Multi-modal object detection via contextual foreground regions

By Iain Rodger[⋆†], Barry Connor[†], Neil M. Robertson[⋆]

[⋆]Visionlab, Heriot-Watt University, Edinburgh, UK. [†]Thales, Glasgow, UK

Abstract

We present a novel method for classifying objects in a static-cam surveillance scenario using colour-thermal imagery. The solution is applicable to the 24-hour surveillance problem and relies on exploiting scene-specific foreground context to determine regions of interest, leading to robust object detection. Moreover, a unified Bayesian framework is employed allowing the graceful exchange of contextual information across modes. This overcomes the dependence on individual sensor management as illumination conditions vary with time, a frequently occurring issue in outdoor surveillance when day transitions to night etc.

1. Introduction

Utilising video camera technology for surveillance and monitoring applications is commonplace and evident across a spectrum of sectors. This includes government agencies in a defence capacity to civilian usage, such as bolstering security around the home etc [Dockstader (2003), Boult (2004)]. Traditionally, surveillance systems would be reliant upon a human operator to perform effective object recognition and scene comprehension tasks. An operator would have to perform this task across a multi-camera system, requiring alertness in order to be efficacious. Whilst the capability of human vision is yet to be surpassed by computers, a potential weakness in the described system is the human-element. Recent advances in computer vision have addressed this issue by creating detection algorithms that relieve some of the burden from operators [Feris (2011)]. The research field of computer vision produces algorithms that, ultimately, aim to replicate or improve upon the human-visual system. These algorithms should be capable of exhibiting human-like performance but at computer-like speeds. This is the main ambition of computer vision, as a whole, hopes to realise and while a great deal of work has been directed towards creating such algorithms, it remains a long way off due to the inherently difficult nature of the problem [Porikli (2013)].

The task of effective object detection and classification in computer vision is complex due to a variety of factors. Such challenges include the wide range of object appearances under differing poses, positions and scales exhibited in the real world. Another difficulty is dynamic illumination which affects how objects appear under varying light conditions [Hager (1998)]. Suppose traditional surveillance events are likely to occur during day and night-time, it is imperative that corresponding detection algorithms can operate in these circumstances. Of course, it is dependent on what constitutes a *traditional surveillance event*. For the purposes of this research work we consider such an event to be an outdoor environment where objects within have no direct or limited control over illumination. Furthermore, image data would be collected from appropriate day and night sensors in a static-cam set-up. Possible surveillance scenarios include a covert reconnaissance mission or a border control environment [Boult (2004)]. A day sensor is essentially any modern colour camera. In contrast, a night sensor has to be capable of *seeing* in the dark and be unaffected by varying light conditions. An example highlighting the benefit of thermal imagery in low light level conditions is presented in Figure 1.

Thermal-based cameras, known as Thermal Imagers (TIs), are sensitive to Electromagnetic Radiation (EMR) in the Infra-Red (IR) domain and meet these requirements. However, IR sensors are under utilised compared to optical-band sensors. This is determined by several factors. Although there has been a notable reduction in cost factor for producing a quality TI, prices remain significant and present a barrier to wider use. Moreover, a price barrier coupled with night-time imaging capability places thermal sensing as a pursuit of mainly military and governmental organisations [*L*ee (2010)]. State-of-the-art TIs supply rich and textured imagery of target scenes providing effective surveillance capability at night or through fog and smoke. The focus of this work will explore problems relevant to 24-hour surveillance using a colour-band and Long Wave Infra-Red, (LWIR), camera.

(a)

(b)

Figure 1: Two pairs of colour-thermal imagery illustrating the benefit of IR sensing. Subfigure *(a)* presents a night-time surveillance scene where a colour camera cannot effectively perform. A thermal sensor, however, shows the presence of a person very clearly [*L*ee (2010)]. Subfigure *(b)* illustrates an LWIR sensor piercing through a foggy scene clearly showing the presence of a hot body in the far-ground that cannot be seen in the colour imagery [Rankin (2011)].

Problem Motivation: If we consider the knowledge that TIs are relatively uncommon, for the reasons mentioned above, there is a clear effect within computer vision research. Namely there is a corresponding lack of research available that focuses on object detection methods in the thermal domain, *relative* to the colour-band domain. This leads to the specific problem of classifying objects in thermal imagery, without a trained classifier, which is addressed in this paper. We hypothesise that inexpensive object classifications can be obtained in thermal imagery, via contextual knowledge transfer from colour domain detection information. Trained colour-based detectors are used to build FG context along with a recent segmentation method to create Foreground (FG) regions, corresponding to FG object activity. Foreground region knowledge acts as scene context and is exploited in the thermal domain.

