Low Resolution Person Detection with a Moving Thermal Infrared Camera by Hot Spot Classification

Michael Teutsch[1], Thomas Müller[1], Marco Huber[2], and Jürgen Beyerer[1]

[1]Fraunhofer IOSB, Karlsruhe, Germany [2]AGT International, Darmstadt, Germany

{michael.teutsch, thomas.mueller, juergen.beyerer}@iosb.fraunhofer.de

marco.huber@ieee.org

Abstract

In many visual surveillance applications the task of person detection and localization can be solved easier by using thermal long-wave infrared (LWIR) cameras which are less affected by changing illumination or background texture than visual-optical cameras. Especially in outdoor scenes where usually only few hot spots appear in thermal infrared imagery, humans can be detected more reliably due to their prominent infrared signature. We propose a two-stage person recognition approach for LWIR images: (1) the application of Maximally Stable Extremal Regions (MSER) to detect hot spots instead of background subtraction or sliding window and (2) the verification of the detected hot spots using a Discrete Cosine Transform (DCT) based descriptor and a modified Random Naïve Bayes (RNB) classifier. The main contributions are the novel modified RNB classifier and the generality of our method. We achieve high detection rates for several different LWIR datasets with low resolution videos in real-time. While many papers in this topic are dealing with strong constraints such as considering only one dataset, assuming a stationary camera, or detecting only moving persons, we aim at avoiding such constraints to make our approach applicable with moving platforms such as Unmanned Ground Vehicles (UGV).

1. Introduction

Person detection and localization is an important part of many camera-based safety and security applications such as search and rescue, surveillance, reconnaissance, or driver assistance. However, achieving a high rate of correct detections with only few false positives or negatives at the same time is still a challenge due to low resolution, changing background, or runtime requirements. Furthermore, when visual-optical cameras are used, strong variation in illumination, background, and human appearance complicate the problem even more, which leads to complex solutions using very high dimensional feature spaces to find and select the few discriminative features among them [4, 19, 34]. Thermal long-wave infrared (LWIR) cameras can provide a better fundament for person detection especially in complex outdoor scenarios with masking background texture or lack of illumination. In such scenarios the thermal signature of persons is more prominent compared to the visual-optical signature [16]. An example coming from the *Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS)* dataset *OSU Color-Thermal Database* [9] is shown in Fig. 1. Although there is some variation in thermal signatures, too, this variation is smaller even across different cameras and datasets compared to visual-optical images. Gradient-based methods such as HOGs [7] try to normalize visual-optical signatures but this leads again to complex approaches when aiming to detect humans reliably in low resolution [4, 34].

Figure 1. Motivation: person in visual-optical and LWIR image.

In this paper, we propose an approach for person detection and localization with a moving thermal infrared camera mounted on an Unmanned Ground Vehicle (UGV). In contrast to many other papers [5, 6, 8, 9, 13, 22, 33, 35] we avoid constraints such as the assumption of a stationary camera or the detection of persons in motion only. Moving and stationary persons can be recognized in real-time in low resolution. Our approach consists of two stages: in the first one, we use Maximally Stable Extremal Regions (MSER) [24] to detect hot spots and normalize the scale to 16×32 pixels. In the second one, we distinguish between

persons and clutter using machine learning algorithms. We evaluate various state-of-the-art descriptor/classifier combinations and propose a Discrete Cosine Transform (DCT) based descriptor together with a novel modified Random Naïve Bayes (RNB) classifier providing better classification performance and generality compared to SVM, AdaBoost, and Random Forest. Besides our own dataset with a moving camera, we evaluate three OTCBVS datasets [8, 9, 25] acquired by stationary cameras. The main contributions compared to previous work lie in the first-time analysis of promising descriptors with LWIR images which perform well in the visual-optical domain [10, 19], the modified RNB classifier, and the generality of our approach to perform well on three out of four different datasets.

The remainder of the paper is organized as follows: literature related to human detection in LWIR images is reviewed in Section 2. Hot spot detection is described in Section 3 and classification in Section 4. Experimental results are given in Section 5. We conclude in Section 6.

