2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

# Combining Motion and Appearance for Scene Segmentation

Paulo Vinicius Koerich Borges, Peyman Moghadam

*Abstract*— **Image segmentation is a key topic in computer vision, serving as a pre-step in a number of robotics tasks, including object recognition, obstacle avoidance and topological localization. In the literature, image segmentation has been employed as auxiliary information in order to improve optical flow performance. In this work, an alternative approach is proposed, in which optical flow information is used to aid image segmentation, aiming at scene understading for mobile robots. The proposed system performs dense optical flow analysis, followed by clustering of the optical flow vectors in a four dimensional space (formed by the $x$ and $y$ positions, angle and magnitude of each vector). Results from the clustering are used as 'seeds' in the segmentation process, performed by watershed segmentation in our implementation. In addition, the flow 'image' is combined with the original image, generating an image better suited for watershed segmentation, reducing the local minima effect often seen in this type of segmentation algorithms. The main pipeline considers the use of multi-modality cameras (visible and thermal-infrared). Since they see substantially different information, multi-modality further improves the amount of features of the resulting flows. Experimental results in urban and semi-urban scenarios with efficient segmentation illustrate the applicability of the method.**

## I. INTRODUCTION

The perception and understanding of objects is an important capability of robots which must interact with their surroundings. Object perception is particularly important in outdoor mobile navigation, where a rich amount of elements populate the environment. A key—and perhaps the most significant—part in generic object perception is the segmentation stage. Bottom-up segmentation of natural images is challenging and it is a generally ill-posed and unconstrained problem, which has led to the development of a number of different approaches [1], [2]. In this paper, we propose a joint image segmentation method, using cameras in both the visible and thermal-infrared modalities. The main contribution lies in using optical flow information to assist in the segmentation process, combining flow information from both visible and thermal-infrared imaging. Rather than focusing on image segmentation for human interpretation *only* (which should analyze only visible spectrum images), we consider the problem of scene understanding for a robot or autonomous vehicle. Therefore, elements that would not be necessarily separated with human assisted segmentation (in case the image quality is poor) should still be segmented and identified by the robot when the goal is to achieve optimal environment awareness. In this sense, the use of thermal-infrared imaging is beneficial, capturing more of the structure of the environment with information that is not necessarily obvious in the visible domain, depending on lighting conditions and illumination angle.

Two main aspects are addressed. First, many segmentation algorithms benefit from the use of 'seeds', which can be provided manually by a user or can be embedded in the segmentation framework. This is the case in watersheds [3], for example, which is the segmentation approach used in this work. Therefore, clustering of the optical flow vector in a four dimensional space (since each vector contains 2-D position, magnitude and angle information) can provide efficient seeds that improve the segmentation process. We employ dense optical flow, which analyzes the full image and does not rely on isolated points of interest, as is the case in sparse optical flow. Second, in order to avoid over segmentation, also common in watershed or mean-shift segmentation [4], we propose the linear combination of the original image with the 'flow' image. Although to a limited extent, the flow information is correlated to depth. Therefore, when the flow and the image are combined, the segmentation is improved since different depths generally imply different objects.

In contrast to other works which employ optical flow for *layered motion* segmentation [5], [6], we consider the full image image segmentation problem, jointly combining motion and appearance. Similarly to [7], we consider optical flow and image segmentation together. In [7], however, the approach focuses on using hierarchical image segmentation in order to find better optical flow under large displacements. To validate the approach, we perform several experiments using data from road in urban and unstructured areas. The results indicate that on average, applying the proposed method improves the segmentation performance, both in single modality or using the joint thermal-infrared and visible alternative.

This paper is organized as follows. Section II addresses related work, contrasting it with the method proposed. Section III presents the proposed method, which is addressed from a multi-modal perspective in Section IV. Section V presents experimental results on several outdoor datasets. Relevant conclusions and future work are discussed in Section VI.

## II. RELATED WORK

Segmenting different objects and parts of an image or video is a long-standing topic in robotic vision. Proposed

This work was supported by the CSIRO Computational Informatics, by the CSIRO Minerals Down Under Flagship, and by Australia's Endeavour Executive Award program.

The authors are with the Autonomous Systems, Computational Informatics-CSIRO, Brisbane, Australia. Paulo Borges is an Adjunct Lecturer and Peyman Moghadam is an Adjunct Fellow at the School of Information Technology and Electrical Engineering, University of Queensland. E-mails: *firstname.lastname*@csiro.au

methods include saliency detection, semantic region identification (sky, road, etc), and trained object detection. Beyond single images, the problem can be approached by identifying repeated patterns among pairs or sets of unlabelled images [8], [9]. Because unknown parts of any frame may present the repeated pattern, iterative refinement methods [8] or graph-based segmentation of detected objects [9] can be applied.