Our goal is to show this approach may be suited to a 24-hour surveillance problem within certain constraints, where we can employ and exploit the vast research available for object detection in colour imagery. This allows a TI to be employed effectively without the burden of training additional classifiers for IR images etc. We will demonstrate this using a publicly available dataset and a self-collected, colour-thermal dataset for the purposes of the experiment. The methods used to build FG context are explored in Section 2 and how this is transferred across modalities is discussed in Section 3. Following this are details of all experiments carried out in Section 4, including data acquisition and reported significant results in corresponding subsections. Lastly a summary of findings is presented in Section 5.

2. Utilising Foreground Context

Generally the methodology presented contains three central elements to complete the stated aims. The first task is to create FG regions for a scene given detection information, using both thermal and colour data. Following this is the extraction of features or *blobs* in IR imagery. Lastly a domain knowledge transfer for the determined FG regions allows the extracted thermal features to be classified, without a trained classifier. In

essence this is a short summary of the entire process. This section will only focus on the underlying mechanics of building FG regions.

To approach the framed problem, foreground regions must first be constructed from observations. Two superpixel-based segmentation methods are considered to achieve this. The more complex of these, explored fully in Section 2.1, originates from a recent algorithm that utilises FG context to segment improved background regions in scenes showing dense object clutter [Rodger (2015)]. However, this algorithm is slightly modified to suit the nature of the colour-thermal problem. The second method employed presents a simpler approach towards the creation of FG regions corresponding to areas of activity within a scene, which is described fully in Section 2.2. Let us explore each method respectively.

2.1. *Modified Background Recovery*

The algorithm presented by [Rodger (2015)] extracts the underlying area for FG objects in an image, where said objects are namely people, and can be summarised as follows. For an image I_i of a video sequence, a set of n_i superpixel regions $SP_i = \{SP_{i,1}, ..., SP_{i,n_i}\}$ is obtained. A superpixel is simply an amalgamation of pixels that are similar in colour or close spatially. The reasoning behind this is to gain a reduction in dimensionality whilst retaining important image structure. The superpixel method used by [Rodger (2015)] is the well-known Simple Linear Iterative Clustering algorithm [Achanta (2012)], which we also elect to utilise. For set SP_i we must consider a set of adjacent or neighbouring superpixel regions, $A(SP_i)$. For every pair of SP regions, given as $(x, y) \in A(SP_i)$, a corresponding dissimilarity score can be computed using the Bhattacharyya distance metric [Aherne (1998)], B_{dist}. This function operates on normalised histograms for each colour channel, where normalised histograms provide a discrete estimate of each superpixels Probability Density Function (pdf). The dissimilarity distance B_{dist} for each paring $(x, y) \in A(SP_i)$ is used to populate a dissimilarity matrix, whose main diagonal would be zero and will exhibit symmetry.

To achieve this we adapt the previous algorithm of [Rodger (2015)] in the following ways to suit our multi-modal surveillance problem. Firstly, the computation of a dissimilarity matrix using B_{dist} needs to incorporate pixel information from thermal imagery. This is easily achieved by simply including the additional IR channel information. We can determine a normalised histogram H using all pixels bound within each superpixel, for information channels, c. The normalised histogram for each channel is given by H_C . Considering we are dealing with colour-thermal data, c will be of the form $RGBT$. Obviously RGB relates to colour and is interchangeable regarding chosen colour space, whereby T is thermal/IR data. Thus the distance between two superpixel distributions, SP_x and SP_y, can be given by the following equation:

$$D_H(SP_x, SP_y, C_{RGBT}) = B_{dist}(H_C(SP_x), H_C(SP_y)) \tag{2.1}$$

where D_H is the dissimilarity (or distance) between each superpixel, calculated for each information channel and multiplied through. As previously mentioned, the resulting distances between superpixels is used to populate a dissimilarity matrix BD for each SP relative to every other SP, as described by the pairing $(x, y) \in A(SP_i)$. The creation of BD matrix forms the basis for region merging and is a key process, hence the need to incorporate an additional channel for IR.

