Human Pose Classification within the Context of Near-IR Imagery Tracking

Jiwan Han, Anna Gaszczak, Ryszard Maciol, Stuart E. Barnes, Toby P. Breckon
School of Engineering, Cranfield University, Bedfordshire, UK

ABSTRACT

We address the challenge of human behaviour analysis within automated image understanding. Whilst prior work concentrates on this task within visible-band (EO) imagery, by contrast we target basic human pose classification in thermal-band (infrared, IR) imagery. By leveraging the key advantages of limb localization this imagery offers we target two distinct human pose classification problems of varying complexity: 1) identifying passive or active individuals within the scene and 2) the identification of individuals potentially carrying weapons. Both approaches use a discrete set of features capturing body pose characteristics from which a range of machine learning techniques are then employed for final classification. Significant success is shown on these challenging tasks over a wide range of environmental conditions within the wider context of automated human target tracking in thermal-band (IR) imagery.

1. INTRODUCTION

Human behaviour analysis remains a major challenge within automated image understanding. To date a wide range of prior work concentrates on the challenges of human pose recovery and understanding within visible-band (EO) imagery with an eventual aim of reliable behaviour understanding [1–4]. Current research in pedestrian/human visual behaviour classification (and related activity recognition) is also largely related to the complexity of visual-band recognition for a complex and varied set of discrete activities (e.g. walking, meeting, loitering) in a complex visual environment [5–7]. Performance beyond basic activities is poor and often non real-time [7].

Here, by contrast we target basic human pose classification in thermal-band (IR) imagery with a view to the prioritization and assessment of activity within an automated surveillance context [8–10]. Unlike recent work in the field [2, 4], we leverage the key advantages of thermal imagery in the localization of limb positions. Despite some limited work on thermal-band limb localization [11, 12], prior work explicitly dealing with thermal-band (IR) imagery within a automated surveillance context is presently largely focused on human detection [8, 10, 13–15] and tracking [9, 16]. Here,

Figure 1. People Detection in thermal-band (IR) imagery using combined HOG features and SVM classification

leveraging both efforts in thermal-band person detection [10, 15] and subsequent limb localization [11, 12], we consider overall human subject pose classification targeting two varying pose classification problems:- 1) the identification of passive or active individuals within the scene and 2) the identification of individuals carrying weapons (or similar large objects).

These problems pose two very relevant challenges within automated image understanding that feed directly into the common Automatic Target Detection (ATD) pipeline which is commonplace in many autonomous sensing and surveillance tasks [15, 17, 18]. We present this work within the wider context of automated target tracking and reporting (Section 3, [10]).

The task of detecting active (i.e. *task* active) or passive individuals, with respect to body pose, within a given environment is motivated by the need to prioritize detected human targets within such a context for either further automated processing (e.g. face detection/extraction/recognition) or review by a human operator. This task can readily be considered in the general case as being both one of behavioural subtlety and the nuances of human behaviour. Whilst this makes automated classification challenging due to its subjective nature, here refine this problem to consider whether a given individual is active or passive toward the sensor or sensing platform encountered. Although challenging, a human observer can readily classify example in-

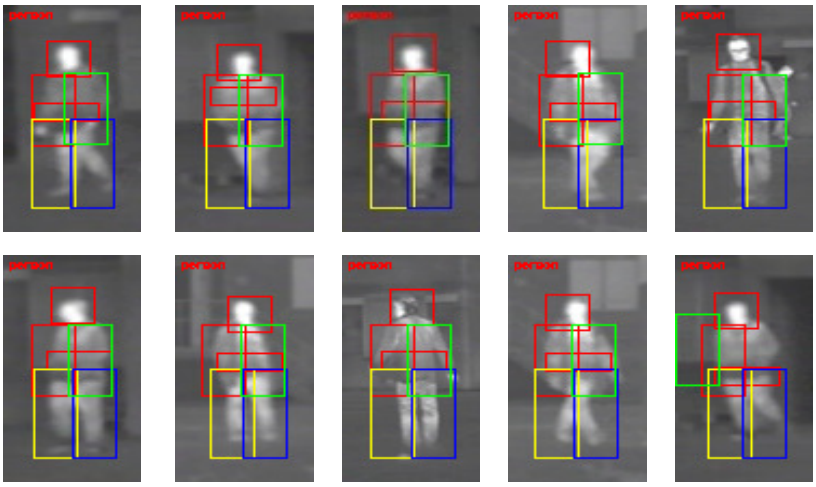

Figure 2. Person sub-region analysis for active/passive state determination

stances (e.g. Figure 3) into such cases, albeit somewhat subjectively. Here, by using hand annotated data sets to inform the generation of our decision boundary (i.e. machine learning) we aim to assess to what extend this difficult task can be performed automatically. To overcome its subjectivity and reliance on nuance we thus rely on human observers to form an informed training set for our task. In this way we capture an expert view of this *{active, passive}* divide within a given context.

