently as long as they have similar layouts such as human body ROIs. This will be furthermore demonstrated in a practical application in section 6.5.

6.4 LSS-based multimodal ROI registration

In this section, we describe our novel multimodal ROI registration method using LSS. For a pair of thermal and visible video frames, our goal is to register the ROIs belonging to moving people in a scene in which they may be temporary stationary for few frames. Our method addresses registration of multiple people merged into one ROI with different levels of occlusion and with partially erroneous foreground segmentation for realistic thermal-visible videos of



(a) (b)

Figure 6.10 Accumulated frequencies (AF) using window size of 40×130 : (a) *TexturelessFar*, (b) *TexturelessNear*.

a close range scene. We assume that each person at each instant lies approximately within one depth plane in the scene. Therefore, we propose that a natural way for estimating depth planes related to multiple moving people is by applying motion segmentation on foreground pixels with the assumption that each motion segment belongs to one person in the scene, but more than one motion segment may belong to a person.

We define the multimodal image registration as multiple labeling sub-problems. Then, we use the disparity voting matching approach to register each individual motion segment rather than a whole foreground blob. Let MS be the set of motion segments belonging to moving people in the scene, and D be a set of labels corresponding to disparities. Our registration method assigns a label $d_k \in D$ in the interval $[d_{min}, ..., d_{max}]$ to each pixel of a motion segment $ms_i \in MS$. Thus, our registration method has two main parts : 1) motion segmentation that divides the registration problem as multiple labeling sub-problems and 2) disparity assignment which assigns disparity to each segment. The two parts of our method are described in the subsequent sections.

6.4.1 Motion segmentation

Our motion segmentation has three steps. Firstly, we extract foreground pixels using the background subtraction method proposed in (Shoushtarian and Bez (2005)). Any background subtraction method with a reasonable amount of error is applicable. Secondly, we compute the motion vector field for foreground pixels using an optical flow method based on block-matching (Ogale and Aloimonos (2007)). To speed up the process, the optical flow is only computed for regions inside the bounding boxes of the union of the foreground masks of two consecutive frames $t-1$ and t, instead of the whole image. Thirdly, we apply the mean-shift segmentation method proposed in (Comaniciu and Meer (1999)) for segmenting the motion vector fields computed in the previous step and computing a mean velocity vector for computed segments. Mean-shift segmentation is applied on (2+2) feature point dimensions, where two dimensions are related to spatial dimensions (horizontal and vertical directions) and the two others are related to the two motion vector components in x and y directions. Applying motion segmentation on ROIs results in a set of motion segments S defined as

$$S = \{sm_1, .., sm_i, .., sm_m\}\,. \tag{6.8}$$

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

$$\hat{m}_i = \frac{\sum_{p \in sm_i} m(p)}{|sm_i|}, \tag{6.9}$$

where $m(p)$ is the motion vector of pixel p. Figure 6.11 shows the motion segmentation

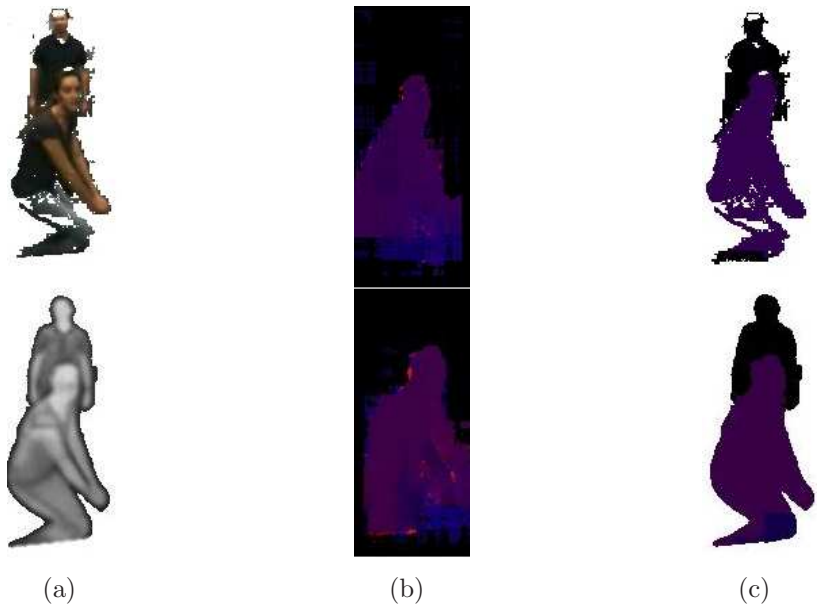


Figure 6.11 (a) Visible and thermal foreground images, (b) motion field vectors, and (c) motion segmentation results (depth segments).

results of one temporary stationary and one moving occluded people. In this figure, motion vectors are visualized by a mapping to HSV color space. Applying motion segmentation on foreground pixels enables us to determine also a depth segment associated to temporary stationary person for which its mean velocity vector is zero. Since in most indoor videos, the motion segmentation of thermal images are more accurate compared to visible images due to less partial ROI misdetection error, we perform motion segmentation for thermal images and we register the thermal motion segments on visible foreground images. However, it could also be done the opposite way.

6.4.2 Disparity assignment

At this step, we assign disparity to each motion segment individually. We use a disparity voting matching approach similar to the one that was previously proposed by Krotosky and Trivedi (Krotosky and Trivedi (2007)). DV matching assigns one single disparity to all the pixels of a column of matching regions. However, different disparities can be assigned to two neighboring columns. Krotosky and Trivedi DV method uses MI as similarity metric and is performed on whole foreground blobs. Their method is able to resolve the horizontal part of an occlusion, but fails to assign correct disparity for the vertical part of an occlusion (in this case, the pixels of a column for a region associated to vertically occluded people

should be assigned to different disparity) (see fig. 6.13). To solve this problem, we propose performing DV on each motion segment, which increase the probability of processing each person individually.. Moreover, based on our previous experiments, we use the informative LSS descriptors as similarity measure.

LSS-based DV algorithm -For each $sm_i \in S$, we build a disparity voting matrix of DV_i of size $(N, d_{max} - d_1 + 1)$ where N is the number of pixels of sm_i and $[d_1 - d_{max}]$ is a preset disparity range. This procedure is performed by shifting column by column $W_{l,j}$ on the reference image for all the columns $j \in s_i$, then doing window matching, the same as we previously described in 6.3.2. Then, for each d_{min} computed by window matching, a vote is added to $DV_i(p_l, d_{min})$ for all $p_l \in (W_{l,j} \cap s_i)$. Since the width of windows are m pixels wide, we have m votes for each pixel belonging to s_i. Finally, the disparity map DM_i is computed as,

$$DM_i(p_l) = argmax_{_d}(DV_i(p_l, d)), \tag{6.10}$$

6.5 Experimental validation and discussion

We have assessed our registration method with two videos of up to 5 people with different clothing, various poses, distances to cameras, and with different level of occlusions. In these experiments, we used the same experimental setup as described previously in section 6.3.1. The first test video was captured during summer with people have lighter clothes on and with a fair amount of textures inside human ROIs in thermal and visible images. The background subtraction errors were mostly misdetection errors. Our second test video was captured during winter with people wearing winter clothes, which results in many textures inside human body ROIs, specifically in the thermal images. The background subtraction results in our second video include both misdetection errors and falsely detected region as foreground. Our disparity range was [5,50] pixels. Fig. 6.12 illustrates successful registrations with our method in the winter video for three frames of people in different levels of occlusions.

6.5.1 Comparative evaluation of our matching and DV matching algorithm

In order to demonstrate the accuracy improvement of our method compared to the state-of-the-art disparity voting algorithm (DV) in (Krotosky and Trivedi (2007)) in handling occlusions, we quantitatively compared our disparity results using motion segmentation and the results of DV using for both LSS as similarity measure. We generated ground-truth disparities by manually segmenting and registering regions of foreground for each frame. Fig. 6.13 illustrates the comparison with ground-truth. Column (a) ground-truth disparity, column

Figure 6.12 Registration results of foregrounds using imperfect background subtraction with false positive and false negative errors.

(b) disparity estimation of DV matching using LSS as similarity measure (LSS+DV), column (c) disparity estimation of our proposed method (LSS+MS+DV), and column (d) illustrate the associated sums of disparity errors. Results in the first and second rows illustrate cases when two people in two different depths in the scene are in occlusion. LSS+DV method fails to assign correct different disparities to the columns containing pixels related to more than one individual since based on a WTA approach, a single disparity is assigned to all the pixels of each column. However, LSS+MS+DV succeeds in assigning accurately different disparities to the two human body ROIs since the DV was applied to each motion segment individually. Accordingly, in fig. 6.13 (d), the first and second rows correspond to the sum of disparity errors of the columns corresponding to two occluded people is much higher for LSS+DV method compared to LSS+MS+DV method.

To register merged objects in a single region, DV makes no assumptions about the assignment of pixels to individual objects and assigns a single disparity to each column inside a ROI based on a maximization of the number of votes. In their matching approach, if a column of pixels belongs to different objects at different depth in the scene, the vote only goes for one of them based on WTA approach. However, in our registration method, motion segmentation gives a reasonable estimate of moving regions belonging to people in the scene, and applying the DV matching on each motion segment gives more accurate results since it is less probable that pixels in one column belongs to more than one object. Therefore, in the worst case, even with erroneous motion segmentation, our method will have at minimum the same accuracy as the DV algorithm.

Fig. 6.13 last row is related to multiple occluded people. Although LSS+MS+DV registration results are not perfect because few small motion segments resulting from over segmentation were not matched correctly, still the results are more accurate than LSS+DV registration results . Accordingly, in Fig. 6.13 (d), last row, there are higher sums of disparity

error for columns related to vertical occlusion for LSS+DV compared to LSS+MS+DV. However, it is noticeable that in some columns, LSS+MS+DV has slightly higher errors caused by small motion segments misalignment.

Fig. 6.14 illustrates registration results with (LSS+MS+DV) and without motion segmentation (LS+DV), and using LSS as similarity measure. It is observable, that for LSS+DV method, the object misalignments happen where there are vertical occlusions while our method performs accurately.

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

In order to demonstrate the improvement of our LSS-based registration method (LSS+MS+DV) compared to the state-of-the-art MI-based registration method (MI+DV) proposed by Krotosky and Trivedi (Krotosky and Trivedi (2007)), we qualitatively and quantitatively compared the two methods. Fig. 6.16 illustrates four examples of the disparity computation and the image registration results obtained using the two methods for our summer video. Note that our results are more accurate, especially for occlusions. Fig. 6.17 illustrates four examples of our winter video. Note that MI+DV results are dramatically poorer. These results demonstrate that for videos where there are falsely detected region as foreground and high differences of patterns inside human body ROIs, MI is not a reliable similarity measure. Oppositely, LSS performs very well, except for few misalignments which occur for very small motion segments.

