Mutual Foreground Segmentation with Multispectral Stereo Pairs

Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau
LITIV lab., Dept. of Computer & Software Eng.
Polytechnique Montréal
Montréal, QC, Canada
{pierre-luc.st-charles,gabilodeau}@polymtl.ca

Robert Bergevin
LVSN - REPARTI
Université Laval
Québec City, QC, Canada
robert.bergevin@gel.ulaval.ca

Abstract

The foreground-background segmentation of video sequences is a low-level process commonly used in machine vision, and highly valued in video content analysis and smart surveillance applications. Its efficacy directly relies on the contrast between objects observed by the sensor. In this work, we study how the combination of sensors operating in the long-wavelength infrared (LWIR) and visible spectra can improve the performance of foreground-background segmentation methods. As opposed to a classic visible spectrum stereo pair, this multispectral pair is more adequate for foreground object segmentation since it reduces the odds of observing low-contrast regions simultaneously in both images. We show that by alternately minimizing a stereo disparity energy and a binary segmentation energy with dynamic priors, we can drastically improve the results of a traditional video segmentation approach applied to each sensor individually. Our implementation is freely available online for anyone wishing to recreate our results.

1. Introduction

Video segmentation and stereo matching are two classic tasks that have recently gained more attention in the context of computer vision outside the visible spectrum. The goal of video segmentation is to partition images based on a semantic analysis of their content, or using appearance and/or motion cues to isolate regions of interest (i.e. the scene "foreground"). On the other hand, stereo matching allows visual data captured from two different sensors to be registered to the same coordinate system, and simultaneously provides depth information about the observed scene via the depth-disparity relation [15]. These tasks have been studied extensively on visible spectrum datasets [26,33,34,42], and are now more commonly considered on data captured using unconventional sensor types.

In this work, our focus lies on foreground-background video segmentation based on foreground motion partitioning and/or background modeling [6,18,36]. While this family of segmentation approaches is not usually affected when applied to unconventional imagery types, it still ultimately fails when the contrast between foreground and background regions is poor, or when the observed scene is affected by large-scale variations in imaging conditions (e.g. illumination fluctuations). Our goal is to address these long-standing problems in video segmentation by combining the visual information of two different types of sensors. In the context of video surveillance, the combination of complementary image spectra (or modalities) has the potential of drastically improving the overall performance of the detection and analytics components of a system under difficult imaging conditions (as shown in Fig. 1). This goal however implies the registration of multimodal data sources, which is not triv-

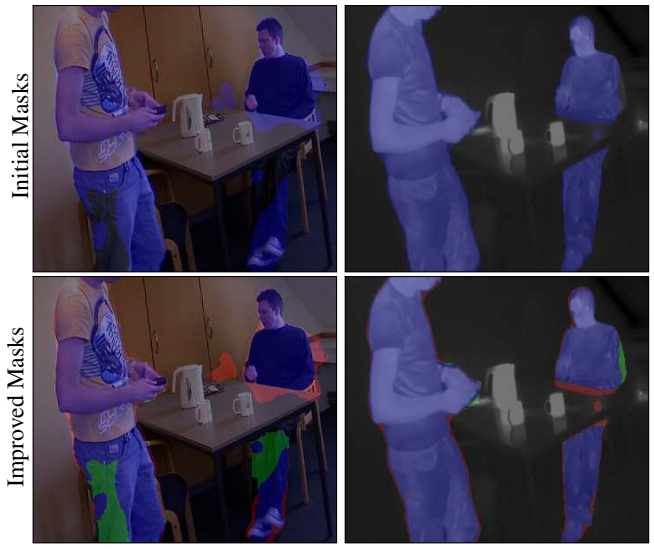

Figure 1. Example of mutual foreground segmentation improvement for a visible-LWIR stereo pair. The initial foreground masks obtained via background subtraction are shown in the top row as regions highlighted in blue. These are refined almost ideally using our proposed approach in the bottow row, where green shows added foreground pixels, and red removed foreground pixels.

ial if a beam splitter cannot be used with the sensors, or if depth information about the observed targets must be kept.

Unlike traditional stereo registration methods that operate on images of the same spectrum band, a multimodal method must account for cases where region matching does not provide reliable results despite favorable imaging conditions (i.e. good object contrast, no occlusions). This is due to the fact that the physical phenomena captured by sensors may be different in each imaging modality. For example, in visible and near-infrared (NIR) imagery, surfaces are generally perceived by their capacity to reflect light, whereas in long-wavelength infrared (LWIR) imagery, their appearance is generally tied to their temperature. There is obviously no clear link between our visual representation of a surface and its temperature, which makes the stereo matching problem for image content registration much more challenging [5].

We propose in this paper a new method that improves the foreground-background segmentation masks provided by a traditional video segmentation method while simultaneously solving the multimodal stereo registration problem using appearance and shape motion cues. We formulate both tasks using Conditional Random Field (CRF) models over the input image pixels, and use their partial segmentation and disparity estimation results as dynamic priors for the iterative minimization of their respective stereo matching and binary partitioning energies. This novel approach essentially solves the "*chicken and egg*" dilemma caused by having to use shape contours to estimate stereo disparities, and stereo disparity maps to correct shape contours. We demonstrate the effectiveness of our proposed method by evaluating it on the VAP dataset [25], and show that our approach improves the overall F-Measure obtained by a state-of-the-art video segmentation method by over 12%. We have made our full implementation available online for future comparisons and for anyone wishing to replicate our results[1].