The second task, weapon pose detection, relates both to the same motivation in autonomous sensing and surveillance tasks but is less subtle in the its formulation. Here we consider weapon pose detection for individuals carrying significant rifle-type firearms in the general sense. By contrast, a state which is nearly always obvious to the human observer without reference to nuance or subtlety. Both pose classification tasks are strongly correlated in both their application and in their complexity. Whilst the former can be considered a subtle task of (perhaps) detected nuances or general motion (i.e. *task* active) indicators, the latter should be a more obvious task of detecting particular pose characteristics indicating an individual armed with a significant weapon. Here we tackle each with varying complexities of feature-space employing a dense high-dimensional representation for the *{active, passive}* determination task and a simple low-dimensional feature space for the weapon pose detection task (i.e. *{armed, unarmed}*). Both are presented as discrete tasks (and results) for use within the context of a general automated target tracking and reporting capability (Section 3).

Active/passive classification is targeted using Histogram of Oriented Gradient (HOG) [19] features descriptors extracted over a set of image sub-regions identified earlier detections. Support Vector Machine classification is used achieving approximately 85% true positive on this highly challenging problem. By contrast, a range of classification approaches are compared for automatic weapon pose detection based on a low complexity geometrical feature descriptor extracted from prior localization of the limbs within the thermal image instance. A successful detection (true positive) rate of approximately 93% is achieved for this second task. Both pose classification tasks have applications in automated visual surveillance [5, 7, 8, 10, 20, 21], autonomous wide area search [15, 17, 18] and wider thermal-band (IR) imagery understanding tasks [22, 23].

2. HUMAN POSE CLASSIFICATION

We present our human pose classification tackling both of the outlined tasks, namely active/passive state determination (Section 2.2) and weapon pose detection (Section 2.3), within the context of automated visual tracking (Section 3) based on people detection for initial in-scene localization (Section 2.1).

2.1 People Detection

To facilitate subsequent pose classification, an initial localization of our human targets within the scene is performed using an adaptation of the highly successful visual-band pedestrian detection approach of [19]. The approaches uses Histogram of Oriented Gradient (HOG) features together with Support Vector Machine (SVM) based classification for people detection under a range of conditions.

The HOG descriptor is based on histograms of oriented gradient responses in a local region around a

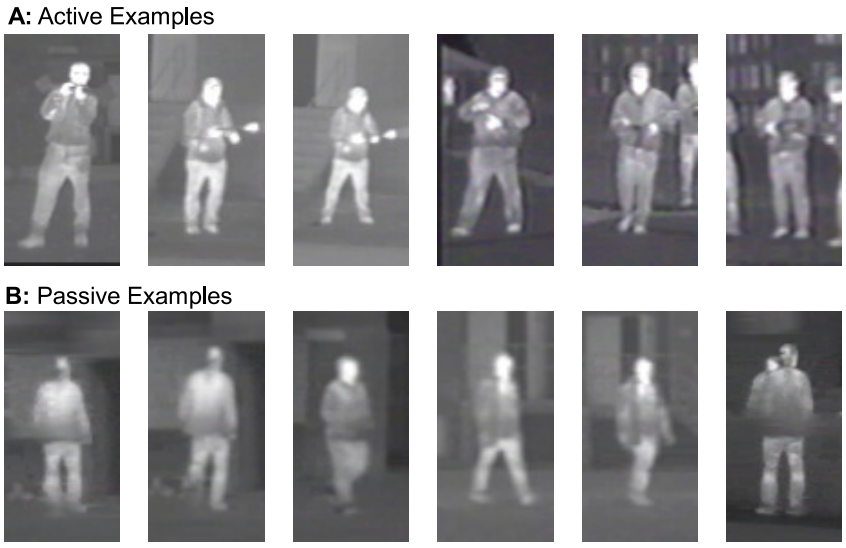

Figure 3. Example labeled training data examples active/passive state determination

given pixel of interest. Here a rectangular block, pixel dimension $b \times b$, is divided into $n \times n$ (sub-)cells and for each cell a histogram of gradient orientation is computed (quantised into H histogram bins for each cell, weighted by gradient magnitude). The histograms for all cells are then concatenated and normalised to represent the HOG descriptor for a given block (i.e. associated pixel location). For image gradient computation centred gradient filters $[-1, 0, 1]$ and $[-1, 0, 1]^T$ are used as per [19]. By re-sampling each localized image region to $w \times h = 64 \times 128$, we then compute the global HOG descriptor of this localized region using a block stride, $s = 8$ ($H = 9$, $n = 4$, $b = 16$), to form the input to the SVM classifier. This results in a 3780 dimensional HOG descriptor as $H \times n \times ((\lfloor \frac{w}{s} \rfloor - 1)(\lfloor \frac{h}{s} \rfloor - 1))$ where $(\lfloor \frac{w}{s} \rfloor - 1)$ is the number of blocks horizontally and $(\lfloor \frac{h}{s} \rfloor - 1)$ the number vertically within the $w \times h$ image region. We perform classifier training using a Radial Basis Function (RBF) kernel, with grid-based kernel parameter optimization, within a cross-validation based training regime. Training uses a data-set of approximately 2000 positive examples (people) and twice as many negative (non-people) examples randomly selected from the same source imagery. Detection is performed using a multi-scale sliding window approach with real-time performance available based on either using a cascaded approach [24] or GPU-based implementation [25].

