Uni-Modal Versus Joint Segmentation for Region-Based Image Fusion

J. J. Lewis, S. G. Nikolov, C. N. Canagarajah and D. R. Bull
Centre for Communications Research
University of Bristol
Bristol, BS8 1UB, UK
{John.Lewis, Stavri.Nikolov, Nishan.Canagarajah, Dave.Bull}@bristol.ac.uk

A. Toet
TNO Human Factors
Kampweg 5
3769 DE Soesterburg
The Netherlands
Toet@tm.tno.nl

Abstract - ***A number of segmentation techniques are compared with regard to their usefulness for region-based image and video fusion. In order to achieve this, a new multi-sensor data set is introduced containing a variety of infra-red, visible and pixel fused images together with manually produced "ground truth" segmentations. This enables the objective comparison of joint and unimodal segmentation techniques. A clear advantage to using joint segmentation over unimodal segmentation, when dealing with sets of multi-modal images, is shown. The relevance of these results to region-based image fusion is confirmed with task-based analysis and a quantitative comparison of the fused images produced using the various segmentation algorithms.***

Keywords: human segmentation, multi-modal segmentation, evaluation of segmentation, region-based, image fusion

1 Introduction

Segmentation aims to divide an image into perceptually homogeneous regions [1]. While many segmentation techniques have been developed, there is no general solution to this problem (yet). It is useful and necessary to be able to compare the quality of different segmentation techniques. Zhang [2] broadly divides segmentation evaluation methods into two categories: *Analytical Methods* which directly examine the theory behind segmentation algorithms; and *Empirical methods* which measure the quality of the segmentation results. Empirical methods are further split into *Goodness methods:* based on some desirable properties of the segmented image; and *Discrepancy Methods:* where segmentations are compared to some "ideal" segmentation or ground truth. While they require no *a priori* information, the major drawback of empirical goodness measures is the difficulty in selecting a measure that works well with many different types of images and a segmentation algorithm can be produced specifically to perform well under that measure.

Discrepancy empirical methods require *a priori* knowledge in the form of a "gold standard" segmentation. In practice the ideal segmentation is usually not known for natural images. The human visual system is, however, good at segmenting an image into various regions based on a variety of cues such as texture and colour. The Berkeley Segmentation Database [3, 4] has 12000 human segmentations of 1000 images, both colour and grey scale, from the Corel image data base. While the human segmentations differed to some degree they were found to have a considerable consistency.

There are various sensors available that produce sets of co-registered images (or images that have been registered) from multi-sensor and multi-modality instruments. This paper addresses the question of whether it is better to segment them individually or as a set.

1.1 Segmentation for Region-Based Image Fusion

The majority of applications of fusion are interested in features within the image, not in the actual pixels. Thus, recent work in image fusion has led to the development of region-based algorithms, for example [5, 6], which initially segment a set of multi-modal images and then fuse these images region-by-region as opposed to the more traditional pixel-by-pixel approach. There are a number of perceived advantages of this, including:

- Fusion rules are based on salient regions of an image rather than arbitrary pixels;
- Regions with certain properties can be manipulated to improve the usefulness of the fused image; and
- Fusion is based on salient regions rather than individual pixels reducing sensitivity to noise, blurring effects and mis-registration [6].

The quality of the segmentation algorithm is of vital importance to the fusion process as errors in the segmentation could lead to important features being distorted or missed altogether. For the correct features to be present in the fused image, ideally, the segmentation algorithm should have the following properties:

- Produce consistent (good) results on a variety of images from different modalities;

- Features should be segmented as single separate regions; and
- As small a number of regions as possible should be created, as the time taken to compute the fused image increases with the number of regions.

2 Automatic Segmentation

In this study we have considered three segmentation algorithms. Two are available for free download over the Internet: JSEG and a pyramid-based algorithm as part of the OpenCV library. The third algorithm was developed at the University of Bristol. These are briefly described in the following sections. Where possible the default parameters were used for the segmentations.

