International Journal of Computer Vision (2019) 127:1044–1062
https://doi.org/10.1007/s11263-018-01141-5

Online Mutual Foreground Segmentation for Multispectral Stereo Videos

Pierre-Luc St-Charles[1] · Guillaume-Alexandre Bilodeau[1] · Robert Bergevin[2]

Received: 20 February 2018 / Accepted: 14 December 2018 / Published online: 9 January 2019

Abstract

The segmentation of video sequences into foreground and background regions is a low-level process commonly used in video content analysis and smart surveillance applications. Using a multispectral camera setup can improve this process by providing more diverse data to help identify objects despite adverse imaging conditions. The registration of several data sources is however not trivial if the appearance of objects produced by each sensor differs substantially. This problem is further complicated when parallax effects cannot be ignored when using close-range stereo pairs. In this work, we present a new method to simultaneously tackle multispectral segmentation and stereo registration. Using an iterative procedure, we estimate the labeling result for one problem using the provisional result of the other. Our approach is based on the alternating minimization of two energy functions that are linked through the use of dynamic priors. We rely on the integration of shape and appearance cues to find proper multispectral correspondences, and to properly segment objects in low contrast regions. We also formulate our model as a frame processing pipeline using higher order terms to improve the temporal coherence of our results. Our method is evaluated under different configurations on multiple multispectral datasets, and our implementation is available online.

Keywords Video object segmentation · Cosegmentation · Multispectral imagery · Energy minimization · Video signal processing

1 Introduction

The detection and segmentation of objects of interest based on motion analysis in video sequences is a fundamental early vision task. In the context of video surveillance and intelligent environments, objects of interest (or "foreground" objects) are disruptors that temporarily break the natural state of the observed scene (the "background"). Several types of approaches exist to classify image regions as being "of interest" based on this criteria (see Bouwmans 2014; Perazzi et al. 2016). While these all have different qualities, they suffer from the same fundamental drawback: if the contrast between an observed object and the background becomes too low, our ability to detect and segment it automatically deteriorates. This problem is not specific to the visible light spectrum, as this camouflaging can occur with any imaging modality.

However, interestingly, the phenomena describing the appearance of an object and the conditions under which it becomes harder to identify are rarely shared across several imaging modalities. This is especially true when considering for example the visible and Long-Wavelength Infrared (LWIR) spectra, as the correlation between the temperature of an object and its visible appearance is very weak (see Bilodeau et al. 2011). We show an example of this in Fig. 1. In fact, many surveillance systems rely on the complementarity of these two imaging modalities to detect abnormal events: the visible spectrum can easily identify large objects near ambient temperatures (e.g. vehicles), and the LWIR spectrum can easily identify objects that exhibit abnormal temperatures (e.g. animals, engine parts).

Communicated by Scharstein.

✉ Pierre-Luc St-Charles
pierre-luc.st-charles@polymtl.ca

Guillaume-Alexandre Bilodeau
guillaume-alexandre.bilodeau@polymtl.ca

Robert Bergevin
robert.bergevin@gel.ulaval.ca

[1] Polytechnique Montréal, 2900, boul. Édouard-Montpetit, Montreal, QC, Canada

[2] Université Laval, 2325, rue de l'Université, Quebec, QC, Canada

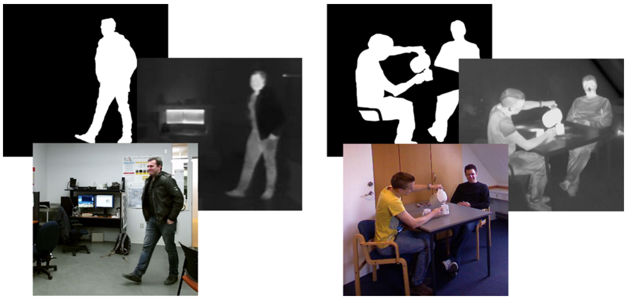

Fig. 1 Examples of mutual foreground segmentation in low contrast conditions for RGB-LWIR image pairs. On the left, the person is only partly perceptible in the LWIR spectrum due to a winter coat, but is clearly perceptible in the visible spectrum. The opposite is true on the right, where legs are hard to perceive in the visible spectrum, but easy to perceive in the LWIR spectrum

Integrating data captured from different spectral bands to attain benefits in recognition tasks is however not trivial. If the optical axes of the sensors are not already aligned using a beam splitter, a registration method has to be used to bring data points back into a common coordinate system. The image registration problem has been thoroughly studied for identical sensor pairs, but multispectral registration is fundamentally more challenging (c.f. Zitová and Flusser 2003). Since the appearance of objects cannot be directly relied upon to find local correspondences, higher level image features such as edges have to be used instead. These are typically harder to compute, and often result in a loss of registration accuracy at the pixel level when parallax effects are not negligible.

Past research has focused mostly on the problems of binary (or foreground-background) segmentation and multispectral image fusion/registration as separate issues. Yet, holistic approaches such as the ones of Torabi et al. (2012) and Zhao and Sen-Ching (2014) can outperform combinations of distinct methods on identical tasks. These holistic approaches first optimize registration using foreground object contours or trajectories as high-level features, and then use integrated image data to improve their segmentation. Solving both problems at once would be more beneficial, but this goal implies a "*chicken-and-egg*" dilemma: the result of one task is needed to obtain the other. An ideal holistic method should thus adopt an iterative optimization approach to resolve this issue. In the case of video sequences, proposed solutions should also consider the temporal redundancy of data to improve their performance. Finally, in the context of surveillance applications, the entire process should function without any human supervision, and allow frame pairs to be processed one at a time.

In this paper, we propose a holistic method to address both segmentation and registration problems by inferring their solutions alternately using move-making algorithms on a set of conditional random fields. We use self-similarity descriptors and shape cues to find proper pixel-level matches across imaging modalities in non-planar scenes, and integrate image data to improve foreground-background partitioning. This integration is achieved by iteratively refining local color models and shape contour positions while continuously realigning data sources. Our two goals are formulated as distinct energy minimization problems, and we use provisional inference results as dynamic priors to converge to a global solution. We also rely on dynamic temporal connections updated via motion cues to improve segmentation coherence over long image sequences.

Our principled bottom-up approach requires no human intervention, and relies on no prior knowledge of the foreground objects' nature. Our models are formulated so that imaging modalities can be combined without assumptions about their specific characteristics, as image regions containing discriminative data are automatically identified. This power of discrimination is exploited to scale the importance of each imaging modality when registering and integrating pixel-level data. It is also used to speed up shape contour evolution in low contrast regions by reducing penalties for label discontinuities when the other view possesses strong intensity gradients in its corresponding regions. Besides, we tackle foreground-background segmentation in the general case of video surveillance, meaning we assume the scene might contain multiple foreground objects at different depths and scales, and that they might not always be moving. This differs significantly from traditional cosegmentation methods, as we make no assumption regarding the distribution of foreground and background regions in the observed scene.

Through our experiments, we show that our primary goal, mutual foreground segmentation, can be achieved efficiently despite low contrast and other adverse conditions in both visible and LWIR images. Performance evaluations show that our approach outperforms both supervised and unsupervised monocular segmentation methods in terms of F_1 score on the VAP dataset of Palmero et al. (2016). Compared to the recent video segmentation method of St-Charles et al. (2016), our method improves its average F_1 score by 13%, from 0.766 to 0.866. To help future benchmarking on this task, we offer a new multispectral video dataset for the simultaneous evaluation of registration and segmentation performance.[1] Finally, we also offer our source code and testing framework online.[2]

Note that our method was previously introduced (St-Charles et al. 2017). Here, beyond presenting an extended description of our approach, we introduce a new spatiotemporal term to our model and study its effect on segmentation accuracy, we conduct an ablation study and test the sensitivity of our main parameters individually, we present new

[1] http://www.polymtl.ca/litiv/vid/index.php.

[2] https://github.com/plstcharles/litiv.

experiments on two pre-existing datasets, we introduce a new non-planar RGB-LWIR video dataset, and we provide a benchmark for the evaluation of segmentation and stereo registration on this new dataset. Our source code and annotations have been made available online for future works tackling a similar problem.

The paper is organized as follows. In Sect. 2, we present previous works related to our multispectral mutual segmentation problem, and highlight major differences. In Sect. 3, we describe our dual modeling approach, inference strategies, and implementation details. In Sect. 4, we present parameter and configuration studies, and evaluation results on three publicly available datasets. Lastly, we conclude with some remarks in Sect. 5.

2 Previous Work

The problem of foreground-background segmentation in images is difficult to tackle without some assumptions or constraints. Monocular segmentation solutions typically rely on visual saliency hypotheses (e.g. single foreground object roughly focused) or human supervision to obtain good results (Arbelaez et al. 2011; Rother et al. 2004). The same problem in the temporal domain (i.e. on image sequences) is easier to address due to the additional assumptions that can be made regarding object or scene motion.

Multiple families of methods exist in video segmentation; the main ones are listed here. Background subtraction methods work by building a model representing the background under the assumption that the camera is static. These methods then perform one-class pixel classification to label all outliers as foreground without supervision (Bouwmans 2014). These methods are favored in cases where foreground objects can temporarily become immobile, as they will retain their labeling for some time. Other video object segmentation approaches instead extend the concept of visual saliency into the temporal domain using highly connected graph structures (Perazzi et al. 2016). These approaches can usually be applied to sequences with changing viewpoints, but are computationally more demanding. Finally, motion clustering methods exist that rely on optical flow or trajectory points partitioning to identify image regions that behave differently from their surroundings (Tron and Vidal 2007). The strong link between motion partitioning and video object segmentation has also become a focus in recent years (Jain et al. 2017; Cheng et al. 2017). Also, in semi-supervised settings, approaches based on end-to-end neural networks have also become increasingly popular for single object video segmentation (Cheng et al. 2017; Caelles et al. 2017).

Foreground-background segmentation can become easier if multiple images of the object(s) of interest are available. Two families of methods have been developed for this circumstance: cosegmentation methods and mutual segmentation methods. Cosegmentation methods typically rely on visual saliency assumptions (e.g. shared foreground appearance and low background correlation across different views), and assume a single object is targeted and shared throughout all views (Rother et al. 2006; Zhu et al. 2016). Interestingly, cosegmentation methods can also work with different object instances from the same object category (Vicente et al. 2011). On the other hand, mutual segmentation methods typically assume that the same object instance is observed from multiple viewpoints, and optimize the geometric consistency of the extracted foreground region (Djelouah et al. 2015; Jeong et al. 2017; Riklin-Raviv et al. 2008). Our work falls into this second family of methods, as we assume the use of a synchronized stereo pair for data capture.