An example of detection using the approach is shown in Figure 1 where we can see the successful localization of human targets within the scene based on this combined HOG/SVM people detection approach. This robust multi-scale detection approach is capable of the detection of both isolated people (Figure 1 top), congregating groups (Figure 1 bottom left), people under mild variation in orientation (Figure 1 top right) and people under varying scale (distance to camera, Figure 1 top right / bottom right). In general this use of an approach based on [19, 25] could alternatively be replaced by a feature-point driven approach (as per [8–10]) or cascaded Haar-like features (as per [15]).

2.2 Active / Passive State Determination

The proposed approach for active/passive state determination of a given human target (localized via Section 2.1) similarly uses Histogram of Oriented Gradient (HOG) features together with Support Vector Machine (SVM) classification. However, the identified sub-image, I_{sub} containing the human target is firstly processed to localize set of six rectangular limb sub-regions, *{head, torso-left, torso-right, lower-left-body, lower-right-body, torso-central}*, over which the HOG features are later extracted.

These sub-regions, illustrated in Figure 2, are localised using a set of SVM classifiers, one per limb region type l, that perform exhaustive pixel-wise search, over a down-scaled $width \times height$ $=64 \times 128$ version of the sub-image, I_{sub}. A feature vector, $\vec{v_i}$, constructed of concatenated raw pixel values and corresponding Laplacian filter response at each pixel location, p_i, for a given patch size, $r \times c$, such that $\vec{v_i} = \{p_i, ..., p_{rc}\} \cap \{f_{laplace}(p_i), ..., f_{laplace}(p_{rc})\}$ where $f_{laplace}()$ is the 2D Laplacian response at a given pixel location using a 3×3 filter kernel [26]. This feature vector forms the input to one of the six SVM classifiers,

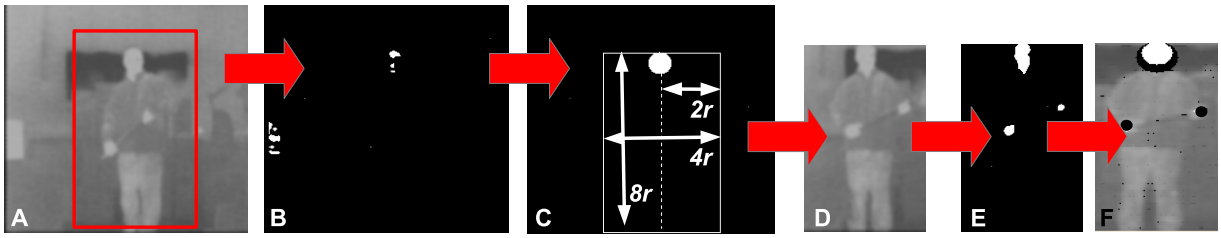

Figure 4. Limb position identification within thermal-band (IR) imagery

$\{SVM_l()\}$for $\{l\} = \{head, torso-left, torso-right, , lower-left-body, lower-right-body, torso-central\}$ corresponding to each limb type (class l). The patch size, $r_l \times c_l$, and hence dimension of feature vector $\vec{v_i}$, for each corresponding limb class, l, is $\{l\} = \{head = 20 \times 20, torso-left = 30 \times 40$, $torso-right = 30 \times 40, lower-left-body = 20 \times 50, lower-right-body = 20 \times 50, torso-central = 30 \times 10\}$. The pixel location, p_l, reporting the strongest positive class response (i.e. furthest from the SVM classification margin [27]) is identified as the location of limb type l such that $p_l = \underset{p_i \epsilon I_{sub}}{max}(SVM_l(\vec{v_i}))$. This set of SVM classifiers is each separately trained on a set of 900 positive samples and 1000 negative samples, for each limb type, using a RBF kernel, with grid-based kernel parameter optimization, within a cross-validation based training regime. Despite the exhaustive search of this classifier set, the use of down-sampling on initial sub-image sub-imageI_{sub} combined with simple feature construction and small patch sizes $r_l \times c_l$ still facilitate real-time performance. As can be seen from Figure 2 this results in sub-region alignment that both adapts to the configuration of target limbs within the scene but occasionally results in false localization (e.g. Figure 2 , lower right - *lower-right-leg*). A feature descriptor is now extracted from each identified limb sub-region (patch), at position p_l for limb l within the initial human target region I_{sub} (Figure 2), and concatenated to form a global descriptor for overall pose classification of the human target.