2.1 OpenCV: Pyramid Segmentation

This algorithm, part of the OpenCV library, uses multi-dimensional pyramids to produce segmentations in near real-time. After performing the pyramid transform on the image, the links between any pixel, a, on level i and its candidate father pixel b on the adjacent level are established if $p(c(a), c(b)) < threshold_1$. After the connected components are defined, they are joined into several clusters. Any two segments A and B belong to the same cluster, if $p(c(A), c(B)) < threshold_2$ where $p(c_1, c_2) = |c_1 - c_2|$. There may be more than one connected component per cluster. The values of $threshold_1 = 130$ and $threshold_2 = 30$ were chosen in this study as they gave relatively good results across the data set. A set of labelled boundary regions are produced. The regions produced are not necessarily connected and the edges tend to be very noisy. Some post processing was required in order to produce the region boundary map needed for comparison with other algorithms. New region labels are extracted from the regional maxima of the histogram of original label values. A clean label map is produced by thresholding the original map with these values.

2.2 JSEG

The JSEG segmentation algorithm was developed by University of California at Santa Barbara [7]. The algorithm, an unsupervised segmentation of colour - texture regions in images and video is described in [8]. It essentially works in two steps: colour quantisation and spatial segmentation. As the number of colours required to effectively segment an image is much less than that needed to display the image, colours are grouped into colour classes and replaced by colour class labels. The resulting colour class map is used to perform spatial segmentation. The spatial segmentation occurs in two parts. Firstly, a criterion for a "good" segmentation is applied to local windows in the class map resulting in a *j-image*. This shows possible boundary locations. A region growing sequence is used to produce the segmented image from the multi-scale j-images. The JSEG algorithm automatically works out variable values (such as number of quantisation levels, number of scales and thresholds for the merge algorithm) and these values are used in this paper. However, segmentations can generally be improved for a particular image by manually selecting these values.

2.3 Combined Morphological-Spectral Unsupervised Image Segmentation

An adapted version of the combined morphological-spectral unsupervised image segmentation algorithm (UoB_Uni), described in [1], was used, enabling it to handle sets of multi-modal images. The algorithm works in two stages. The first stage produces an initial segmentation by using both textured and colour cues. The detail coefficients of the Dual-Tree Complex Wavelet Transform (DT-CWT) are used to process texture. The gradient function is applied to all levels and orientations of the DT-CWT coefficients and up-sampled to be combined with the gradient of the intensity information to give a perceptual gradient. The larger gradients indicate possible edge locations. The watershed transform of the perceptual gradient gives an initial segmentation. The second stage uses these primitive regions to produce a graph representation of the image which is processed using a spectral clustering technique. A segmented IR and visible image is shown in Figure 1.

2.4 Joint vs. Unimodal Segmentation

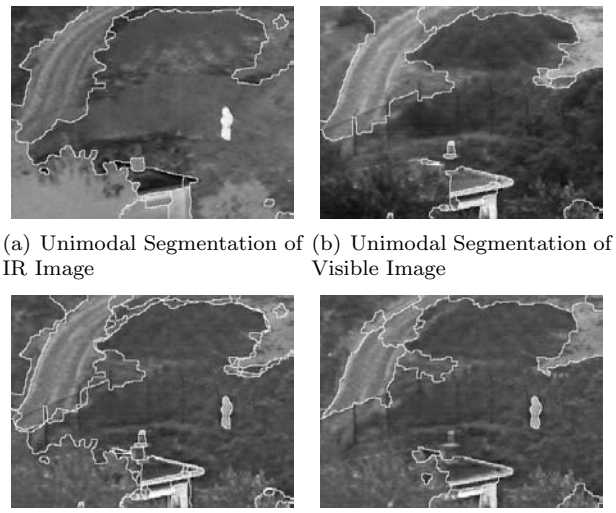
Until now, we have considered segmenting images separately, only using the information from an individual image to produce a single independent segmentation map. This *separate segmentation*, σ, of N images $I_1 \dots I_N$ produces the segmentations $S_1 \dots S_N$ where:

$$S_1 = \sigma(I_1), \dots, \; S_N = \sigma(I_N) \tag{1}$$

However, fusion tasks deal with sets of two or more images of the same scene containing complimentary information. A weak region in one image may correspond to a strong region in another image. There is a potential advantage of using information from all images to produce a single segmentation map for all images in the set since they can depict the same object or scene in different ways. This process, called *Joint Segmentation*, has been used in papers such as [5] and is shown below:

$$S_{joint} = \sigma(I_1 \dots I_N) \tag{2}$$

UoB_Uni was adapted to perform joint segmentation (UoB_Jnt). The effects of segmenting the images in different ways are shown in Figure 1. In particular, the inefficient segmentation given by union of the two unimodal segmentation maps, which is necessary in order to fuse the images is shown in Section 1(c). In general, jointly segmented images work better for fusion. This is because the segmentation map will contain a minimum number of regions to represent all the features in the scene most efficiently. A problem can occur for separately segmented images, where different images have different features or features