In contrast to single frames or an assorted set of snapshots, long video sequences, offer significant temporal consistency elements. Video object/region segmentation (in a spatial sense) frequently follows an interactive or supervised approach. With interactive methods, the user is required to annotate object or region boundaries in certain key frames. This information is then propagated to other frames while errors can be manually adjusted [10], [11]. Semi-automated tracking-based techniques attempt to decrease the amount of supervision by assuming manual segmentation on the initial frame only [12], [13].

Unfortunately, all methods above demand user input for indicating areas of interest, and are therefore user dependent. Bottom-up approaches, on the other hand, can segment frame regions in a video in a fully automated way, using cues such as motion and appearance similarity. Motion segmentation techniques cluster pixels in video images applying bottom-up motion cues. Common approaches include performing segmentation considering a spatiotemporal video volume [14], or starting with an image segmentation per frame and then matching segmented regions in neighbouring frames [15], [16]. Early works have also used watersheds and motion for segmentation [17], however with focus on moving objects on static backgrounds. Recently, two segmentation methods using motion which are closely related to the work in this paper have been proposed [18], [19]. In those methods, tracking is performed to form long-term motion trajectories, followed by affine motion clustering of these trajectories, which is then used as input for segmentation. Other works performed pure *motion* segmentation based on two-frame optical flow, focusing more on motion layer extraction rather than generic image segmentation [5], [6]. Layering methods work well in traversal views with well defined planes, but have limited performance in more continuous vanishing views, such as a front vehicle mounted cameras. Extending the analysis to 3-D, it is possible to perform structure from motion and combine the 3-D point-cloud with appearance in order to detect specific classes of objects [20], [21]. Using machine learning, training is performed for the classification of typical road elements (sky, road, cars, tree, building, etc). In contrast to the works above, in this paper we perform segmentation considering not only the pixel domain but also the flow domain assuming a moving camera, as explained in the following sections.

## III. Proposed Method

The basic pipeline for the proposed method is shown in Figure 1. The main idea is to use optical flow vectors as an additional source of information in a image segmentation framework. In our implementation, we use a gradient-based watershed algorithm [22] for segmentation, due to its generally good performance. We extend the traditional watershed implementation to consider not only the luminance information in the gradient based watershed, but also the topological optical flow information, obtained from the difference between consecutive frames. Optical flow information is applied in two ways: (i) by combining (weighted summation) the optical flow "image" $\mathbf{F}$ and the original image $\mathbf{V}$ into a third image $\mathbf{S}$, which is then segmented, and (ii) by clustering the optical flow vectors and using the clusters as seeds (represented by the set $\mathbf{\Phi}$) in the segmentation process. Figure 2 illustrates the topological structure of the flow representation corresponding to the visible spectrum image in Figure 4. This figure indicates that the optical flow topology is potentially applicable to a watershed segmentation framework, presenting even better defined "hills" and "valleys" than normal images, therefore being less prone to oversegmentation, as discussed in Section III-B. More details about each of the modules is given next.

### A. Dense Optical Flow

Although generally not as stable over time as their sparse counterpart, dense optical flow is preferred because it provides a more comprehensive indication of the flow over the whole frame, instead of relying only on feature points. Among dense optical flow algorithms, the three dimensional (3D) structure tensor methods [23], [24] have gained popularity due to their noise robustness and low systematic error. Therefore, we employ this technique in this paper. In order to work satisfactorily, 3D structure tensor techniques consider the brightness change constraint, which assumes that a given feature will have constant luminance from one frame to the next, and it will only undergo local translations. In order to regularize the results, it builds the tensor for each pixel element within its surrounding pixels, where the local optical flow is considered invariant. The optical flow estimation is then converted to an eigenvalue analysis problem. Because the locally constant optical flow assumption is often not met in real applications, the 3D structure tensor technique can be extended to use the affine motion model. In this case the tensor is defined by projecting the image into a second degree polynomial and integrating the affine model into the tensor, where a linear system framework solves for the affine parameters [24].

### B. Joint Segmentation

The joint segmentation stage combines the input image and its corresponding optical flow representation into a single matrix (this operation is represented by the block $\triangledown$ in Figure 1) , which is then applied to the watershed segmentation process. Let $\mathbf{V}$ be the input image and $\mathbf{F}$ represent the flow image corresponding to $\mathbf{V}$. The two images are combined according to a weighted sum given in (1), which depends on a confidence level $\alpha$ of each flow vector. This confidence measure estimates the correctness of each displacement vector. Therefore, vectors with a lower $\alpha$ have

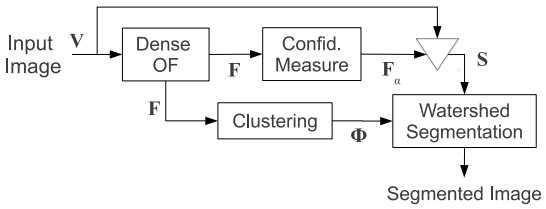



Fig. 1: Block diagram illustrating the basic pipeline of the proposed method. The segmentation stage considers a multi-channel input, consisting of the raw input image and the dense optical flow representation in the corresponding frame. It also considers seed locations provided by the clustering of the optical flow vectors.