Previous mutual segmentation methods have typically focused on single-spectrum imaging (Riklin-Raviv et al. 2008; Ju et al. 2015; Bleyer et al. 2011), or have used depth sensors to solve the registration problem and to provide a range-based solution for foreground object detection (Jeong et al. 2017; Djelouah et al. 2015; Zhang et al. 2016). Of these, our proposed method is closest to the work of Riklin-Raviv et al. (2008), who termed the idea of “mutual segmentation” for objects in visible image pairs. Their approach addresses the uncertainty of object boundary localization under occlusions and noise by iteratively optimizing active contours without supervision. Their use of a biased shape term however entails that a free parameter directly controls the elimination of ambiguous shape segments in the image pair. In our work, we avoid this parameterization issue by relying on local saliency and self-refining color models to automatically integrate multiple view data. Our object contours then expand and contract until they naturally converge. Besides, the method of Riklin-Raviv et al. (2008) considers that all images are related only by planar projective homographies, and thus it cannot handle parallax issues in 3D scenes. This latter problem was addressed by Ju et al. (2015), who also proposed a contour-based modeling approach for mutual foreground segmentation in stereo pairs. This more recent approach however relies on the assumption that near-perfect foreground contours obtained via human supervision are available in at least one of the views. Lastly, the work of Bleyer et al. (2011) is also somewhat related to ours: they tackle disparity (or parallax) estimation for calibrated stereo pairs using a piecewise planar model based on object segmentation. However, their main goal is scene-wide data registration, which is very computationally demanding. According to Tippetts et al. (2016), processing an image pair took the method about 20 min. In our case, we only focus on the registration and segmentation of foreground objects classified as such in a video surveillance mindset. This makes our proposed approach much more lightweight and applicable to real data streams.

The use of multispectral data (other than RGBD) has been mostly neglected in the context of mutual segmentation or cosegmentation due to the registration problem. As stated before, this difficulty is due to the (typically) low correlation between the appearances of objects in different spectral bands (see Zitová and Flusser 2003). Beam splitters can be used to avoid the registration problem altogether (Bienkowski et al. 2012; Hwang et al. 2015). These setups are however very delicate, and they induce color distortions. Moreover, the elimination of parallax also prevents the recovery of depth information from the scene.

In practice, if the chosen spectral bands are not too distant in terms of their imaging characteristics (e.g. visible light and near-infrared), modern image descriptors and similarity measures can be used to find local correspondences with varying degrees of success (see Pinggera et al. 2012). These "close" spectrum pairs are however less interesting to integrate in machine vision systems due to their resemblance. On the other hand, traditional appearance-based matching approaches suffer when distant spectrum pairs are selected; see for example the study done for visible (RGB) and Long-Wavelength Infrared (LWIR) pairs by Bilodeau et al. (2014). Multispectral registration thus has to rely on higher level features that encapsulate raw object appearance in order to find proper local correspondences. In the recent literature, some have relied on edge matching in local neighborhoods (Coiras et al. 2000; Mouats and Aouf 2013) or in Hough space (Pistarelli et al. 2013) to resolve this problem. Edge-based approaches are however more suited to man-made environments, and underperform in more general settings (e.g. open terrain) where large intensity gradients are rarer or more weakly correlated between imaging modalities.

Other works have instead addressed the registration problem in the temporal domain by adopting motion-based cues (Torabi et al. 2012; Zhao and Sen-Ching 2014; Nguyen et al. 2016), which is more similar to our approach. In the work of Torabi et al. (2012), the trajectories of foreground objects are used for high-level registration based on the idea that position and motion are fully independent of appearance. In the works of Zhao and Sen-Ching (2014) and Nguyen et al. (2016), foreground shapes obtained via background subtraction are used for contour matching. This latter strategy has been shown to be more pixel-accurate for the registration of foreground objects, but it still depends strongly on the performance of the segmentation method used. In our proposed method, we address this problem by combining contour-based registration and segmentation into a global optimization framework.

Finally, as for the combination of multispectral registration and segmentation, we can highlight the existence of a few papers. Torabi et al. (2012) propose a solution based on object-wise planar registration, and improve segmentation masks obtained via background subtraction by combining multispectral data using a sum-rule approach. Zhao and Sen-Ching (2014) also rely on object-wise planar registration, and use multiple object trackers to improve the results of parallel segmentors *a posteriori*. In this case, the methods are run in cascade to resolve the "*chicken-and-egg*" optimization dilemma stated earlier. The strategies of Torabi et al. (2012) and Zhao and Sen-Ching (2014) do not handle occlusions well due to their high level registration approach, and only provide a single-pass improvement to the segmentation results of a given frame pair. Palmero et al. (2016) introduced a human body segmentation method for trimodal (RGBD-LWIR) image sequences based on feature fusion using a random forest classifier. They also avoid pixel-level registration by predefining a set of homographies to use at runtime based on detected foreground object depth. Davis and Sharma (2007) proposed a dual background subtraction model and contour extraction technique to improve RGB-LWIR foreground fusion based on local visual saliency evaluation. Similarly, Li et al. (2017) proposed a background subtraction method based on the low-rank decomposition of integrated RGB-LWIR pairs to improve foreground segmentation in a global framework. The main shortcoming of these latter two works is that they only handle planar scenes (i.e. scenes where parallax issues are negligible) using a single predefined homography. To the best of our knowledge, no method has previously been proposed to tackle multispectral non-planar registration and mutual foreground segmentation simultaneously.

3 Proposed Approach

Our approach can be described based on its two main components: the stereo matching model for disparity (or parallax) estimation on epipolar lines, described in Sect. 3.1, and the shape matching model for binary image segmentation, described in Sect. 3.2. These two models are conditional random fields formulated as discrete energy functions that tackle the multispectral registration and segmentation problems in an integrated fashion. Our energy functions are minimized alternately using move-making algorithms, as described in Sect. 3.3. The flowchart in Fig. 2 illustrates our approach.

We begin with an introduction of the general terms and notation used in this section. Given a set of rectified images $\mathcal{I} = \{I_k\}$ (with $k = \{0, 1\}$ in the case of a stereo pair), the disparity label space $\mathcal{L}_D = \{0, \ldots, d_{\max}\}$, and the background-foreground label space $\mathcal{L}_S = \{0, 1\}$, our goal is to find the optimal pixel-wise disparity and segmentation labelings $\mathcal{D} = \{\mathcal{D}_k\}$ and $\mathcal{S} = \{\mathcal{S}_k\}$ such that:

$$\mathcal{D}_k = \text{argmin}_{\mathcal{D}_k} E_k^{\text{stereo}}(\mathcal{D}_k), \tag{1}$$

$$\mathcal{S}_k = \text{argmin}_{\mathcal{S}_k} E_k^{\text{segm}}(\mathcal{S}_k), \tag{2}$$

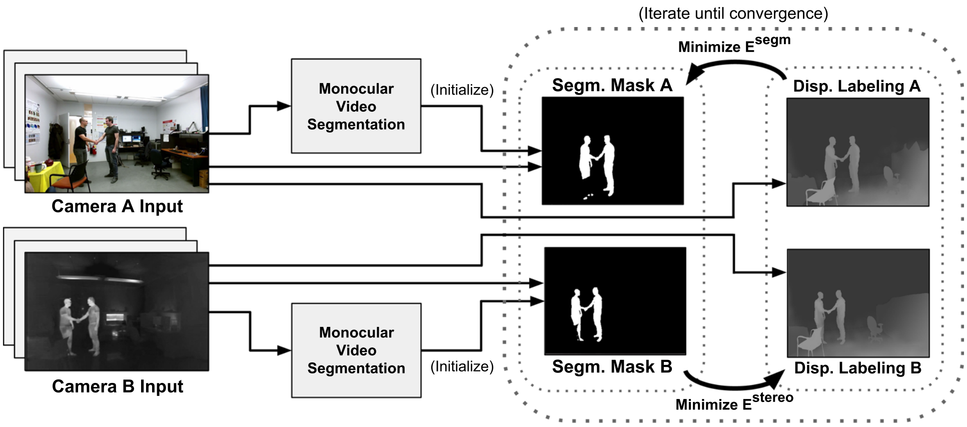

Fig. 2 Flowchart of the proposed method. A monocular video segmentation method is first used to initialize segmentation masks for both cameras individually. Then, the energies of the stereo and segmentation models (described in Sects. 3.1 and 3.2, respectively) are alternately minimized until a proper global solution is reached. The output of our method then consists of the refined segmentation masks of the input frames, and of the reciprocal disparity labelings computed for both cameras

where $D_k = \{d_p : p \in I_k, d_p \in \mathcal{L}_D\}$ is a disparity labeling, $S_k = \{s_p : p \in I_k, s_p \in \mathcal{L}_S\}$ is a segmentation labeling (or mask), and where the energy cost functions E_k^{stereo} and E_k^{segm} are described in Sects. 3.1 and 3.2, respectively. For now, note that these functions are linked through their estimation results, $\mathcal{D}_k$ and $\mathcal{S}_k$, which are used as dynamic priors throughout the minimization. In other words, disparity labels d_p for each pixel p in I_k are used in E_k^{segm} for appearance data integration, and segmentation labels s_p are used in E_k^{stereo} to improve stereo matching. Lastly, note that we sometimes omit the k subscript in the following subsections to simplify the notation, as most equations only deal with one image of the stereo pair at a time.

3.1 Stereo Registration Model

We tackle the multispectral stereo registration problem for non-planar scenes using a sliding window strategy for pixel matching. This search for correspondences is limited to an horizontal axis on the image plane due to epipolar geometry constraints. These constraints restrict the disparity (or parallax) between the 2D projections of an observed 3D object point to one dimension (see Hartley and Zisserman 2003). In short, given the intrinsic and extrinsic parameters of the stereo pair obtained via calibration, we can rectify the input images. This forces the corresponding projection of a 2D point in one view to be located somewhere on the same horizontal line in the other view. While calibration does require human intervention, it is a one-time effort generally accepted in an unsupervised system. It could also be replaced by an automatic approach (e.g. Nguyen et al. 2016).

For a pixel-wise disparity label map D, we define its energy (or cost) to be minimized as

$$E^{\text{stereo}}(D) = E^{\text{appearance}}(D) + E^{\text{shape}}(D) + E^{\text{uniqueness}}(D) + E^{\text{smooth1}}(D). \quad (3)$$

Each term in this cost function is crafted to promote a desired property of the output disparity labeling, and is described in detail in the following paragraphs. The first three terms are unary costs summed over all pixels of the image. The appearance and shape terms evaluate the local affinity between a pixel p and its corresponding pixel shifted by d_p in the other view. The uniqueness term penalizes multiple matches with p in the other view. The last term is a sum of pairwise smoothness costs used to penalize irregular disparities in uniform image regions. Note that in order to maximize processing speed for image pair sequences, we keep our stereo model simple. Our results would undoubtedly improve with second-order terms such as those of Woodford et al. (2009) or Kohli et al. (2009), but at an important increase in computational complexity. Moreover, since we only focus on the registration of foreground objects, higher-order surface smoothness priors are not as important here.