2. Related Work

We limit our review of related work to approaches for person recognition in thermal infrared images. We do not consider near infrared (NIR), which is popular in driver assistance systems [17] but needs active illumination. Fusion of visual-optical and infrared images is not discussed here but promising as long as the visual-optical images are of good quality [9, 22]. Person tracking is not reviewed, too.

Many approaches for person recognition in thermal infrared images are dealing with strong constraints such as considering only one specific dataset, assuming a stationary camera, or detecting only moving persons [5, 6, 8, 9, 13, 22, 23, 33, 35]. The authors use own datasets for their experiments [21, 31, 33, 36] or the OTCBVS benchmark datasets: the OSU Thermal Pedestrian Database [5, 6, 8, 13, 23, 26, 31, 33, 35], the thermal IR subset of OSU Color-Thermal Database [13, 21, 22, 33], and the Terravic Motion IR Database [26]. Regions of interest (ROIs) are detected either with background subtraction [5, 6, 8, 9, 13, 22, 33, 35], keypoint detectors [21], sliding window [23, 26, 36], or thresholding methods such as MSER [31]. All approaches except of background subtraction can be applied with a moving camera.

These ROIs can be verified using machine learning algorithms. Chen *et al*. [5] apply a method for unsupervised category learning called fuzzy Adaptive Resonance Theory (ART) neural networks to separate object and background pixelwise. Davis and Keck [8] calculate gradient magnitudes in four different directions and automatically learn a weighted filter bank using AdaBoost classifier. Li *et al*. [23] use a combination of Histogram of Oriented Gradient (HOG) features with geometric constraints and a linear Support Vector Machine (SVM) classifier. Dai *et al*. [6] sequentially exploit shape features, which are classified with an SVM, and appearance features, which are used for PCA-based localization. Leykin *et al*. [22] model and recognize human motion after tracking. Miezianko and Pokrajac [26] use histogram ratios of second-order gradient model patches and a linear SVM. Teutsch and Müller [31] propose a descriptor based on Hu moments, central moments, and Haralick features followed by SVM classification. Jüngling and Arens [21] use Speeded Up Robust Features (SURF) to detect and classify body parts and assemble them with an Implicit Shape Model (ISM). Zhang *et al*. [36] compare Edgelet and HOG features classified with either cascaded AdaBoost or cascaded SVMs. The evaluation shows similar results on both visual-optical and infrared videos. Finally, Zhang *et al*. [35] use level set-based contour tracking.

3. Hot Spot Detection

According to Section 2, three methods are applicable for hot spot detection: keypoint detection, sliding window, and MSER. Keypoint detection is prone to detect only few or no keypoints for low resolution objects. This leads to partial or missed detections. The sliding window approach is time-consuming especially when many different object scale levels are considered. Thus, we choose MSER detection due to its low computational demand [28] and ability to detect persons as a whole in high or low resolution.

MSERs are the result of a blob detection method based on thresholding and connected component labeling [24]. The application of MSER detection in this paper follows the assumption that the body temperature of persons in LWIR images is generally higher than the temperature of the surrounding background. This is true for many outdoor scenarios. Additional MSERs will be detected for background hot spots such as warm engines, open windows and doors, or areas heated up by the sun. Depending on the number and size of hot spots, merged detections will appear affecting the human blob shape. We use the following hierarchical MSER approach in order to handle such merged detections of either several persons or persons with background.

In general, the human appearance in LWIR images is not homogeneously bright due to clothes and other effects. Instead, there are smooth gray-value transitions inside the human blob and in case of a merge also to surrounding bright background regions. When the gray-values are not totally homogenous inside the merge region, the merge can be resolved by considering multiple detected MSERs for each hot spot. This is possible since an MSER detection does not have a global optimum but several local optima i^* [24]. The detection result is a chain of MSERs where each successor is a superset of the predecessors. Each MSER can belong to a body part, a whole person or a merged region. Thus, the MSERs associated with all possible solutions i^* are consid-

ered for hot spot classification to find the MSERs that most likely contain a person. Teutsch and Müller [31] allow only one MSER per hot spot resulting in a high rate of missed persons due to background merges.