Initially feature approaches using either Laplacian pre-filtered pixel map [26] rescaled to a common dimension (as per [28]) and normalized histograms (as per [29]) were investigated for this purpose but with limited success. Instead we again use a HOG descriptor for each limb sub-region by first rescaling each one to $w \times h = 30 \times 40$, and computing the descriptor using a block stride, $s = 8$ ($H = 9$, $n = 4$, $b = 16$), as per Section 2.1. This results in a 288 dimension HOG descriptor for each sub-region ($H \times n \times ((\lfloor \frac{w}{s} \rfloor - 1)(\lfloor \frac{h}{s} \rfloor - 1))$ as per discussion of Section 2.1) which are then concatenated over the six sub-regions to form a 1728 dimension input to the SVM classifier. As per Section 2.1, we perform classifier training with grid-based parameter optimization for a Radial Basis Function (RBF) kernel under a cross-validation based training regime.

Training is performed using a sub-set of the positive (people) imagery set used for initial detection manually labeled as *{active, passive}*. Examples are shown in Figure 3 where we can see a set of both active examples (Figure 3A) and corresponding passive examples (Figure 3B). Given the nature of the problem-space for classification this labeling is somewhat subjective for border-line cases as can be appreciated from the set illustrated in Figure 3. Results for this task are presented in Section 4.1.

2.3 Weapon Pose Detection

Here weapon pose detection based on the pose classification of a given human target (localized via Section 2.1) is carried using out a contrasting approach based on inter-limb geometrical features available from limb position identification within the upper torso (Figure 4). These form the basis for a feature vector representation which is then classified using one of either k Nearest Neighbour (kNN), Neural Network or Naive Bayesian classification [27].

The identified sub-image containing the human target (Section 2.1) is processed using a multi-level thresholding technique to essentially find a set of Maximally Stable Extremal Regions (MSER) using a variant of [30]. It is initially processed using a Gaussian filter ($width = 7, \sigma = 1.55$, [26]) for noise removal (Figure 4A). Subsequently, following a similar approach for shape extraction used in [31], an initial seed threshold, t_0, is selected as the first cumulative histogram entry of the sub-image, C_i, greater than the mean histogram entry, m, over the sub-image, $t_0 = \underset{i}{min}\{C_i \geq m\}$. Upon application of this threshold, flood filling is subsequently used to grow thresholded regions by including connected pixels within $\tau_0 \pm \alpha$ in a region growing fashion ($\alpha \approx 10$). This is followed by morphological closing

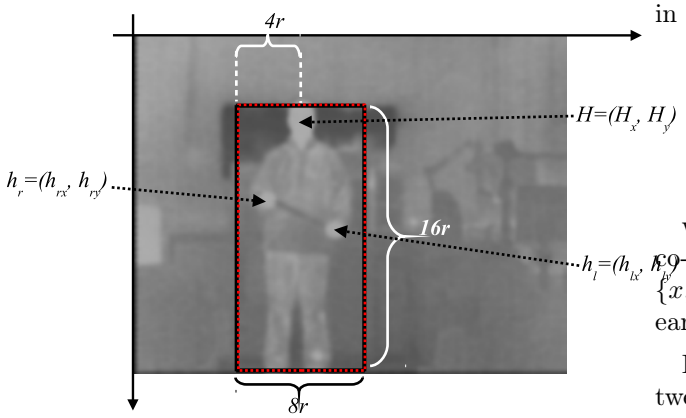

Figure 5. Identified limb geometry used for feature vector extraction

to improve region cohesiveness [26]. Contour extraction [32] is performed over the set of remaining regions to firstly identify a set of candidate head regions with an area less than threshold H_{area_1} and shape compactness measure [26] less than $H_{compactness_2}$(Figure 4B). If no such regions are present, this process is repeated iteratively with lower threshold, $t_k = C_{i-k}$, for $k = \{1..n\}$ whilst $t_k \geq t_{min}$. If several head candidate regions are identified within the sub-image, the upper-most occurring region with a shape compactness measure closest to 1 (perfect circle case) is selected (Figure 4C). Identified candidate head regions are approximated by fitting an extremal bounding circle, radius r, enclosing the concave candidate region within (Figure 4C). A tighter region of interest of dimension $width \times height = 8r \times 16r$ with the head positioned at topmost centre,$(4r, r)$ (Figure 4C/D), is defined for identification of the hands. The same process used for head candidate detection is now repeated to identify candidate hand regions below the head position in this region of interest. The two largest regions are identified as hands based on threshold criteria for the area of extremal enclosing circle of the region, H_{area_2} and shape compactness measure [26] less than $H_{compactness_2}$(Figure 4E & F). Where only one such hand region is detected both hand positions are assumed to be the same (e.g. clasped).