(a) Unimodal Segmentation of IR Image

(b) Unimodal Segmentation of Visible Image

(c) Unimodal Segmentations Overlaid on Region Fused Image

(d) Joint Segmentation Overlaid on Region Fused Image

Figure 1: Unimodal and Joint Segmentations

which appear as slightly different sizes in different modalities. Where regions partially overlap, if the overlapped region is incorrectly dealt with, artefacts will be introduced and the extra regions created to deal with the overlap will increase the time taken to fuse the images. Joint segmentation can overcome some of the problems of noise and other inaccuracies in an image to produce a more reliable segmentation. However, thus far, assessment of these results has been subjective [5] - based on the researchers opinions. A key aim of the new study is to quantify this objectively.

3 The Multi-Sensor Image Segmentation Data Set

Evaluation of segmentation results is traditionally very subjective. In order to quantify this objectively, as mentioned in Section 1, the Berkeley Segmentation data set [4] has provided a useful assessment tool for the comparison of segmentations of natural images. We have attempted to produce a similar facility for the assessment of multi-modal image segmentation, albeit on a smaller scale. Our aim is to provide an objective method for the assessment of multi-modal image segmentation and in addition to better understand how people segment multi-modal and fused images.

3.1 Image Database

Eleven varied sets of registered IR and visible images (14 IR and 11 grey scale visible images) were chosen giving both noisy, clean, cluttered and uncluttered images. All images are publicly available through the `ImageFusion.org` website [9]. These images were fused with three pixel-based fusion algorithms using the following transforms: Contrast Pyramid (PYR); Discrete Wavelet Transform (DWT); and the Dual-Tree Complex Wavelet Transform (CWT). The fusion methods are further described in [5]. This gave a total of 58 images to manually segment. A selection of these images are shown in Figures 3 and 4.

The images were pseudo-randomly distributed according to the following rules:

- Each subject sees only one image from each visible/IR/fused set (i.e. either the visible, IR or a fused image);
- Each subject should segment at least one visible image, one IR image and one fused image;
- An image should not be allocated for a second time until all available images have been allocated once; and not for a third time until all available images have been allocated twice; etc.

Thus, all images in the database were segmented a similar number of times and a subject's previous work did not influence their future work. As a result 5-6 human segmentations were produced for each image.

3.2 Segmentation Tool

A slightly modified version of the Java tool used to create the Berkeley Segmentation Database was used in our study to manually segment the images. The interface, shown in Figure 2, allows users to easily create segments by drawing on the image, split regions and adjust the boundaries between regions.

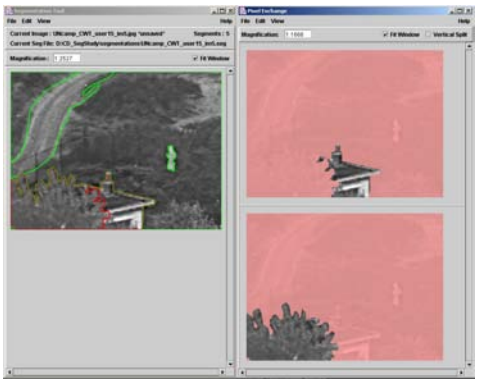


Figure 2: Segmentation Tool

3.3 Experimental Setup

63 (predominantly undergraduate students; 27 female; 34 male) with normal or corrected to normal vision were employed for one hour to manually segment five images each. Two sessions were run with over thirty subjects at each. The average subject's age was 21.3 years (standard deviation = 2.7 years). A mixture of CRT (36) and TFT (25) screens were used. Optical mice were used throughout. Details on subjects eyesight and normal computer usage were also recorded.