4. Hot Spot Classification

The aim of classification is to verify the detected hot spots. Not only persons but also open doors and windows, vehicle engines, or animals will be detected. Machine learning algorithms are used to distinguish between persons and clutter. In order to achieve reliable results fast, appropriate features for setting up an object descriptor and a well-fitting classifier providing high generality are necessary.

4.1. Features

In surveillance applications, persons can usually be far away from the camera leading to low resolution. We consider this by scaling all hot spots to 16×32 pixels for size and appearance normalization and using features that are suitable for such low resolution appearances. We could introduce different scaling levels with respect to the object distance, but we want to assess the suitability of our considered descriptor/classifier combinations especially for the difficult case of low resolution. Since there can be dozens of MSERs in one image, we focus on descriptors that are very fast to calculate but promising regarding the results they achieved in their field of application:

- Moments [31]: This descriptor is a feature mix of invariant moments consisting of Hu moments, central moments, and Haralick features. The idea is to capture the holistic character of the object blob shape in the hot spots. The descriptor size is 178.

- Discrete Cosine Transform (DCT) [12]: Since halo effects [5, 13, 35] and motion blur [31] appear regularly and can affect the object appearance, we try to handle that with a local DCT-based descriptor. DCT is calculated in 8×8 pixel blocks and the first 21 DCT coefficients are kept. The block stride is 4 pixels to have overlapping blocks. The final descriptor is set up by concatenation of the DCT coefficients of each block. Other block sizes and strides led to worse performance than the chosen parameters. The descriptor size is 441.

- Histograms of Oriented Gradients (HOG) [7]: HOGs are expected to suffer from halo effects and motion blur. We achieved best results by using 8×8 blocks with 4 pixels block stride and 9 histogram bins. The descriptor size is 756.

- Integral Channel Features (ICF) [10]: Gradient magnitudes are calculated and subdivided in seven gradient orientation images. While the first image contains all magnitudes, the other six images contain only the magnitudes of specific gradient orientations. Local sums are calculated in randomly picked rectangular regions along all seven images and concatenated to set up the descriptor. These local sums are called *first-order features* [10]. We achieved better results compared to conventional Haar features. The descriptor size is 2000.

- Multi-LBP [19]: Local Binary Patterns (LBP) are calculated in four different quantizations. We choose 16×16 pixels for block size and 8 for block stride since very small blocks as proposed by Heng *et al*. [19] caused too strong locality leading to worse generality. In the original paper, the combination of Multi-LBP together with a ShrinkBoost classifier achieved very good results for the visual-optical low resolution Daimler-Chrysler pedestrian classification dataset [27]. The descriptor size is 3072.

4.2. Classification

Besides the evaluation of state-of-the-art classifiers such as SVM, Boosting [15], and Bagging [2], we analyze the modified version of a Random Naïve Bayes (RNB) classifier [29]. There are two motivations: (1) handling the merge of multiple hot spots and (2) achieving higher generality across different datasets or in case of slight appearance variation. Merged multiple hot spots of persons and clutter are in many cases still separable by MSER but lead to an appearance where some body parts are not observable anymore [35]. Another problem when dealing with new samples coming from the same or different cameras is that not all features may still fit to the learned model. Slight changes in appearance or pose can make most features still fit but few of them not fit to the model anymore. This can lead to worse classification performance if the feature space is considered as a whole (e.g., SVMs). Decision Trees as used in classification meta-algorithms such as bagging or boosting provide better generality since features are considered separately. However, non-optimal depth values can lead to overfitting or underfitting, feature selection and splitting values may be biased, and there is an oversensitivity to the training set, to irrelevant attributes, and to noise [30].