Based on these identified limb positions a number of geometric features are extracted using the identified limb geometry illustrated in Figure 5. Firstly we extract the hand positions normalised for both torso and image dimensions, $\vec{h_{l_{normalized}}}$and $\vec{h}_{r_{normalized}}$ as follows in Eqn. 1:-

$$\begin{aligned} \vec{h_{l_{normlized}}} &= (\frac{H_x - h_{lx}}{16r}, \frac{H_y - h_{ly}}{16r}) \\ \vec{h_{r_{normalised}}} &= (\frac{H_x - h_{rx}}{16r}, \frac{H_y - h_{ry}}{16r}) \end{aligned} \quad (1)$$

where elements H_i, h_{ri} and h_{li} refer to the position co-ordinates of the head ($\vec{H}$) and hands $\{\vec{h_r},\vec{h_l}\}$ for $i = \{x, y\}$ (see Figure 5) and r to the head radius defined earlier.

Furthermore we define the Euclidean distance between the identified hand positions as (Eqn.2):

$$d = \frac{\sqrt{(h_{lx} - h_{rx})^2 + (h_{ly} - h_{ry})^2}}{16r} \quad (2)$$

Finally considering the triangle formed between the three identified limb positions of head, $\vec{H}$, and both hands $\{\vec{h_r},\vec{h_l}\}$ (Figure 5) we can define an additional feature based on the ratio of the inscribed circle radius, r_{inside}, (i.e. maximal possible circle inside triangle $\triangle(\vec{H}, \vec{h_r}, \vec{h_l})$) to that of the circumscribed circle, R (i.e. circle intersecting points $\{\vec{H}, \vec{h_r}, \vec{h_l}\}$). Following from the definition in [33], this is presented as follows:

$$\rho = \frac{r_{inside}}{R} = \frac{2A^2}{abc(a+b+c)} \quad (3)$$

where A, a, b, c are defined as follows from our normalised positions (Eqn. 1) as:

$$\begin{aligned} A &= | \, h_{l_{normlized}x} h_{r_{normlized}y} - h_{r_{normlized}x} h_{l_{normlized}y} \, | \\ a &= \sqrt{(h_{l_{normlized}x})^2 + (h_{l_{normlized}y})^2} \\ b &= \sqrt{(h_{r_{normlized}x})^2 + (h_{r_{normlized}y})^2} \\ c &= \sqrt{(h_{l_{normalised}x} - h_{r_{normalised}x})^2} \\ &\quad \overline{+(h_{l_{normalised}y} - h_{r_{normalised}y})^2} \end{aligned} \quad (4)$$

A full derivation of Eqn. 3 with reference to Eqn. 4 is presented in [34].

Overall from these feature definitions we arrive at a short simple geometric feature vector, $v_{feature_6} = \{h_{l_{normlized}x}, h_{l_{normlized}y}, h_{r_{normlized}x}, h_{r_{normlized}y}, d, \rho\}$ in $\mathbb{R}^6$. Given the strong correlation of d to our other features we train our three classification approaches (kNN, Neural Network and Naive Bayes) over this feature space and additionally with d omitted in $\mathbb{R}^5$ ($v_{feature_5} = \{h_{l_{normlized}x}, h_{l_{normlized}y}, h_{r_{normlized}x}, h_{r_{normlized}y}, \rho\}$).

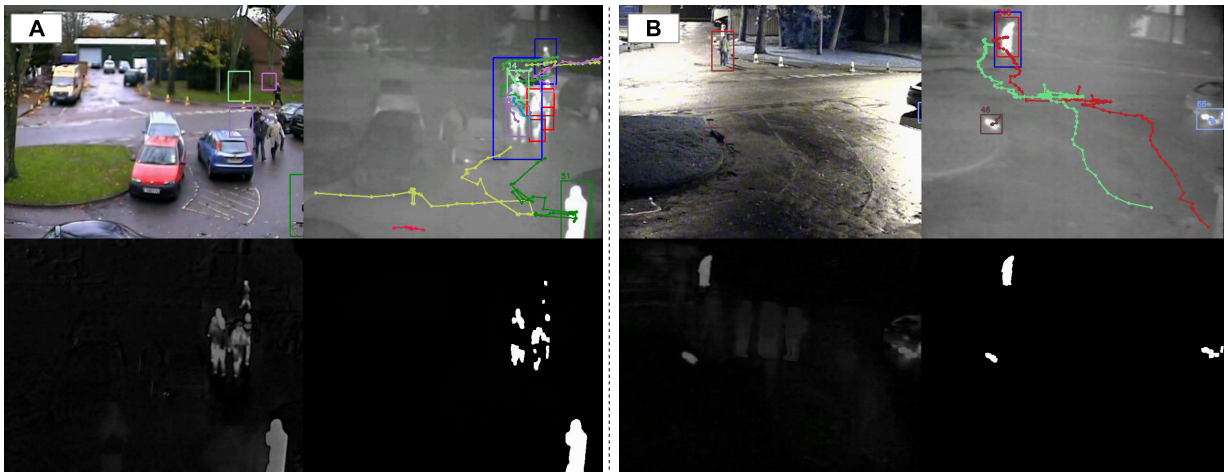

Figure 6. Multi-person in-scene tracking based on combined detection and optic flow based motion tracking.