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

$$E = 1 - \frac{N_{v \cap t}}{N_t}, \tag{6.11}$$

where $N_{v \cap t}$ is the number of overlapping aligned thermal foreground pixels on visible foreground pixels and N_t is the number of thermal foreground pixels. The best performance with zero overlapping error is when all the thermal pixels on the reference image have corresponding visible pixels on the second image. Note that our registration results are aligned thermal on visible images. This evaluation measure includes the background subtraction errors and also ignores misaligned thermal pixels which have falsely matched visible foreground pixels. However, since for both methods the background subtraction errors are included in the overlapping error, the differences between the two methods errors are still a good indicator for comparing overall registration accuracies for a large numbers of frames. Fig 6.15 illustrates the overlapping error using our LSS+MS+DV and MI+DV (Krotosky and Trivedi (2007))

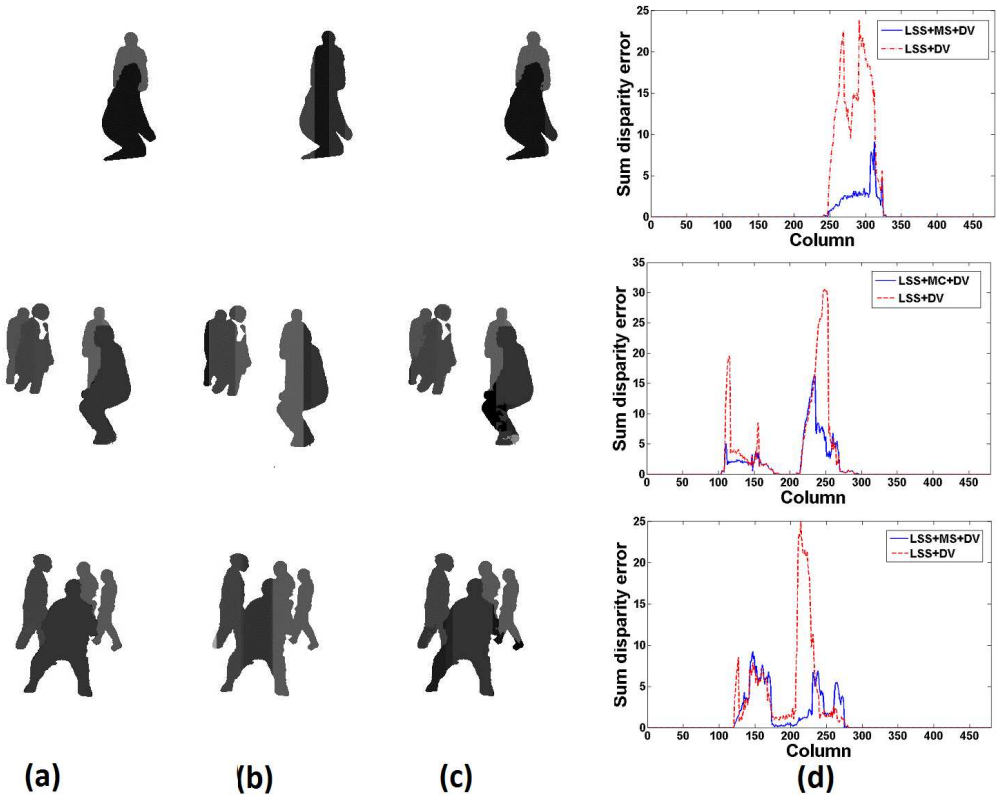


Figure 6.13 Comparison of LSS-based DV method and our proposed disparity assignment method(a) Ground truth, (b) LSS+DV, (c) LSS+MS+DV, and (d) Sum disparity errors over columns.

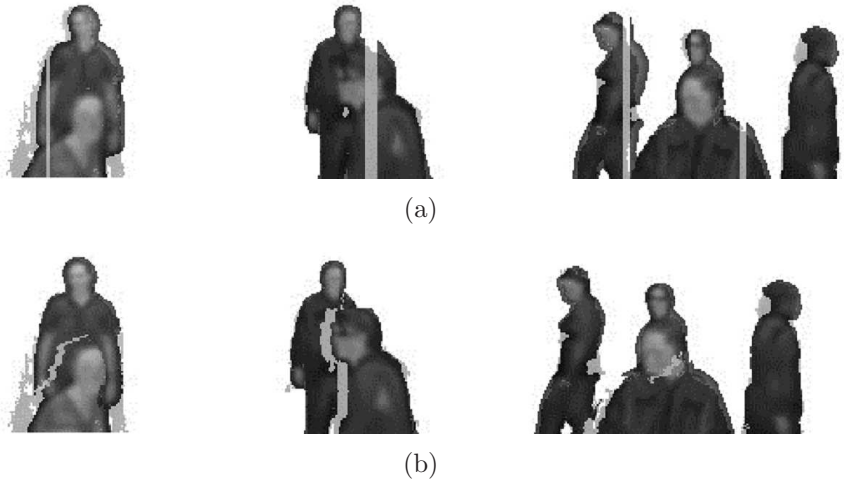
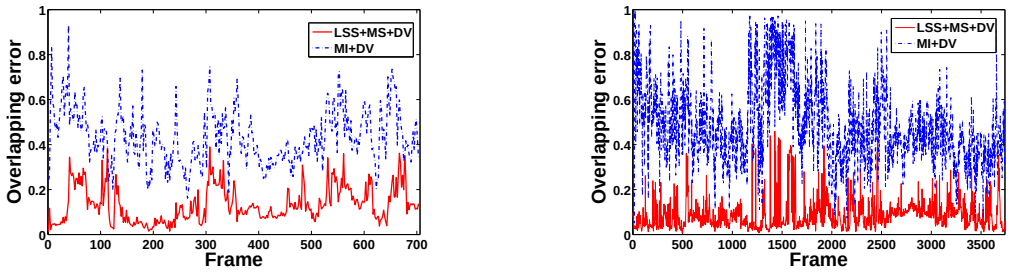


Figure 6.14 Comparison of LSS+DV and LSS+MS+DV detailed registration : (a) LSS+DV registration and (b) LSS+MS+DV registration.



(a) (b)

Figure 6.15 Overlapping error : (a) Summer video (702 frames), (b) Winter video (3740 frames)

methods for summer and winter videos. The difference of mean overlapping error of the two methods over all frames for the summer video is 0.3007 and for the winter video, it is 0.4049. These results demonstrate that our method performs much better and more consistently compared to MI+DV (Krotosky and Trivedi (2007)) method, especially for winter video in accordance with our qualitative results and previous discussions.

6.6 Conclusion

In this paper, we applied LSS as a multimodal dense stereo correspondence measure and shown its advantages compared to MI, the most commonly used multimodal stereo correspondence measure in the state-of-the-art for human monitoring applications. We also proposed an LSS-based registration method, which addresses the accurate registration of regions associated to occluded people in different depths in the scene. In our results, we have shown the improvement of our registration method over the DV method proposed by (Krotosky and Trivedi (2007)). Moreover, we have shown that our method significantly outperforms the state-of-the-art MI-based registration method in (Krotosky and Trivedi (2007)). As future direction for this work, we are working on improving the motion segmentation results to obtain more accurate segments and to avoid over segmentation.

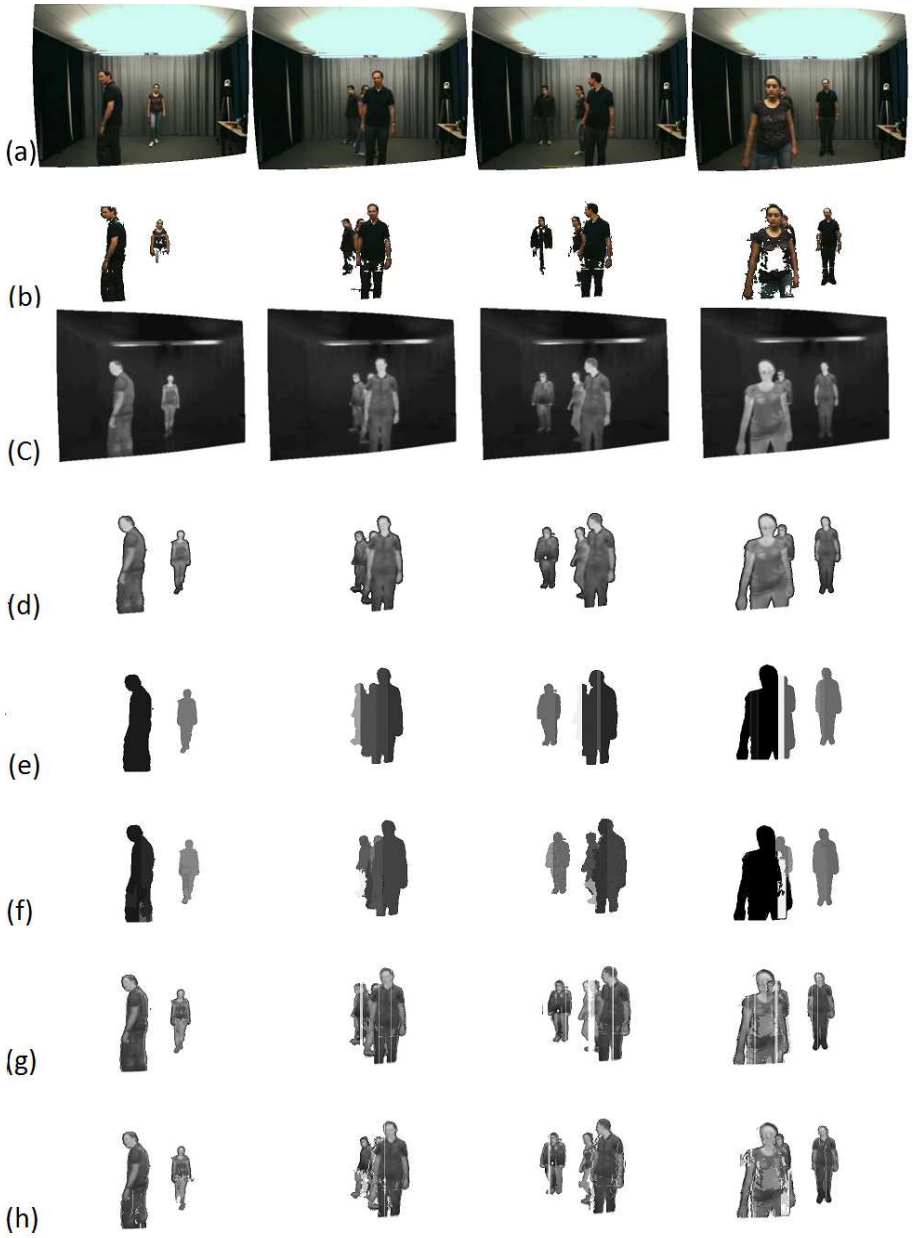


Figure 6.16 Comparison of MI+DV method in (Krotosky and Trivedi (2007)) and our proposed method LSS+MS+DV for our summer video using imperfect foreground segmentation (mainly misdetection). (a) visible image, (b) visible foreground segmentation, (c) thermal image, (d) thermal foreground segmentation, (e) MI+DV disparity image, (f) LSS+MS+DV disparity image, (g) MI+DV registration, and (h) LSS+MS+DV registration.

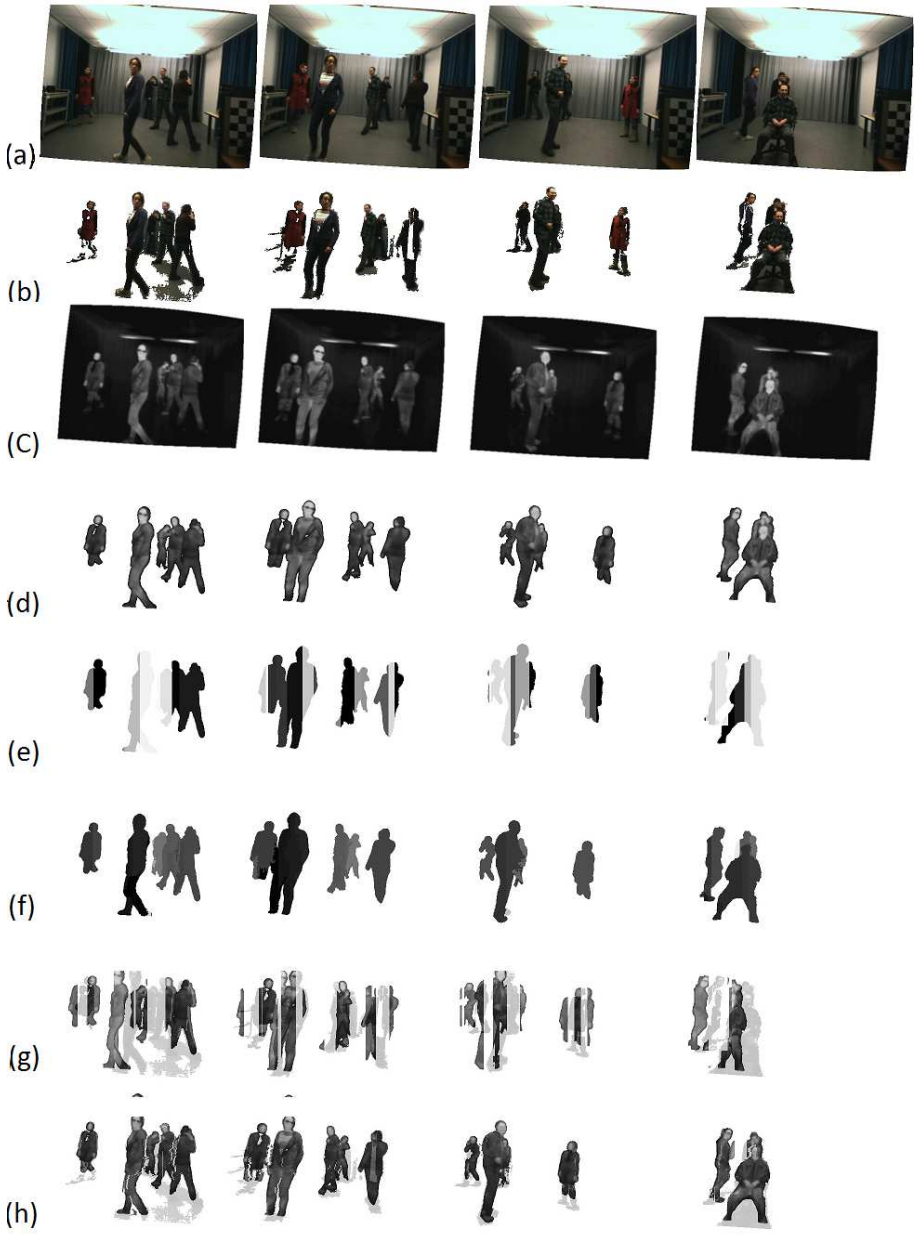


Figure 6.17 Comparison of MI+DV method in (Krotosky and Trivedi (2007)) and our proposed method LSS+MS+DV for our winter video using imperfect foreground segmentation (false detection and misdetection). (a) visible image, (b) visible foreground segmentation, (c) thermal image, (d) thermal foreground segmentation, (e) MI+DV disparity image, (f) LSS+MS+DV disparity image, (g) MI+DV registration, and (h) LSS+MS+DV registration.