2. Related Work

Below, we primarily discuss works related to segmentation and stereo registration in multimodal images. For an overview of video segmentation based on motion detection and/or background modeling, we refer readers to the surveys of [8,18]. For a comprehensive overview of two-viewpoint stereo registration methods in the visible spectrum, we refer readers to [33]. Finally, for more information on CRFs and their uses in image segmentation and cosegmentation, we refer to the seminal works of [7,13] and the survey of [45].

Multimodal stereo registration has been studied extensively in the past, but accurate region matching still remains a challenge today [46]. While using a hardware solution such as a beam splitter [17,43] can eliminate this problem completely, the total loss of scene depth information incurred by this approach is rarely desired. Most software registration solutions proposed in the literature thus far deal with modalities that are not too "distant" in regard to their imaging characteristics, e.g. visible and NIR. These solutions are however much less effective for modality pairs such as visible-LWIR, in which object appearance is much less correlated, as stated before. The studies of [5,27] demonstrate that traditional gradient-based descriptors and local similarity measures do not always provide ideal matching performance for stereo registration with LWIR imagery. Even modern descriptors based on self-correlation response encoding [19] are not ideal in this setting. Nonetheless, HOG [10,27], LSS [35,40], MI [23,28] and DASC [19] are still often seen as the most reliable means to find dense correspondences via appearance in multimodal stereo pairs.

An alternative approach in multimodal stereo registration is based on the extraction of correspondences using motion cues from foreground shapes instead of appearance cues. This approach assumes that the sensors capture images continuously, and that the objects of interest for registration are those moving in the observed scene (as is typically the case in video surveillance setups). Using motion to detect and isolate these objects is thus possible, as the scene's background motion is typically either known, or null. Registration can then be achieved using correspondences found on the contours of objects segmented via motion layers [39,44] or from shape trajectories [9,41]. In both cases, even large differences in the imaging characteristics of the sensors do not hinder the registration process, and the trajectories or contours obtained for both images should be fairly similar, as long as the contrast with the background is sufficient in each modality. The main disadvantage of this approach is the overall lower number of correspondences that can be found in a given image pair, leading most methods to rely on sparse disparity maps or simplified planar registration models [37]. Also, relying on point correspondences found via shape contours or trajectories makes occlusion handling much more difficult. In our proposed method, we combine stereo matching via appearance cues (using gradient-based descriptors) and shape motion cues (using object contour descriptors) to provide highly accurate and robust disparity maps.

Mutual foreground segmentation, or more generally image cosegmentation, has been studied in very different contexts over the years [45]. The work of [32] coined cosegmentation as the simultaneous partitioning of several images sharing similar semantic content in the same modality, but without any constraint on viewpoint or object instances. This task is therefore much more generic than what

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

we focus on here, as we assume a camera setup where registration is possible. The specific case that we address is more akin to mutual segmentation as described by Riklin *et al.* [29], in which fragmented segments caused by occlusions or low contrast regions were corrected using a coplanar contour template taken from another viewpoint. Besides, most cosegmentation methods rely on the implicit assumption that all images contain and share a unique instance of the same foreground class, with varying cluttered backgrounds. On the other hand, our proposed method is able to segment multiple visually distinct foreground objects observed simultaneously in the same scene, without supervision.

Finally, some authors have studied unsupervised multimodal mutual foreground segmentation before. In [41] and [44], similar methods using per-blob planar registration and multiple object tracking are proposed to cleanup the segmentation masks of two background subtraction models *a posteriori*. This approach however does not handle occlusions directly, and requires several instances of segmentation and tracking algorithms tuned for each modality to run simultaneously in order to solve the "*chicken and egg*" stereo registration problem using shape contour points. In [24], a mutual segmentation method based on low-rank representation model is proposed which properly exploits the complementarity of grayscale-visible and LWIR imagery. It however does not address the registration problem, meaning that it can only be applied to preregistered planar scenes. Lastly, a method for pedestrian segmentation was proposed by Palmero *et al.* in [25] based on trimodal (visible-LWIR-depth) feature fusion using a learning-based approach. This method however also requires a trained model of the scene for planar registration.