Training is performed using approximately 10% of the data samples used in Section 2.2, manually labeled as *{armed, unarmed}* within a k-fold cross-validation framework. Varying values of parameter k within the kNN classifier are investigated with the best performing results presented (Table 1). The neural network is defined using a 3 layer and 7 hidden node topology with a sigmoid activation function which is trained over 1000 iterations using backpropogation [27]. Results for this task are presented in Section 4.2.

3. TRACKING CONTEXT

Our pose classification work is presented within the context of an automated visual surveillance scenario (as illustrated in Figure 6) where generic foreground objects are first identified based on a scene change detection approach (static camera case), classified using the approach outlined in Section 2.1 to be then classified based on pose using the approaches of Section 2.2 / 2.3. In Figure 6 we see two side-by-side examples (Figure 6 A & B) this scenario where both visible-band (top-left image) and thermal-band (top-right image) are used in each. The lower images of Figure 6 (A & B) show the foreground scene objects identified by scene change detection for each example in each of the visible and thermal band respectively. Within this context we use either a classical Mixture of Gaussian (MoG) based adaptive background model [10, 35, 36] to obtain foreground object separation (Figure 6 A) or alternatively scene segmentation based on optic flow [20] (Figure 6 B). The later has the distinct advantage of object separation by motion characteristics in the case of co-occurring scene objects and has also been shown to be robust to camera motion in other work [15].

This set of identified foreground objects are filtered to isolate objects corresponding to human targets using the approach outlined in Section 2.1. Each scene object is tracked using an optic flow driven tracking approach similar to [20] with data association between multiple scene targets handled using the classical Hungarian data association algorithm [37]. In Figure 6, each foreground object is shown with a varying coloured bounding box whilst those confirmed as human targets are shown by red bounding boxes. In the dual-band visible/thermal experimental setup illustrated in Figure 6 all processing is performed on the thermal-band (IR) imagery with the corresponding visible-band (EO) present for visualization purposes only.

Here, individual (per image frame) target occurrences present the input to the pose classification approaches outlined in Section 2.2 & 2.3 facilitating both temporal filtering and consistent reported via [10].

4. RESULTS

We illustrate our results for human pose classification for both the active/passive state (Section 2.2) and weapon pose detection (Section 2.3) tasks over specific test image sets captured using an un-cooled infrared camera (*Thermoteknix Miricle 307k*, spectral range: 8-12μm) as per Section 3 / Figure 6 context.

4.1 Active / Passive State Determination

Testing is performed over an extensive data set of thermal imagery gathered both from static camera surveillance (e.g. Figure 6, [10]) and mobile sensing platforms (via [15]). Overall a successful true positive classification rate of 85% is achieved for this binary two class

Figure 7. Successful (true positive) active/passive state classification

Figure 8. Successful (true positive) active/passive state classification

Figure 9. Unsuccessful (false positive) active/passive state classification

problem with the corresponding false positive classifications being split approximately evenly between the two possible classes. Examples of successful classification is shown on some detailed examples in Figure 7 where we can see a number of active examples including targets holding weapons (Figure 7 left) and passive targets (Figure 7 right). Further examples are shown over a larger subset of the data in Figure 8 where we can also clearly see the variation in ambient thermal conditions within the data sets in use. The subtly of the differentiation and potential ambiguity is clearly apparent from these examples (Figure 8). Conversely, a number of unsuccessful false positive detections are shown in Figure 9 illustrating the difficultly of the task in hand.

4.2 Weapon Pose Detection

The results of our weapon pose detection approach are shown in Table 1 for both feature sets, $v_{feature_6}$and $v_{feature_5}$, where we can generally see a largely consistent performance over both feature sets across all three classification approaches. Here we present the True Positive (TP, "weapon"), True Negative (TN, "not weapon"), False Positive (FP, "false weapon detection") and False Negative (FN, "weapon present but missed") in addition to the accuracy and precision measures which are defined as follows (Eqn. 5 & 6):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