3.1.1 Appearance and Shape Terms

These two terms convey the cost of matching an image patch centered on a pixel $p \in I$ to another one in the second view which is offset according to its disparity label d_p. The terms are both defined as

$$E^{\{\text{appearance, shape}\}}(D) = \sum_{p \in I} \mathcal{A}\big(p, r(p, d_p)\big) \cdot \mathcal{W}(p), \quad (4)$$

where $r(p, d_p)$ returns the pixel location in the other view obtained by shifting p by d_p on its epipolar line, $\mathcal{A}(p, q)$ encodes the affinity cost for matching descriptor patches centered at p and q in each image, and $\mathcal{W}(p)$ encodes the saliency coefficient for pixel p (detailed further down). For the appearance term, the affinity cost map $\mathcal{A}$ is obtained by densely computing local image descriptors over I_0 and I_1, and by matching them using L2 distance in 15x15 patches to dampen noise. As stated in Sect. 2, classic appearance-based descriptors are not ideal for wide spectrum pairs such as RGB-LWIR. To address this issue, we used Dense Adaptive Self-Correlation descriptors (DASC; Kim et al. 2015), which are based on self-similarity measures. We also tested the Local Self-Similarity descriptor (LSS; Shechtman and Irani 2007) during our preliminary experiments, and found a slight decrease in terms of overall registration performance. For the shape term, we densely compute Shape Context descriptors (Belongie et al. 2002) over S_0 and S_1, which are the provisional segmentation masks. We then match these descriptors using the same approach as for the appearance term to obtain the shape affinity cost map $\mathcal{A}$. Our hypothesis here is that the combination of these two types of descriptors can provide better matching results than either one alone. However, remember that multispectral matches are often unreliable due to non-discriminative descriptors in uniform image regions or in regions with very low multispectral correlation. To avoid increasing pixel matching penalties in such cases, we multiply the affinity cost by a local saliency coefficient. In both the appearance and shape terms, this local saliency coefficient for a given pixel p is defined as

$$\mathcal{W}(p) = \max\left\{\mathcal{H}\left(\left[\mathcal{A}\left(p, r(p,d)\right)\forall d \in \mathcal{L}_D\right]\right), \mathcal{H}\left(K(p)\right)\right\}, \tag{5}$$

where $K(p)$ returns the matrix of local descriptors in the patch centered on pixel p, and $\mathcal{H}(\cdot)$ computes the sparseness metric of Hoyer (2004) over a vector or matrix. This metric returns a value $\in [0, 1]$, meaning $\mathcal{W}(p)$ is also in that interval. In simple terms, if all affinity values are uniform (i.e. all disparity offsets have the same cost), and if the local patch's descriptor bins are all uniform, then $\mathcal{W}(p)$ will take a low value. In turn, this will lower the cost for d_p evaluated through the affinity map $\mathcal{A}$, and make local labeling depend more on neighboring decisions through the smoothness term. A simplified case of this is illustrated in Fig. 3. Besides, note that in E^{shape}, we nullify the saliency outside foreground regions to avoid influencing background disparity estimation around object contours. We can assume that disparity estimation for background regions will be less accurate due to this missing term contribution, but since we focus on the registration of foreground shapes, this is inconsequential. We study the individual contributions of the appearance and shape terms to the overall performance of our approach in Sect. 4.

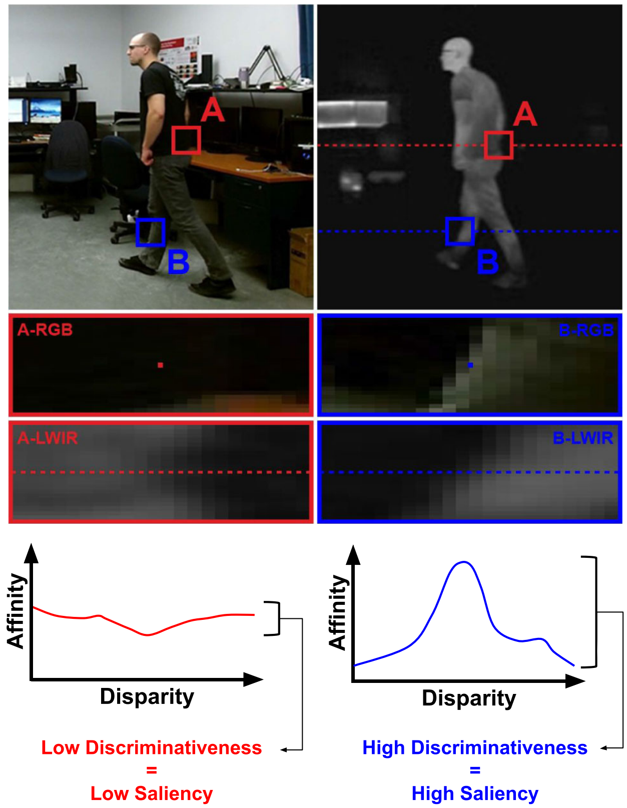

Fig. 3 Simplified case of saliency evaluation during a correspondence search on an epipolar line. On the left, for the "A" pair, low contrast in one image leads to roughly uniform affinity scores and matching costs, which translate into a low local saliency value. On the right, for the "B" pair, good contrast leads to varied affinity scores and matching costs, and a high local saliency value

3.1.2 Uniqueness Term

This unary term is used to penalize having multiple epipolar correspondences tied to the same pixel. This helps spread and equalize disparity labels in occluded and weakly discriminative image regions. Our formulation for this term is different from the classic mutual exclusion constraint proposed by Kolmogorov and Zabih (2001), which assigns an infinite cost to all extra correspondences found for a pixel p. Instead, we rely on a soft constraint that permits many-to-one correspondences with gradually increasing costs. This strategy allows our stereo model to temporarily stack extra correspondences during label swaps if the extra cost is worth absorbing. This translates into faster and larger label moves in the early steps of our inference approach, and redistribution of extra correspondence costs over future iterations. Since our method only requires a rough registration of foreground shapes to start properly segmenting them, this allow us to bootstrap the segmentation model without spending too much time on disparity estimation. We define the uniqueness cost for a pixel p as

$$\mathcal{U}(p) = \begin{cases} \sum_{n=1}^{N(p)-1} \frac{w \cdot n}{w+n-1} & \text{if } N(p) > 1 \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where $N(p)$ returns p's current correspondence count with pixels in the other view, and w is a small weight (we used $w = 3$ in our tests). For this to work, we need to keep track of pixel correspondence counts ($N(p)$) as latent variables in our model. However, since we use a move-making strategy for model inference, many correspondences might be removed in a single iteration. This makes the total cost of a move over several pixels hard to predict with (6) due to its nonlinearity. To solve this problem, we define our uniqueness term as

$$E^{\text{uniqu.}}(D) = \lambda_u \cdot \sum_{p \in I} \left(\frac{-\mathcal{U}\big(r(p,d'_p)\big)}{N\big(r(p,d'_p)\big)} + \frac{w \cdot N\big(r(p,d_p)\big)}{w+N\big(r(p,d_p)\big)-1} \right), \tag{7}$$

where d'_p is the previous disparity label of p, and λ_u is a fixed scaling factor. Note that we specify the values used for important factors such as λ_u in Sect. 3.3, and test their contribution to overall performance in Sect. 4.4. The formulation behind (7) provides the worst-case energy variation between two labeling states, and guarantees that estimated label update costs provided to the move-making algorithm will always be similar but greater than the evaluated costs once the full move is complete. The left term of the sum corresponds to the energy refunded if a previous pixel correspondence is broken, and the right term corresponds to an increase due to a new correspondence. The approximation of the true energy variation is required so that the optimizer always minimizes (3), which lets us avoid having to fall back to an older labeling state if the total energy increases. We further explain why this estimation is needed via an example in Fig. 4.

3.1.3 Smoothness Term

Lastly, we rely on a classic truncated pairwise (first-order) smoothness term to enforce the spatial coherence of our model. This term penalizes cases where neighboring pixels have irregular disparity labels despite being located in a roughly uniform image region, as described by a weak local gradient magnitude. If the gradient detected between the two pixels is instead strong, the penalty is lowered, as object edges more likely correspond with breaks in labeling. We define this term as

$$E^{\text{smooth1}}(D) = \lambda_{s1} \cdot \sum_{\langle p,q \rangle \in \mathcal{N}} \min\Big(|d_p - d_q|, 10\Big)^2 \cdot G^s_I(p,q), \tag{8}$$

with

$$G^s_I(p,q) = \max\left(\exp\left(1 - \frac{|\nabla I(p,q)|}{g}\right) - 0.5,\ 0\right), \tag{9}$$

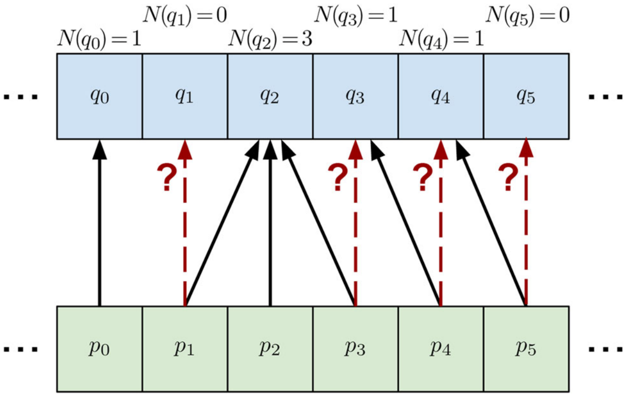

Fig. 4 Example showing why an approximation of the uniqueness cost variation must be used under a move-making optimization approach. The two rows are epipolar lines whose pixels have to be matched individually. Already established correspondences are shown with solid black arrows, and move proposals are shown with dashed red arrows. The proposals all originate from a single disparity label for each move, which in this case is $d = 0$, meaning $r(p_x, d) = q_x$. Here, the move operation could lower $N(q_2)$ (and thus lower the total energy) by reassociating p_1 with q_1 and p_3 with q_3, but the energy variation induced by these swaps cannot be predetermined exactly. It will depend on how many links with q_2 are broken during the move (i.e. one or two) due to the other terms, and whether p_4 is still linked with q_4 afterwards, and so on (Color figure online)

where λ_{s1} is a fixed scaling factor, $\mathcal{N}$ is the set of first order cliques in the graph model, $\nabla I(p,q)$ returns the normalized local image gradient magnitude between pixels p and q of image I, and g is a constant value defining the expected object contour gradient magnitude (also specified in Sect. 3.3). The truncation value (10 is used) allows large discontinuities to occur by capping the maximum smoothness penalty.