3. Proposed Method

Our approach can be split into two major components: the stereo matching CRF model for disparity estimation, and the shape matching CRF model for foreground-background segmentation. Both tasks are formulated as discrete energy minimization problems that are iteratively solved using fusion moves, as described in Section 4. In more formal terms, given a set of rectified images $\boldsymbol{\mathcal{I}} = \{I_k\}$ (with $k = \{0, 1\}$ in our stereo case), the disparity label space $\mathcal{L}_D = \{0, \dots, d_{\text{max}}\}$, and the background-foreground label space $\mathcal{L}_S = \{0, 1\}$, our goal is to find the optimal pixel-wise disparity and segmentation labelings $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_k\}$ and $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_k\}$ such that:

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

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

where $D_k = \{d_{p,k} : p \in I_k, d_{p,k} \in \mathcal{L}_D\}$ is a disparity labeling, $S_k = \{s_{p,k} : p \in I_k, s_{p,k} \in \mathcal{L}_S\}$ is a segmentaton labeling, and where the energy cost functions E_k^{stereo} and E_k^{segm} are described in Sections 3.1 and 3.2, respectively. These two functions are linked through their estimation results ($\mathcal{D}_k$ and $\mathcal{S}_k$) which are used as dynamic priors. More specifically, disparity labels $d_{p,k}$ for each pixel p in I_k are used as priors to improve inter-modality shape consistency in (12) and (13). Additionally, shape descriptors are recomputed in S_0 and S_1 after every segmentation update, and the affinity between these descriptors is used in (4) to improve stereo matching in foreground regions. Besides, note that we sometimes omit the k subscript in the following sections to simplify the notation, as most equations only deal with one image of the stereo pair at a time.

3.1. Stereo Energy

Since we are working with rectified image pairs, registration can be formulated without loss of generality as a 1D search for matches on epipolar lines [15]. Calculating the disparity (or offset) between the locations of each pixel p in I_0 and its best match in I_1 is our ultimate goal in this section, as it will allow us to properly overlay and improve foreground shapes in the next section. We define the energy cost for a disparity labeling configuration D as

$$\begin{aligned} E^{\text{stereo}}(D) = E^{\text{appearance}}(D) + E^{\text{shape}}(D) \\ + E^{\text{smooth1}}(D) + E^{\text{uniqueness}}(D). \end{aligned} \tag{3}$$

Each term in this cost function promotes a property of the desired output labeling: the appearance and shape terms help find adequate matches based on inter-modal cues, the smoothness term penalizes inconsistency in disparity labeling, and the uniqueness term penalizes multiple stereo matches to or from a unique pixel location.

Appearance and shape terms. These two unary terms are very similar in nature, as they express the cost of matching two image patches of the stereo pair. These are both defined as

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

where $r(p, d_p)$ returns the location obtained by shifting pixel p by disparity 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 match saliency at pixel p. DASC descriptors [19] are densely computed over I_0 and I_1 for the appearance term, while Shape Context descriptors [4] are densely computed over S_0 and S_1 for the shape term. The patches used in the affinity map $\mathcal{A}$ are 15x15, and their affinity cost is computed by accumulating the L2 distances between the normalized descriptors.

Finally, the saliency map $\mathcal{W}$ is defined as

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

where $K(p)$ returns the matrix of descriptors in the patch around pixel p, and $\mathcal{H}(\cdot)$ computes Hoyer's sparseness metric [16] over a vector or matrix. The reasoning behind using saliency is that, as stated before, multimodal matches are often unreliable. Here, 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 the cost of assigning a disparity label to p will be greatly lowered due to $\mathcal{W}(p)$ (which is always $\in [0,1]$). Besides, for the shape term, we also nullify the saliency outside foreground shapes to avoid influencing background disparity estimation near foreground object contours. We can assume that disparity estimation for background regions will be less accurate due to this missing term contribution, but since we primarily focus on the registration of foreground shapes, this is inconsequential.

Smoothness term. We impose a truncated pairwise smoothness constraint on the disparity labeling produced by our solution. The idea here is that neighboring pixels should have similar disparity labels, especially if the local image gradient between them is small, while object edges should still be sharp. We define this smoothness term as

$$E^{\text{smooth1}}(D) = \lambda_{s1} \cdot \sum_{\langle p,q\rangle \in \mathcal{N}} \min\left(|d_p - d_q|, 10\right)^2 \cdot G_k(p,q), \quad (6)$$

with

$$G_k(p,q) = \exp\left(1 - \frac{\nabla I_k(p,q)}{g}\right), \quad (7)$$

and where λ_{s1} is a fixed cost scaling factor, $\mathcal{N}$ is the set of first order cliques in the graph model, $\nabla I_k(p,q)$ returns the normalized local image gradient intensity between pixels p and q of image I_k, and g is a constant value which defines the expected object contour gradient intensity (we used g=32). The truncation value (10 is used) allows large discontinuities to occur by capping the maximum smoothness penality.

Uniqueness term. The purpose of this term is to penalize stereo associations to pixels which are already matched elsewhere, thus helping spread disparities in regions with very little salient information. Unlike the mutual exclusion constraint of the uniqueness term proposed in [20] (i.e. infinite cost beyond the first match), our term encodes a soft constraint, meaning that many-to-one correspondences are allowed, but at a cost. The advantage of this approach is that the labeling can evolve faster during early inference steps, as matches can be temporarily "stacked" on individual pixels to allow larger label moves. Over time however, matches are automatically "unstacked" by our optimizer to regain the added cost. The very large majority of pixels end up having a single match once the solution converges, which typically happens in fewer iterations using this approach. We define the uniqueness cost incurred by pixel p as

$$\mathcal{U}(p) = \begin{cases} \sum_{n=1}^{N(p)-1} \frac{w \cdot n}{w+n-1} & \text{if } N(p) > 1 \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