Initially, the subjects were shown two example images and given basic training on how to use the segmentation tool. Images that were significantly different from those in the database were used for the training. Similar to the instructions given to produce the Berkeley Segmentation data set [4] the instructions were kept intentionally general. This was so as not to lead the subject to produce a specific type of segmentation. Thus, variations in segmentations were due to differences in perception and not to some other aspect of the experimental set up. The instructions were to:

> *Divide each image into pieces, most important pieces first, where each piece represents a distinguished thing in the image. The number of things in each image is completely up to you. Something between 2 and 20 is usually reasonable. Take care and try and be as accurate as possible.*

3.4 The Human Segmentations

The segmentations produced by humans give the "ground truth" data. These were generally good although a few subjects either struggled to use the tool or failed completely to understand the task. Hence some 20 (out of 315) obviously erroneous segmentations were removed. Some example segmentations are given in Figures 3 and 4. In addition to these segmentations, this data base has been supplemented with a set of segmentations by an "expert", one of the authors.

3.5 Segmentation Error Measures

In order to evaluate the consistency of the human segmentations and to compare automatic segmentations with human segmentations, some method to measure the consistency between segmentation maps was required. We adopted the approach used by the Berkeley Segmentation Data Set team [3, 10], described briefly here. Precision-recall graphs are used for comparison where [10]: *Precision*, P, is the fraction of detections that are true positives rather than false positives; and *Recall*, R, is the fraction of true positives that are detected rather than missed. An *F-measure* captures the trade off between precision and recall as their weighted harmonic mean:

$$F = PR/(\alpha R + (1-\alpha)P) \qquad (3)$$

where α is the relative cost of R and P, set to 0.5 for this work.

For the P-R graph to be produced, the number of true and false detections must be computed. A segmentation is compared to each human segmentation for a particular image in turn and the scores averaged to give a single P, R and F value for each image. The correspondence is computed as a minimum cost bipartite assignment problem. The weight between a boundary pixel of an automatic segmentation and a human segmentation is proportional to their relative distance. All matches beyond some threshold are determined to be non-hits. Advantages of this measure are that it tolerates localization errors and finds explicit correspondences only (i.e. multiple detections are penalized).

The precision-recall graph for the human segmentations (comparing the consistency between different human subjects) is shown in Figure 5. This shows a fairly high level of consistency between segmentations.

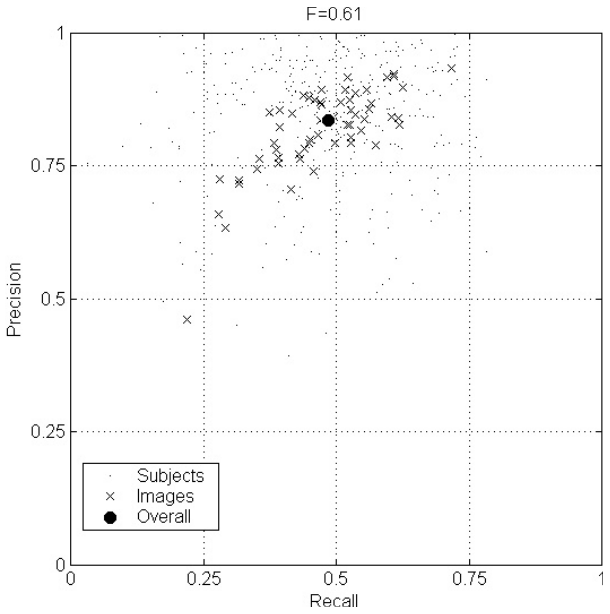


Figure 5: Precision-Recall Graph for the Human Segmentations Showing a Fair Degree of Consistence Between Human Subjects

4 Analysis of Segmentations with Human "Ground Truth"

Two comparisons were carried out: between the joint segmentations and the unimodal segmentations of the input images; and (as one cannot generate a joint segmentation of a fused image) between the joint segmentations of the input images and the respective unimodal segmentations of the fused images. To perform the comparisons, the quality (precision and recall values) of each type of segmentation is calculated by comparing the automatic and manual segmentations of the image sets.

4.1 Comparison of Joint vs Unimodal Segmentation of Input Images

Figure 6 clearly shows that joint segmentation (F = 0.57) out-performs unimodal segmentation (F = 0.38) for performing segmentation on sets of multi-modal images when using human segmentations as a ground truth. This is as expected as the strong features in one multi-modal image will compliment those in another image where the feature is weaker.