CHAPTER 7

A LSS-BASED REGISTRATION OF STEREO THERMAL AND VISIBLE VIDEOS USING BELIEF PROPAGATION FOR HUMAN MONITORING

Abstract

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

7.1 Introduction

A fundamental issue associated to close-range multispectral imaging is accurately registering corresponding information and features of images with dramatic visual differences, such as thermal and color images. In a thermal-visible unsupervised visual surveillance system that monitors a close-range scene, matching corresponding features in a pair of visible and thermal videos has some specific difficulties. People in the field of view of the cameras are of various sizes, in various poses, clothes, distances to cameras, and at different levels of occlusion. They might have colorful/textured clothes that are visible in color images, but not in thermal images. On the other hand, there might be some textures observable in thermal images caused by the amount of emitted energy from different parts of the human body that are not visible in a color image. Due to the high differences between thermal and visible image characteristics, the only viable registration approach is partial image ROI registration. In this approach, matching is performed on the observable targets in both spectrums (like people) rather than the entire scene using a dense stereo correspondence algorithm.

Classical dense two-frame stereo matching computes a dense disparity map for image pixels using known camera configuration. Stereo matching is a well-studied subject for unimodal imaging system. An extensive taxonomy of two-frame stereo correspondence algorithms is presented in (Scharstein and Szeliski (2002)). However, this subject is new for multimodal visual surveillance applications. We summarize the problems associated to multimodal dense stereo as follows :

- **Dissimilar patterns**.This problem is specific to multimodal dense stereo. It is caused by the different types of image modalities. The corresponding regions in two images might be differently textured or one textured while the corresponding one is homogenous.
- **Depth discontinuities**. This difficulty is caused by segmentation results that contain two or more merged objects at different depths in the scene. In this case, correct disparities might be significantly different between neighboring pixels located on the depth boundaries.
- **Occlusions**. Some pixels in one view might be occluded in the other view. Therefore they should not be matched with pixels in the other view.

The global optimization approach has many advantages for stereo vision. It can explicitly encode various visual image cues (e.g. color segmentation) that are inferred from scene structure in the stereo model as smoothness assumptions to elegantly handle depth discontinuities, occlusions, and non-informative pixels caused by dissimilar patterns (corresponding pixels that do not contain similar visual information). However, applying global optimization to multimodal stereo problem is challenging since most similarity measures, which are used for color images, are not viable for multimodal images. In our previous works, local self-similarity (LSS) (Shechtman and Irani (2007)) was integrated into a local stereo correspondence method and its strengths were compared to MI and several other viable similarity metrics in the context of visual surveillance systems (Torabi and Bilodeau (2011); Torabi *et al.* (2011)). This paper has two significant new contributions. First, we integrated LSS as viable similarity feature in a global optimization correspondence approach, and second, we formulated a multimodal stereo matching in a Markov Random Fields (MRFs) framework using color and motion information as smoothness assumptions for partial image ROI registration.

The rest of the paper is organized as follows : The review of related works is presented in section 7.2. In section 7.3, we describe the strengths of LSS as a viable image feature for matching thermal and visible images. In section 7.4, the overview of our registration system is presented, and, in section 7.5 the detail description of each step of our algorithm is described. Our experiments shown in section 7.6 demonstrate that our method is effective

and efficient for video surveillance applications. Finally, in section 7.7, we conclude this paper by describing the advantages and limitations of our algorithms.

7.2 Related Works

In the thermal-visible video surveillance research context, the majority of the image registration approaches are related to global image registration that globally transform a reference image on the second image. Krotosky and Trivedi give a comparative survey of multimodal registration approaches (Krotosky and Trivedi (2007)). Global transformation approaches, either extract low-level image features such as edge features (Coiras *et al.* (2000)), or temporal-spatial features such as object trajectories (Torabi *et al.* (2010, 2012)) to estimate a transformation matrix that transforms one image on another with the assumption that all the objects in the scene approximately lie in one depth plane. A few works in literature cover a video registration method appropriate for close-range people monitoring. These methods have been categorized as partial image ROI registration (Krotosky and Trivedi (2007)).

In previous partial image registration approaches excluding ours (Torabi and Bilodeau (2011); Torabi *et al.* (2011)), MI is the only similarity measure used in local dense correspondence algorithm for human monitoring applications (Krotosky and Trivedi (2007); Chen *et al.* (2003); Egnal (2000)). The accuracy of MI as a similarity metric is directly affected by the MI window sizes. For unsupervised human monitoring applications, obtaining appropriate MI window sizes for the registration of multimodal pairs of images containing multiple people with various sizes, poses, distances to cameras, and different levels of occlusion is quite challenging. In the video surveillance context, Chen *et al.* proposed a MI-based registration method for pairs of thermal and visible images that matches windows on foreground regions in the two images with the assumption that each window contains one single depth plane (Chen *et al.* (2003)). In their method, the problem of depth discontinuity inside an ROI was not addressed. Later, Krotosky and Trivedi proposed a MI-based disparity voting matching approach (Krotosky and Trivedi (2007)). Their method, for each ROI column, computes the number of votes related to each disparity and assigns a disparity with maximum votes. Their method theoretically considers depth discontinuities that may occur between neighboring columns, but it ignores vertical depth discontinuity where the pixels on a column belong to multiple depths. For example, two people with different heights, where the shorter person is in front of the taller one. To the best of our knowledge, in our context of visual surveillance, all the existing methods for multimodal stereo matching are local correspondence approach.

Recent global stereo algorithms have achieved impressive results by modeling disparity image as Markov Random Field (MRF) and determining disparities simultaneously by ap-

plying energy minimization method such as belief propagation (Sun *et al.* (2003); Felzenszwalb and Huttenlocher (2006); Yang *et al.* (2009b)), and graph cuts (GT) (Boykov *et al.* (2001); Bleyer and Gelautz (2007)). Tappen and Freeman have shown that GC and BP produce comparable results using identical MRF parameters (Tappen and Freeman (2003)). Sun *et al.* proposed a probabilistic framework to integrate into BP model, additional information (e.g., segmentation) as soft constraints (Sun *et al.* (2003)). Moreover, they have shown that the powerful message passing technique of BP deals elegantly with textureless regions and depth discontinuity problems. Later, Felzenszwalb and Huttenlocher proposed an efficient BP algorithm that dramatically reduced the computational time (Felzenszwalb and Huttenlocher (2006)). Their method is interesting for time sensitive applications like video surveillance. More recently, different extension of this efficient BP was proposed in several works (Yang *et al.* (2010); Klaus *et al.* (2006)).

In our previous works, we have shown that local Self-Similarity (LSS), as a similarity measure, is viable for thermal-visible image matching and outperforms MI, especially for matching corresponding regions that are differently textured (high differences) in thermal and visible images (Torabi and Bilodeau (2011); Torabi *et al.* (2011)). We also proposed a LSS-based local stereo correspondence approach for close-range multimodal video surveillance applications (Torabi and Bilodeau (2011)). In this work, we adopt LSS as similarity measure in an energy minimization stereo model using the efficient BP model (Felzenszwalb and Huttenlocher (2006)).

7.3 LSS For Multimodal Image Registration

Local Self Similarities (LSS) is an image visual feature that has been proposed by Shechtman and Irani (Shechtman and Irani (2007)) and has been previously applied to problems such as object categorization, image classification, pedestrian detection, and object detection (Walk *et al.* (2010); Yang *et al.* (2009a); Vedaldi *et al.* (2009)). LSS describes statistical co-occurrence of small image patch (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 patch centered at pixel p and all possible patches in a larger surrounding image region. SSD is normalized by the maximum value of the small image patch intensity variance and noise (a constant that corresponds to acceptable photometric variations in color or illumination). It is defined as

$$S_p(x,y) = exp(-\frac{SSD_p(x,y)}{max(var_{noise}, var_{patch})}). \tag{7.1}$$

Then, the correlation surface is transformed into a log-polar representation partitioned into e.g. 80 bins (20 angles and 4 radial intervals). The LSS descriptor is defined by selecting the maximal value of each bin that results in a descriptor with 80 entries. LSS has two interesting characteristics for our application : 1) LSS is computed separately as a set of descriptors in one individual image and then it is compared between pair of images. In contrast, MI is computed directly between the two images. This characteristic makes LSS viable to be used in a global correspondence approach. 2) The measurement unit for LSS is a small image patch that contains more meaningful patterns compared to a pixel as used for MI computation. This property makes LSS describing layout accurately without being too sensitive to detailed texture variances. For multimodal human ROI matching, where human body have similar layouts in both modalities but they are not identical in textural appearance, LSS is a powerful descriptor.

In our application, before matching the LSS descriptors between pair of thermal and visible images, we discard the non-informative ones using a simple method. Non-informative descriptors are the ones that do not contain any self-similarities (e. g. the center of a small image patch is salient) and the ones that contain high self-similarities (a homogenous region with a uniform texture/color). A descriptor is salient if all its bin's values are smaller than a threshold. The homogeneity is detected using the sparseness measure of (Hoyer and Dayan (2004)). Discarding non-informative descriptors is like an implicit segmentation or edge detection, which increases the discriminative power of the LSS measure and avoids ambiguous matching. It is important to note that the remaining informative descriptors still form a denser collection compared to sparse interest points. Figure 7.1 shows pixels having informative descriptors (white pixels) for a pair of thermal and visible images. The regions belonging to the human body boundaries and some image patterns are the informative regions.

(a) (b)

Figure 7.1 Informative LSS descriptors. (a) Visible image and informative LSS descriptors (b) Thermal image and informative LSS descriptors.