where, $N(p)$ returns the number of matches held by pixel p, and w is a small weighting constant (we used w=3). We thus keep track of pixel association counts as latent variables in our model. However, since we rely on large label moves to solve the inference problem, many correspondences may be removed in a single iteration, making the total cost of a move over several pixels hard to predict with (8) due to its nonlinearity. To solve this problem, we define our uniqueness cost for a given labeling as

$$E^{\text{uniqu.}}(D) = \lambda_u \cdot \sum_{p \in I} \frac{-\mathcal{U}(r(p,\tilde{d}_p))}{N(r(p,\tilde{d}_p))} + \frac{w \cdot N(r(p,d_p))}{w + N(r(p,d_p)) - 1}, \quad (9)$$

where $\tilde{d}_p$ indicates the last disparity label given to p, and λ_u is a cost scaling factor. This formulation guarantees that estimated pixel move costs provided to the solver will always be greater or equal to their real costs once the full move is complete, but remain fairly similar.

Finally, note that we initialize the disparity labeling of our model by naively minimizing our energy function while considering only the unary terms $E^{\text{appearance}}$ and E^{shape}. This results in disparities roughly equivalent to those obtained using a naive winner-takes-all sliding window matching approach, which is a good enough starting point for rapid convergence.

3.2. Segmentation Energy

Our ultimate goal is to refine the binary segmentation masks obtained using a traditional video segmentation method. We opted for the method described in [38], for which the implementation was available online. Thus, for each frame in the analyzed sequence, we initialize the labeling within our model using the masks generated by that method. Then, we define our overall energy cost for a given foreground-background labeling S as

$$E^{\text{segm}}(S) = E^{\text{color}}(S) + E^{\text{contour}}(S) + E^{\text{smooth2}}(S). \quad (10)$$

Each term in this cost function once again promotes a property of the desired output labeling: the color term maximizes the separation between foreground and background color distributions, the contour term penalizes shape mismatches between the input masks, and the smoothness term penalizes label discontinuities away from local image gradients.

Color term. We use two Gaussian mixtures with $K = 5$ components to build the full foreground and background color appearance models for each image of the stereo pair, similar to the approach used in [30] for interactive image segmentation. In our case however, we rely on our initial labeling mask and its evolution over each resegmentation iteration to refine our mixture models. We define our color term 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 how well a pixel color x fits a Gaussian mixture model with component weights $\boldsymbol{\beta}$, means $\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}$. Note that subscripts to Gaussian mixture parameters in (11) indicate either foreground (1) or background (0) model. These parameters are initialized using *k-means*, and refitted after every minimization step using the latest binary partitioning.

Contour term. This term's role is to combine foreground shapes across the two captured images, and it is ultimately responsible for the elimination of erroneously classified blobs and for the reconnection of shape fragments. It does so simply by using foreground and background distance transform maps based on prior segmentation masks to penalize the labeling of pixels far from their respective shape contours. We reuse subscript k here to properly highlight the contribution of each modality to this term, which is defined as

$$E_k^{\text{contour}}(S_k) = \lambda_c \cdot \sum_{p \in I_k} \begin{cases} F_k(p) + 0.5 \cdot F_{k'}\big(r(p, d_p)\big) & \text{if } s_{p,k} = 1 \\ B_k(p) + 0.5 \cdot B_{k'}\big(r(p, d_p)\big) & \text{otherwise} \end{cases}, \tag{12}$$

where λ_c is a cost scaling factor, k' is the opposite index of k in the stereo pair, $F_k(p)$ returns the Euclidean distance between pixel p and the closest foreground pixel present in the previous segmentation mask $\tilde{S}_k$ of sensor k, and similarly for $B_k(p)$ with background pixels. Note here that the inter-modal cost contribution is scaled by half, meaning that shape contours will slightly prefer sticking to their own previous results. This improves the stability of the segmentation while optimizing, reducing the risks of eliminating relevant shape fragments too rapidly.

Smoothness term. This last term serves the same purpose as (6), i.e. it penalizes label discontinuities everywhere except for regions where local image gradients are strong. Its formulation in the segmentation energy is also very similar to (6), but we reuse the inter-modality contribution idea of (12), and apply it this time to the gradient scaling factor. We define it as

$$E_k^{\text{smooth2}}(S_k) = \lambda_{s2} \cdot \sum_{\langle p,q \rangle \in \mathcal{N}} \Big(s_{p,k} \oplus s_{q,k}\Big) \cdot \Big(G_k(p,q) + 0.5 \cdot G_{k'}(p', q')\Big), \tag{13}$$

where λ_{s2} is a fixed cost scaling factor, $\oplus$ is the XOR operator, p' is a shorthand for $r(p, d_p)$, and q' is a shorthand for $r(q, d_q)$. The inter-modality contribution here allows shape contours from one modality to "snap" onto edges that are only present in the other, making it possible to expand and relocate contours inside low contrast image regions.