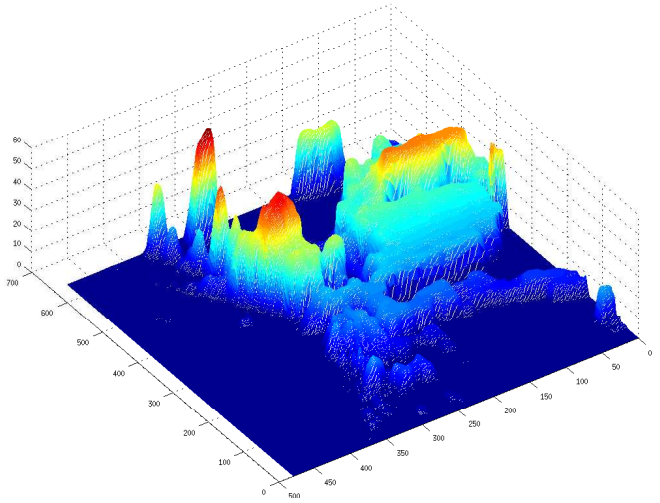


Fig. 2: 3D representation of the flow magnitude, illustrating well defined "hills" and "valleys". The corresponding visible image that generated this flow is shown in Figure 4.

a reduced influence in the final segmentation. To estimate the confidence measure, we employ an intuitive method based on linear subspace projections [25], obtaining "eigenflows" from the original optical flow signal. In summary, the confidence measure is estimated given the assumption that the better the displacement vectors can be reconstructed, the more reliable they are around their neighborhood. Therefore, the reconstruction error of the flow vector is used as a confidence measure.

The final watershed segmentation is performed over "image" $\mathbf{S}$, whose $i$-th element is given by

$$s_i = \frac{1}{1+\alpha_i}(v_i + \alpha_i |f_i|) \tag{1}$$

where $\alpha_i$ is the confidence measure for the $i$-th optical flow vector in $\mathbf{F}$, and $v_i$ and $f_i$ are the $i$-th element of $\mathbf{V}$ and $\mathbf{F}$, respectively. The operator $|\cdot|$ represents the magnitude of the flow vector. The experiments in Section V show the results of the combined segmentation on $\mathbf{S}$.

### C. Outlier Removal and Clustering

As mentioned, in addition to being included as a topological region in the segmentation process, the flow vectors can also be used as efficient seeds in the segmentation. An important step once the dense optical flow estimation is performed is to apply RANSAC [26] to reduce the influence of noisy vectors. Assuming an approximation that the cameras move on a surface plane (since they are mounted on a vehicle) the great majority of correct optic flow lines follow the same direction as their neighbours. Therefore, outlier removal is done by simply eliminating optic flow vectors whose direction differ in a given amount from the average optic flow direction.

This reduces the impact of outliers that potentially cause false local minima in the watershed segmentation.

The optical flow vectors in the camera are rarely induced from one object only. In a normal scene there are several objects, each inducing optical flow in the image. Therefore, it is possible to cluster the optical flow vectors according to their source.

Within a frame, each induced optical flow vector is described by its $x$ position, $y$ position, angle $\theta$, and magnitude $A$. Let $\mathbf{D}_f$ represent the description of the optical flow vectors within the given frame, $f$. For this frame, the optical flow information is given as

$$\mathbf{D}_f = \begin{bmatrix} x_1 & y_1 & \theta_1 & A_1 \\ x_2 & y_2 & \theta_2 & A_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_K & y_K & \theta_K & A_K \end{bmatrix}, \tag{2}$$

where $K$ is the number of flow vectors in the frame. We cluster the optical flow vectors according to their location, magnitude and angle by employing Gaussian Mixture Model (GMM) in a 4 dimensional space. The center (mean) of each cluster is projected onto the 2-D image space is used as a seed in the segmentation process. A mixture model is a probabilistic model for representing the presence of sub-populations within an overall population. Formally, a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. GMM's are efficient in clustering within normal distributions. With the reasonable assumption that the distribution of the magnitudes of the induced optical flow vectors is normal [27], GMM's can be employed to cluster the data according to their source. The number of clusters is defined using a standard methodology in which an expectation-maximization algorithm estimates the finite mixture models corresponding to each number of clusters considered and using Bayesian inference criterion to select the number of mixture components, which is then set as the number of clusters [28]. This approach usually gives a number of clusters between 20 and 60 in typical outdoors images.

## IV. Multi-Modal Segmentation

Visible cameras are common imaging sensing devices that have been extensively exploited for robot perception, navigation and localization. However, due to the limited range of the spectrum (visible spectrum range from $0.4 - 0.7\mu$m) that these imaging sensors operate at, they have been

restricted by changing atmospheric, weather and illumination conditions in challenging environments. Recently, there has been increasing interest in using alternative imagery sensing modalities for robotics applications that are more robust to environmental conditions. Alternative image sensing modalities (e.g., near-infrared, thermal-infrared) can sense the environment at various electromagnetic wavelengths beyond the visible spectrum. Moreover, information from multiple modalities can be integrated to enhance the scene perception and understanding.