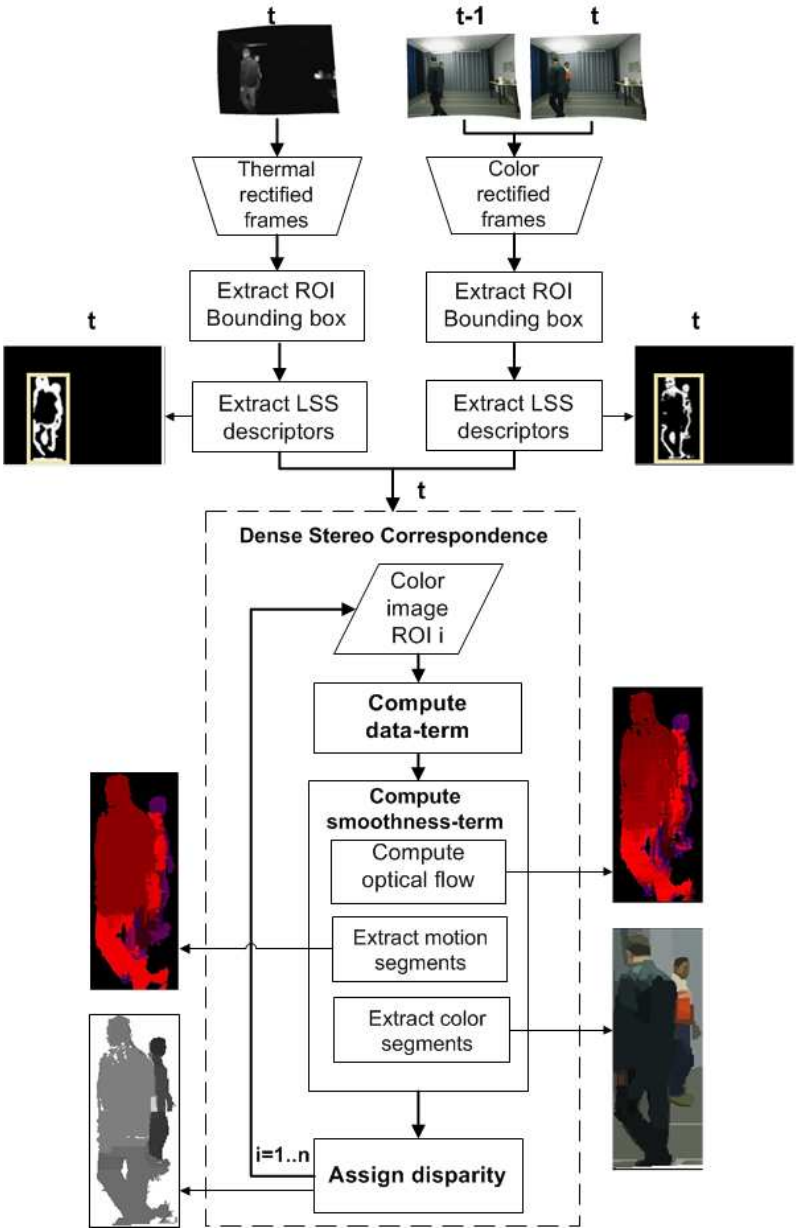


Figure 7.2 Block diagram of thermal-visible dense stereo matching algorithms augmented with input images, intermediate and disparity image results.

7.4 Overview Of Our Approach

Our registration algorithm is designed for video surveillance systems where the input data is a pair of synchronized thermal and visible videos. In our algorithmic design, it is feasible to add a new module for higher level processing, such as tracking. However, in this work, we only focus on the registration algorithm. The overall algorithm consists of several steps as shown in figure 7.2. At each time step t, the input data of our system is a rectified pair of thermal and visible frames at t and rectified visible frame at $t-1$. For the visible spectrum, two consecutive frames are needed to compute the optical flow in a later step of our algorithm. Due to the high differences in imaging characteristics of thermal and visible sensors, our registration is focused on the pixels that correspond to ROIs. As the first step of our algorithm, we extract image ROIs on pair of thermal and visible images using a background subtraction method (Shoushtarian and Bez (2005)). Each image ROI is defined by its bounding box. The registration is applied on the pixels inside the box. In the thermal spectrum, a bounding box is surrounding a foreground region at time t. In the visible image, a bounding box is surrounding overlapping foreground regions at time $t-1$ and t. In this way, for efficiency, the optical flow computations (later step) are performed only inside the visible image bounding box. The next step is extracting LSS descriptors for foreground pixels inside the bounding boxes at frame t. In figure 7.2, the image results of this step show pixels with informative LSS in white and non-informative ones in black (informative pixels are determined using the method described in section 7.3).

The main body of our registration algorithm begins after LSS feature extraction. Registration is done by matching visible ROIs on thermal ROIs. The reason for matching visible ROIs on thermal ROIs is that for color image, both color and motion cues are available to be used as complementary image cues in our registration model. However, for thermal image, the color cue is not defined. In our matching strategy, each bounding box on visible image is viewed as a smaller image. Registration is done separately for each bounding box. Disparities are assigned to all pixels inside a box using a global optimization that minimizes an energy function which is described in details in the following sections. Our energy function consists of a data term and a smoothness term. The data term is computed based on self-similarities matching between pixels that contain informative LSS descriptors. The smoothness term is computed using motion and color cues of pixels inside a bounding box in the visible image. To extract the motion cues, we compute the optical flow using a state-of-the-art method (Ogale and Aloimonos (2007)). Then, we use mean-shift segmentation to cluster the motion vector fields extracted in the previous step (Comaniciu and Meer (1999)). To visualize the optical flow and segmentation images, we mapped the motion vector fields to HSV color

system. To extract the color cues, we apply the same mean-shift segmentation on pixel intensities to compute the color segmentation. Figure 7.2 shows results of optical flow, motion segmentation, and color segmentation. Finally, the disparities are assigned to pixels inside the bounding box using an efficient belief propagation method (Felzenszwalb and Huttenlocher (2006)).

7.5 Detailed Description

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

7.5.1 Thermal-Visible Stereo Model

We formulate the registration as a multi-labeling problem (we use the notation from (Felzenszwalb and Huttenlocher (2006))). We assume that P is the set of all pixels inside the image bounding box and that L is a set of labels, which are disparity quantities in our problem. A labeling f assigns a label $f_p \in L$ to each pixel $p \in P$. We model our stereo matching using a Markov Random Field (MRF) framework and estimate the quality of labeling using an energy function defined as,

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{(p,q) \in N} V(f_p, f_q). \tag{7.2}$$

where D_p is data term (cost of assigning label f_p to pixel p), V is the smoothness term (cost of assigning labels f_p and f_q to two neighboring pixels p and q), and N are edges (neighborhood system) in the image graph. In our image graph, we use a four-connected neighborhood system.

7.5.2 Data Term

The data term only encodes the similarity distance of informative LSS descriptors on matching thermal and visible pixels for a preset disparity range. The distance is basically the $L1$ distance between two informative LSS descriptors on a pair of thermal and visible images.

$$D_p(f_p) = \begin{cases} L1(p_l, p_r) & \text{if } p_l, p_r \in \text{informative} \\ 1 & \text{otherwise} \end{cases}, \tag{7.3}$$

where p_l is the LSS descriptor of pixel p inside bounding box on the visible image and p_r is the LSS descriptor of matching pixel of p on the corresponding row of thermal image by disparity offset f_p. In our data-term, if two matching pixels are containing informative LSS descriptors (more details section 7.3) ; we compute a normalized $L1$ distance as data term. Otherwise we, simply assign the maximum possible value for data term since matching is not defined if one of the pixels either on thermal or visible does not contain an informative descriptor. Then, we map the data term to values between $[0-255]$ as pixel intensity interval values.

7.5.3 Smoothness Term

In our stereo model for pair of thermal-visible videos, the smoothness term has a crucial role for passing the influence of messages from pixels with informative LSS far away to non-informative ones, while the influence in the depth discontinuous regions should fall off quickly (Note that we used a belief propagation based energy minimization that is based on an iterative message passing between neighboring pixels in the image graph). For this reason, we incorporated visual cues including motion and color segmentation in the stereo model as soft constraint to accurately determine disparities. The main advantage of this approach rather than a segment-based stereo algorithm such as (Klaus *et al.* (2006)), which assumes that depth discontinuity occurs on the boundary of segmented regions as a hard constraint, is that messages are still passed between segmented regions ; therefore it is more robust to incorrect segmentation results. In the following, we describe how we incorporate motion and color in our smoothness term.

Motion

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

We apply a simple two-frame motion segmentation using two consecutive color image frames $t-1$ and t. Firstly, we compute the motion vector field for all pixels (including foreground and background) inside the window of an ROI using an optical flow method based on block-matching (Ogale and Aloimonos (2007)). Second, we apply the mean-shift segmenta-

Figure 7.3 (a) Image window (b) Foreground (c) Optical flow (d) Motion segments.

tion method proposed in (Comaniciu and Meer (1999)) (on foreground pixels) for segmenting the motion vector field computed in the previous step, and for assigning a mean velocity vector to each segment. We apply motion segmentation only on foreground regions inside the image window at frame t in order to extract also a segment associated to temporary stationary person for which its mean velocity vector is zero. Mean-shift segmentation is applied on (2+2) feature point dimensions, where two dimensions are related to spatial dimensions and the two others are related to the two motion vector components in x and y directions. Figure 7.3 shows the motion segmentation result of three merged people in one ROI where two people are moving and the other one is temporary stationary. In order to visualize the motion segments, motion vectors are mapped to HSV color space. Our motion segmentation results in a set of regions $SM = \{sm_1, .., sm_i, .., sm_m\}$ inside the image window. Each motion segment sm_i, itself, is a set of foreground pixels labeled with a motion vector field value, which is the mean of motion vector fields belonging to the pixels inside the segment.

There are three difficulties associated with motion segmentation. First, an image ROI belonging to objects closer to the camera might be too over-segmented and fragmented into several motion segments. Second, imperfect foreground segmentation causes some pixels inside an ROI not being assigned to any motion segments. Figure 7.4(a) and (b) show an example of over segmentation ; (c) and (d) an example of imperfect background subtraction. Third, the occluded pixels (occluded pixels are obtained by method proposed in (Ogale and Aloimonos (2007)) at frame $t-1$, which are visible at frame t, have no defined motion vectors. This last difficulty causes inaccurate motion segment boundaries that do not correspond to actual depth discontinuities in the image. Figure 7.5 shows an example of motion segmentation where the motion segment boundaries are inaccurate due to the existing occluded pixels. Applying motion segmentation on foreground regions eliminates those occluded pixels which are part of background. However, those which are inside an ROI containing two people like in our example, cause inaccurate motion segment boundaries. In order to avoid inaccurate disparity assignment caused by imperfect motion segmentation, we apply color segmentation

as a complementary visual cue.

Color

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

Color segmentation is done using the same mean-shift segmentation that we applied for motion segmentation (Comaniciu and Meer (1999)). In figure 7.2, the color segmentation block shows an example of our segmentation. We use RGB color system to represent the color segments. We also use an over segmentation to avoid merging color regions belonging to more than one people.

Integrating Multiple Cues

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

$$V(f_p, f_q) = \begin{cases} \alpha|f_p - f_q| & \text{if } \; p,q \in MS \; \wedge \; p,q \notin O \\ \beta|f_p - f_q| & \text{elseif } \; p,q \in CS \\ |f_p - f_q| & \text{otherwise} \end{cases} . \tag{7.4}$$

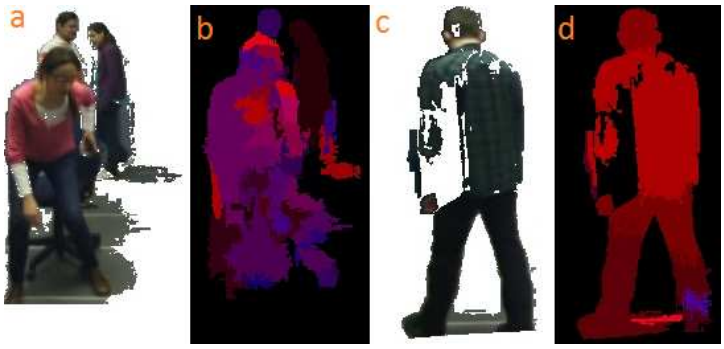


Figure 7.4 (a) Foreground visible,(b) motion segmentation, example of over-segmentation, (c)Foreground visible, (d) Motion segmentation, example of misdetected regions

Figure 7.5 (a) Foreground visible (b) Optical flow (c) Motion segmentation (d) Occluded pixels (white pixels).

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

7.5.4 Disparity Assignment

In our algorithm, an optimal labeling with minimum energy is approximated using the efficient loopy belief propagation proposed by Fezenswalb and Huttenlocher (Felzenszwalb and Huttenlocher (2006)). Their method substantially reduces the complexity time of belief propagation approach from $O(nk^2T)$ to $O(nkT)$, where n is the number of pixels (nodes), k is number of possible disparities (labels), and T is the number of iteration. For stereo problem modeled in term of posteriori probabilities, BP algorithm is used for performing inference on MRFs by applying the max-product algorithm (Sun *et al.* (2003)). The equivalent computation used in (Felzenszwalb and Huttenlocher (2006)) is negative-log probabilities, where the max-product becomes min-sum and the energy function definition (equation 7.2) can be used directly.

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

$$m^i_{pq}(f_q) = Min_{f_p}\left(V(f_p, f_q) + D_p(f_p) + \sum_{r \in N(p)-q} m^{i-1}_{rp}(f_p)\right), \tag{7.5}$$

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

$$b_p(f_p) = D_p(f_p) + \sum_{q \in N(p)} m^N_{qp}(f_p). \tag{7.6}$$

Finally, the disparity (label) which individually is assigned to each pixel p is the label with minimum value in final belief vector. In our implementation of efficient BP (Felzenszwalb and Huttenlocher (2006)), we used two of their techniques to speed up the processing time. First, by using their message updating that reduces the computational complexity from $O(k^2)$ to linear time $O(k)$. Second, by using their alternating message updating techniques for bipartite graph (like an image grid), which reduces the number of update message in each iteration to half. More details can be found in (Felzenszwalb and Huttenlocher (2006)).