3.2 Segmentation Model

Our segmentation model's role is to integrate multispectral image data so that foreground objects can be properly segmented in both views, even in low contrast imaging conditions. Our model also needs to be lightweight enough so that cost updates and inference is fast, as shape priors are continuously modified. Since our goal is to build an unsupervised approach, we initialize the priors described below using the approximate masks provided by a monocular segmentation method (i.e. the one of St-Charles et al. 2016). This method was chosen because it can detect multiple foreground objects at once, and it can keep segmenting them at least partially if they become immobile. In Sect. 4, we show that our method works even when an initialization mask is provided for only one of the two views.

We describe the energy cost of a pixel-wise segmentation proposal S as

$$\begin{aligned} E^{\text{segm}}(S) = E^{\text{color}}(S) &+ E^{\text{contour}}(S) \\ &+ E^{\text{smooth2}}(S) + E^{\text{temp}}(S) \end{aligned}. \tag{10}$$

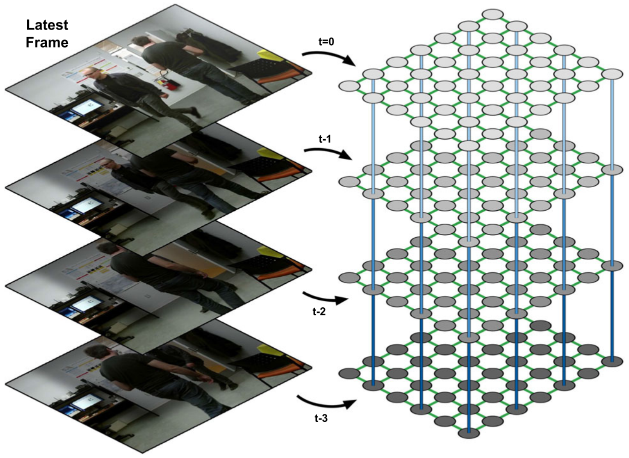

Fig. 5 Illustration of the simplified frame layering used in our segmentation model for temporal labeling refinement. In green, the first-order cliques that form E^{smooth2} are used to enforce spatial coherence in every layer. In blue, the higher order cliques that form E^{temp} are used to enforce temporal coherence across layers. Note that due to foreground motion, these cliques would not all be linked to the same underlying nodes; in reality, the links are dictated by image realignment based on optical flow (Color figure online)

Once again, the terms of this cost function are defined so that various characteristics expected of the segmentation masks can be promoted. The first two terms are unary costs summed over all pixels, and their role is to influence local segmentation decisions based on image data. The color data term maximizes the separation between the color distributions of foreground and background pixels, while the contour data term penalizes shape mismatches between the views based on distance transforms. The third term is a pairwise smoothness sum similar to (8), and is used to penalize labeling irregularities in uniform image regions. Lastly, the temporal term is a sum of higher order clique costs used to enforce temporal labeling coherence. These terms are all described in the following paragraphs. Note that due to the presence of the higher order temporal term in (10), our model is built as a multi-layer lattice, as illustrated in Fig. 5. The top layer's nodes correspond to the pixels of the latest frame of the video sequence, and lower layers' nodes correspond to the pixels of older frames. This effectively creates a pipeline where segmentation masks can be improved over time based on new image data. We discuss the improvement achieved using this approach with various pipeline depths in Sect. 4.

3.2.1 Color Term

We define the cost for this unary term using a color mixture model for each modality of the stereo pair. We employ the classic approach of Rother et al. (2004) which relies on Gaussian mixture models to represent foreground and background regions. These models can provide us with the probability that a pixel belongs to the background or foreground based on its color value. In our implementation, we use six mixture components, and use our initial and updated segmentation masks to refine our models after each iteration, in each frame. We define the color cost of all pixels as

$$E^{\text{color}}(S)=\sum_{p\in I}\begin{cases}-\log\Big(h(I_p;\boldsymbol{\beta_1},\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})\Big) & \text{if } s_p=1\\ -\log\Big(h(I_p;\boldsymbol{\beta_0},\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})\Big) & \text{otherwise}\end{cases}, \tag{11}$$

where $h(x;\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})$ returns the relative likelihood that the pixel color x fits a Gaussian mixture model with component weights $\boldsymbol{\beta}$, means $\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}$. Note that the parameter subscripts in (11) indicate that either the foreground or background model is used based on s_p. These parameters are initialized using *k-means*, and refitted after every minimization step using the new estimated segmentation masks.

3.2.2 Contour Term

Next, we define another data term that penalizes label swaps far from shape boundaries, and that combines these boundaries across the stereo pair. Its value is computed using shape distance transforms: first, we build maps in which each pixel is assigned its Euclidean distance to the closest pre-existing foreground or background pixel in the current view. We then use these maps to deduce the label update costs for each pixel in our graph, considering a mix of distances in both views at once (note the use of subscript k below). More specifically, we define our contour term as

$$E_k^{\text{cont.}}(S_k)=\lambda_c\sum_{p\in I_k}\begin{cases}F_k\big(p\big)+\lambda_m\cdot F_{k'}\big(r(p,d_p)\big) & \text{if } s_p=1\\ B_k\big(p\big)+\lambda_m\cdot B_{k'}\big(r(p,d_p)\big) & \text{otherwise}\end{cases}, \tag{12}$$

where λ_c and λ_m are fixed scaling factors, k' is the opposite index of k in the stereo pair, $F_k(p)$ returns a nonlinear distance cost (described below) for pixel p based on its distance to the closest foreground pixel in the previous segmentation of view k, and similarly for $B_k(p)$ with background pixels. Note in (12) that λ_m scales the term's multispectral cost contribution. During our tests, we give it a value $\in]0,1[$, meaning that shape contours will prefer sticking to their own previous results. This improves the stability of the segmentation while optimizing, reducing the risk of eliminating relevant shape fragments too rapidly. For the nonlinear distance cost function behind $F_k(p)$ and $B_k(p)$, we use an exponential to increase the contrast between close range and long range contour overlaps. More specifically, we use a relation of the form

$$\text{distance-cost}(p) \propto \frac{1}{\exp\big(-t(p)\big)}, \tag{13}$$

where $t(p)$ returns the actual Euclidean distance between p and its nearest pixel with a foreground or background label in the previous inference result, depending on the current value of s_p. The contour term's main responsibility is to control the evolution of object contours over several optimization passes. The multispectral contribution allows contours to be modified in regions where only one modality contributes meaningful information. The simple formulation of our contour term also avoids the needless filling of cavities, and it makes no assumption on the foreground-to-background ratio in the images.

3.2.3 Smoothness Term

This pairwise term is similar to the one used in (8); its role is to penalize label discontinuities everywhere except for image regions where local gradients are strong. In this case, however, we reuse the multispectral contribution idea of (12), and apply it to the gradient scaling factor. We define this term as

$$E_k^{\text{smooth2}}(S_k) = \lambda_{\text{s2}} \cdot \sum_{\langle p,q\rangle \in \mathcal{N}_k} \Big(s_p \oplus s_q\Big) \cdot \Big(G^s_{I_k}(p,q) + \lambda_m \cdot G^s_{I_{k'}}(p',q')\Big), \tag{14}$$

where λ_{s2} is a fixed scaling factor, $\oplus$ is the XOR operator, p' is a shorthand for $r(p, d_p)$, and q' is a shorthand for $r(q, d_q)$. In (14), the right-hand parentheses group returns the gradient coefficient with its multispectral contribution, and the left-hand group returns 1 or 0 based on whether a label discontinuity is found. As before, the use of local image gradients helps "snap" these discontinuities to real object contours. However, the multispectral contribution allows shape contours to settle in uniform regions if the other view possesses a strong local gradient there. Paired with the contour term, this allows our model to properly expand and contract shape boundaries across image modalities. We study the effect of λ_m on the performance of our method in Sect. 4.

3.2.4 Temporal Term

Lastly, we present the formulation and role of our temporal term. Unlike the other terms presented so far, this term is based on higher order cliques that are composed at runtime, and updated for each frame. The role of these cliques is to enforce spatiotemporal labeling coherence despite foreground object motion. Our graph structure can be visualized as a stack of analyzed frames; this structure is shown in Fig. 5. While the depth (or layer count) of this stack is predetermined, its temporal cliques are composed based on node realignments provided by optical flow maps. This allows cliques to remain attached to the same object part despite movement, and thus enforce labeling smoothness across frames. We compute optical flow maps using the method proposed by Kroeger et al. (2016). As for the cost term itself: given $\mathcal{C}$, the set of all temporal cliques in our model, and using the subscript l to identify different temporal layers in these cliques, we define it as

$$E^{\text{temp}}(D) = \lambda_{\text{s2}} \cdot \sum_{c \in \mathcal{C}} \sum_{l=1}^{L-1} \big(s_{c,l} \oplus s_{c,(l+1)}\big) \cdot G^t\big(c, l\big), \tag{15}$$

where λ_{s2} is the same scaling factor as in (14), L is the pipeline's depth in frames, $s_{c,l}$ returns the label of the lth node in clique c, and $G^t(c,l)$ returns a scaling factor for clique c at layer l (described next). Overall, this term is similar to the pairwise smoothness terms described earlier, but it can link more nodes together. However, instead of scaling costs via local image gradients, we rely on temporal image gradients in $G^t(c,l)$. These new gradient values are obtained by computing the absolute color differences between pixels of consecutive frames realigned using optical flow maps. Strong color differences are indicative of uncertain regions where consistency costs should be reduced due to occlusions or bad optical flow estimation. Here, similarly to (9), we define the new gradient scale term as

$$G^t(c,l) = \max\left(\exp\left(1 - \frac{|i_{c,l} - i_{c,(l+1)}|}{g}\right) - 0.5,\ 0\right), \tag{16}$$

where $i_{c,l}$ returns the color value of the lth node in clique c. We study the contribution of this new term in Sect. 4 given various layer count configurations.

3.3 Inference and Implementation Details

Simultaneously minimizing the cost functions defined in (3) and (10) is not trivial. Both functions rely on each other's provisional results as dynamic priors, and (10) contains a higher order term. A simple cost function can typically be minimized iteratively using a move-making algorithm such as α-expansion (Boykov et al. 2001) that returns a local minimum within a known factor of the global minimum. In our case, the dynamic weights and links used to connect our two cost functions cause their global objective to be updated each time a new labeling is obtained for either half of the model. This means that the global minimum of our model is always changing, and that reaching it is difficult. Instead, we focus on converging to a local minimum in each function by alternating label move operations. Recently proposed move-making algorithms can deal with higher-order terms and dynamic priors without having to resort to a move-and-check or rollback strategy (c.f. Lempitsky et al. 2010; Kappes et al. 2013).