Thermal-infrared imagery, which captures radiation emitted from the surfaces of objects relative to their temperature (long-wave infrared (LWIR) wavelengths from $7 - 14\mu$m), is an alternative sensing modality that has been recently employed for several robotics perception applications [29], [30]. Thermal-infrared and visible images have very different statistics and power spectra [31]. Thermal-infrared cameras have several advantages compared to visible-spectrum cameras. Thermal-infrared cameras detect radiation emitted from the scene without using an external illumination source. Hence, it is invariant to lightning conditions and can operate over nights and in total darkness. These cameras are more robust in challenging environmental conditions such as presence of fog, or dust. Thermal-infrared images contain less high-frequency textual information (less clutter) compared to visible spectrum; thus can be employed to improve scene segmentation and understanding.

The advantages of using thermal-infrared images are more evident in the outdoors robotics application scenario, where stronger temperature variation is usually present. In addition, visible spectrum cameras often suffer from over or under exposure outdoors, depending on illumination conditions. In these cases, thermal-infrared cameras can often identify objects which are not always seen in the visible spectrum.

Therefore, we combine information from these alternative sensing modalities to provide enhanced scene segmentation and richer environment models. The proposed approach discussed in previous section can be extended to multi-modal images, and a full framework combining both modalities is proposed.

In this case, the processing pipeline is represented by the diagram shown in Figure 3. The visible and thermal-infrared images are assumed registered and calibrated (as discussed in Section V-A), which are fed into the dense optical flow algorithm [23]. In our implementation, the combination represented by the triangle ($\triangledown$) in Figure 3 is done by averaging the two images, although other techniques can be used [32]. Figure 4 illustrates the concept, showing two images, thermal-infrared and visible, and their corresponding optical flow vectors. Figure 5 shows the difference in optical flow energy for the thermal-infrared and visible modalities.

For the gathering and pre-processing the thermal-infrared data, we use the algorithm developed by Vidas et al. [33]. Modern thermal-infrared cameras usually consist of 14-bit monochromatic images. In natural environments, however, the intensity range is typically much smaller than 14-bits, and the histogram normalization to a conventional 8-bit image

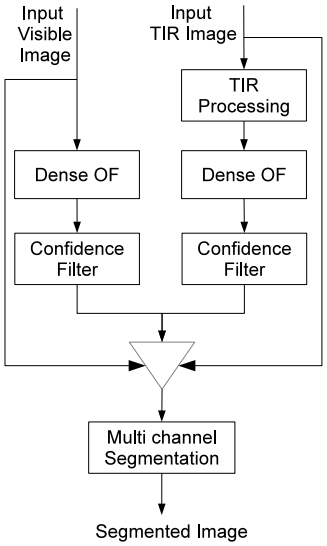
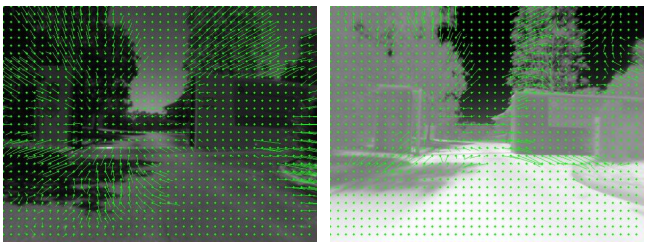
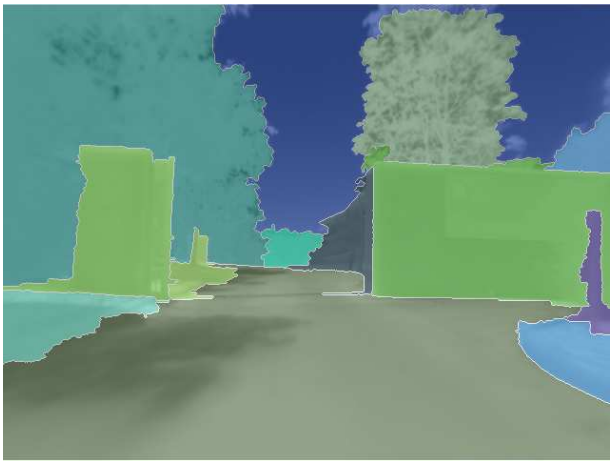



Fig. 3: Block diagram illustrating the full processing chain, combining the visible and thermal-infrared (TIR) images.



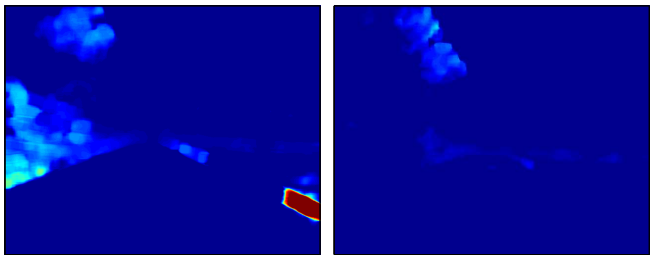
(a) Visible image flow. (b) Thermal-infrared image flow.