From Table 1 we see that the neural network marginally produces the highest (correct) True Positive (TP) detection for weapons with the lowest (missed item) False Negative (FN) performance on $v_{feature_6}$. It is however, marginally outperformed in terms of True Negative (FN) by the Bayes approach which also exhibits desirable low False Negative (FN) performance. The accuracy and precision statistics for all three classification approaches (Table 1) on either feature set illustrate only marginal performance differences over the entire set. In general weapon detection performance (i.e. TP) can be reported as being high (up to 95%) and the likelihood a weapon is missed as low (5%) but all of the approaches suffer from significant false positive reporting (FP). This could be attributed to the simplicity of the feature sets in use and their reliance on quite basic limb localization in the first instance. A subset of the successful pose classification results using this approach is shown in Figure 10 where we see the successful determination of both armed (Figure 10, true positive = lower) and unarmed individuals (Figure 10, true negative = upper). Here (Figure 10) the identified positions of the limbs have been labeled with white circular overlays to help illustrate the reference positions from which the key geometrical features are being derived from (Section 4.1). Again, as per the earlier active/passive results (Section 4.1), we can see the considerable ambiguity that remains in this pose classification task relying upon such geometrical features alone (Figure 10 upper/lower). This is reflected within the performance results for this task given in Table 1 and areas of poor performance can be largely attributed to the limitation of the feature space with respect to this task.

5. CONCLUSIONS

Overall, we present results for two varying human pose classification tasks within the context of tracked individuals in thermal-band (IR) imagery extending prior work that focused soley on [8, 10, 13–15] and tracking [9, 16]. The results presented for the subtle active/passive classification task using a dense feature-space are somewhat similar to those achieved on the more prominent weapon pose detection task using a significantly smaller feature space. Both tasks illustrate the potential for human pose classification within thermal-band IR imagery in contrast to contemporary work in visual-band pose estimation [1–4] where the limitations of limb localization yield far lower second-order results when used as an input to behaviour classification tasks. Here we compare two approaches for limb localization, following either a multiple SVM classifier or MSER-driven approach, that notably extends the earlier thermal-band work of [11, 12] in this area. Overall, although the resultant true positive pose classification rates remain high for the tasks here, significant future work is required on reducing false positive reporting. Within the tracking context, following the work of [10], consideration within spatio-temporal feature space may aid in overcoming these issues.

Future work will investigate both the use of spatio-temporal features (e.g. 3D feature point representations [38]), recent advanced in real-time saliency detection [39] for application to thermal-band contour features [14] and the use of cross-spectral stereo [23].

The authors gratefully acknowledge the support of Stellar Research Services, Selex Galileo and HEFCE in this research activity.

Classifier	Feature Set	TP	TN	FP	FN	Accuracy	Precision
kNN ($k = 3$)	$v_{feature_6}$	92.5%	73.33%	26.67%	7.50%	0.81	0.70
Neural Network	$v_{feature_6}$	**95.0%**	68.33%	31.67%	**5.00%**	0.79	0.67
Naive Bayes	$v_{feature_6}$	92.5%	73.33%	26.67%	7.50%	0.81	0.70
kNN ($k = 19$)	$v_{feature_5}$	87.5%	73.33%	26.67%	12.50%	0.79	0.69
Neural Network	$v_{feature_5}$	92.5%	61.67%	38.33%	7.50%	0.74	0.62
Naive Bayes	$v_{feature_5}$	92.5%	**75.00%**	**25.00%**	7.50%	**0.82**	**0.71**

Table 1. Weapon pose detection performance
(TP = True Positive, TN = True Negative, FP= False Positive, FN = False Negative)

Figure 10. Successful (true positive) weapon pose classification (upper unarmed, lower armed with weapons)

REFERENCES

1. A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 44–58, Jan. 2006.
2. V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. Inf. Conf. Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, June 2008.
3. S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 1465–1472, IEEE, June 2011.
4. M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images," *International Journal of Computer Vision*, vol. 99, pp. 190–214, Mar. 2012.
5. E. Maggio and A. Cavallaro, *Video tracking: theory and practice*. Wiley, 2011.
6. A. Wiliem, V. Madasu, W. Boles, and P. Yarlagadda, "A suspicious behaviour detection using a context space model for smart surveillance systems," *Computer Vision and Image Understanding*, vol. 116, pp. 194–209, Feb. 2012.
7. S. Gong and T. Xiang, *Visual Analysis of Behaviour From Pixels to Semantics*. London: Springer, 2011.
8. B. Besbes, A. Rogozan, and A. Bensrhair, "Pedestrian recognition based on hierarchical codebook of SURF features in visible and infrared images," in *Proc. Intelligent Vehicles Symp.*, pp. 156–161, IEEE, June 2010.
9. J. Wang, D. Chen, H. Chen, and J. Yang, "On pedestrian detection and tracking in infrared videos," *Pattern Recognition Letters*, vol. 33, pp. 775–785, Apr. 2012.
10. T. Breckon, J. Han, and J. Richardson, "Consistency in muti-modal automated target detection using temporally filtered reporting," in *Proc. SPIE Electro-Optical Remote Sensing, Photonic Technologies, and Applications VI*, vol. 8542, pp. 23:1–23:12, November 2012.
11. S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima, "Real-time estimation of human body posture from monocular thermal images," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 15–20, 1997.
12. D. J. Bankman and T. M. Neighoff, "Pattern recognition for detection of human heads in infrared images," *Optical Engineering*, vol. 47, no. 4, p. 46404, 2008.
13. J. W. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *Proc. Int. Conf. Pattern Recognition*, vol. 4, pp. 713–716, 2004.
14. J. W. Davis and V. Sharma, "Background-subtraction in thermal imagery using contour saliency," *Int. Journal of Computer Vision*, vol. 71, no. 2, pp. 161–181, 2007.