However, to reach a local minimum in both functions simultaneously, the terms have to be carefully designed so that the cost functions can converge under roughly similar conditions. We achieve this as anticipated using shape contours: these tend to settle on the maxima in gradient intensity maps that correspond to object boundaries, and can be easily matched across image modalities despite some local shape variability. In practice, our optimization strategy converges once the target objects in the scene (roughly identified via the initialization masks) are properly covered by foreground segments that are registered between the two views. This convergence also happens without having to use a decaying metaparameter to force a solution after a fixed number of iterations.

We rely on the move-making algorithms of Fix et al. (2011) and Fix et al. (2014) for the inference of our stereo and segmentation models, respectively. Both are modified for use in a dynamic graph structure. While faster inference solutions do exist, these were deemed fast enough for our experiments, even without having to parallelize label moves. In both cases, our move proposals only consist of uniform labeling maps, meaning our inference approach is fairly similar to α-expansion. We build our graphical models in C++ using the OpenGM library (Andres et al. 2012), and reuse the same structure for all frames in a video, updating only the composition of temporal factors in (15) as required. We settled for these two generic optimizers to show that the formulation of our models is not tied to the optimization approach we use.

We tackle the alternating minimization of energies (3) and (10) for each frame of a video by first minimizing the stereo model's energy using unary terms only, or by realigning its previous disparity labeling result via optical flow. Simultaneously, the segmentation model is initialized using the masks provided by an unsupervised monocular method, as stated earlier. Then, segmentation and disparity label moves are iteratively computed in small batches until no more moves in $\mathcal{L}_S$ can reduce the energy of (10). This typically happens after less than three passes over the disparity label space ($\mathcal{L}_D$), and less than 50 moves in the segmentation label space ($\mathcal{L}_S$), the exact number depending on the quality of the initialization. For reference, with our baseline implementation, this is equivalent to approximately 30 seconds worth of processing time on a single core of a 3.7 GHz Intel i7-8700K processor for a VGA-sized image pair. This processing time seems to scale in a roughly linear fashion with respect to the number of pixels in the analyzed images.

As for the free parameters listed earlier, we use the following configuration for our tests in the next section:

- Stereo model uniqueness term weight: $\lambda_u = 0.4$
- Stereo model smoothness term weight: $\lambda_{s1} = 0.001$
- Expected object contour gradient intensity: $g = 30$
- Segmentation model contour term weight: $\lambda_c = 7$
- Segmentation model smoothness weight: $\lambda_{s2} = 7$
- Multispectral contribution term weight: $\lambda_m = 0.5$

The values listed above have been empirically found to provide good overall segmentation performance on a small subset of our test data via grid search. As previously noted, we study the effect of several of these parameters on the overall performance of our method in the next section. For optical flow and DASC descriptors computations, we kept the default parameters provided by their original authors. For Shape Context computations, we used 50 pixel-wide descriptors with 10 angular bins and 3 radial bins. For the depth of our frame processing pipeline, we used two temporal layers (i.e. the current frame and the previous one), as adding more did not improve overall performance significantly over the extra processing cost; this is discussed in Sect. 4.4. Finally, to reduce the computational cost when using higher order terms in our segmentation model, we use a stride of two pixels when creating the temporal cliques used in (15). For more implementation details, we refer the reader to our source code.[3]

4 Experiments

In this section, we first discuss our evaluation methodology, and then present evaluation results for mutual segmentation and stereo registration. Since close-range (non-planar) multispectral video datasets are quite uncommon in the literature, we had to adapt existing datasets to our problem. For multispectral mutual segmentation, we rely on a modified version of the VAP trimodal dataset of Palmero et al. (2016); the modifications we made are detailed in Sect. 4.2. For stereo registration, we rely on the benchmark of Bilodeau et al. (2014). We follow up with an ablation study of our method in which we remove key terms from our energy functions, and then study the effect of tuning key parameters of these terms. Finally, we provide evaluation results for both segmentation and stereo registration on a newly captured and annotated RGB-LWIR dataset for future comparisons.

4.1 Evaluation Methodology

Since our primary goal is mutual foreground segmentation, we employ binary classification metrics for the first part of our evaluation. Commonly used metrics in the context of video segmentation are Precision (Pr), Recall (Re), and F_1 score. These are based on three types of pixel-wise classification result counts, namely True Positives (TP), False Positives (FP), and False Negatives (FN). These metrics are defined as

[3] https://github.com/plstcharles/litiv.

Table 1 Evaluation results on the multispectral video segmentation dataset of Palmero et al. (2016)

Method	Metric	Scene 1		Scene 3		Overall		
		Visible	LWIR	Visible	LWIR	Visible	LWIR	Average
St-Charles et al. (2016) (unsupervised)	*Pr*	0.820	0.755	0.716	0.514	0.768	0.635	0.701
	Re	0.810	**0.975**	0.688	**0.969**	0.749	**0.972**	0.861
	F_1	0.815	0.851	0.702	0.672	0.758	0.762	0.760
Palmero et al. (2016) (semi-supervised)	*Pr*	–	–	**0.817**	0.777	–	–	–
	Re	–	–	0.568	0.564	–	–	–
	F_1	–	–	0.670	0.654	–	–	–
Rother et al. (2004) (GrabCut; supervised)	*Pr*	0.685	0.808	0.653	**0.847**	0.669	**0.828**	0.748
	Re	0.759	0.896	**0.929**	0.916	0.844	0.906	0.875
	F_1	0.721	0.850	0.767	**0.880**	0.744	**0.865**	0.804
Proposed method (unsupervised)	*Pr*	**0.894**	**0.860**	0.788	0.749	**0.841**	0.804	**0.821**
	Re	**0.902**	0.901	0.918	0.937	**0.910**	0.919	**0.914**
	F_1	**0.898**	**0.880**	**0.848**	0.833	**0.873**	0.857	**0.866**

Bold results are the best in that category across all methods

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{18}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{19}$$

In all three cases, higher values indicate better performance. The F_1 score corresponds to the harmonic mean of the precision and recall scores. We use it as an overall indicator of binary segmentation performance, as it was shown in the work of Goyette et al. (2012) to be strongly correlated with the final ranking of methods on a large binary segmentation dataset based on numerous other metrics.

Our second goal is to evaluate stereo registration performance. For this, we employ the strategy of the Middlebury dataset (Scharstein et al. 2014), and report the percentage of pixels labeled with disparity errors larger than some fixed distance thresholds (in pixels). We also report average frame-wide pixel disparity errors, noted $\bar{d}_{\text{err}}$ below. In this case, lower values indicate better performance.

4.2 VAP 2016 Dataset

For this first part of our evaluation, we adapted the dataset of Palmero et al. (2016) to our needs. This dataset was originally intended for the trimodal (RGBD-LWIR) detection and segmentation of people in images, and it is provided as a set of videos. It consists of 5724 image triplets split into three scenes, with their associated groundtruth foreground-background segmentation masks. We obtained the calibration data used by the original authors to roughly register scene contents via homographies, and rectified all RGB and LWIR image pairs using the OpenCV calibration toolbox. The depth images were left unused during all our experiments, and the second scene was removed due to missing calibration data. Finally, to avoid skewing the performance evaluation by continuously segmenting empty frames or frames with purely static and/or unoccluded foreground regions, we manually selected a subset of groundtruth masks for our experiments. These masks were picked at a rate of roughly 2 Hz from all originally available masks while focusing on time spans with people interacting.

We present the segmentation performance of our proposed method, as well as the performance of baseline video and image segmentation methods in Table 1. We could not evaluate the performance of the works listed in the last paragraph of Sect. 2 that simultaneously tackle segmentation and registration due to a lack of open-source code and datasets. Besides, comparing our results to those of other methods that assume single-spectrum data or planar scenes would also be unfair. For the video segmentation baseline, we rely on the method of St-Charles et al. (2016), which is fully unsupervised. We use the method's default parameters from its original implementation, and process each spectrum individually. For the image segmentation baseline, we rely on the GrabCut method of Rother et al. (2004), and provide this method with manually defined bounding boxes for all foreground objects. We used OpenCV's GrabCut implementation, and ran five iterations per image. Finally, we provide partial results for the method of Palmero et al. (2016) that were obtained using the original predictions provided by the authors.

We can observe that our proposed method outperforms the unsupervised video segmentation approach of St-Charles et al. (2016) in terms of overall F_1 score by a margin of 0.1, equal to a relative improvement of over 13%. This confirms that our approach can properly integrate multispectral information through stereo registration in order to improve segmentation performance beyond that of a state-of-the-art