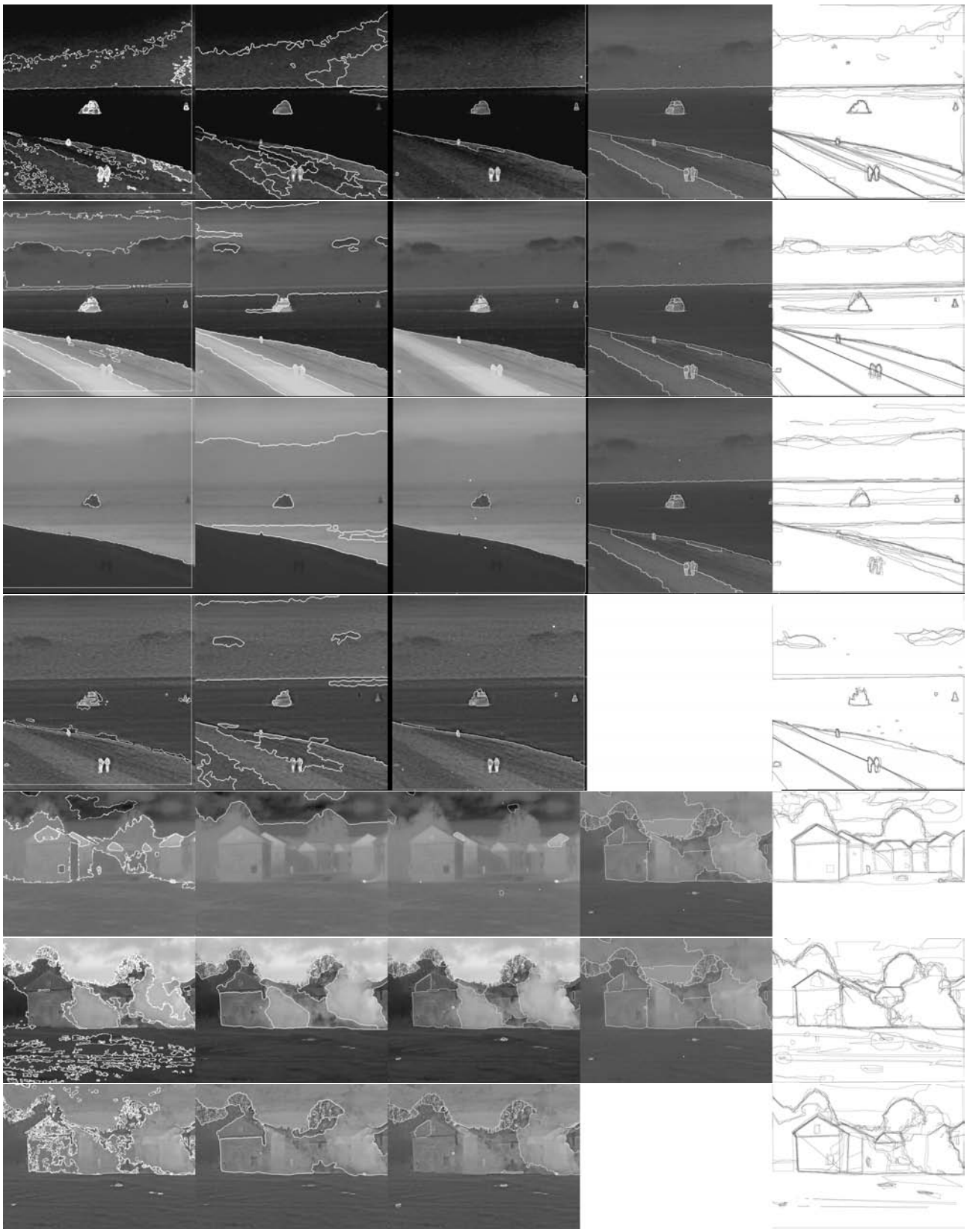


Figure 3: Segmentations: L→R: OpenCV, JSEG, UoB_Uni, UoB_Jnt, Human (Darker edges indicate more subjects contributed to it); T→B:"Sea" IR1, "Sea" IR2, "Sea" Visible, "Sea" CWT Fused; "Octec" IR, "Octec" Visible, "Octec" CWT Fused. The joint segmentations (UoB_Jnt) are produced from the visible and IR images of the set - joint segmentations are not possible from fused images.