4. Implementation Details

We optimize our two energy functions using fusion moves via QPBO [14,31] based on Fix *et al.*'s generalized approach from [11], adapted to be used in the OpenGM framework [3] without parallelization. While this strategy initially allowed rapid prototyping of our energy functions, we could have used a more constrained and much faster max-flow optimization strategy instead (e.g. FastPD [22] or SoSPD [12]) since all pairwise terms in our two models also verify the submodularity test [21]. As stated before, we alternate between the minimization of (3) and (10) in order to continuously improve their respective priors. Inference termination is automatically reached when no more label moves in $\mathcal{L}_D$ or $\mathcal{L}_S$ can reduce the energies of (3) or (10). This typically happens after between 100 to 500 iterations, depending on the quality of the initialization labelings. For reference, with our naive optimization implementation, 500 iterations took approximately 130 seconds when computed on a 3rd generation Intel i7 processor at 3.4 GHz.

For the fixed parameters presented in the previous section, we empirically determined that $\lambda_{s1} = 0.0025$, $\lambda_u = 1$, $\lambda_c = 10$, and $\lambda_{s2} = 10$ offered adequate performance, and that finer tuning would not affect the final results much more. For more implementation details, we invite the reader to refer to our code online (the link is provided in Section 1).

5. Evaluation

For our experiments, we adapted the dataset of Palmero *et al.* [25] to our needs, which was originally intended for trimodal (visible-LWIR-depth) video segmentation. This dataset consists of 5724 frame triplets split into three scenes, with for each frame a groundtruth foreground-background segmentation mask. We obtained the calibration images they used to learn their planar registration model, and rectified all visible-LWIR image pairs using the OpenCV [2] calibration toolbox. The depth images were left unused during all our experiments, and the second scene of the dataset had to be removed due to missing calibration data. Finally, the groundtruth masks we used for our evaluation were manually selected from the dataset at approximately 2 Hz while focusing on intervals with visible actors interacting. This is done to avoid skewing the evaluation results by continuously segmenting empty frames or frames with purely static and unoccluded foreground regions.

We could not compare our method to its closest match

Category	*Pr*	*Re*	*FM*
Scene 1 (visible)	0.819	0.810	0.815
Scene 1 (LWIR)	0.755	**0.975**	0.851
Scene 3 (visible)	0.716	0.688	0.702
Scene 3 (LWIR)	0.514	**0.969**	0.671
Average (visible)	0.768	0.749	0.759
Average (LWIR)	0.634	**0.972**	0.761
Average (Scene 1)	0.787	0.892	0.833
Average (Scene 3)	0.615	0.828	0.686
Average (Overall)	0.701	0.860	0.759

Table 1. Baseline results obtained using the PAWCS method of [38] on the recalibrated stereo pairs from the VAP segmentation dataset [25].

Category	*Pr*	*Re*	*FM*
Scene 1 (visible)	**0.875**	**0.931**	**0.902**
Scene 1 (LWIR)	**0.845**	0.922	**0.881**
Scene 3 (visible)	**0.723**	**0.942**	**0.818**
Scene 3 (LWIR)	**0.691**	0.958	**0.803**
Average (visible)	**0.799**	**0.936**	**0.860**
Average (LWIR)	**0.768**	0.940	**0.842**
Average (Scene 1)	**0.860**	**0.926**	**0.892**
Average (Scene 3)	**0.707**	**0.950**	**0.810**
Average (Overall)	**0.784**	**0.938**	**0.851**

Table 2. Improved results obtained using the proposed mutual segmentation method on the recalibrated stereo pairs from the VAP segmentation dataset [25].

in the literature ([44]) as the authors' dataset is not public, their source code is not available, and some of their implementation details are missing (e.g. tracking & background subtraction parameters). Instead, we compare our mutual segmentation results to the results obtained by a traditional video segmentation approach applied to each video stream individually (as they also did in [44]). In our case however, we opted for a more modern baseline, i.e. the method of [38]. According to [1], this method is the top performer in unsupervised foreground-background segmentation for static camera viewpoints, and one of its main advantages is that it does not require fine tuning for each video sequence. Besides, we could not obtain the results of [25] to compare their trimodal segmentation to our new bimodal approach on their modified dataset. To ease future comparisons with this work, we make this modified dataset available online[2].

We rely on three binary classification metrics commonly used to quantitatively evaluate the performance of binary segmentation methods, namely Precision (*Pr*), Recall (*Re*), and F-Measure (*FM*). These are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{16}$$

where TP, TN, FP and FN are respectively the *True Positives*, *True Negatives*, *False Positives*, and *False Negatives* classification counts obtained by pixel-wise comparisons of our segmentation masks with the groundtruth masks.

We present the quantitative results obtained using [38] (Table 1) as well as our proposed method (Table 2). Bold entries represent the best result obtained between the two.

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

We can observe that our method outperforms [38] for all Precision and F-Measure entries, and for six out of nine Recall entries. This demonstrates that our proposed approach can most of the time improve the segmentation results of a state-of-the-art method without introducing more false positive or negative pixel labelings. For the categories where our Recall measures are lower, we can observe that [38]'s precision is very low, meaning that it was probably detecting far too many foreground regions that our approach had to eliminate. By contrast, our precision in those categories is fairly good, meaning our method eliminated most of these false positives, but also created some false negatives in the process. The best F-Measure improvement is achieved for LWIR imagery in Scene 3 (19.7% increase over the original result), whereas the overall average F-Measure improvement for both scenes is 12.1%. Finally, we show in Figures 2 and 3 typical segmentation improvements obtained for various frame pairs of the dataset, from which we can see that our method is often able to find the optimal middle ground between two very noisy initial segmentation masks.