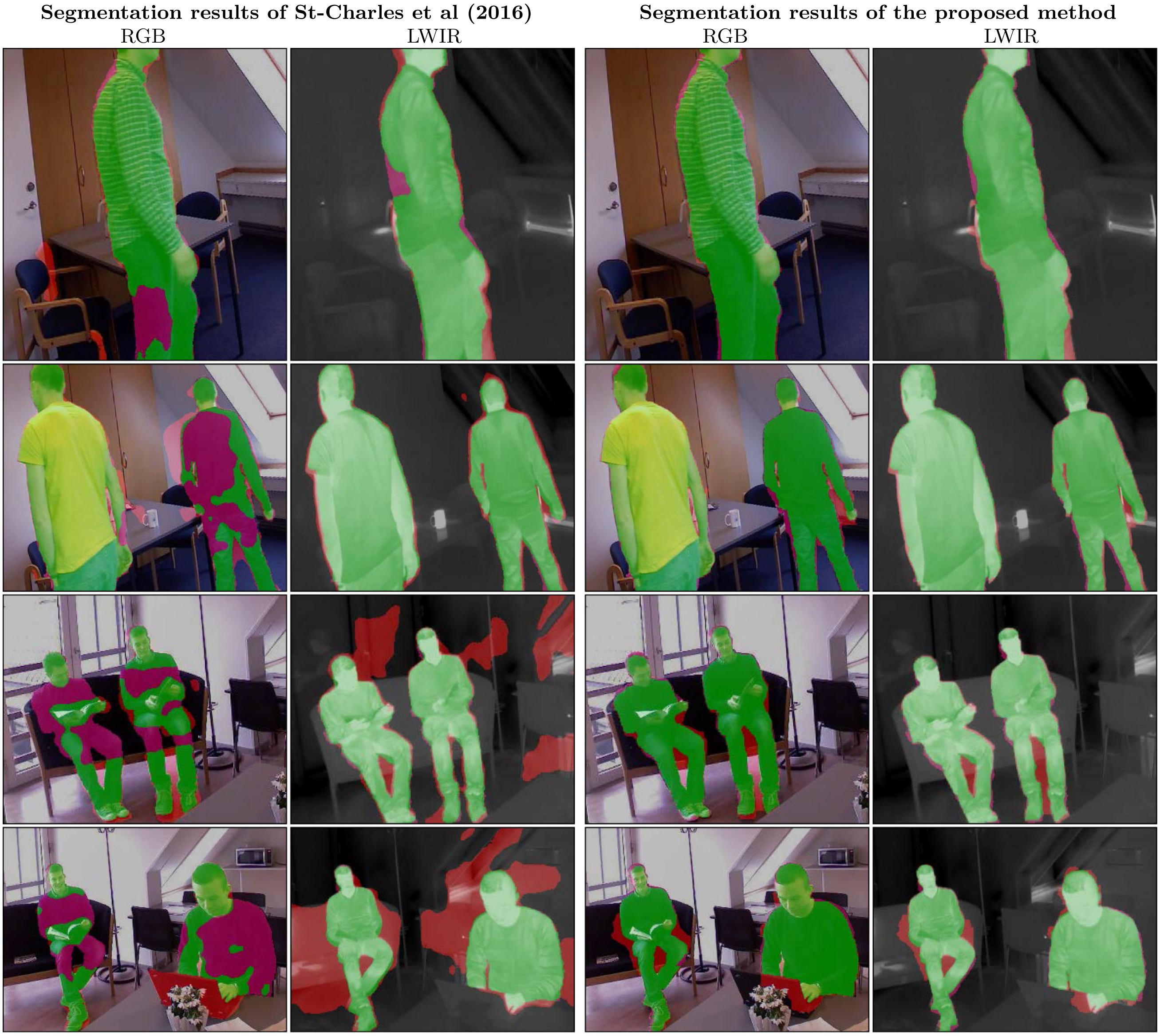
Segmentation results of St-Charles et al (2016)
RGB
LWIR
Segmentation results of the proposed method
RGB
LWIR

Fig. 6 Examples of typical segmentation results from the VAP dataset of Palmero et al. (2016); the left two columns show the segmentation masks obtained via the method of St-Charles et al. (2016) and used to initialize our method, and the right two columns show our final segmentation masks. Image regions properly classified as foreground are highlighted in green over the original images, while regions highlighted in orange and magenta show false positives and false negatives, respectively. Images have been cropped to show more details (Color figure online)

monocular method. Interestingly, our proposed method even outperforms the supervised image segmentation approach of Rother et al. (2004), which relies on manual annotations to pinpoint all foreground objects in every frame. This can be explained by the fact that foreground objects in this dataset have better contrast in the LWIR spectrum than in the visible spectrum, and because our approach propagates this contrast information across the stereo pair. Additionally, our method outperforms the semi-supervised approach of Palmero et al. (2016) in Scene 3 despite having to estimate full disparity maps for stereo registration, and without requiring training. Finally, we show in Fig. 6 some qualitative results for this dataset. The last row of this figure presents an interesting case: in this frame pair, the initial foreground masks provided to our method both contain important errors in different regions, but the output is excellent. This shows that despite not having a proper foreground shape template, the real underlying shape can be found and extracted correctly via our iterative process.

4.3 Bilodeau et al. 2014 Dataset

We now evaluate our proposed method's stereo registration accuracy using the benchmark dataset of Bilodeau et al.

(2014). This dataset was originally intended for the evaluation of image descriptors and similarity measures in the context of multispectral stereo matching, once again provided as a set of videos. It consists of 5390 RGB-LWIR frame pairs split into three scenes, with over 25,000 sparse correspondences annotated on visible foreground objects.

As stated before, we evaluate performance on this dataset by analyzing the accuracy of disparity labelings. Unfortunately, previous works tackling multispectral registration have often relied on their own foreground overlap ratios to assess their performance (e.g. Nguyen et al. 2016), meaning comparisons here are impossible. Here, to provide a reusable evaluation baseline, we compare our results to those obtained using a sliding window patch-matching approach, similar to the strategy used by Bilodeau et al. (2014). In short, local disparity labels are assigned based on the best match (or smallest distance) found between image patches in a winner-takes-all fashion. To describe the similarity between these image patches, we rely on descriptors, namely LSS (Shechtman and Irani 2007) and DASC (Kim et al. 2015), and on Mutual Information scores (MI; Maes et al. 1997). Note that for these experiments, we used the same metaparameters (e.g. patch size, bin counts) as those used by our own method, or translated them to be roughly equivalent. Also, for fairness, we relied on the same smoothness term we used in our own method (E^{smooth1}) to regularize the patch matching disparity estimation results. Finally, to highlight the issue of applying traditional stereo registration methods on multispectral datasets, we evaluate the block matching algorithm of K. Konolige implemented in OpenCV. These results are presented in Table 2.

We can note that our proposed method performs very well compared to the baseline methods. Unsurprisingly, OpenCV's block matching method fails on this dataset as it tries to compare image textures directly across the pair, despite their low correlation. The approaches based on self-similarity descriptors (LSS, DASC) and mutual information perform slightly better, but still produce highly inaccurate results. On average, above 50% of all the evaluated points are labeled with disparities at least four pixels off from the groundtruth. On the other hand, our approach manages to label 51.8% of all evaluated points within a single pixel of the groundtruth, and provides an average disparity error of only 3.21 pixels. Note however that while this performance is good enough for our primary task (mutual foreground segmentation), it is still far from the current state-of-the-art in single-spectrum stereo registration. For example, on the Middlebury dataset (Scharstein et al. 2014), top-performing methods typically label less than 20% of all points with a disparity error larger than a single pixel. This highlights the difficulty of multispectral stereo registration.

Table 2 Evaluation results on the multispectral video registration dataset of Bilodeau et al. (2014)

Method	Video 1			Video 2			Video 3			Overall		
	% err. > 1px	% err. > 4px	$\bar{d}_{\text{err}}$	% err. > 1px	% err. > 4px	$\bar{d}_{\text{err}}$	% err. > 1px	% err. > 4px	$\bar{d}_{\text{err}}$	% err. > 1px	% err. > 4px	$\bar{d}_{\text{err}}$
OpenCV's Block Matcher	95.5	95.3	27.51	99.8	99.7	38.99	99.8	99.6	34.76	98.3	98.2	33.75
LSS Sliding Window	69.1	45.7	8.50	89.4	77.3	9.87	73.6	36.0	7.50	77.4	53.0	8.62
MI Sliding Window	82.0	62.5	10.20	86.9	61.4	9.78	84.4	63.2	10.08	84.4	62.4	10.02
DASC Sliding Window	79.2	55.0	8.94	77.4	55.9	9.11	73.5	42.1	6.88	76.7	51.0	8.31
Proposed method	**47.7**	**17.3**	**3.28**	**55.6**	**25.4**	**3.11**	**52.0**	**17.5**	**3.26**	**51.8**	**20.1**	**3.21**

Bold results are the best in that category across all methods

Table 3 Overall performance for various configurations of the proposed method on the datasets of Palmero et al. (2016); Bilodeau et al. (2014)

Method configuration	$\bar{d}_{\text{err}}$	F_1
No shape term $\left(E^{\text{shape}}\right)$	8.71	0.860
No appearance term $\left(E^{\text{appearance}}\right)$	3.69	0.851
No saliency maps $(\mathcal{W})$	3.47	0.856
No uniqueness term $\left(E^{\text{uniqueness}}\right)$	3.33	0.865
No color term $\left(E^{\text{color}}\right)$	3.46	0.822
No contour term $\left(E^{\text{contour}}\right)$	4.16	0.624
No temporal term $\left(E^{\text{temporal}}\right)$	3.29	0.855
No initial LWIR segm. mask	10.82	0.820
No initial visible segm. mask	8.32	0.800
Default configuration	3.21	0.866

4.4 Parameters and Ablation Study

In this section, we study the behavior of our method when key terms and parameters are modified from the default configuration listed in Sect. 3.3 on the two previously introduced datasets. First, we perform an ablation study to determine which energy terms are the most important in our models; this study is presented in Table 3.

According to the F_1 scores, modifying the stereo energy formulation only has a small effect on segmentation performance. On the other hand, removing the color or contour terms from the segmentation energy has larger impacts, and the latter of the two is the most important contributor to overall performance. As for the registration performance, the shape term seems to be the most important, but all terms contribute to the overall performance of the method. The positive contribution of both appearance and shape terms also confirms the hypothesis set in Sect. 3.1. Besides, interestingly, when our model is initialized in only one of the two modalities using approximative masks, its segmentation performance is still at least as good as GrabCut's (as reported in Table 1). This highlights the robustness of our approach, and shows that it can perform well even in adverse initialization conditions.

Next, we show the effect of parameter tuning. The segmentation and registration performance for our proposed method in terms of overall F_1 score and average disparity error ($\bar{d}_{\text{err}}$, in pixels) is presented for various configurations in Fig. 7. Note that we roughly tuned our method with segmentation performance as a priority to obtain our default configuration. Nonetheless, registration performance is usually near-optimal or stable around the same parameter values. In general, we can note that the choice of parameters does not seem to drastically alter our method's performance, as both metrics fairly remain stable over large value intervals.