7.6 Experiments

7.6.1 Experimental setup

We tested our method using visible-thermal synchronized videos of a $5m \times 5m$ room at a fixed temperature of 24 $°C$. The videos were recorded by stationary thermal and visible cameras with baselines of $10cm$ and $13cm$. The videos include up to five people moving throughout the scene. People have colorful, thick, or light clothes, which appear differently textured in thermal and visible images. Moreover, they may also carry objects, such as a bag that is only visible in one image modality. Figure 7.6 shows our camera setup and examples of calibration images in visible and thermal.

In order to simplify the matching to a $1D$ search, the thermal and visible cameras were

(a) (b) (c)

Figure 7.6 (a) Camera setup. The halogen lights behind the cameras are used for calibration, (b) visible calibration image and (c) thermal calibration image.

calibrated using the standard method described in (Heikkila and Silven (1997)) and implemented in the camera calibration toolbox of MATLAB. Since in the thermal images, the calibration checkboard pattern is not visible at room temperature, we illuminated the scene using high intensity halogen bulbs placed behind the two cameras. In this way, the dark squares of the checkboard absorb more energy and appear visually brighter than the while squares in the thermal images.

Figures 7.7 and 7.8 illustrate two examples of successful registration of visible image on thermal foreground images using our algorithm. At the same time, these two figures illustrate the benefit of combining thermal and visible information. People are at different depth levels and with different clothing (such as wearing scarf or jacket). Background subtraction is imperfect and includes false positive (shadows) and false negative (partial misdetections) errors. In figure 7.7, a person carries a hot pot that is clearly distinguishable in thermal image, but not as easy to detect in the visible image. In figure 7.8, a person is carrying a bag at room temperature, and hence is not detected in the thermal image. Our global optimization approach has successfully estimated correct disparity for the bag region since it is connected

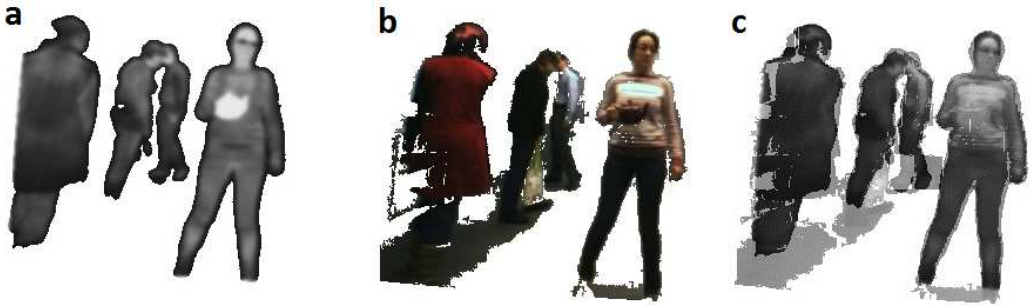


Figure 7.7 Detailed registration a person carrying a hot pot. (a) Foreground thermal image, (b) Foreground background image, and (c) Registration of visible image on thermal image.

to the person region in the image. However, using a Winner Take All (WTA) matching approach, such as $MI + DV$ (Krotosky and Trivedi (2007)), estimating correct disparity is not obvious and is limited to matching window sizes.

In order to assess our registration for video surveillance applications, we compared our proposed Local Self Similarity based Belief Propagation algorithm ($LSS{+}BP$) with the state-of-the-art Mutual Information based Disparity Voting algorithm ($MI + DV$) in (Krotosky and Trivedi (2007)) and with our previous work, Local Self Similarity based registration using DV matching ($LSS + DV$) in (Torabi and Bilodeau (2011)). We focus on two main aspects that demonstrate the efficiency of our method compared to previous works : 1) depth discontinuity handling of occluding/occluded people, and 2) the effect of different disparity ranges, whether small or large, on the registration performance.

In the following sections, we present our comparative evaluation regarding these two aspects.

7.6.2 Evaluation Of Disparity And Registration Accuracy For Occlusions

In order to demonstrate the disparity accuracy improvement of our matching approach compared to state-of-the-art DV matching approaches (Krotosky and Trivedi (2007); Torabi and Bilodeau (2011)) for occlusion handling, we quantitatively compared the disparity results of our proposed BP and of DV. In order to perform a fair comparison, we use LSS as similarity measure in the two approaches. We generated ground-truth disparities by manually segmenting and registering regions of foreground of each pair of images.

Figure 7.9 illustrates the comparison of $LSS + BP$ and $LSS + DV$ disparity results with ground-truth. Results in the first and second rows illustrate examples where two people at two different depths in the scene appear in a single region. The third row shows an example where multiple people are in occlusion and where object segmentation is erroneous. $LSS{+}DV$

Figure 7.8 Detailed registration of a person carrying a bag. (a) Foreground thermal image, (b) Foreground background image, and (c) Registration of visible image on thermal image.

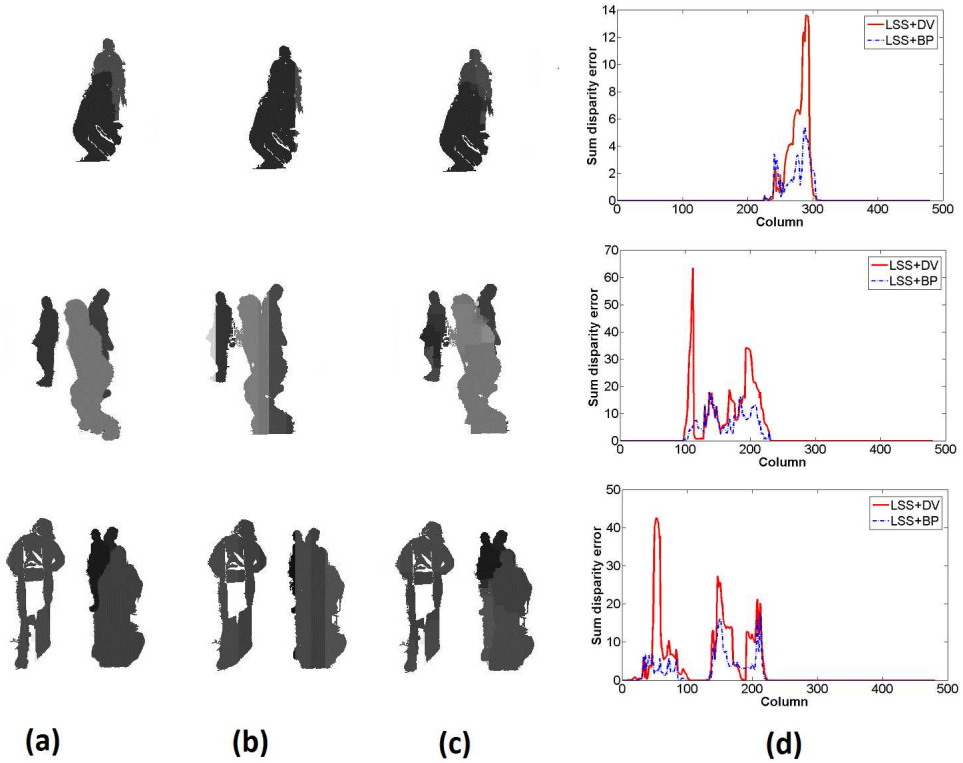


Figure 7.9 Comparison of the disparity accuracy of $LSS + DV$ and $LSS + BP$ methods :(a) ground-truth, (b) $LSS + DV$ disparity map, (c) $LSS + BP$ disparity map, and (d) Sum of disparity errors at each image column.

method fails to assign correct different disparities to the columns containing pixels related to more than one disparity level. In order to register people merged in a single region, DV method makes no assumptions about the assignment of pixels to individual person and assigns a single disparity to each column inside an ROI, based on a maximization of the number of votes. If pixels on a column of image belong to different objects at different depth in the scene, the vote only goes for one of them based on WTA approach. However, $LSS + BP$ succeeds in assigning accurately different disparities to the two human body ROIs using a belief propagation global optimization, where the color and motion cues were integrated as soft constraint in an energy function $LSS+BP$ gives a reasonable estimate of moving regions belonging to people in the scene. Accordingly, in Figure 7.9 (d), the sum of disparity errors of the columns corresponding to occluded people is in general higher for $LSS + DV$ method compared to LSS+BP method. However, in a few number of columns in three plots, LSS+BP has a slightly higher sum of disparity error.

Figure 7.10 illustrates detailed registration of three video frames of people at different levels of occlusion using $LSS + BP$ and $LSS + DV$ methods for a relatively large disparity range between $[5 - 50]$ pixels. In these examples, $LSS + DV$ fails to accurately register pixels related to depth discontinuity regions. In the following, we discuss the effect of a wide disparity range for WTA local matching approach such as DV compared to our proposed algorithm.

7.6.3 Evaluation Of Registration Accuracy Using Different Disparity Ranges

In this part of our experiments, we compared the registration results of $MI + DV$ (Krotosky and Trivedi (2007)), $LSS + DV$ (Torabi and Bilodeau (2011)), and our proposed $LSS + BP$ for two videos using disparity ranges of $[2 - 20]$ pixels and $[5 - 50]$ pixels where in both videos, up to five people are walking throughout the scene. In order to perform a fair comparison, both videos are recorded in the same room with similar environmental factors but for one video, the camera baseline is $10cm$ and for the other one it is $13cm$. In order to perform a quantitative evaluation of the registration performance of the algorithms, we defined an overlapping error that gives an estimate of the registration errors. The overlapping error is defined as,

$$E = 1 - \frac{N_{v \cap t}}{N_v}, \tag{7.7}$$

where $N_{v \cap t}$ is the number of overlapped thermal and visible foreground pixels and N_t is the number of visible foreground pixels. The best performance with zero overlapping error is when all the visible pixels on the reference image have corresponding thermal pixels on the second image (we register the visible on the thermal image). This evaluation measure includes the

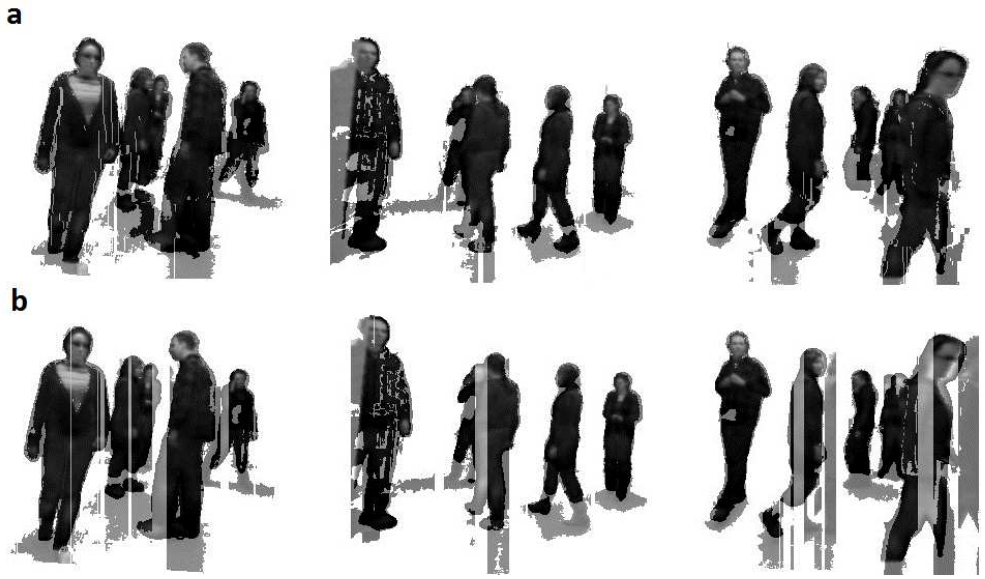


Figure 7.10 Comparison of $LSS + DV$ and $LSS + BP$ methods registration accuracy (large disparity range of $[5-50]$ pixels) :(a) $LSS+BP$ detailed registration, (b) $LSS+DV$ detailed registration.