Crucially the merging process is affected by the incorporation of context within the scene, from the observed presence of FG objects. The context is determined by detecting and tracking objects of interest within a scene, fitting ellipsoid kernels to tracks and convolving these with merged superpixel regions. The Epanechnikov kernel [Epanechnikov (1969)] is chosen for this purpose and is determined by the kernel function $K_w = \frac{3}{4}(1 - u^2)$, where $|u| \leq 1$. After altering the scores present in BD with the mapped kernel, a merging function is required to determine which SP regions should be combined. A global threshold T_g to control the merging process can be determined using Eqn. 2.2. If we let Y denote the indexes of the columns of matrix BD, W_t as a weighting factor and xy are corresponding matrix elements, T_g is given by:

$$T_g = mean(\underset{y \in Y}{min}(BD_{xy})) \times W_t \tag{2.2}$$

which feeds directly into merging function U:

$$U(SP_x, SP_y) = \begin{cases} 1, & \text{if } BD_{xy} < T_g \text{ and } SP_{xy} \text{ adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

If regions from set SP_i meet these conditions then the merging function U simply joins the distinct regions, eliminating inter boundaries and the new region is relabelled. This process of exploiting FG context ultimately allows underlying areas that are populated by people, or other objects of interest, in an image to be recovered. Extracted regions will then correspond to likely object activity. A final smoothing step is also performed using an adapted Bayesian framework to carry forward elements of previous segmentations. The output is a normal segmentation label map and a binarised map which exhibits two distinct layers. From the referenced work these are the background-foreground and total background layers.

Moving on, we are also interested in more than one object class, namely cars and people. A capable detector should therefore be utilised to this end. Any suitably effective detector can, in effect, be slotted in. We choose Aggregate Channel Features [dollár (2014)] to detect pedestrians and SubCat [Ohn-Bar (2015)] to provide vehicle detections. These detectors are needed to create corresponding object trajectories, achieved by implementing a simple tracker that links detection points between frames. The subsequent tracks are then used to influence the region merging process giving mapped FG areas. Lastly, the final modification from the prior art involved focuses on the background-foreground layer computation. This binary array is adapted to provide probabilities over the FG region, so that pixel xy is not solely binary but instead exists in the range $[0, 1] = \{xy \in \mathbb{R} | 0 \leq xy \leq 1\}$.

We deem the adaptation as a foreground-background (FGBG) map. Originally the binary map is determined every frame with only some information being carried forward to future segmentations. However, the adapted method accumulates each layer over the whole process so that each pixel location in the map is added over the sequence. If we let I be the binarised FGBG image array, then the complete output Z, from adding each FGBG layer for a sequence of n frames, can be simply given by:

$$Z = norm(\sum_{1}^{n} I_n) \tag{2.4}$$

where the usual rules of matrix addition apply. To elaborate each element in FGBG I is added to the corresponding element in I_n, while output Z must share matrix dimensions with those added. Additionally array Z is normalised so array elements exist over $[0, 1]$. This approach differs in that previously, only a small amount of information from subsequent frames was carried over for the final segmentation smoothing stage. Instead we retain the whole layer and effectively stack them over time, allowing a more complete picture of the FGBG layer to be garnered from accumulating each array. Ultimately this region map corresponds to an object's likely area of activity, or conversely inactivity, which will be the basis of classifying thermal signatures via contextual foreground regions.

2.2. *Observing Superpixel Variance*

By contrast the algorithm described below is relatively simpler from that shown in Section 2.1. It relies on observing the variance of image pixels, within superpixel patches, over a video sequence for both colour and thermal imagery. The key underlying assumption is that high variance corresponds with foreground object activity, for a stationary surveillance scene. Once the most active superpixels are determined for each modality, these can be merged to form a SP variance layer. The remaining step is to accumulate detections over the sequence and create confidence maps for each object class detected. The detection bounding boxes are mapped to the SP variance layer and if it lies with a region of activity (high variance), it is convolved with a Gaussian kernel to give a spread of confidence scores in range $[0, 1]$ for each bounding box region. If the detection lies in a low variance region, the kernel confidence score is halved for the mapped bounding box. This process leads to a FGBG region map after normalisation. Again, this FGBG layer has areas corresponding to object class activity, akin to the FGBG layer given in Section 2.1.