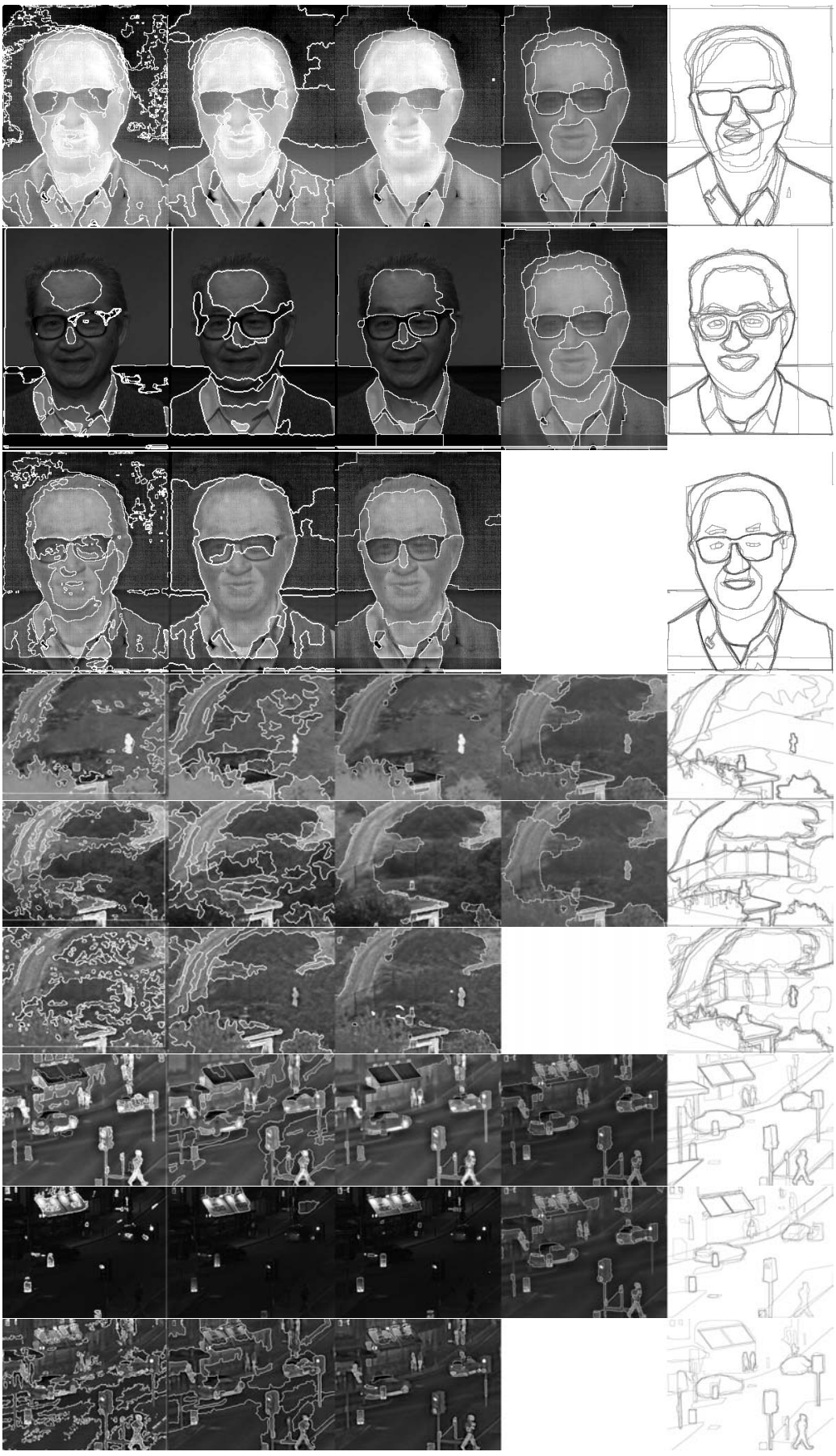


Figure 4: Segmentations: L→R: OpenCV, JSEG, UoB_Uni, UoB_Jnt, Human; T→B: "Glasses" IR, "Glasses" Visible, "Glasses" CWT Fused, "UN Camp" IR, "UN Camp" Visible, "UN Camp" CWT Fused, "Queens RD" IR, "Queens RD" Visible and "Queens RD" CWT Fused. The joint segmentations (UoB_Jnt) are produced from the visible and IR images of the set - joint segmentations are not possible from fused images.