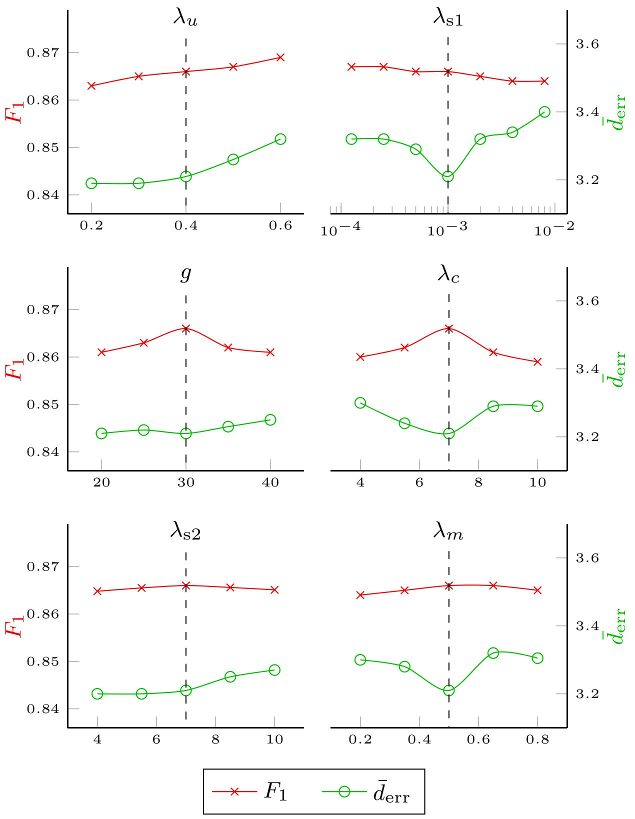

Fig. 7 Overall performance for various parameter values of the proposed method on the datasets of Palmero et al. (2016); Bilodeau et al. (2014). The default configuration of each parameter is shown with the dashed line. Remember that for F_1, higher is better, and for $\bar{d}_{\text{err}}$, lower is better

Table 4 Overall segmentation performance for various temporal pipeline depths on the dataset of Palmero et al. (2016)

Method configuration	*Pr*	*Re*	F_1
2 Layers, real-time	0.817	0.910	0.863
3 Layers, real-time	0.821	0.915	0.866
4 Layers, real-time	0.825	0.918	0.867
5 Layers, real-time	0.826	0.918	0.868
2 Layers, deferred	0.821	0.914	0.866
3 Layers, deferred	0.824	0.920	0.870
4 Layers, deferred	0.827	0.921	0.870
5 Layers, deferred	0.826	0.919	0.868
No temporal term	0.801	0.919	0.855

Finally, in Table 4, we evaluate our approach configured with different temporal pipeline depths, and while allowing deferred output or not. The notion of "pipeline depth" here corresponds to the number of edges in the higher order temporal terms introduced in Sect. 3.2. Deferred segmentation outputs are masks generated by our method with the added latency of the full pipeline, meaning the results are evaluated with a delay equal to the pipeline depth. These masks

Table 5 Evaluation results for the proposed method on our newly captured multispectral video dataset

Evaluation type (method)	Metric	Video 1		Video 2		Video 3		Overall		
		Visible	LWIR	Visible	LWIR	Visible	LWIR	Visible	LWIR	Average
Segmentation (St-Charles et al. 2016)	Pr	0.933	0.716	0.938	0.763	0.935	0.821	**0.935**	0.767	0.851
	Re	0.721	0.997	0.834	0.938	0.750	0.996	0.768	**0.977**	**0.872**
	F_1	0.813	0.834	0.883	0.841	0.832	0.900	0.843	0.858	0.851
Segmentation (Proposed)	Pr	0.883	0.937	0.874	0.923	0.921	0.942	0.893	**0.934**	**0.910**
	Re	0.776	0.842	0.818	0.783	0.850	0.849	**0.815**	0.825	0.820
	F_1	0.826	0.887	0.845	0.878	0.884	0.893	**0.852**	**0.876**	**0.864**
Registration (DASC Sliding Window)	% err. $>$ 1px	90.6	88.6	92.1	90.8	88.5	87.2	90.4	88.8	89.6
	% err. $>$ 2px	85.2	81.9	86.6	83.7	81.9	80.2	84.6	81.9	83.3
	% err. $>$ 4px	75.5	71.6	78.6	74.2	72.3	70.0	75.5	71.9	73.7
	$\bar{d}_{\text{err}}$	30.26	21.90	31.22	29.11	26.48	23.34	29.32	24.79	27.05
Registration (Proposed)	% err. $>$ 1px	75.0	74.5	76.3	76.5	68.7	69.9	**73.3**	**73.6**	**73.5**
	% err. $>$ 2px	59.5	59.2	63.4	63.5	53.3	54.3	**58.7**	**59.0**	**58.8**
	% err. $>$ 4px	43.8	43.8	46.8	47.0	32.0	32.4	**40.9**	**41.1**	**41.0**
	$\bar{d}_{\text{err}}$	26.47	22.12	14.43	14.97	9.00	9.06	**16.63**	**15.38**	**16.01**

Bold results are the best in their respective evaluation category

are thus allowed more iterations in our graphical model, and benefit from more temporal information (i.e. past and future frame data). On the other hand, the real-time segmentation outputs are the masks generated by our method for all new image pairs, provided without delay. From these results, we can note that the difference between deferred and real-time output is surprisingly small. This means that our model's temporal inertia allows it to smooth out shape variations without having to peek at future frame data, which is useful for real-time surveillance systems. Besides, the overall improvements obtained by using more than two temporal layers is marginal, as more temporally consistent results also entail that some relevant shape fragments around non-rigid objects are discarded. Finally, note that using more layers results in an important increase in computational complexity: using four layers roughly triples the time required for model inference compared to the default configuration.

4.5 LITIV 2018 Dataset

To help others compare their work on multispectral segmentation and registration, we developed and annotated a new dataset. We recorded video sequences using a stereo pair composed of a Kinect v2 for Windows (at Full HD resolution) and a FLIR A40 LWIR camera (at QVGA resolution). The sensors were roughly aligned on a fixed baseline support (approximately 50 centimeters apart) and synchronized via software to capture frame pairs at 30 Hz. Calibration data for image rectification was obtained by capturing snapshots of a foam core checkerboard pattern heated using halogen lamps to make it visible in LWIR images. For the annotations, we simultaneously recorded depth and user segmentation masks provided by the Kinect SDK, and transformed this data into foreground-background segmentation masks, adding manual touch-ups where needed. Stereo correspondences were also manually annotated like in the work of Bilodeau et al. (2014) to allow an approximate evaluation of registration performance in foreground image regions. In total, this dataset contains over 6000 frame pairs split into three videos, and its groundtruth is composed of 866 binary segmentation masks and 15,182 point correspondences roughly distributed among frames with visible foreground. As for the capture conditions, we deliberately recorded sequences with both strong and weak contrast between foreground and background regions in the two image modalities. More specifically, we used two different temperature calibrations to make individuals more or less perceptible in LWIR images, we introduced some cluttered background in part of the visible images, and we had people carry and exchange objects that modify their appearance in both spectral bands. Overall, this dataset should be more challenging than already available RGB-LWIR video datasets. The fact that it also allows the simultaneous evaluation of foreground segmentation and stereo registration also makes it quite unique in the current literature.

We have made this new dataset available online along with our modified version of the VAP dataset for other authors.[4] Our Kinect's raw data which includes depth images and mapping information is also provided for those interested in trimodal segmentation tasks.

[4] http://www.polymtl.ca/litiv/vid/index.php.

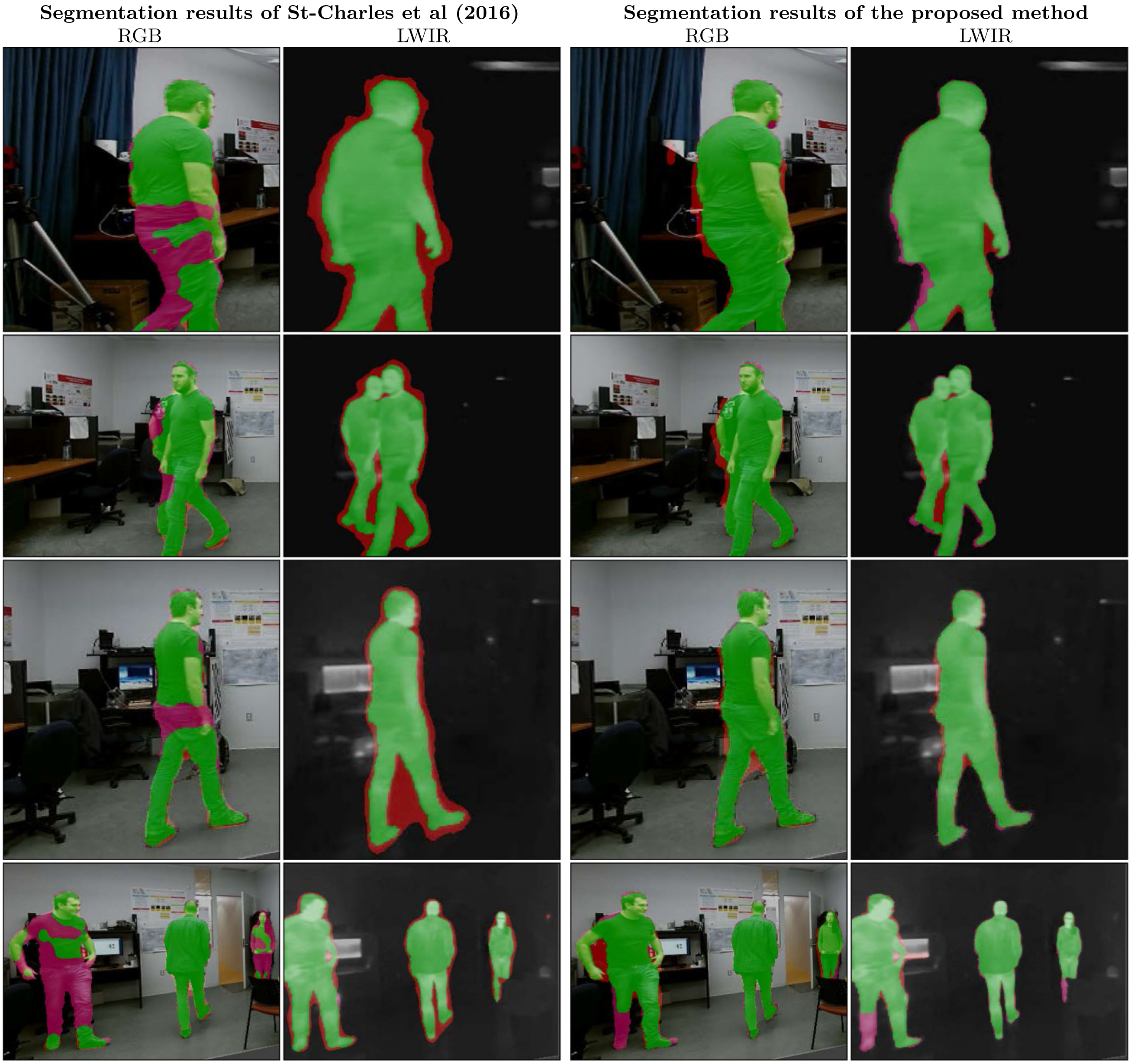
Segmentation results of St-Charles et al (2016)
RGB
LWIR
Segmentation results of the proposed method
RGB
LWIR

Fig. 8 Examples of typical segmentation results from our newly captured dataset; the left two columns show the segmentation masks obtained via St-Charles et al. (2016) and used to initialize our method, and the right two columns show our final segmentation masks. Image regions properly classified as foreground are highlighted in green over the original images, while regions highlighted in orange and magenta show false positives and false negatives, respectively. Images have been cropped to show more details (Color figure online)

We offer our proposed method's results on this new dataset as a baseline for future comparisons in Table 5. We can note that compared to the other two datasets, segmentation results here are still good, but registration errors are much higher. This is primarily due to the fact that our camera baseline is very large (≈ 50 cm), which leads to high disparities for close-range objects (over 150 pixels in some cases), and because our images are higher resolution than those of Bilodeau et al. (2014). Also, we can note that registration errors are higher in the first video sequence: this is caused by the loss of some small foreground segments near image borders which were annotated with correspondences. As for the segmentation results, there are cases where foreground objects are only partly detected, which results in slightly lower Recall scores in some videos. Nonetheless, these results show that our method is capable of segmentating foreground objects in difficult imaging conditions. Finally, we present qualitative segmentation results for this dataset in Fig. 8. We can notice in the bottom row a case where segmentation errors were propagated from the visible image to the infrared one (i.e. two legs are falsely annotated as background). In short, our model can sometimes settle object boundaries in the wrong region due to occlusions in one of the views, or when strong gradients within the object happen to fit the contour model better than the object's real boundaries. A typical example of this is when a person occludes a computer monitor while wearing a shirt that is similarly colored: our model will tend to merge the monitor's contour with the person's blob. This rarely happens in practice, as a very close match in terms of visual appearance and thermal signature is required. Furthermore, as seen from the overall F_1 results in Table 5, our new method outperforms the previous segmentation method in both image modalities.