To describe the method mathematically we must again consider the initial superpixel set. The first colour image in a sequence of length t is superpixellated, using SLIC algorithm, into a set of n_i regions $SP_i = \{SP_{i,1}, ..., SP_{i,n_i}\}$. Given we only need one superpixel representation to proceed, $i = 1$ in this case. The resulting label map can be used to observe pixel variance for *both* colour and thermal imagery, given both

modes of imagery should share dimensions. Pixels xy_t underlying each superpixel will then have an associated label, allowing pixels belonging to each superpixel to be read and stored in vector form. Let this be shown as $xy_t \in SP_n = A_n^t$, where each vector of pixels A is determined by the set label n and image sequence length t. For example, if the number of superpixels desired was 50 for an image sequence of length 100, then we would have 50 vectors of image pixels for each of the 100 images per modality or channel. Thus we can calculate the variance occurring for each superpixel per image by:

$$G_c^t = var(A_n^t) \tag{2.5}$$

where G_c^t is the array of n superpixel variances through a sequence of images of length t and the variance function var is defined as:

$$\frac{1}{N-1}\sum_{i=1}^{N}|B_i - \mu|^2 \tag{2.6}$$

where B is a variable vector of N scalar observations and μ is the mean of B.

From this it is easy to determine the superpixels exhibiting the most variance across the whole scene, by simply employing Equation 2.6 again but this time to array G for each channel. This produces a vector of length n for each channel c indicating variance observed within each superpixel over the entire sequence. Let the vector of SP variances per channel be deemed SPV_c. The next step is to perform a simple mean threshold to eliminate any SPs that are showing low to no signs of change over time. The final process is to combine information from the vector SPV_c to form a predicate for merging superpixels. If we let X_c be a set of integers corresponding to the most varying superpixels, determined above, for each channel c, then we can obtain a set of integers M that indicate which superpixels should be merged to form the FGBG region map. For instance if only 2 channels are used for this method, equivalent to IR and the blue colour channel for example, then the set of superpixels to merge can be obtained by performing a *union* of sets:

$$M = X_1 \cup X_2, \qquad X_{12} \in n_1 \tag{2.7}$$

where n_1 is set of integers for number of superpixels used. By merging superpixels in set M we create a SP variance layer from simply observing pixel value variance through time. However, no object class information is incorporated with the SP variance map unlike the method explained in Section 2.1. To achieve this a trained object classifier in the colour domain must again be utilised.

The basic principle is to store and accumulate any object class detections for incorporation with the FGBG layer. For each detection made per colour image, a corresponding bounding box will exist. These detection bounding boxes are then mapped to the SP variance layer, illustrated in Figure 2 where active regions are shown as white and low variance regions are black. The bounding box area can be convolved with a Gaussian kernel to provide it with a confidence score. For each convolved detection box, if it lies within an active region the scores remain unchanged. However, if the detection lies out-with the active regions then the scores are simply halved. This process remains the same regardless of class and is to reflect the uncertainty of making a detection in a region that has not exhibited much variance. The aggregation of these mapped and convolved detections, for each object class, leads to the creation of a FGBG map after a defined length of observation. Again the score aggregated map must be normalised so array elements exist in the range $[0, 1]$.

The end results from the methods explored in this Section and corresponding Subsections is a confidence region map for object classes, based upon observations for a given time period. The algorithmic framework implemented to create the FGBG maps for each object class operate under the assumptions explained previously. Namely the problem is posed as a static surveillance scenario where classifications in the thermal domain can be achieved without having a thermal classifier. This is achieved by utilising a trained detector in the colour domain and transferring the observed knowledge to be fully exploited. The methods outlined above provide the mechanism to effectively build up *prior* information that can be carried forward to help detect thermal blobs, the details of which can be explored more fully in Section 3.

Figure 2: Person detection bounding box is mapped to SP variance layer and convolved with Gaussian kernel to aggregate and build confidence map, which is the Foreground-Background layer. The resulting FGBG layer will be utilised to classify thermal signatures without a classifier at a later stage.

3. Transfer Knowledge Between Modes

The classification process for thermal features requires two stages and works under the assumption that the previously available colour signal is now unavailable. Thus all objects have to be identified without a trained classifier, purely in the thermal domain. The first stage is to extract thermal features from the target imagery which are inputs for the second stage, classification via a Bayesian framework using the obtained FGBG maps. The algorithm of choice to achieve thermal feature extraction is Maximally Stable Extremal Regions (MSER) [Matas (2002)], specifically the *VLFeat* implementation.