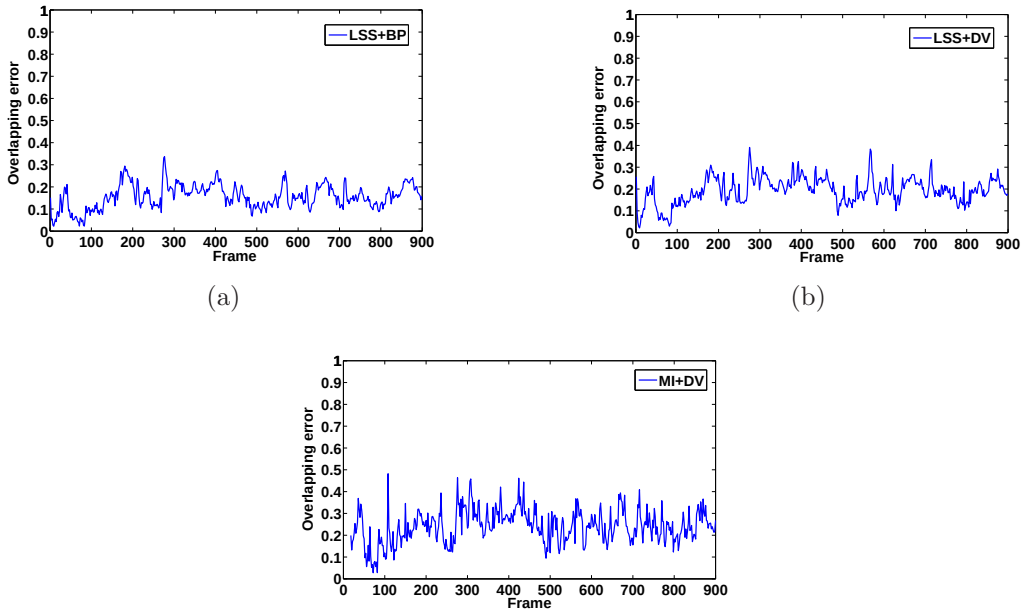


Figure 7.11 Overlapping error using disparity range $[2-20]$: (a) LSS+BP , (b) LSS+DV, and (c) MI+DV.

background subtraction errors and also ignores misaligned visible pixels inside foreground regions of thermal image. However, since for the three methods, the background subtraction errors are included in the overlapping error, the differences between the overlapping errors are still good indicators for comparing overall registration accuracies for a large numbers of frames.

Figure 7.11(a) illustrates the overlapping errors over 900 video frames. For DV methods, we used matching window size of 30 pixels wide that we experimentally found to have the minimum mean overlapping errors among the three size of 10, 20, and 30 pixels. The mean overlapping error of $MI + DV$ is 0.24, $LSS + DV$ is 0.19, and $LSS + BP$ has the minimum error among the three methods which is 0.15. $LSS + DV$ has the second place and $MI + DV$ is the least accurate. However, the three methods have reasonable overlapping errors and are stable over 900 frames, considering the background subtraction errors as well. The standard deviation (*std*) value of $LSS + BP$ is 0.05, $LSS + DV$ is 0.06, and $MI + DV$ is 0.07. Again, $LSS + BP$ has the most stable performance.

Figure 7.12(a) illustrates the overlapping errors over 4000 video frames. For DV methods, we used matching window size of 30 pixels. The mean overlapping error of $MI + DV$ is 0.49, $LSS + DV$ is 0.25, and $LSS + BP$ is 0.20. Similarly to the previous experiment, $LSS + BP$ has the minimum error among three methods, $LSS + DV$ has the second place, and $MI + DV$ is the least accurate. The *std* value of $LSS + BP$ is 0.07, $LSS + DV$ is 0.25, and $MI + DV$ is 0.18. It is should be noted that for all three methods, overlapping errors have increased. However, compared to the other video, it is observable that the mean overlapping error of DV methods, especially $MI + DV$ significantly increased. Moreover, they have a larger number of overlapping error outliers (large *std*) compared to the previous video, which shows some performance instabilities over the whole video. Furthermore, $LSS + DV$ performs better than $MI + DV$. This shows that LSS used as similarity metric is a more robust feature for multimodal matching compared MI in the case of visible and infrared images. $BP + LSS$ was less influenced by the change of disparity range.

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

Figure 7.13 shows four examples of tested video frames using a disparity range of $[2-20]$. For these video frames, figure 7.14 illustrates qualitatively the resulting disparity maps, and

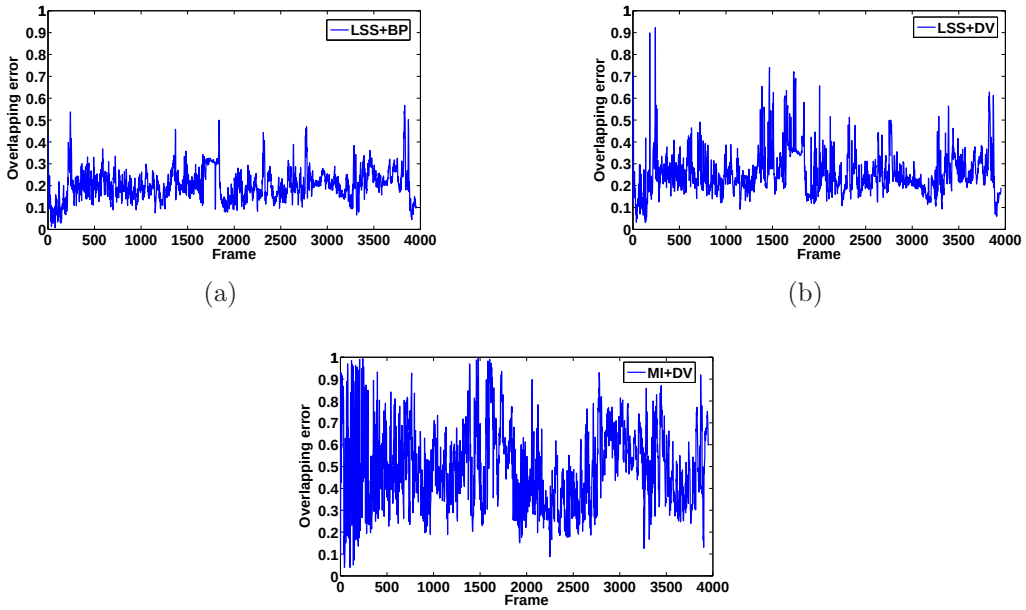


Figure 7.12 Overlapping error using a disparity range of $[5-50]$: (a) LSS+BP , (b) LSS+DV, and (c) MI+DV.

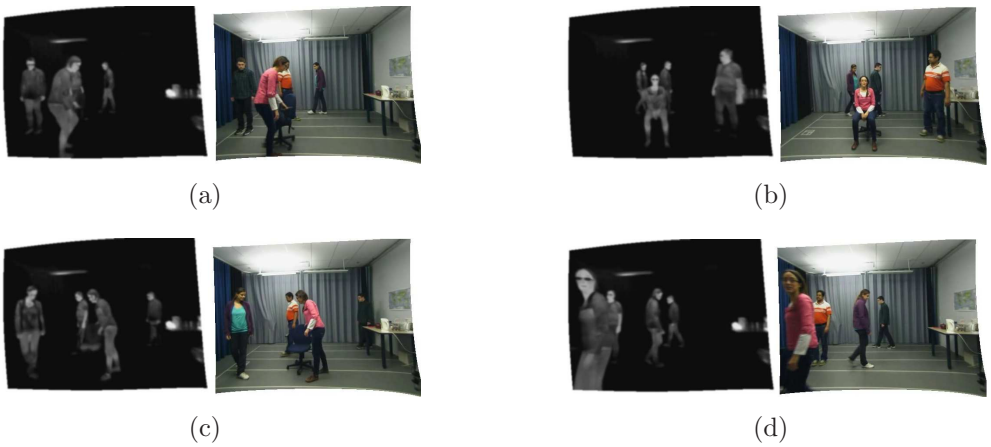


Figure 7.13 Example of Tested video frames of video with a disparity range of [2-20].

registrations of visible foreground image on thermal foreground image using $LSS + BP$, $LSS + DV$, and $MI + DV$. Figure 7.14, rows (d) and (e) show the disparity maps for the DV methods. In both methods, disparity assignments are inaccurate for depth discontinuity regions. However, $LSS + DV$ results in more accurate disparity map. Figure 7.14, rows (c) shows the disparity map of $LSS + BP$ method. It has more accurate results, especially for depth discontinuity regions. However, the last column shows some color and motion over-segmentation for the person close to the camera that results in less smooth disparity map inside the human body ROI compared to the farther objects.

7.7 Conclusions

In this paper, we proposed a stereo model for thermal-visible partial ROI registration using an efficient belief propagation algorithm that outperforms previous state-of-the-art stereo registration designed for close range video surveillance applications. We have tested our methods on two indoor videos, over 4900 frames. Our results demonstrate that our method assigns more accurate disparity to pixels related to depth discontinuity regions and that it is more stable for large disparity range compared to previous works (Krotosky and Trivedi (2007); Torabi and Bilodeau (2011)).

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

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

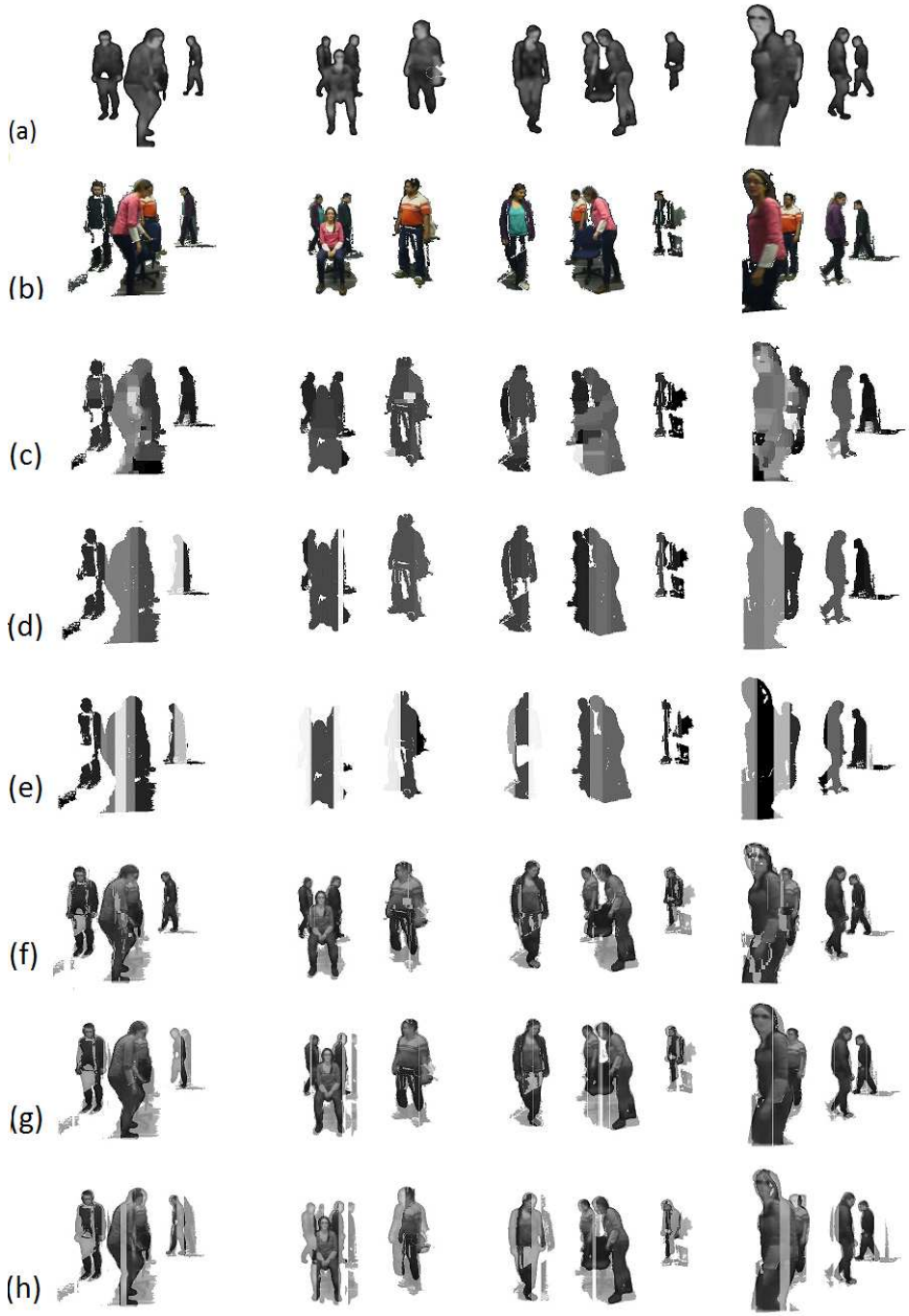


Figure 7.14 Qualitative Comparison : (a) thermal foreground image, (b) visible foreground image (c) disparity map $LSS{+}BP$, (d) disparity map $LSS{+}DV$, (e) disparity map $MI{+}DV$, (f) registration $LSS + BP$, (g) registration $LSS + DV$, (h) registration $MI + DV$.