6. Conclusion

We introduced a novel approach for the mutual segmentation of foreground objects observed using a multimodal stereo pair that also fully addresses the data registration problem. Through our experiments, we demonstrated that our method successfully combines motion and appearance cues from visible and LWIR imagery to improve stereo matching and to find object contours despite unfavorable imaging conditions. Our evaluation shows that our solution vastly outperforms a state-of-the-art video segmentation method based on background modeling applied to individual video streams. Moreover, we did not need to use specific tuning or special heuristics to handle each image modality, meaning that our method could be applied to any

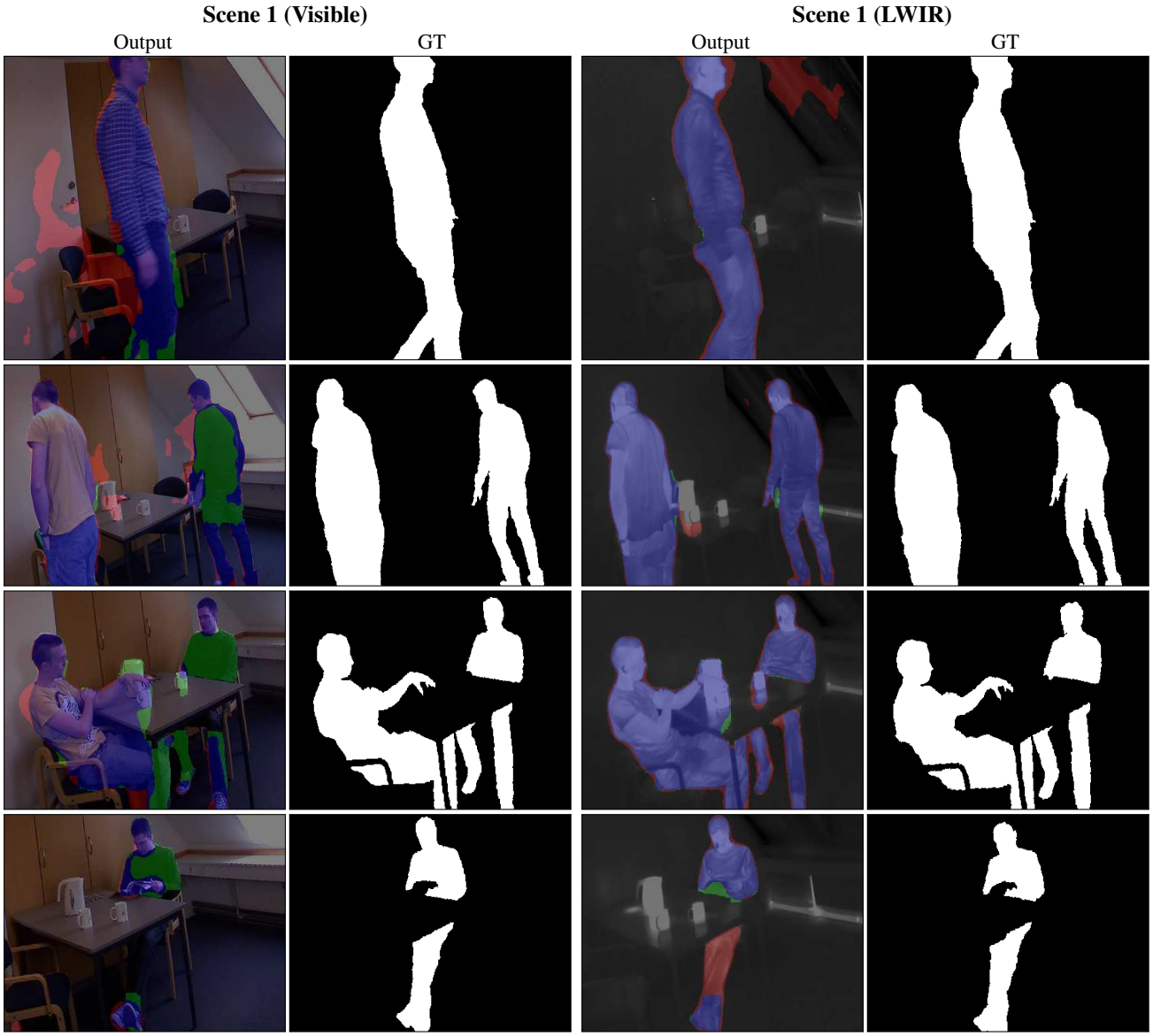

Figure 2. Visualization of segmentation improvements achieved over [38] using our proposed approach in Scene 1 of the VAP dataset [25]. Each pair of columns shows the improved segmentation masks as well as the groundtruth for a given modality. Blue image regions indicate unmodified foreground segmentation, red indicates removed foreground, and green indicates added foreground. Images have been cropped to show more details. The bottom row of the LWIR columns shows a case where many foreground labels are wrongly removed.

multimodal stereo pair. Our method however does not yet consider the content overlap between consecutive frames, meaning segmentation masks and stereo disparity maps are not linked temporally or reused for model reinitialization.