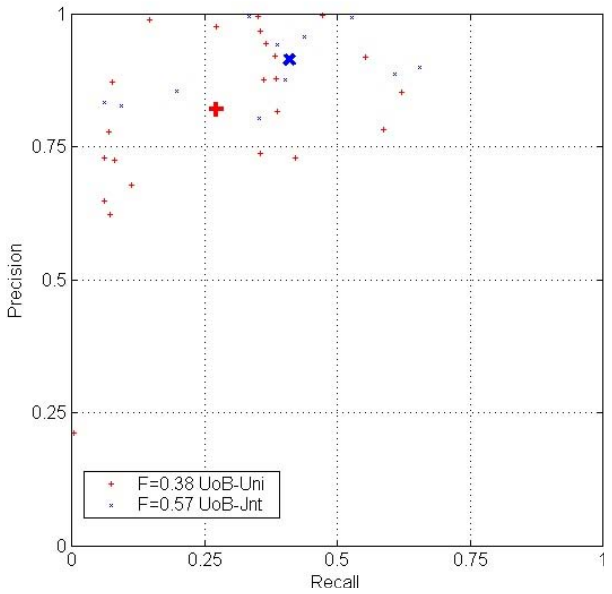


Figure 6: Precision-Recall Graph for the Comparison of the Joint (of the input images) and UoB_Uni Segmentations (of the input images)

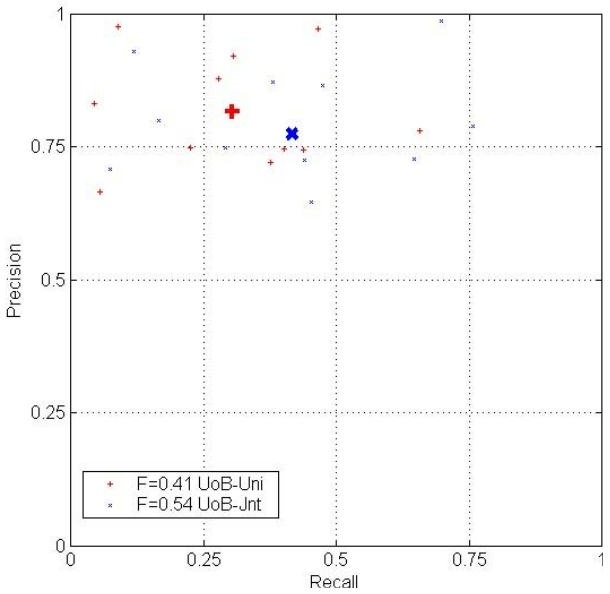


Figure 7: Precision-Recall Graph for the Comparison of the Joint (of the input images) and UoB_Uni Image Segmentations (of the fused images)

4.2 Comparison of Joint vs Unimodal Segmentation for Fused Images

Figure 7 shows that joint segmentations produced from the sets of input images (F = 0.54) out-perform the unimodal segmentations produced from the fused images (F = 0.41) This means that to segment a fused image, a better segmentation will be produced by a joint segmentation of the set of input images rather than actually segmenting the fused image. This is because the joint segmentation is based on information from both input images rather than just the fused image as with unimodal segmentation. This may be because information, useful to the segmentation process, is lost during the fusion process.

5 Analysis of Segmentations for Region-Based Fusion

Section 4 shows the advantages of using joint segmentation over unimodal segmentation and the effects of using different segmentation algorithms based on comparison with human segmentations. In this section, we consider the quality and usefulness of the fused images produced by using the region-based algorithm with the different segmentations.

5.1 Quantitative Fused Image Assessment

We have shown previously [5] that region-based fusion is at least as good as pixel-based fusion and in earlier sections of this paper that joint segmentation (UoB_Jnt) outperforms unimodal segmentation. But does good multi-modal segmentation translate to better fusion results? Two image fusion metrics, the *Xydeas and Petrovic* ($Q^{AB/F}$) [11] and the *Piella and Heijmans (IQM)* [12] were applied to the region-based fused images in the data set, using the four segmentation methods described in this paper. The results are given in the Table 5.1. The $Q^{AB/F}$ metric places the algorithms with UoB_Jnt performing best, followed by UoB_Uni, JSEG and OpenCV giving the worst results. The IQM metric rates the JSEG, UoB_Uni and UoB_Jnt methods above OpenCV but with little to pick between them. From these results, we conclude that region-based fusion using joint segmentation performs at least as well as, if not better than using unimodal segmentation, in terms of the fused images it produces.