CHAPTER 8

GENERAL DISCUSSION

8.1 On The Registration Of Far-Range Videos

For long-range scenes, where the objects are far, our proposed image registration is a global registration method based on the assumption that the objects lie approximately on one depth plane in the scene. Such an assumption is valid either when only one object moves throughout the scene in a single plane, or when the captured scene is much farther than the distances between moving objects, in the case where multiple people are moving in the scene (long-range video). For this thesis, we consider multiple people are moving throughout a long-range scene. In such a case, the cameras are placed relatively far from imaged scene. The people can thus be considered in the same plane in the scene.

The global image registration approach requires a number of sparse corresponding image features between thermal and visible images to estimate a homography that globally transforms one image on another one. The main advantage of this approach is its efficiency in terms of its computational time which makes it interesting for online video surveillance applications. In a global image registration, two problems should be solved : 1) detecting viable image feature for matching thermal and visible images, and 2) matching features and estimating the homography. One of the most important characteristics of our method compared to the state-of-the-art methods is its performance for significantly different zoom settings between the thermal and visible cameras. In fact, we consider that object scales may vary significantly during the video and in some case, extracting low level features inside object regions may get difficult due the small size of objects. Therefore, we used the spatio-temporal information of the scene that is object trajectory points, and performed sequence-to-sequence trajectory matching rather than a low-level image-to-image matching. Our feature detection approach raises another problem, which is object trajectory computation. In the literature, the few works that applies trajectory-based image matching, assume that object trajectories are computed in an offline process which is not practical for online applications. Moreover, the accuracy of computed trajectories in both thermal and visible videos has a crucial effect on the image registration result. Using independently thermal and visible videos for trajectory computation might result in inaccurate and disconnected trajectories in challenging scenarios. Our approach to handle this difficulty regarding the trajectory computation is an important contribution of our method compared to the state-of-the-art. In fact, in our al-

gorithm, registration and tracking are performed simultaneously in an iterative scheme. In other words, the whole process is online and no offline processing for object trajectories computation is required. The iterative scheme improves the quality of the trajectories as shown in the experiments of our article (section 4).

The other advantage of our method is that for scenes where the planar assumption is not completely valid due to either the distance of the cameras from the scene or the angle of the view of the camera from the scene, our iterative registration method estimates the most accurate transformation matrix related to the current position of people in the current frame. Our matching process is a RANSAC-based method with matching criteria of the overlapping foreground regions and overlapping trajectory points. The limitation of the overlapping criterion is that it also includes background subtraction error that might be misleading for the registration. In our experimental results, we have shown that in general this criterion improves the overall registration results over thousands of video frames.

8.2 On The Choice Of An Appropriate Feature For The Registration Of Close-Range Visible And Thermal Videos

Next, our work focused on the problem of partial image ROI registration for close-range scenes where the assumption of planar homography is not valid and multiple objects may exist in the scene, each being at a different distance from the cameras. This field of study is not well documented in literature, especially for video surveillance applications.

In the context of thermal-visible video registration, there is no work in the literature that compares various LIDs and similarity measures for registration purposes especially for human monitoring applications. In the related state-of-the-art, only MI (classic multimodal similarity measure) is applied for matching between thermal and visible images using a local stereo correspondence approach. The shortcomings of MI in challenging human monitoring scenarios were the main motivation for us to evaluate other LIDs and similarity measures for this task. We studied comparatively various descriptors and measures in challenging human monitoring scenarios for matching thermal and visible human ROIs. Our comparisons were carried fairly using the same object segmentation, parameters, and matching window sizes throughout all our experiments. We have determined that LSS, as a similarity measure, outperforms other LIDs and similarity measures including MI. The property of LSS, which makes it interesting for our application, is that the basic unit for measuring internal joint statistics is a small image patch that captures more meaningful image patterns than individual pixels as used in MI computation ; therefore it is more robust for small differences in shape boundaries which in our case is human body shape in thermal and visible images. Also detecting informative

descriptors is a useful tool to match corresponding differently textured regions since thermal and visible pixels can be matched only if they both contain informative descriptors. In this way, we only focus on matching the similar patterns. LSS might fail for partial image ROI registration, if its descriptor window size (surrounding window in descriptor computation) is too big and contains large amount of background especially in color image where the background is textured as well.

To test further LSS, we integrate it inside a DV method proposed in the state-of-the-art. Our results have demonstrated that using identical DV parameters for registration, our proposed LSS-based registration outperforms the similar state-of-the-art MI-based registration approach Krotosky and Trivedi (2007). This result was reproduced on several human monitoring scenarios including people with different scales, poses, and clothing.

8.3 On The Advantages and Limitations Of Using Motion Segmentation

In order to improve the state-of-the-art DV matching strategy for assigning accurate disparities on depth discontinuity regions (where multiple people are merged in a one image region), we proposed using motion segment as visual cue to segment merged region to motion segments, then performing DV on each motion segment separately. The idea behind proposed motion segmentation is that our registration targets are humans moving in different directions or possibly temporary stationary humans. Therefore motion segment is a good estimate of disparity layer existing in the image. Applying DV matching on a motion segment, results in a more accurate registration of occluded people compared to the standard DV approach. The problem associated with depth-layer (motion segment) estimation using motion segmentation is over-segmentation, where a ROI is segmented to several small motion segments due to the close distance of a human target to the camera. Moreover, motion segmentation is also influenced by video frame rate where for low video frame rates ; it might not perform accurately and give as a result imperfect motion segments.

8.4 On Considering Stereo Matching As A Global Stereo Correspondence Problem

The problem with the local stereo correspondence approach is that it is influenced directly by the size of the local region (matching window) that is used to determine the disparity for matching regions between two images. There is always a trade-off between choosing larger windows for matching evidence, and smaller windows for the precision and details needed for an accurate registration. In the literature, all the existing registration methods, in the context of multimodal video surveillance, are local stereo correspondence approach. The main reason

is the difficulty of choosing a viable similarity measure for matching thermal and visible images that can be incorporated inside an energy function. MI is the only applied similarity measure in the related state-of-the-art multimodal registration methods and it is only suited for the local stereo correspondence approach.

Our proposed multimodal similarity measure, LSS, similarly to MI, computes statistical co-occurrence of pixel intensities. However LSS, unlike MI, is firstly computed and extracted from an individual image as a descriptor and then compared between pairs of images. This property of LSS makes it suitable for a global stereo correspondence approach. In this thesis, we proposed a global optimization approach for the stereo thermal-visible videos. In the context of multimodal video surveillance, to the best of our knowledge, our method is the first global stereo correspondence approach. The main characteristic of our method is that we applied motion segmentation as a soft constraint in the smoothness term of our proposed energy minimization function rather than applying motion segment as a hard constraint like the one we proposed in our LSS-based local correspondence algorithm (section 6). By the soft constraint, we mean that even the pixels inside a motion segment are encouraged to be assigned to the same disparity value, but still there are the messages passing through the neighboring segments (via neighboring pixels belonging to the different motion segments) to globally minimize the energy over all the segments simultaneously.

Moreover, we used color segmentation as the complementary visual cue integrated in the smoothness-term of our energy function to recover the shortcomings of motion cue. In this way, we handle accurate disparity assignment of occluded people more elegantly compared to a local stereo correspondence approach. Also, we have demonstrated that for a global correspondence, registration errors increase less by increasing the number of people in the scene and having a larger disparity range compared to local approach. Although a global correspondence approach is much more stable and robust to larger disparity range, still tuning the disparity assignment costs (α and β in chapter 7) are influenced by the disparity range. Smaller costs allow two neighboring pixels to be assigned to two different disparities more easily compared to the high cost values. Therefore, in our method these values are determined experimentally for the different disparity ranges.

Another issue in applying global stereo correspondence approach for multimodal video surveillance system is the frame rate of the input videos. In our research, we have processed two videos, one with a frame rate of 7 FPS and another one with a frame rate of 20 FPS. We found out that using a higher frame rate increase the accuracy of the motion segmentation that we integrated as a visual cue in the smoothness-term. The main reason is that people move around the scene with different speeds. At the low frame rate, the camera does not capture accurately all the movements of a person who moves fast. That consequently results

in a large number of occluded pixels between two consecutive frames (more details about occluded pixels in section 7) which reduces the performance of the smoothness-term. Even that for the occluded pixels color cue will be used as soft constraint, still the performance decreases because in general the color cue is less reliable compared to the motion.

One limitation of our proposed multimodal stereo model compared to a unimodal stereo model is that the data-term is sparser since we use informative LSS descriptors as matching feature while unimodal stereo model uses pixel intensity for all the pixels as a simple viable feature. In a unimodal stereo, color over-segmentation is sufficient as soft constraint. However in our case, using color segmentation is not sufficient in our smoothness term since small segments might not include any informative data-term and result in an inaccurate over-segmented disparity map. Even by using motion and color in our smoothness-term, our disparity map results still do not reach the same level of smoothness and accuracy as a unimodal stereo model. Moreover, using color segmentation as complementary visual cue limits our method to register color image on thermal image as the opposite is not possible.

A general limitation of both our local and global stereo correspondence methods is concerns our multimodal camera calibration. Due to the differences of thermal and visible cameras, the camera calibration is a hard task and is not as accurate as camera calibration using two visible cameras with identical lenses. Therefore, more care and a large number of calibrating images are required to estimate the intrinsic and extrinsic camera parameters. However, we could reach to the accuracy required for image rectification.

CHAPTER 9

CONCLUSION

In our research, we have studied the problem of video registration for multimodal video surveillance systems. Our thesis includes registration approaches that are appropriate for both long-range and close-range human monitoring application domains.

For long-range human monitoring, we have proposed a complete system, which performs image registration, sensor fusion, and multiple people tracking iteratively using a feedback scheme. Our proposed system is applicable to online video surveillance applications. The proposed methods resulted in a journal article that is published in the journal of *Computer Vision and Image Understanding* and is included in section 4 of this thesis. For this part of our thesis, our main contributions are :

- Designing a long-range multimodal people monitoring system appropriate for online applications.
- Proposing a feedback scheme between system's modules that result in the overall improvement of the whole system compared to the similar system using an offline trajectory computation process.

For close-range human monitoring, we have proposed LSS as a viable similarity measure for matching thermal and visible images. We have compared this measure with the state-of-the-art viable LIDs and similarity measures and we have shown that LSS is the most accurate measure among them for thermal-visible human ROI registration. This performance evaluation resulted in a journal paper that we have submitted to *Pattern Recognition Letters* and have included in the section 5 of this thesis. For this part of our thesis, our main contributions are :

- Evaluating various local image descriptors and similarity measures for the partial ROI image registration.
- Introducing LSS as most robust similarity measure for matching thermal and visible human body ROI registration.

Furthermore, we have proposed two partial image ROI registration approaches which both produce dense disparity maps of foreground pixels of one image to be used to register them on the second image. The first one is a LSS-based local dense stereo correspondence method that solves the problem of depth discontinuity related to occluded people, by estimating motion segments and using a WTA voting approach to assign disparities to each motion segment. This method, similarly to all the WTA window-based matching approaches, has

certain level of limitations concerning the accuracy of disparity assignment to the foreground pixels, especially for those pixels related to the depth-discontinuity regions in image. The proposed local dense stereo correspondence method is an improvement over a state-of-the-art DV method that resulted in a journal article submitted to *Pattern Recognition*. We have been included this article in the section 6 of this thesis. For this part of our thesis, our main contributions are :

- Proposing a LSS-based correspondence method for thermal-visible video registration.
- Handling the problem of depth discontinuity by integrating motion segmentation to estimate depth-layers in the scene.