The Naïve Bayes (NB) classifier can provide good classification performance and generality across different datasets even when the assumption of conditional independence of the used features is obviously violated by a wide margin [11]. Actually, we think it is an advantage that NB considers features independently: even if few features do not fit to the model at all, the classifier decision may still be correct since these features will cause low likelihoods for both person and clutter and, thus, do not significantly affect the classification decision. Instead, the classifier will focus more on the features fitting to its model. Note that this is happening on-line.

The NB classification decision is given by:

$$\text{class}_{\text{NB}}(\mathbf{f}) = \arg\max_i \{P(c_i) \cdot \prod_{j=1}^{n} P(f_j \mid c_i)\} \quad (1)$$

where $\mathbf{f} = (f_1, \ldots, f_n)^T$ is the feature vector, $P(c_i)$ is the prior probability for class c_i with $i \in \{0, 1\}$ and $P(f_j \mid c_i)$ is the likelihood for feature f_j given class c_i. The product $\prod$ of these likelihoods is based on the naïve assumption that the features f_j of a descriptor are conditionally independent. Many different likelihood models can be used dependent on the distribution of the training samples for each feature. Since standard distributions such as Gaussian or Log-Gaussian do not fit well to the distributions of many of our used features, we achieved best results by using normalized, smoothed class-conditional histograms $h_j^{c_i}$ as likelihood model for each feature f_j. In order to weaken the violated assumption of conditional independence of the features, Independent Component Analysis (ICA) [20] can be applied to the feature vector prior to classification. Since the unsupervised training of ICA can lead to worse class separability when using the transformed feature vectors, Bressan and Vitria [3] propose to use a Class-Conditional ICA (CC-ICA). We adopt this idea leading to the following formalization:

$$\text{class}_{\text{NB}}(\mathbf{W}_i\mathbf{f}) = \arg\max_i \{P(c_i) \cdot \prod_{j=1}^{n} h_j^{c_i}(f_j^{\mathbf{W}_i})\} \quad (2)$$

where $\mathbf{W}_i$ is the class-conditional unmixing matrix calculated by FastICA [20] and $f_j^{\mathbf{W}_i}$ denotes the feature f_j transformed by $\mathbf{W}_i$.

When it comes to the idea of using NB as weak classifier for classification meta-algorithms, it is recommended not to use it for AdaBoost [32] but for approaches similar to Random Forest [29, 18]. We use a Random Forest framework with few adoptions from boosting as seen in Algorithm 1. For training of each weak classifier, typical RNB or Random Forest meta-algorithms use a random selection of features, bootstrap aggregation for selection of training samples, and majority voting for the final decision [2]. The Out-Of-Bag (OOB) set of not selected training samples for each weak classifier can be used to reject the current classifier if it is too weak. We use these approaches but add some novel ideas: we train and apply CC-ICA for each weak classifier. Furthermore, we do not use majority voting but the sum of weighted posteriors for the overall decision. The weight w_b for each classifer NB_b is calculated similar to AdaBoost but by using the OOB set only instead of the whole training data. While majority voting would cause a discrete decision function, the posteriors P_b induce a continuous decision function. Since each object usually gets multiple detected MSERs, non-maximum suppression is applied to find the best hypothesis. If two bounding boxes show strong overlap, the one with the higher posterior sum is kept. For some datasets, the performance is slightly increased if previously false classified samples were added randomly to the current bootstrap. A typical parameter setup is $B = 1000$ weak classifiers, $m = 10$ features per weak classifier, $t = 0.6$ as acceptance threshold (see Algorithm 1), and $P(c_i) = 0.5$ for $i \in \{0, 1\}$ as prior probability.