Table 5.1: Comparison of Fusion Methods

Segmentation Method	$Q^{AB/F}$		IQM	
	Mean	σ	Mean	σ
OpenCV	0.5313	0.0974	0.6207	0.1199
JSEG	0.5692	0.0889	0.7023	0.1107
UoB_Uni	0.5770	0.0910	0.7072	0.1146
UoB_Jnt	0.5837	0.0812	0.7070	0.1119

5.2 Task Based Analysis

In Section 1.1, a requirement of the segmentation algorithm is given that each feature should correspond to an individual segment. We defined a task of finding the human figures in the "Sea", "Octec", "UnCamp", "Dune" and "Trees" image sets. For the figure to be detected it must be segmented correctly. Table 5.2 shows whether the algorithms managed to segment the

figure(s) out of the 36 images test images. The results are split into figure detection on the input images and on the fused images. The unimodal segmentations of each image (both input and fused) images were used. The single joint segmentation for all input images was also used for each input image as well as the fused images. Again, the joint segmentations clearly outperform the unimodal segmentations. It should be noted that the JSEG segmentation results are produced using the default settings and could be improved by tuning the algorithm to this task.

Table 5.2: Comparison of Segmentation Methods in Terms of Successful Figure Detections

Seg. method	OpenCV	JSEG	UoB_Uni	UoB_Jnt
Input Images				
Detected	7	3	6	13
Missed	8	12	9	2
Fused Images				
Detected	12	2	16	18
Missed	9	19	5	3

6 Conclusions

In this paper a new data set was introduced enabling the comparison of segmentation methods for registered multi-modal images. Three unimodal segmentation algorithms were used OpenCV, JSEG and UoB_Uni and in addition to these, UoB_Jnt, an adaptation of UoB_Uni allowing joint segmentation. A clear advantage to using joint segmentations over unimodal segmentation when dealing with sets of multi-modal images was shown both in terms of the segmentation quality and the fused image quality. The relevance of these results to region-based image fusion was confirmed with task based analysis and a quantitative comparison of the fused images produced using the various segmentation algorithms. The human segmentations produced by this study will be made available, on the `ImageFusion.org` site [9].

Acknowledgements

This work has been partially funded by the UK MOD Data and Information Fusion Defence Technology Centre. The original "UN Camp", "Trees", "Dune" and "Sea" IR and visible images are kindly supplied by TNO Human Factors Research Institute and the Octec images by David Dwyer of Octec Ltd. These images are available online at `ImageFusion.org`. The "Face" images are taken from the Human Identification at a Distance data set, produced by Equinox Corp. available at `equinoxsensors.com`.

References

[1] R. O'Callaghan and D. R. Bull. Combined morphological-spectral unsupervised image segmentation. *IEEE Transactions on Image Processing*, 14(1):49–62, 2005.

[2] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.

[3] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Eighth International Conference on Computer Vision (ICCV)*, volume 2, page 416, 2001.

[4] The Berkeley Segmentation Data Set. `http://www.cs.berkeley.edu/projects/vision/grouping/segbench/`, 2005. viewed October 2005.

[5] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and C. N. Canagarajah. Pixel- and region-based image fusion using complex wavelets. In *Information Fusion, Special Issue on Image Fusion: Advances in the State of the Art*. Elsevier, (in press), 2005.

[6] G. Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion*, 4:259–280, 2003.

[7] Y. Deng and B.S. Manjunath. Project: JSEG - Segmentation of color-texture regions in images and video. `http://vision.ece.ucsb.edu/segmentation/jseg/`, 2004. viewed Jan 2004.

[8] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001.

[9] The Online Resource for Research in Image Fusion (ImageFusion.org). `www.ImageFusion.org`, 2005. viewed December 2005.

[10] D. R. Martin, C. C. Faowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, colour and texture cues. *IEEE Transactions on Pattern Analysis and machine Intelligence*, 26(5):530–549, 2004.

[11] V. Petrovic and C. Xydeas. On the effects of sensor noise in pixel-level image fusion performance. In *Proceedings of the Third International Conference on Image Fusion*, volume 2, pages 14–19, Paris, France, 2000.

[12] G. Piella and H. Heijmans. A new quality metric for image fusion. In *International Conference on Image Processing, ICIP*, pages 173–176, Barcelona, Spain, 2003.