5 Conclusion

We have presented a new method for simultaneous multispectral foreground segmentation and stereo registration, and validated its capabilities on several datasets. Our approach is based on the alternating minimization of two linked energy functions that integrate multispectral shape and appearance cues. We have shown that both segmentation masks and disparity maps can simultaneously converge to good local minima without any human supervision. Furthermore, with the help of higher order factors, we achieve strong temporal coherence in our segmentation results by linking consecutive video frames inside our graphical models. To make the comparison of methods tackling this problem easier in the future, we provide our full implementation online, as well as a newly created multispectral dataset for evaluation.

If supporting large stereo baselines is unnecessary, the method could use a stronger constraint on multispectral contour similarity to improve coherence between views. Besides, explicit occlusion handling in our stereo model would further improve overall performance on the current datasets. Our model could also be generalized to provide instance-level segmentation by using a separate foreground appearance model for each object. Finally, a three-way energy minimization solution tackling foreground segmentation, stereo registration, and optical flow could be designed based on our current inference approach.

Acknowledgements This work was supported by NSERC, by FRQ-NT team Grant No. 2014-PR-172083, and by REPARTI (Regroupement pour l'étude des environnements partagés intelligents répartis) FRQ-NT strategic cluster. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X GPU used for this research. We also thank Chris Holmberg Bahnsen who provided us with the full calibration data needed to rectify the stereo pairs of the VAP trimodal segmentation dataset.

References

Andres, B., Beier, T., & Kappes, J. (2012). OpenGM: A C++ library for discrete graphical models. CoRR abs/1206.0111, http://arxiv.org/abs/1206.0111

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 898–916. https://doi.org/10.1109/TPAMI.2010.161.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 509–522. https://doi.org/10.1109/34.993558.

Bienkowski, L., Homma, C., Eisler, K., & Boller, C. (2012). Hybrid camera and real-view thermography for nondestructive evaluation. *Quantitative Infrared Thermography, 254*.

Bilodeau, G. A., Torabi, A., & Morin, F. (2011). Visible and infrared image registration using trajectories and composite foreground images. *Image and Vision Computing*, *29*(1), 41–50. https://doi.org/10.1016/j.imavis.2010.08.002.

Bilodeau, G. A., Torabi, A., St-Charles, P. L., & Riahi, D. (2014). Thermal-visible registration of human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, *64*, 79–86. https://doi.org/10.1016/j.infrared.2014.02.005.

Bleyer, M., Rother, C., Kohli, P., Scharstein, D., & Sinha, S. (2011). Object stereo—joint stereo matching and object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3081–3088). https://doi.org/10.1109/CVPR.2011.5995581.

Bouwmans, T. (2014). Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*, *11*, 31–66. https://doi.org/10.1016/j.cosrev.2014.04.001.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(11), 1222–1239. https://doi.org/10.1109/34.969114.

Caelles, S., Maninis, K. K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., & Van Gool, L. (2017). One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Cheng, J., Tsai, Y. H., Wang, S., & Yang, M. H. (2017). SegFlow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*.

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

Davis, J. W., & Sharma, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, *106*(2–3), 162–182. https://doi.org/10.1016/j.cviu.2006.06.010.

Djelouah, A., Franco, J. S., Boyer, E., Le Clerc, F., & Perez, P. (2015). Sparse multi-view consistency for object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *99*, 1. https://doi.org/10.1109/TPAMI.2014.2385704.

Fix, A., Gruber, A., Boros, E., & Zabih, R. (2011). A graph cut algorithm for higher-order markov random fields. In *Proceedings of the IEEE international conference on computer vision* (pp. 1020–1027).

Fix, A., Wang, C., & Zabih, R. (2014). A primal-dual algorithm for higher-order multilabel markov random fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1138–1145).

Goyette, N., Jodoin, P. M., Porikli, F., Konrad, J., & Ishwar, P. (2012). Changedetection.net: A new change detection benchmark dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1–8). https://doi.org/10.1109/CVPRW.2012.6238919.

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision* (2nd ed.). New York, NY: Cambridge University Press.

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

Hwang, S., Park, J., Kim, N., Choi, Y., & So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1037–1045).

Jain, S.D., Xiong, B., & Grauman, K. (2017) Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Jeong, S., Lee, J., Kim, B., Kim, Y., & Noh, J. (2017). Object segmentation ensuring consistency across multi-viewpoint images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ju, R., Ren, T., & Wu, G. (2015). Stereosnakes: contour based consistent object extraction for stereo images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1724–1732).

Kappes, J., Andres, B., Hamprecht, F., Schnorr, C., Nowozin, S., Batra, D., et al. (2013). A comparative study of modern inference techniques for discrete energy minimization problems. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1328–1335).

Kim, S., Min, D., Ham, B., Ryu, S., Do, M. N., & Sohn, K. (2015) DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2103–2112).

Kohli, P., Ladický, L., & Torr, P. H. (2009). Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, *82*(3), 302–324. https://doi.org/10.1007/s11263-008-0202-0.

Kolmogorov, V., & Zabih, R. (2001). Computing visual correspondence with occlusions using graph cuts. *Proceedings of the IEEE Conference on Computer Vision*, *2*, 508–515.

Kroeger, T., Timofte, R., Dai, D., Van Gool, L. (2016). Fast optical flow using dense inverse search. In *Proceedings of European conference on computer vision* (pp. 471–488).

Lempitsky, V., Rother, C., Roth, S., & Blake, A. (2010). Fusion moves for markov random field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(8), 1392–1405.

Li, C., Wang, X., Zhang, L., Tang, J., Wu, H., & Lin, L. (2017). Weighted low-rank decomposition for robust grayscale-thermal foreground detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(4), 725–738.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, *16*(2), 187–198.

Mouats, T., & Aouf, N. (2013). Multimodal stereo correspondence based on phase congruency and edge histogram descriptor. In *Proceedings of 16th international conference on information fusion* (pp. 1981–1987).

Nguyen, D. L., St-Charles, P. L., & Bilodeau, G. A. (2016). Non-planar infrared-visible registration for uncalibrated stereo pairs. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 63–71).

Palmero, C., Clapés, A., Bahnsen, C., Møgelmose, A., Moeslund, T. B., & Escalera, S. (2016). Multi-modal rgb-depth-thermal human body segmentation. *International Journal of Computer Vision*, *118*(2), 217–239. https://doi.org/10.1007/s11263-016-0901-x.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 724–732).

Pinggera, P., Breckon, T., & Bischof, H. (2012). On cross-spectral stereo matching using dense gradient features. In *Proceedings of British machine vision conference*. https://doi.org/10.5244/C.26.103

Pistarelli, M. D., Sappa, A. D., & Toledo, R. (2013) Multispectral stereo image correspondence. In *Computer analysis of images and patterns* (pp 217–224). New York: Springer.https://doi.org/10.1007/978-3-642-40246-3_27

Riklin-Raviv, T., Sochen, N., & Kiryati, N. (2008). Shape-based mutual segmentation. *International Journal of Computer Vision*, *79*(3), 231–245. https://doi.org/10.1007/s11263-007-0115-3.

Rother, C., Kolmogorov, V., & Blake, A. (2004). "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, *23*(3), 309–314. https://doi.org/10.1145/1015706.1015720.

Rother, C., Minka, T., Blake, A., & Kolmogorov, V. (2006). Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 993–1000). https://doi.org/10.1109/CVPR.2006.91.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., & Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. In *Proceedings of German conference pattern recognition* (pp. 31–42). https://doi.org/10.1007/978-3-319-11752-2_3

Shechtman, E., & Irani, M. (2007). Matching local self-similarities across images and videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8). https://doi.org/10.1109/CVPR.2007.383198.

St-Charles, P. L., Bilodeau, G. A., & Bergevin, R. (2016). Universal background subtraction using word consensus models. *IEEE Transactions on Image Processing*, *25*(10), 4768–4781.

St-Charles, P. L., Bilodeau, G. A., & Bergevin, R. (2017). Mutual foreground segmentation with multispectral stereo pairs. In *Proceedings of the IEEE conference on computer vision workshops*.

Tippetts, B., Lee, D. J., Lillywhite, K., & Archibald, J. (2016). Review of stereo vision algorithms and their suitability for resource-limited systems. *Journal of Real-Time Image Processing*, *11*(1), 5–25.

Torabi, A., Massé, G., & 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*(2), 210–221. https://doi.org/10.1016/j.cviu.2011.10.006.

Tron, R., & Vidal, R. (2007). A benchmark for the comparison of 3-d motion segmentation algorithms. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Vicente, S., Rother, C., & Kolmogorov, V. (2011). Object cosegmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2217–2224). https://doi.org/10.1109/CVPR.2011.5995530.

Woodford, O., Torr, P., Reid, I., & Fitzgibbon, A. (2009). Global stereo reconstruction under second-order smoothness priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2115–2128.

Zhang, C., Li, Z., Cai, R., Chao, H., & Rui, Y. (2016). Joint multiview segmentation and localization of RGB-D images using depth-induced silhouette consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4031–4039).

Zhao, J., & Sen-Ching, S. C. (2014). Human segmentation by geometrically fusing visible-light and thermal imageries. *Multimedia Tools and Applications*, *73*(1), 61–89.

Zhu, H., Meng, F., Cai, J., & Lu, S. (2016). Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *The Journal of Visual Communication and Image Representation*, *34*, 12–27.

Zitová, B., & Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, *21*(11), 977–1000. https://doi.org/10.1016/S0262-8856(03)00137-9.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.