The second proposed method is a LSS-based global dense stereo correspondence method. Our approach for performing an accurate disparity assignment is using a global optimization method that globally assigns the disparities to foreground pixels using LSS as a similarity measure and visual cues including motion and color as soft constraints. The global optimization is performed using an efficient BP method. This global method is more accurate in disparity assignment compared to the previous local method, especially for disparity assignment of the depth-discontinuity region in the image. The proposed global dense stereo correspondence algorithm results in a journal paper that has been submitted to *Computer Vision and Image Understanding* and has been included in the section 7 of this thesis. For this part of our thesis, our main contributions are :

- Integrating LSS as a similarity measure and the motion as visual cue in an energy function for a global stereo correspondence method.
- Improving the motion segmentation shortcomings for depth-layer estimation by adding color cue as a supplementary visual cue.

9.1 Future works

We are concluding this thesis by presenting some applications and possible improvements of our methods for both long-range and close-range human monitoring application domains.

1. Our global stereo correspondence method can be improved by automatically adjust the smoothness constant values (α and β), by integrating the registration method in a multimodal tracking system using a feedback scheme. In fact, using the estimated disparities of the previous frame could be used as a prior to improve the disparity assignment for occluded people in the current frame. For example, the people that are occluded in current frame and had few pixels differences in estimated disparities at previous frame (close people in the scene), the discontinuity cost should be higher than people that are occluded and far from each other in the scene.

2. For the long-range monitoring, our proposed complete multimodal video surveillance system can be augmented by a specialized sensor fusion for specific targeted environmental factors to construct a multimodal object model as input data that helps to improve tracking results. Such a multimodal people tracking system can be used to simplify further analysis like trajectory pattern analysis to detect suspicious object trajectories for security reasons.
3. For close-range monitoring, our registered data could be used as input data for a visual diagnosis system for medical applications. For medical applications, the combination of thermal and visible data allows the rich information provided by visible cameras to be used to assist the search of thermal patterns in regions of interest on the thermal images to detect the inflammation regions for some disease diagnosis.
4. Another application of our registered thermal and visible ROI retrieved from a close-range scene is in object categorization and in analyzing the interaction of humans with other objects in their environment with the ultimate goal of building a human-machine interface that responds to different human behaviors. Also registered data can be used in a human behavioral analysis system for a video surveillance system specialized for elderly people monitoring (safety applications).

REFERENCES

BAY, H., TUYTELAARS, T. and GOOL, L. V. (2006). Surf : Speeded up robust features. *In ECCV*. 404–417.

BILODEAU, G., ST-ONGE, P. and GARNIER, R. (2011a). Silhouette-based features for visible-infrared registration. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. 68 –73.

BILODEAU, G. A., TORABI, A. and MORIN, F. (2011b). Visible and infrared image registration using trajectories and composite foreground images. *Image Vision Comput.*, 29, 41–50.

BLEYER, M. and GELAUTZ, M. (2007). Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions. *Image Commun.*, 22, 127–143.

BOYKOV, Y., VEKSLER, O. and ZABIH, R. (2001). Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23, 1222 –1239.

CALONDER, M., LEPETIT, V., STRECHA, C. and FUA, P. (2010). BRIEF : Binary Robust Independent Elementary Features. *European Conference on Computer Vision*.

CASPI, Y., SIMAKOV, D. and IRANI, M. (2006). Feature-based sequence-to-sequence matching. *Int. J. Comput. Vision*, 68, 53–64.

CHEN, H.-M., VARSHNEY, P. and SLAMANI, M.-A. (2003). On registration of regions of interest (roi) in video sequences. *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2003)*. 313 – 318.

CHRISTENSEN, H. and PHILIPS, P. (2002). Empirical evaluation methods in computer vision. eds. World Scientific Publishing Co.

COIRAS, E., SANTAMARIA, J. and MIRAVET, C. (2000). Segment-based registration technique for visual-infrared images. *Optical Engineering*, 39, 282–289.

COLLINS, R., LIPTON, A., FUJIYOSHI, H. and KANADE, T. (2001). Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89, 1456 –1477.

COMANICIU, D. and MEER, P. (1999). Mean shift analysis and applications. *The Proceedings of the Seventh IEEE International Conference on Computer Vision, (ICCV 1999)*. vol. 2, 1197 –1203 vol.2.

CONAIRE, C., O'CONNOR, N., COOKE, E. and SMEATON, A. (2006). Comparison of fusion methods for thermo-visual surveillance tracking. *9th International Conference on Information Fusion*. 1–7.

CONAIRE, C. O., O'CONNOR, N. E. and SMEATON, A. (2008). Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. *Mach. Vision Appl.*, 19, 483–494.

DALAL, N. and TRIGGS, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*. CVPR '05, 886–893.

DAVIS, J. W. and SHARMA, V. (2005). Fusion-based background-subtraction using contour saliency. *CVPR '05 : IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 11–19.

DAVIS, J. W. and SHARMA, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Comput. Vis. Image Underst.*, 106, 162–182.

DENG, Y., YANG, Q., LIN, X. and TANG, X. (2005). A symmetric patch-based correspondence model for occlusion handling. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. vol. 2, 1316 –1322 Vol. 2.

EGNAL, G. (2000). Mutual information as a stereo correspondence measure. *Tech. Rep. MS-CIS-00-20, University of Pennsylvania*.

FELZENSZWALB, P. F. and HUTTENLOCHER, D. P. (2006). Efficient belief propagation for early vision. *Int. J. Comput. Vision*, 70, 41–54.

FOOKES, C., MAEDER, A., SRIDHARAN, S. and COOK, J. (2004). Multi-spectral stereo image matching using mutual information. *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*. 961 – 968.

GIL, A., MOZOS, O. M., BALLESTA, M. and REINOSO, O. (2010). A comparative evaluation of interest point detectors and local descriptors for visual slam. *Mach. Vision Appl.*, 21, 905–920.

HAMMOUD, R. I. (2009). *Augmented Vision Perception in Infrared : Algorithms and Applied Systems*. Springer Publishing Company, Incorporated, première édition.

HAN, J. and BHANU, B. (2003). Detecting moving humans using color and infrared video. *International Conference on Multisensor Fusion*. 228–233.

HAN, J. and BHANU, B. (2007). Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40, 1771 – 1784.

HARTLEY, R. and ZISSERMAN, A. (2003a). *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, UK, seconde édition.

HARTLEY, R. and ZISSERMAN, A. (2003b). *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, UK, seconde édition.

HEIKKILA, J. and SILVEN, O. (1997). A four-step camera calibration procedure with implicit image correction. *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. 1106 –1112.

HONG, L. and CHEN, G. (2004). Segment-based stereo matching using graph cuts. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1, 74–81.

HOYER, P. O. and DAYAN, P. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5, 1457–1469.

IRANI, M. and ANANDAN, P. (1998). Robust multi-sensor image alignment. *ICCV '98 : Proceedings of the Sixth International Conference on Computer Vision*. 959–966.

JU, X., NEBEL, J.-C. and SIEBERT, J. P. (2010). *3D thermography imaging standardization technique for inflammation diagnosis*, SPIE. 266–273.

KLAUS, A., SORMANN, M. and KARNER, K. (2006). Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. vol. 3, 15 –18.

KROTOSKY, S. J. and TRIVEDI, M. M. (2007). Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106, 270 – 287.

KUMAR, P., MITTAL, A. and KUMAR, P. (2010). Addressing uncertainty in multi-modal fusion for improved object detection in dynamic environment. *Information Fusion*, 11, 311 – 324.

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

LEYKIN, A. and HAMMOUD, R. (2006). Robust multi-pedestrian tracking in thermal-visible surveillance videos. *CVPRW '06 : Conference on Computer Vision and Pattern Recognition Workshop*. 136–144.

LI, J. and ALLINSON, N. M. (2008). A comprehensive review of current local features for computer vision. *Neurocomput.*, 71, 1771–1787.

LOWE, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60, 91–110.

MAYORAL, R. and AURNHAMMER, M. (2004). Evaluation of correspondence errors for stereo. *Pattern Recognition, International Conference on*, 4, 104–107.

MIKOLAJCZYK, K. and SCHMID, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27, 1615 –1630.

MORIN, F., TORABI, A. and BILODEAU, G.-A. (2008). Automatic registration of color and infrared videos using trajectories obtained from a multiple object tracking algorithm. *Computer and Robot Vision, Canadian Conference*. 311–318.

O CONAIRE, C., COOKE, E., O'CONNOR, N., MURPHY, N. and SMEARSON, A. (2005). Background modelling in infrared and visible spectrum video for people tracking. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03*.

OGALE, A. and ALOIMONOS, Y. (2007). A roadmap to the integration of early visual modules. *International Journal of Computer Vision*, 72, 9–25.

PLUIM, J., MAINTZ, J. and VIERGEVER, M. (2003). Mutual-information-based registration of medical images : a survey. *Medical Imaging, IEEE Transactions on*, 22, 986 –1004.

SADJADI, F. (2005). Comparative image fusion analysais. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03*. 8.

SARVAIYA, J., PATNAIK, S. and BOMBAYWALA, S. (2009). Image registration by template matching using normalized cross-correlation. *Advances in Computing, Control, Telecommunication Technologies, 2009. ACT '09. International Conference on*. 819 –822.

SCHARSTEIN, D. and SZELISKI, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47, 7–42.

SHECHTMAN, E. and IRANI, M. (2007). Matching local self-similarities across images and videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*. 1 –8.

SHOUSHTARIAN, B. and BEZ, H. E. (2005). A practical adaptive approach for dynamic background subtraction using an invariant colour model and object tracking. *Pattern Recogn. Lett.*, 26, 5–26.

SOCOLINSKY, D. (2007). Design and deployment of visible-thermal biometric surveillance systems. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*. 1 –2.

SUN, J., LI, Y., BING, S. and YEUNG SHUM, K. H. (2005). Symmetric stereo matching for occlusion handling. *In CVPR*. 399–406.

SUN, J., ZHENG, N.-N. and SHUM, H.-Y. (2003). Stereo matching using belief propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25, 787 – 800.

TAPPEN, M. and FREEMAN, W. (2003). Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. 900 –906.

TORABI, A. and BILODEAU, G.-A. (2009). A multiple hypothesis tracking method with fragmentation handling. *Computer and Robot Vision, 2009. CRV '09*. 8–15.

TORABI, A. and BILODEAU, G.-A. (2011). Local self-similarity as a dense stereo correspondence measure for thermal-visible video registration. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. 61 –67.

TORABI, A., MASSE, G. and BILODEAU, G.-A. (2010). Feedback scheme for thermal-visible video registration, sensor fusion, and people tracking. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. 15 –22.

TORABI, A., MASSE, G. and BILODEAU, G.-A. (2012). An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116, 210 – 221.

TORABI, A., NAJAFIANRAZAVI, M. and BILODEAU, G. (2011). A comparative evaluation of multimodal dense stereo correspondence measures. *Robotic and Sensors Environments (ROSE), 2011 IEEE International Symposium on*. 143 –148.

TRIVEDI, M. M., MEMBER, S., CHENG, S. Y., MALCOLM, E., CHILDERS, E. M. C., KROTOSKY, S. J., MEMBER, S. and MEMBER, S. (2004). Occupant posture analysis with stereo and thermal infrared video : Algorithms and experimental evaluation. *IEEE Trans. Veh. Technol*, 53, 1698–1712.

TRUCCO, E. and VERRI, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall PTR.

VEDALDI, A., GULSHAN, V., VARMA, M. and ZISSERMAN, A. (2009). Multiple kernels for object detection. *IEEE 12th International Conference on Computer Vision (ICCV 2009)*. 606 –613.

WALK, S., MAJER, N., SCHINDLER, K. and SCHIELE, B. (2010). New features and insights for pedestrian detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*. 1030–1037.

WEISS, Y. and FREEMAN, W. (2001). On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *Information Theory, IEEE Transactions on*, 47, 736 –744.

YANG, J., LI, Y., TIAN, Y., DUAN, L. and GAO, W. (2009a). Group-sensitive multiple kernel learning for object categorization. *IEEE 12th International Conference on Computer Vision (ICCV 2009)*. 436 –443.

YANG, Q., WANG, L. and AHUJA, N. (2010). A constant-space belief propagation algorithm for stereo matching. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. 1458 –1465.

YANG, Q., WANG, L., YANG, R., STEWENIUS, H. and NISTER, D. (2009b). Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31, 492 –504.

ZHU, Z. and HUANG, T. (2007). Multimodal surveillance : An introduction. *IEEE Conference on Computer Vision and Pattern Recognition*. 1–6.