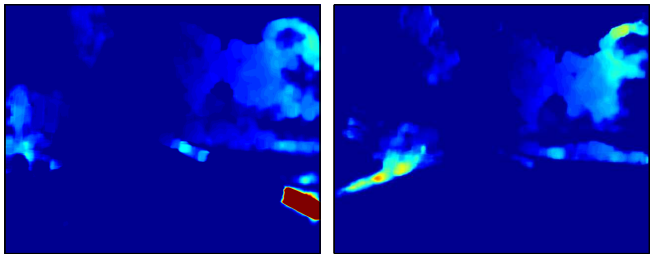
(c) Resulting segmented image.

Fig. 4: Illustration of the flow vectors and the resulting segmented image (overlaid with the original visible image).

(a) Visible Image (b) Thermal-infrared Image



(c) Visible image horizontal flow. (d) Thermal-infrared image horizontal flow.



(e) Visible image vertical flow. (f) Thermal-infrared image vertical flow.

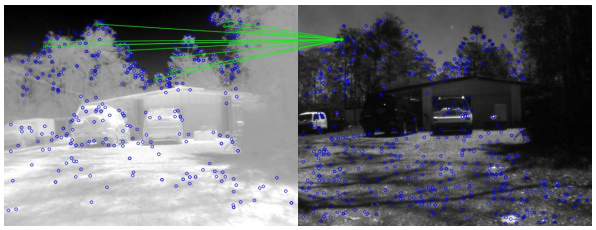
Fig. 5: Image illustrating the difference in optical flow energy for the thermal-infrared and visible modalities.

is an alternative to simplify processing and make it more effective. In this paper, we follow an efficient normalization procedure [33] which involves averaging the minimum and maximum intensity for a frame, and then normalizing the range spanning from 256 less than this average value to 256 greater than this value to the 8-bit interval of [0, 255], according to

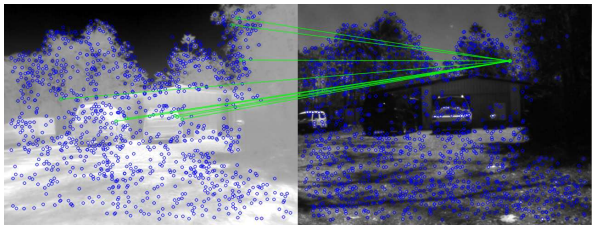
$$p_i = 255 \frac{p_i - \min(\mathbf{P})}{\max(\mathbf{P}) - \min(\mathbf{P})} \quad (3)$$

where $p_i$ represents the $i$-th element in thermal-infrared image $\mathbf{P}$. This scaling corresponds to a quantization factor of 2 and preserves the texture in areas of high and low contrast. Moreover, temporal smoothing is also employed [34] to avoid the normalization mean being shifted by more than $L$ levels relative to the previous frame.

When using both modalities, the inappropriateness of sparse optical flow becomes evident, as there is very small correspondence (particularly with respect to their positions in the image) between the flow vectors in each domain. Therefore, combined key point and variational methods [7] is not an adequate alternative in this case. For illustration purposes, the results in Figure 6 indicate that no correspondence is found between the thermal or visible images, for SURF and SIFT detector/descriptors tests. The green lines indicate the failed homography estimation between the modalities, as expected.



(a) SIFT matching.



(b) SURF matching.

Fig. 6: Examples illustrating the lack of matches between the thermal (left) and visible (right) modalities, for SIFT and SURF detectors/descriptors.

## V. Experiments

In this section we present experimental results, comparing the proposed approach to other traditional segmentation algorithms.

### A. Practical Considerations

For the experiments, a Thermoteknix Miricle 307K thermal-infrared camera and a Basler scA780 visible spectrum camera were used. The cameras were mounted on a car and driven in urban and semi-urban environments in Australia, with roads, trees, traffic lights, and buildings. The approximate route is indicated in Figure 7, covering nearly 6 kilometers. The thermal-infrared camera consists of a long-wave uncooled microbolometer detector sensitive in the $7-14\mu$m range, with resolution of $640 \times 480$. It is able to see objects in the $-20$ to $150\,^{\circ}$C, and it has a NEDT (Noise-Equivalent Differential Temperature) of 85mK. In our implementation, we used the ROS (Robotics Operating System) [35] package developed by Vidas et al [33], which contains a FFMPEG4 based driver for streaming off the thermal camera.[1]. This package also provides an efficient module for the calibration and registration of the thermal-infrared and visible cameras.