MSER: The core idea of using MSER to extract features in LWIR imagery towards the task of pedestrian classification is not a new one, as evidenced by [Teutsch (2014)]. However, we present a completely different approach to their work in a similar context. The MSER algorithm is ultimately a feature / blob detector that aims to extract stable connected components for level sets of a given image. The *stability* of regions is determined by how much variation exists within each binarised component. The algorithm is often chosen due to it's simplicity, robustness, it works for low resolution imagery and has small computational cost. One drawback is that it can often be hard to fine tune, although some parameters present in the *VLFeat* version allow even stable regions to be rejected by size, which can be useful in some scenarios. The MSER algorithm is utilised to extract numerous thermal blobs, for each thermal image frame in a test sequence. Each blob is classified as a belonging to one object class, or not, by utilising previously obtained FGBG maps for each object class in a Bayesian scheme.

Bayesian Framework: We employ Bayes theorem [Papoulis (1984)] to compute the probability of a detected object to exist given the condition of its surrounding region probability. Using the commonly given Bayes proportionality $P(A|B) \propto P(B|A)P(A)$, we can express our posterior probability as $P(Obj|R)$. The object is the extracted MSER blobs and the region R is determined by the earlier obtained FGBG map, relating to each object class. This is shown in Equation 3.1:

$$P(Obj|R) = \frac{P(R|Obj)P(Obj)}{P(R)} \tag{3.1}$$

where $P(Obj)$ is the *prior* information for objects, $P(R|Obj)$ is the likelihood which usually expresses a prediction model for given data and the denominator $P(R)$ is simply a normalising constant.

The prior information is the output FGBG layers from the methods outlined in Section 2, as it is an aggregation of detection observations shown as a confidence map. In other words it is the state of knowledge before the experiment to classify thermal features. Thus, to calculate the posterior $P(Obj|R)$ for an MSER blob belonging to an object class, or vice versa, it is a simple case of choosing a likelihood probability for each object class and summing probability values from the relevant prior FGBG maps.

Every classification task is treated as a binary problem where an MSER blob can either be an object such as a person, or not a person and two probability values must then be computed to reflect this. The correct FGBG map created from specific object class detections must be used to determine the corresponding posterior

Figure 3: Registered colour-thermal image pair from OTCVBS dataset showing two people in a sparsely populated urban environment. [Davis (2007)].

accurately, where the FGBG prior is then inverted to calculate the probability of a thermal blob *not* being an object. Given it is always a binary problem a flat or uninformative likelihood of $P(R|Obj) = [0.5, 0.5]$ can be chosen, at least initially, to produce results that are free from bias.

An uninformative likelihood essentially tells us that a thermal blob belonging to an object class or vice versa is equally likely. Lastly, the probability of prior $P(Obj)$ for each MSER blob is simply an average of corresponding FGBG pixel values at locations where extracted thermal features appear. Once probabilities for every MSER have been obtained and normalised, the highest value indicates if the blob is a specific object class or not. These posterior values for every frame are stored along with corresponding MSER bounding boxes to be evaluated using human ground truth data.

4. Experiments & Results

The described methods for classifying thermal signals in a surveillance scenario, without a trained classifier, are tested using two datasets. For each dataset the FGBG confidence maps are constructed using three differing lengths of observation, for every defined object class. The output probabilities for each object class and associated blob bounding boxes are then used to calculate classification accuracy via human ground truth information.

4.1. *Data Acquisition*

For the experiments undertaken two datasets are employed. Firstly, a publicly available colour-thermal dataset is obtained from the OSU Color-Thermal Database - the OTCBVS dataset [Davis (2007)]. This set of LWIR and colour imagery is already registered spatially and temporally, whilst also being a static-cam surveillance scenario containing a sparse number people as foreground objects. This provides an excellent platform to initially test the algorithms over. Example imagery from the OTCBVS set is illustrated in Figure 3. The second dataset was collected using a low-cost colour camera embedded in a mobile device and a state-of-the-art TI produced by Thales, the Catherine MP [Crawford (2006)]. This set contains lots of clutter and an additional foreground object of interest, the car. It represents more of a real-world urban surveillance scenario in contrast with the relatively *sanitised* OTCBVS dataset.

The Catherine MP LWIR uses an integrated detector cooler assembly which comprises a 640×512, 20μm pitch QWIP array, sensitive to long wave infrared radiation at wavelengths of 8μm to 12μm at a frame rate of 100 Hz. Given the accompanying colour camera has a frame rate of approximately 30 Hz with a larger spatial resolution, any image data collected using these two sensors has to be processed before being useful for experiments. To elaborate the colour and thermal imagery must be registered spatially so they show the same image plane, as well as temporally registered where objects within both images exist at the same point in time.