Algorithm 1 Modified Random Naïve Bayes classifier

for b = 1 to B **do**
  Generate bootstrap $\mathcal{B}$ from training set $\mathcal{T}$
  Choose m features randomly
  Calculate $\mathbf{W}_i^b$ with CC-ICA using $\mathcal{B}$
  Train weak classifier NB_b using $\mathcal{B}$ after CC-ICA
  Calculate error e_b with OOB set $\mathcal{T} \setminus \mathcal{B}$
  Set classifier weight $w_b = \frac{1}{2} \cdot \ln \frac{(1-e_b)}{e_b}$
  if $e_b > t$ **then**
    Reject classifier NB_b
    $b = b - 1$
  end if
end for
return $\arg\max_i \{\sum_{b=1}^{B} w_b \cdot P_b(c_i \mid \mathbf{W}_i^b \mathbf{f})\}$

5. Experimental Results

Since we apply supervised learning of the classifier models, we separate each of the used datasets in disjoint training and test sets. Table 1 provides some information about the four datasets and the sequences used for training. Partial or merged detections of persons do not appear in this table. They were not considered for classifier training since we want to detect complete persons. Sequences 1-3 of the OSU Color-Thermal Pedestrian Database show the same location while sequences 4-6 show a different one. The Terravic Motion IR dataset represents a mixture of very different scenes where sequences 8-13 show the same forest location.

The MSER results calculated for the training and test datasets were labeled manually for person or background MSERs (cf. two right-most columns in Table 1). Figure 2 shows some example person and background MSERs in the four datasets. The appearance of person and background hot spots varies across and inside the datasets.

Since the large amount of background MSERs in OSU Color-Thermal and Terravic Motion IR dataset is caused by the same background hot spots due to the stationary camera, we balance the training sets and randomly pick 25,000 and 20,000 background samples when learning the model.

For each MSER we calculate each of the descriptors described in Section 4.1 and classify it with various classifiers. We use an SVM with Radial Basis Function (RBF) kernel and 3-fold cross validation, Real AdaBoost with 1,000 de-

Table 1. Evaluated datasets with numbers of images and sequences, subsets used for training, and number of (hierarchical) MSERs.

	image size	# images	# seq.	training sequences	# person / # background MSERs	
					training	test
AMROS	360×288	6,742	2	1	1,067 / 4,005	2,066 / 415
OTCBVS OSU Thermal	360×240	286	10	1-3	781 / 361	2,033 / 453
OTCBVS OSU Color-Thermal	320×240	8,544	6	1-3	24,797 / 252,003	14,549 / 255,100
OTCBVS Terravic Motion IR	320×240	25,355	18	8-13	27,823 / 124,891	13,270 / 255,804

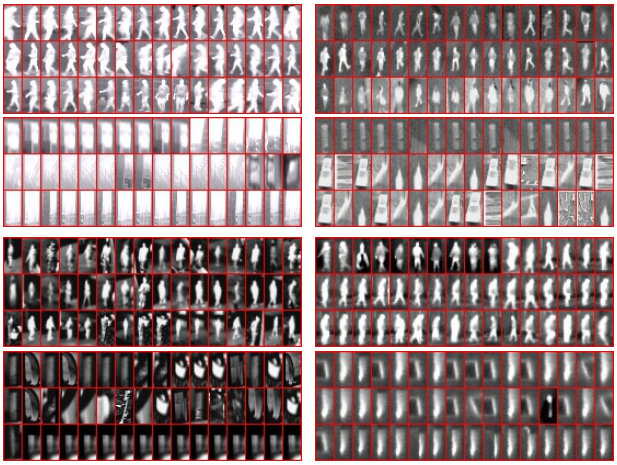

Figure 2. MSERs of persons (upper) and background (lower) for datasets AMROS (upper left), OSU Thermal (upper right), OSU Color-Thermal (lower left), and Terravic Motion IR (lower right).

cision trees of depth 1, Gentle AdaBoost with 1,000 decision trees of depth 3, and Random Trees with 1,000 decision trees. Further classifiers such as Linear SVM or k-Nearest Neighbor (kNN) were tested, too, but performed worse. All classifiers were taken from the OpenCV library [1]. We also evaluate three versions of our modified RNB classifier: without ICA, with standard ICA, and with CC-ICA.