The data was logged using ROS and the resulting frame rate was approximately 14 frames per second. In addition to our own footage, the proposed algorithm was also tested with footage from the Hopkins 155 dataset [36]. Although the

[1]This open-source ROS package is available online for download from http://code.google.com/p/thermalvis-ros-pkg

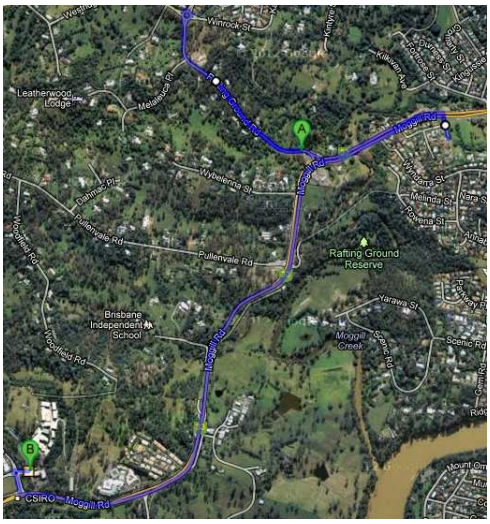



Fig. 7: View of the 6 kilometers driving route used for gathering data.

scope of this dataset is more focused on motion segmentation and does not contain thermal-infrared imaging, we used it in our tests since the dataset annotated with the segmentation ground-truth.

### B. Results

The results presented in the following compare the proposed algorithm with standard segmentation methods. The results are divided in three main parts:

- Standard watershed segmentation, without optical flow information.
- Separate image segmentation for the visible and thermal-infrared modalities, combining image and optical flow information, as discussed in Section III.
- Duo-modality image segmentation, combining visible and thermal-infrared data and their respective optical flow information.

Figures 8 and 9 illustrate the results for plain watershed segmentation (second row) compared to the proposed segmentation obtained by using optical flow information (third row). Finally, the fourth row shows the final result of the full segmentation pipeline, combining both modalities. In both figures, the results illustrate the benefits of using optical flow as well as the multi-modality setup as an alternative to visible cameras only. Several objects of similar characteristics are properly separated, with reduced over segmentation. A similar behavior is observed when applying the proposed approach to the Hopkins 155 dataset, with an example given in Figure 10. To quantify the results on this dataset, we compare our segmentation with the ground-truth provided for *some* objects (since only foreground objects are labeled) by overlapping segmented areas with the annotated foreground objects. For this evaluation, we use the average per-frame pixel error rate [37] given by

$$\epsilon(\mathrm{m}) = \frac{|\mathrm{XOR}(\mathrm{m}, \mathrm{g})|}{N} \tag{4}$$

TABLE I: Quantitative comparison according to (4), considering annotated foreground objects only. Lower values are better. These results are based on ground-truth data from the Hopkins 155 dataset.

| **Method** | **Score** $\epsilon$ |
|---|---|
| Standard Watershed [22] | 2890 |
| Trajectory Analysis [18] | 699 |
| Proposed | 801 |

TABLE II: Average execution time for the main stages of the pipeline. The 'Others' row include the pre-prossessing and summing junctions.

| **Stage** | **Execution Time (ms)** |
|---|---|
| Dense Optical Flow | 16.55 |
| Clustering | 9.90 |
| Flow Confidence Measure | 1.19 |
| Watershed Segmentation | 13.09 |
| Others | $\sim$ 2.00 |
| **Total** | 42.73 |

where m corresponds to each method's segmentation, g is the ground truth, $N$ is the number of images tested and XOR is the exclusive-OR operation. Table I shows the results, comparing the proposed algorithm with standard watersheds [22] and long term motion segmentation [18]. It is important to note that in both cases, the comparison is only indicative, since the first method [22] does not exploit motion and the second method [18] does not embed any still image segmentation analysis. We notice that using this metric, the long term trajectory presents the best results. One reason for this is that the objects used for the segmentation evaluation are foreground objects with motion, which is the focus of the long term trajectory algorithm. In contrast, the qualitative evaluation indicates that for background objects, our proposed method presents better performance, segmenting elements that are on the same plane and distance from the camera, since it also performs still image segmentation analysis.

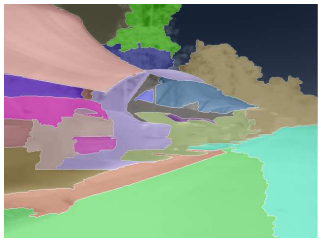
Another positive aspect of the proposed architecture is its relatively low computational complexity, running in real-time at 14 frames per second. Table II shows the average computation times for each part of the algorithm for the image resolution employed, running Linux Ubuntu on a 2.6 GHz Intel Core i7 with 4GB of RAM. The results indicate that with the current implementation it is possible to achieve approximately 23 frames per second. As a comparison, methods that perform motion segmentation using long term trajectories [18], for example, achieved approximately 0.39 frames per second running on the same machine.

## VI. Conclusions

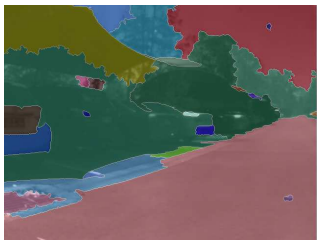
We have proposed a new video image segmentation method which jointly analyzes pixel and optical flow information. In each frame, optical flow "images" are combined with the original images, such that the output image presents better segmentation properties. In addition, optical flow vectors (which are described by position, angle and magnitude information) are clustered in space and the results are used

(a) Input Thermal-infrared image.