To enforce spatial coherence between Catherine MP imagery and colour imagery a straightforward image alignment technique is used. Manually selected control points identifying common features in both images are chosen, which allows a transform matrix to be computed via a geometric mapping process. This transform matrix is then utilised to perform the spatial transformation of imagery. It is acknowledged that much more

Figure 4: Registered colour-thermal image pair from self-collected dataset showing people, cars and clutter in a populated urban environment.

sophisticated and automatic options exist to do this between multi-modal image sets, but for the purposes of this experiment they are not deemed absolutely necessary given a reasonable level of accuracy can be obtained to effectively map FGBG layers from one modality to the other. Temporal registration is a much simpler issue that only requires the difference in frame rates as a ratio. For the Catherine MP to colour camera situation this ratio is ≈ 3.334, so for every 3 frames traversed in the colour sequence, 10 frames have elapsed in the LWIR sequence. This is enforced to create a real world colour-thermal dataset with high-quality, LWIR thermal imagery. An example of this image set is provided in Figure 4.

4.2. *Test Conditions*

For both datasets the image sequences are split into two sets. The first set is used to build FGBG / prior maps for each object class, using *3* different lengths of observation in terms of number of images. The second set of images is used to test the algorithmic framework for classification, under the assumption that the colour information is useless (i.e. night-time), where only thermal imagery is utilised. The test sets are manually ground truthed for both datasets. The OTCBVS contains only people as foreground objects while the collected dataset contains both people and cars. Bounding boxes are generated for these object classes.

The algorithm parameters, such as MSER variations, are kept constant throughout. The only change is the number of superpixels is doubled from 50 SPs for the OTCBVS data to 100 SPs, accounting for the larger resolution present in the collected dataset. A flat likelihood of $[0.5, 0.5]$ is used for all experiments, where probabilities and FGBG maps are generated for each object class via both methods. The length of observation for the OTCBVS set is multiples of 1000 images, with a test set of roughly 1000 images. The higher quality, self-collected dataset uses multiples of 250 images and a test set of approximately 350 images.

4.3. *Classification Performance*

The bounding boxes generated from MSER classification are used to evaluate the overall system accuracy by calculating overlap with ground-truth object bounding boxes. The metric used is *Accuracy* which is calculated in the usual manner:

$$ACCURACY = \frac{\Sigma TP + \Sigma TN}{\Sigma TP + \Sigma TN + \Sigma FP + \Sigma FN} \tag{4.1}$$

where TP is a true positive, TN is a true negative, FP is a false positive and FN is a false negative. For this experiment a TP is where an extracted MSER has a greater probability of being an object (car/person) and also has a greater than 50% overlap with the correct object class bounding box. A TN is where the dominant probability deems the thermal blob as *not an object*, and there is less than 50% overlap with ground-truth bounding box. A FP is obtained when the highest MSER probability classes the blob *as an object* but there is not sufficient overlap with ground-truth boxes. Lastly, a FN is when an ample overlap exists between MSER and ground truth boxes, but the dominant probability deems the blob as *not an object*. This overlap convention is defined in Equation 4.2 and is a common evaluation approach for detection problems [Everingham (2010)].