The results are shown in Table 2. We calculated Receiver Operating Characteristic (ROC) curves for each descriptor/classifier combination for each dataset and display the Area Under Curve (AUC) for a compact presentation of our results. Significant difference in classification performance is given for the first two decimal places of each AUC value, with the third and fourth the results are becoming more and more similar. For each descriptor the best classifier performance is displayed with red AUC values. The best individual result for each dataset is underlined. Across all four datasets, DCT and ICF descriptors perform best while the classifier performance shows some variation. For three out of four datasets the combination of DCT descriptor and RNB classifier with CC-ICA achieves the best individual performance. Furthermore, the results confirm the conclusions of Bressan and Vitria [3] and Fan and Poh [14] that CC-ICA can improve NB classifier performance. For a better visualization of this result, we calculated mean ROC curves and variances across the four datasets for DCT descriptor and the three RNB variants in Fig. 3. CC-ICA +

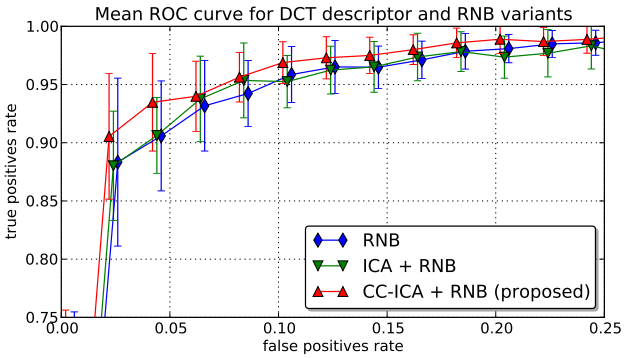

Figure 3. Comparison of Random Naïve Bayes variants.

RNB performs best especially for false positive (FP) rates of less or equal 6 %.

Since the results for OSU Color-Thermal and Terravic Motion IR datasets are significantly worse compared to AMROS and OSU Thermal, we try to improve the performance by adding training samples from other datasets. Therefore, all training and test samples of AMROS (7,553 samples), OSU Thermal (3,628), and Terravic Motion IR (421,788) are added to the training set of OSU Color-Thermal (276,800). To avoid one dataset dominating the whole training set, the subsets coming from OSU Color-Thermal and Terravic Motion IR are limited to 10,000 samples each by random sampling without replacement. The OSU Color-Thermal test set remains the same. We only chose DCT and ICF as descriptors since they provide the best tradeoff between performance and runtime. In Table 3, the results are presented as AUC values together with the AUC change compared to Table 2. While the performance of SVM decreases since it is considering the whole feature space, the performance of AdaBoost, Random Trees, and RNB increases proving the generality of these approaches. The same experiment was conducted for Terravic Motion IR with similar results as seen in Table 3.

The final classification results for DCT descriptor are presented with ROC curves in Fig. 4. We compare SVM, Random Trees, Gentle AdaBoost, and modified RNB with CC-ICA. Our proposed combination of DCT descriptor and modified RNB classifier performs best in three of the four considered datasets. As seen in the ROC curves, more than 90 % of the persons are classified correctly with a FP rate of 1 % for AMROS and OSU Thermal while around 80 % true positives (TPs) are achieved with 1 % false positives

Table 2. Area Under Curve (AUC) comparison for 5 descriptors, 7 classifiers, and 4 datasets.