(b) Input Visible Image.

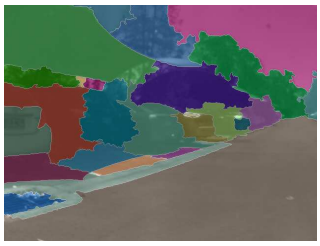


(c) Standard segmentation based only on the thermal-infrared image.

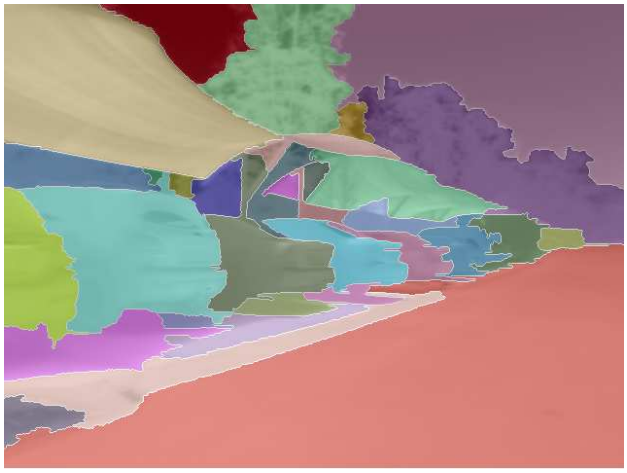


(d) Standard segmentation based only on the visible image.

(e) Segmentation based on the thermal-infrared image and OF.



(f) Segmentation based on the visible image and OF.

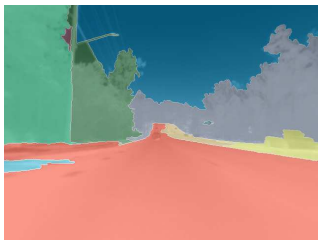


(g) Segmentation based on the visible and thermal-infrared imaging and OF.

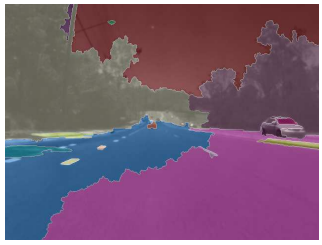
Fig. 8: Segmentation results inside parking lot. Key: OF: Optical Flow.

(a) Input thermal-infrared image.

(b) Input Visible Image.

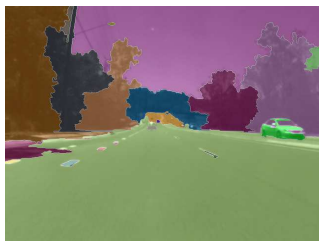


(c) Standard segmentation based only on the thermal-infrared image.

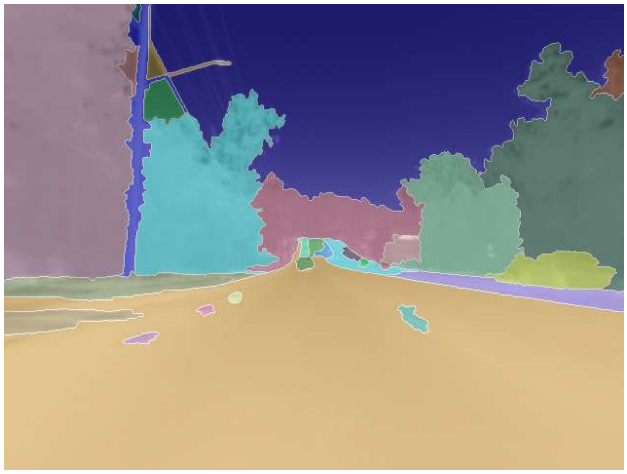


(d) Standard segmentation based only on the visible image.

(e) Segmentation based on the thermal-infrared image and OF.



(f) Segmentation based on the visible image and OF.



(g) Segmentation based on the visible and thermal-infrared imaging and OF.

Fig. 9: Segmentation results on highway. Key: OF: Optical Flow.

as seeds in the segmentation process. We illustrate the application of the method in several kilometers of driving in urban environments and semi-urban roads, with road, trees, traffic lights, signs and other elements. Apart from the main optical flow analysis, we combine information from visible and thermal-infrared cameras. The results indicate that the proposed method improves the segmentation performance compared to other well known methods.

One current limitation is that the system does not determine which modality is performing better at a given frame, or region inside a frame. Therefore, it cannot determine whether thermal-infrared or visible images should have more relevance when both modalities are combined. In this direction, future work includes incorporating a relative quality metric

(a) Flow image on the 'cars9' sequence.

(b) Segmented image from the 'cars9' sequence.

Fig. 10: Segmentation example on the Hopkings 155 dataset.

between the two modalities, in order to perform a more efficient modality combination. In addition, the combination of the images can be done adaptively, according to the global or local contrast of each modality. It is also important to evaluate the approach with other types of segmentation, such as mean-shift algorithms, for example. From an application perspective, future work will also consist of adding the method into a topological localization framework, applied to mobile robots.