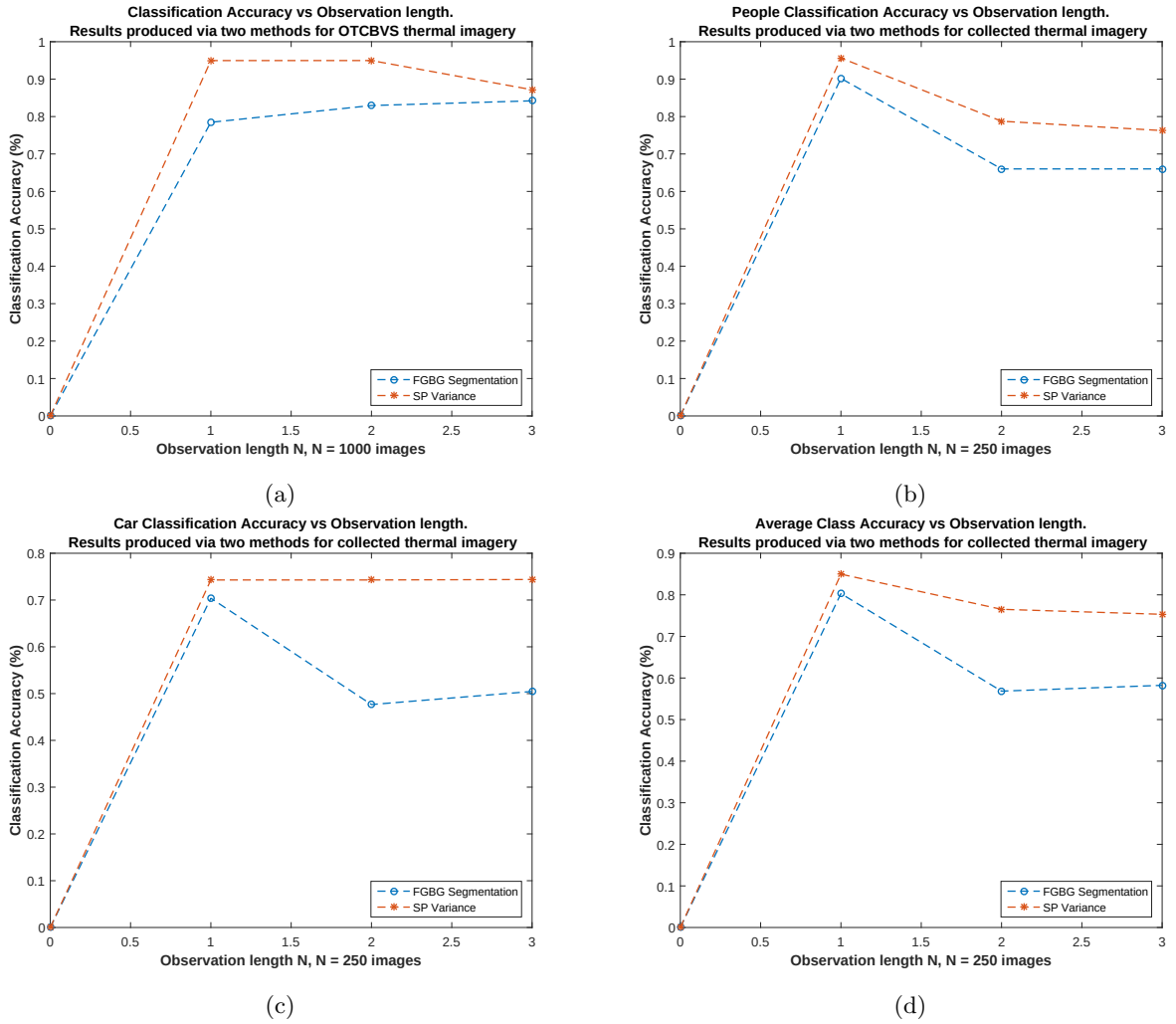

Figure 5: Overall accuracy results for two methods over varying lengths of observation are presented, where tests experiments are performed using two datasets. Subfigure *(a)* shows accuracy results for person classification over OTCBVS dataset. Subfigures *(b)* and *(c)* illustrate person and car classification results respectively over a self-collected dataset. Finally, Subfigure *(d)* gives an averaged class accuracy for two methods over the self-collected dataset.

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{4.2}$$

where $B_p \cap B_{gt}$ is the intersection of MSER bounding boxes B_p and object ground truth bounding box B_{gt}, while $B_p \cup B_{gt}$ is their union. All variations of accuracy results, for each algorithm, over the OTCBVS and self-collected dataset are presented in Figure 5.

The obtained accuracy results show clearly that the simpler method of superpixel variance, combined with ordinary detection information in the colour domain, is an effective method to classify thermal features. Or to a greater extent, it is a more robust and accurate method for building the FGBG prior maps that are used in the MSER classification process, when compared to the complex *Modified Background Recovery* approach. This is true regardless of the object classes examined. It also appears that in most cases the length of observation doesn't appear to have an overwhelmingly positive effect. Initially this seems counter-intuitive as increasing FG object evidence should provide a more accurate platform to classify thermal signatures.

However, there may be several non-obvious factors that mean this isn't the case. For instance, the adapted background recovery will tend to propagate errors through a sequence if the scene isn't densely populated. Thus the longer observation length will only compound this problem. Furthermore, the longer superpixel variance is

calculated through a scene can lead to key areas falling out of the SP variance map later on in the sequence as the scene changes. Perhaps a better implementation would negate very old variance information with new, updated scene information. In any case it is clear that while improvements can be made, the SP Variance method provides a more accurate FGBG map, feeding directly into the Bayesian classification scheme for extracted thermal features.