15. T. Breckon, A. Gaszczak, J. Han, M. Eichner, and S. Barnes, "Multi-modal target detection for autonomous wide area search and surveillance," in *Proc. SPIE Security and Defence: Unmanned/Unattended Sensors and Sensor Networks*, SPIE, September 2013. to appear.
16. M. Yasuno, S. Ryousuke, N. Yasuda, and M. Aoki, "Pedestrian detection and tracking in far infrared images," in *Proc. Int. Conf. Intelligent Transportation Systems*, pp. 182–187, 2005.
17. K. Wahren, I. Cowling, Y. Patel, P. Smith, and T. Breckon, "Development of a two-tier unmanned air system for the MoD grand challenge," in *Proc. 24th International Conference on Unmanned Air Vehicle Systems*, pp. 13.1 – 13.9, March 2009.
18. T. Breckon, S. Barnes, M. Eichner, and K. Wahren, "Autonomous real-time vehicle detection from a medium-level UAV," in *Proc. 24th International Conference on Unmanned Air Vehicle Systems*, pp. 29.1–29.9, March 2009.
19. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 886–893.
20. X. Li and T. Breckon, "Combining motion segmentation and feature based tracking for object classification and anomaly detection," in *Proc. 4th European Conference on Visual Media Production*, pp. I–6, IET, November 2007.
21. A. Gaszczak, T. P. Breckon, and J. W. Han, "Real-time people and vehicle detection from UAV imagery," in *Proc. SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, p. Vol. 7878 Number 78780B, Jan. 2011.
22. M. Magnabosco and T. Breckon, "Cross-spectral visual Simultaneous Localization And Mapping (SLAM) with sensor handover," *Robotics and Autonomous Systems*, vol. 63, pp. 195–208, February 2013.
23. P. Pinggera, T. Breckon, and H. Bischof, "On cross-spectral stereo matching using dense gradient features," in *Proc. British Machine Vision Conference*, pp. 526.1–526.12, September 2012.
24. Q. Zhu, M. Yeh, K. Cheng, and S. Avidan, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 1491–1498, IEEE, 2006.
25. V. Prisacariu and I. Reid, "fastHOG-a real-time GPU implementation of HOG," *Department of Engineering Science, Oxford University, Tech. Rep*, vol. 2310, no. 09, 2009.
26. C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010. ISBN-13: 978-0470844731.
27. C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
28. J. Han, T. Breckon, D. Randell, and G. Landini, "The application of support vector machine classification to detect cell nuclei for automated microscopy," *Machine Vision and Applications*, vol. 23, no. 1, pp. 15–24, 2012.
29. A. Chenebert, T. Breckon, and A. Gaszczak, "A non-temporal texture driven approach to real-time fire detection," in *Proc. International Conference on Image Processing*, pp. 1781–1784, IEEE, September 2011.
30. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.
31. A. Kheyrollahi and T. Breckon, "Automatic real-time road marking recognition using a feature driven approach," *Machine Vision and Applications*, vol. 23, no. 1, pp. 123–133, 2012.
32. S. Suzuki and Others, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.
33. P. Pébay and T. Baker, "Analysis of triangle quality measures," *Mathematics of Computation*, vol. 72, no. 244, pp. 1817–1839, 2003.
34. R. Maciol, "Weapon detection in thermal images using gesture recognition," Master's thesis, School of Engineering, Cranfield University, Bedfordshire, UK, Sept. 2008.
35. D. Hall, J. Nascimento, P. Ribeiro, E. Andrade, P. Moreno, S. Pesnel, T. List, R. Emonet, R. B. Fisher, J. S. Victor, and J. L. Crowley, "Comparison of target detection algorithms using adaptive background models," in *Proc. Int. W'shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 113–120, 2005.
36. Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, vol. 27, no. 7, pp. 773–780, 2006.
37. L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. Int. Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.
38. G. Flitton, T. Breckon, and N. Megherbi, "Object recognition using 3D SIFT in complex CT volumes," in *Proc. British Machine Vision Conference*, pp. 11.1–12, September 2010.
39. I. Katramados and T. Breckon, "Real-time visual saliency by division of gaussians," in *Proc. International Conference on Image Processing*, pp. 1741–1744, IEEE, September 2011.