## REFERENCES


[1] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260–280, 2008.

[2] F. Yi and I. Moon, "Image segmentation: A survey of graph-cut methods," in *Systems and Informatics (ICSAI), 2012 International Conference on*. IEEE, 2012, pp. 1936–1941.

[3] Y. Deng and B. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 8, pp. 800–810, 2001.

[4] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[5] J. Xiao and M. Shah, "Motion layer extraction in the presence of occlusion using graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1644–1659, 2005.

[6] M. Pawan Kumar, P. H. Torr, and A. Zisserman, "Learning layered motion segmentations of video," *International Journal of Computer Vision*, vol. 76, no. 3, pp. 301–319, 2008.

[7] T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. IEEE, 2009, pp. 41–48.

[8] G. Kim and A. Torralba, "Unsupervised detection of regions of interest using iterative link analysis," *MIT Open Access Articles*, 2009.

[9] Y. J. Lee and K. Grauman, "Collect-cut: Segmentation with top-down cues discovered in multi-object images," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3185–3192.

[10] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut: robust video object cutout using localized classifiers," vol. 28, no. 3, p. 70, 2009.

[11] B. L. Price, B. S. Morse, and S. Cohen, "Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues," in *Computer Vision (ICCV), 2009 IEEE International Conference on*. IEEE, 2009, pp. 779–786.

[12] X. Ren and J. Malik, "Tracking as repeated figure/ground segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on*. IEEE, 2007, pp. 1–8.

[13] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label mrf optimization," *Algorithms*, vol. 1, p. 3, 2010.

[14] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2141–2148.

[15] W. Brendel and S. Todorovic, "Video object segmentation by tracking regions," in *Computer Vision (ICCV), 2009 IEEE International Conference on*. IEEE, 2009, pp. 833–840.

[16] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller, "Multiple hypothesis video segmentation from superpixel flows," *Computer Vision (ECCV), 2010 European Conference on*, pp. 268–281, 2010.

[17] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, no. 5, pp. 539–546, 1998.

[18] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," *Computer Vision (ECCV), 2010 European Conference on*, pp. 282–295, 2010.

[19] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, "Track to the future: Spatio-temporal video segmentation with long-range motion cues," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3369–3376.

[20] P. Sturgess, K. Alahari, L. Ladicky, and P. Torr, "Combining appearance and structure from motion features for road scene understanding," in *British Machine Vision Conference*, 2009, pp. 1–11.

[21] S. Vidas, P. Moghadam, and M. Bosse, "3D thermal mapping of building interiors using an RGB-D and thermal camera," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 2303–2310.

[22] I. Vanhamel, I. Pratikakis, and H. Sahli, "Multiscale gradient watersheds of color images," *Image Processing, IEEE Transactions on*, vol. 12, no. 6, pp. 617–626, 2003.

[23] G. Farneback, "Fast and accurate motion estimation using orientation tensors and parametric motion models," in *Pattern Recognition, 2000 IEEE International Conference on*, vol. 1. IEEE, 2000, pp. 135–139.

[24] ——, "Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field," in *Computer Vision (ICCV), 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 171–177.

[25] C. Kondermann, D. Kondermann, B. Jähne, and C. Garbe, "An adaptive confidence measure for optical flows based on linear subspace projections," *Pattern Recognition*, pp. 132–141, 2007.

[26] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[27] N. Nourani-Vatani, P. V. K. Borges, J. M. Roberts, and M. V. Srinivasan, "On the use of optical flow for scene change detection and description," *Journal of Intelligent & Robotic Systems*, pp. 1–30, 2013.

[28] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *The computer journal*, vol. 41, no. 8, pp. 578–588, 1998.

[29] K. Nagatani, K. Otaki, and K. Yoshida, "Three-dimensional thermography mapping for mobile rescue robots," in *Field and Service Robotics*, 2012.

[30] S. Vidas and P. Moghadam, "HeatWave: a handheld 3D thermography system for energy auditing," *Energy and Buildings*, vol. 66, pp. 445–460, Nov. 2013.

[31] N. J. Morris, S. Avidan, W. Matusik, and H. Pfister, "Statistics of infrared images," in *Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on*. IEEE, 2007, pp. 1–7.

[32] S. Varjo, J. Hannuksela, and S. Alenius, "Comparison of near infrared and visible image fusion methods," in *Proc. International Workshop on Applications, Systems and Services for Camera Phone Sensing*, 2011.

[33] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, "A Mask-Based approach for the geometric calibration of Thermal-Infrared cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, pp. 1625–1635, 2012.

[34] S. Vidas and S. Sridharan, "Hand-held monocular SLAM in thermal-infrared," in *Control, Automation, Robotics and Vision, International Conference on*, 2012.

[35] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2, 2009.

[36] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on*. IEEE, 2007, pp. 1–8.

[37] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1995–2002.