5. Conclusion

The main hypothesis put forward is the transfer of contextual knowledge across modes to potentially obtain object classifications, without using a trained classifier. Ultimately the experiments carried out have shown this to be true. Two methods for generating foreground contextual knowledge are presented and the output of these serve as input to a Bayesian classification scheme, which mediates the knowledge transfer between modes. Initially, a complex background segmentation algorithm is adapted to suit the problem posed, which in turn leads to a simpler solution utilising superpixel variance to determine regions of activity. From experiments carried out over two colour-thermal datasets it is readily apparent that the simpler of these methods, combined with the Bayesian classification framework, is better suited to generate contextual foreground regions to aid multi-modal detection. This solution needs further examination and optimisation before being deployed in a real night-time surveillance scenario. Nevertheless, a potential has been identified and it may provide a tractable solution to multi-modal detection and classification problems that rely on individual sensor management.

REFERENCES

Dockstader, S., Berg, M. & Tekalp, M. 2003 Stochastic kinematic modeling and feature extraction for gait analysis *IEEE Transactions on Image Processing* **12.8** 962–976

Boult, T.E., Gao, G., Micheals, R. & Eckmann, M. 2004 Omni-directional visual surveillance *Image and Vision Computing* **22.7** 515–534

Feris, R., Siddiquie, B. & Zhai, Y. 2011 Attribute-based vehicle search in crowded surveillance videos *Proc. International Conference on Multimedia Retrieval*

Porikli, F., Brémond, F., Dockstader, S., Ferryman, J., Hoogs, A., Lovell, B. C., Pankanti, S., Rinner, B., Tu, P. & Venetianer, P. 2013 Video surveillance: past, present, and now the future *IEEE Signal Processing Magazine* **30.3** 190–198

Hager, G. D. & Belhumeur, P. N. 1998 Efficient Region Tracking With Parametric Models of Geometry and Illumination *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20.10** 1025–1039

Aherne, F. J., Thacker, N. A. & Rockett, P. I. 1998 The Bhattacharyya metric as an absolute similarity measure for frequency coded data *Kybernetika* **34.4** 363–368

Rodger, I., Connor, B. & Robertson M. N. 2015 Recovering Background Regions In Videos Of Cluttered Urban Scenes. *IEEE International Conference on Image Processing.* **Proc. To Appear 2015**

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. 2012 SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **34**, 2274–2282.

Lee, D. 2010 Using Thermal Cameras to Secure the Homeland. *Web Article. "Photonics.com"* **Published Feb. 2010**, Accessed 01/08/15

Rankin, A., Huertas A., Matthies L., Bajracharya M., Assad C., Brennan S., Bellutta P. & Sherwin G. W. 2011 Unmanned ground vehicle perception using thermal infrared cameras *SPIE Defense, Security, and Sensing International Society for Optics and Photonics*

Dollár P.,Appel R., Belongie S. & Perona P. 2014 Fast Feature Pyramids for Object Detection *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* **36**, 1532–1545.

Ohn-Bar, E. & Trivedi, M. M. 2014 Learning to Detect Vehicles by Clustering Appearance Patterns *IEEE Transactions on Intelligent Transportation Systems.*

Matas, J., Chum, O., Urban, M. & T. Pajdla 2002 Robust wide baseline stereo from maximally stable extremal regions *Proc. of British Machine Vision Conference*, 384 – 396.

Teutsch, M., Müller, T., Huber, M., & Beyerer, J. 2014 Low Resolution Person Detection with a Moving Thermal Infrared Camera by Hot Spot Classification *Computer Vision and Pattern Recognition Workshop*

Davis, J. & Sharma, V. 2007 Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery *IEEE OTCBVS WS Series Bench; Computer Vision and Image Understanding* **106** 162 – 182

Papoulis, A 1984 Probability, Random Variables and Stochastic Processes *New York: McGraw-Hill* 38–39 & 78–81

Everingham, M., Gool, L.V., Williams, C. KI., Winn, J. & Zisserman, A 2010 The pascal visual object classes (voc) challenge *International journal of computer vision* **88.2** 303–338

Epanechnikov, V. 1969 Non-Parametric Estimation of a Multivariate Probability Density *Theory of Probability & Its Applications* **14.1** 153–158

Crawford, S., Craig, R., Haining, A., Parsons, J., Costart, E., Bois, P., Gauthier, FH & Cocle, O 2006 Thales long-wave advanced IR QWIP cameras *Proc. SPIE 6206* **6206H**