For our future work, we intend to add temporal connectivity constraints to our graphs through higher-order terms in order to improve the spatiotemporal consistency of the segmentation.

Acknowledgment

This work was supported in part by NSERC, 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 also gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X GPU used for this research.

Finally, special thanks to Chris Holmberg Bahnsen, who provided us the full calibration data needed to rectify the stereo pairs of the VAP trimodal segmentation dataset.

References

[1] ChangeDetection.net – a video database for testing change detection algorithms. http://changedetection.net/. Accessed: 2017-06-23.

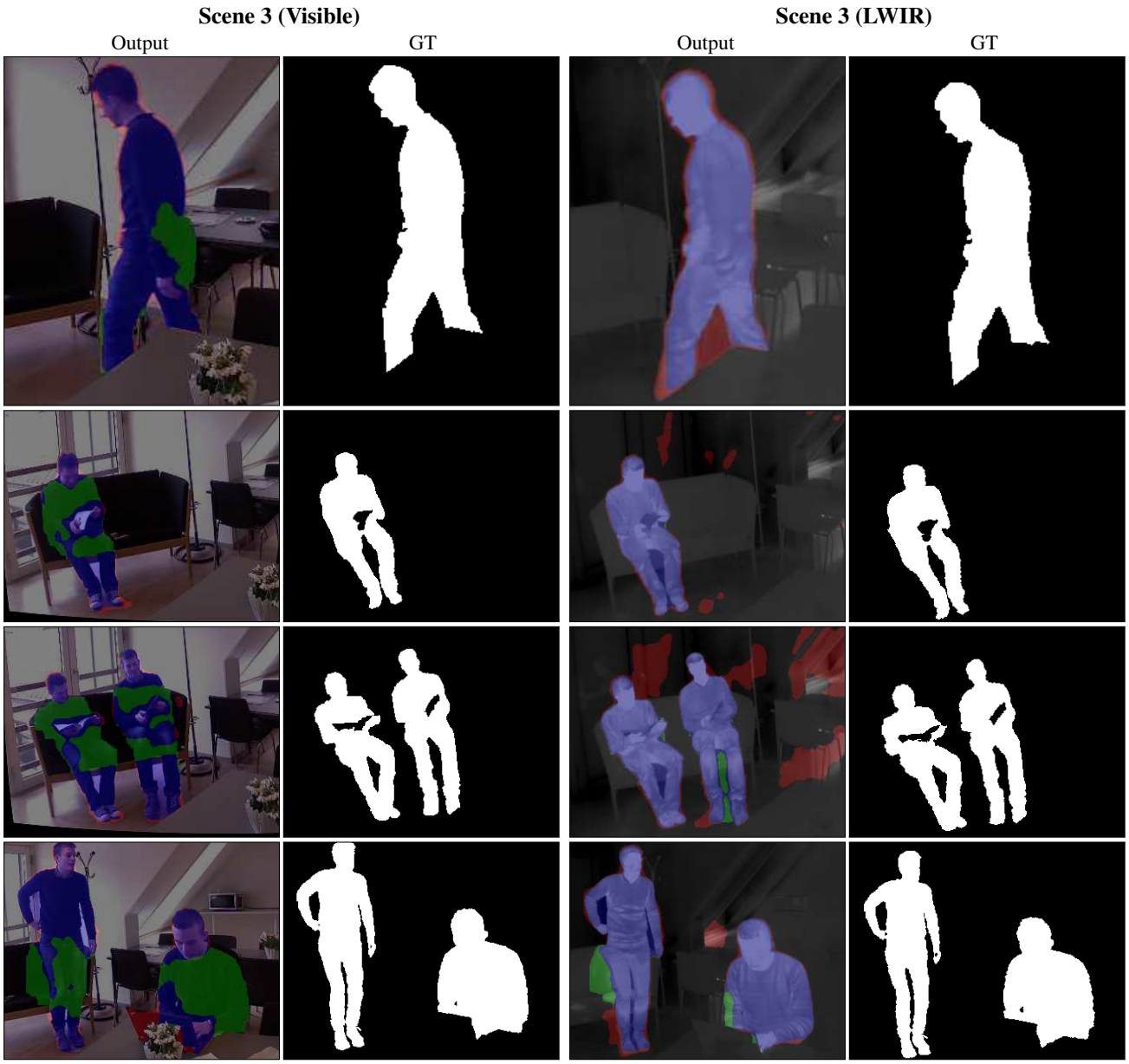

Figure 3. Visualization of segmentation improvements achieved over [38] using our proposed approach in Scene 3 of the VAP dataset [25]. Each pair of columns shows the improved segmentation masks as well as the groundtruth for a given modality. Blue image regions indicate unmodified foreground segmentation, red indicates removed foreground, and green indicates added foreground. Images have been cropped to show more details. The bottom row of both column pairs shows cases where many foreground labels are wrongly added.

[2] Open source computer vision library. https://github.com/opencv/opencv.

[3] B. Andres, T. Beier, and J. Kappes. OpenGM: A C++ library for discrete graphical models. *CoRR*, abs/1206.0111, 2012.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, Apr 2002.

[5] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi. Thermal-visible registration of human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, 64(0):79–86, 2014.

[6] T. Bouwmans. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*, 11:31–66, 2014.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[8] S. Brutzer, B. Hoferlin, and G. Heidemann. Evaluation of background subtraction techniques for video surveillance. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1937–1944, June 2011.

[9] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-

to-sequence matching. In *Proc. 8th European Conf. Comput. Vis. Workshops*, 2002.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 886–893 vol. 1, 2005.

[11] A. Fix, A. Gruber, E. Boros, and R. Zabih. A graph cut algorithm for higher-order markov random fields. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1020–1027, 2011.

[12] A. Fix, C. Wang, and R. Zabih. A primal-dual algorithm for higher-order multilabel markov random fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1138–1145, 2014.

[13] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(2):271–279, 1989.

[14] P. L. Hammer, P. Hansen, and B. Simeone. Roof duality, complementation and persistency in quadratic 0–1 optimization. *Mathematical Programming*, 28(2):121–155, 1984.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[16] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Research*, 5(Nov):1457–1469, 2004.

[17] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1037–1045, 2015.

[18] P.-M. Jodoin, S. Piérard, Y. Wang, and M. Van Droogenbroeck. Overview and benchmarking of motion detection methods. In T. Bouwmans, F. Porikli, B. Hoferlin, and A. Vacavant, editors, *Background Modeling and Foreground Detection for Video Surveillance*, chapter 1. Chapman and Hall/CRC, June 2014.

[19] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2103–2112, 2015.

[20] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proc. IEEE Int. Conf. Comput. Vis.*, volume 2, pages 508–515, 2001.

[21] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):147–159, Feb 2004.

[22] N. Komodakis and G. Tziritas. Approximate labeling via graph cuts based on linear programming. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(8):1436–1453, 2007.

[23] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Comput. Vis. and Image Understanding*, 106(2):270–287, 2007.

[24] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin. Weighted low-rank decomposition for robust grayscale-thermal foreground detection. *IEEE Trans. Circuits Syst. Video Technol.*, 27(4):725–738, 2017.

[25] C. Palmero, A. Clapés, C. Bahnsen, A. Møgelmose, T. B. Moeslund, and S. Escalera. Multi-modal rgb–depth–thermal human body segmentation. *Int. J. Comput. Vis.*, 118(2):217–239, 2016.

[26] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 724–732, June 2016.

[27] P. Pinggera, T. Breckon, and H. Bischof. On cross-spectral stereo matching using dense gradient features. In *Proc. British Mach. Vis. Conf.*, 2012.

[28] J. Pluim, J. Maintz, and M. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans. Med. Imag.*, 22(8):986–1004, 2003.

[29] T. Riklin-Raviv, N. Sochen, and N. Kiryati. Shape-based mutual segmentation. *Int. J. Comput. Vis.*, 79(3):231–245, 2008.

[30] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, Aug. 2004.

[31] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8, 2007.

[32] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 993–1000, June 2006.

[33] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proc. German Conf. Pattern Recognit.*, pages 31–42, 2014.

[34] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[35] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8, 2007.

[36] A. Sobral and A. Vacavant. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Comput. Vis. and Image Understanding*, 122:4–21, 2014.

[37] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. Online multimodal video registration based on shape matching. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 408–413, June 2015.

[38] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. Universal background subtraction using word consensus models. *IEEE Trans. Image Process.*, 25(10):4768–4781, 2016.

[39] X. Sun, T. Xu, J. Zhang, and X. Li. A hierarchical framework combining motion and feature information for infrared-visible video registration. *Sensors*, 17(2):384, 2017.

[40] A. Torabi and G.-A. Bilodeau. Local self-similarity-based registration of human ROIs in pairs of stereo thermal-visible videos. *Pattern Recognit.*, 46(2):578–589, 2013.

[41] A. Torabi, G. Massé, and G.-A. Bilodeau. An iterative integrated framework for thermal-visible image registra-

tion, sensor fusion, and people tracking for video surveillance applications. *Comput. Vis. and Image Understanding*, 116(2):210–221, 2012.

[42] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar. CDnet 2014: An expanded change detection benchmark dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 387–394, June 2014.

[43] A. M. Waxman, D. A. Fay, A. N. Gove, M. Seibert, J. P. Racamato, J. E. Carrick, and E. D. Savoye. Color night vision: fusion of intensified visible and thermal IR imagery. In *Proc. Symp. OE/Aerosp. Sens. and Dual Use Photon.*, pages 58–68, 1995.

[44] J. Zhao and S.-C. S. Cheung. Human segmentation by geometrically fusing visible-light and thermal imageries. *Multimedia Tools and Applicat.*, 73(1):61–89, 2014.

[45] H. Zhu, F. Meng, J. Cai, and S. Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *J. Visual Commun. and Image Represent.*, 34:12–27, 2016.

[46] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vis. Comp.*, 21(11):977–1000, 2003.