dataset	descriptor	classifier						
		SVM	Real AdaBoost	Gentle AdaBoost	Random Trees	RNB	ICA + RNB	CC-ICA + RNB
AMROS	MOMENTS [31]	0.9776	0.9931	0.9893	0.9785	0.9960	0.9979	0.9967
	DCT [12]	0.9870	0.9903	0.9843	0.9729	0.9994	0.9994	0.9995
	HOG [7]	0.9707	0.9713	0.9762	0.9703	0.9614	0.9661	0.9702
	ICF [10]	0.9947	0.9916	0.9949	0.9938	0.9933	0.9942	0.9952
	Multi-LBP [19]	0.9226	0.9349	0.9286	0.9140	0.9661	0.9480	0.9508
OTCBVS OSU Thermal	MOMENTS [31]	0.9685	0.9674	0.9742	0.9716	0.9471	0.9855	0.9910
	DCT [12]	0.9859	0.9920	0.9882	0.9876	0.9916	0.9726	0.9944
	HOG [7]	0.9863	0.9888	0.9916	0.9918	0.9941	0.9923	0.9932
	ICF [10]	0.9909	0.9874	0.9803	0.9346	0.9541	0.9243	0.9676
	Multi-LBP [19]	0.9745	0.9750	0.9696	0.9598	0.9667	0.9486	0.9527
OTCBVS OSU Color-Thermal	MOMENTS [31]	0.9367	0.6828	0.6542	0.6701	0.8856	0.9434	0.9496
	DCT [12]	0.9289	0.9657	0.9827	0.8568	0.9433	0.9559	0.9584
	HOG [7]	0.9946	0.9832	0.9854	0.9318	0.8783	0.8549	0.8957
	ICF [10]	0.9863	0.9891	0.9942	0.9401	0.9344	0.9443	0.9506
	Multi-LBP [19]	0.9725	0.9554	0.9672	0.8105	0.9654	0.9553	0.9584
OTCBVS Terravic Motion IR	MOMENTS [31]	0.7906	0.8606	0.9108	0.8859	0.7520	0.7898	0.8079
	DCT [12]	0.9640	0.9473	0.9419	0.9075	0.9599	0.9675	0.9698
	HOG [7]	0.9224	0.9173	0.9387	0.9141	0.8861	0.8674	0.8935
	ICF [10]	0.9567	0.9523	0.9570	0.9448	0.9631	0.9495	0.9553
	Multi-LBP [19]	0.9648	0.9627	0.9503	0.8899	0.9641	0.9544	0.9670

Table 3. Area Under Curve (AUC) improvement compared to Table 2 for two datasets by training with all four datasets.

dataset	descriptor	classifier						
		SVM	Real AdaBoost	Gentle AdaBoost	Random Trees	RNB	ICA + RNB	CC-ICA + RNB
OTCBVS OSU Color-Thermal (N-Training)	DCT [12]	0.8916 −0.0373	0.9724 +0.0068	0.9858 +0.0031	0.9350 +0.0782	0.9751 +0.0318	0.9744 +0.0185	0.9768 +0.0148
	ICF [10]	0.9858 −0.0005	0.9702 −0.0189	0.9786 −0.0156	0.9486 +0.0085	0.9561 +0.0217	0.9765 +0.322	0.9758 +0.0252
OTCBVS Terravic Motion IR (N-Training)	DCT [12]	0.9571 −0.0069	0.9655 +0.0182	0.9716 +0.0297	0.9305 +0.0230	0.9778 +0.0179	0.9825 +0.0150	0.9830 +0.0132
	ICF [10]	0.9527 −0.0040	0.9667 +0.0144	0.9771 +0.0201	0.9781 +0.0333	0.9586 −0.0045	0.9647 +0.0125	0.9696 +0.0143

for OSU Color-Thermal and Terravic Motion IR. As there are more hot spot merges and motion blur effects in the AMROS dataset compared to the OTCBVS datasets, we assume this to be the reason for the big performance difference of CC-ICA RNB compared to the other classifiers.

The classification performance is sufficient for scenes where about 1-5 % of wrongly classified background MSERs is acceptable. This is not the case for sequences 4-6 of OSU Color-Thermal, the indoor scene (sequence 1), and the forest location (sequences 14-16) of Terravic Motion IR. Here we have up to 100 background MSERs per image. Even with a FP rate of 1 % there are still more FPs than TPs. The combined results for MSER detection and RNB classification for all scenes except for the mentioned ones are presented in Table 4. We achieve high TP rates of 83-98 % with MSER detection and a large amount of FPs in the background. The false negative (FN) rate is 1-17 %. With RNB classification the FN rate is not rising significantly while the FP rate can be reduced to 2-14 %. Figure 5 depicts some results. The left column shows the detections classified as persons (red boxes) and background (green boxes) while the right column visualizes recognized persons only. The example image from OSU Color-Thermal (row 3) shows the problem of a too large amount of background MSERs. The non-optimized processing rate on an Intel Xeon with 3.60 GHz is 10-25 Hz.

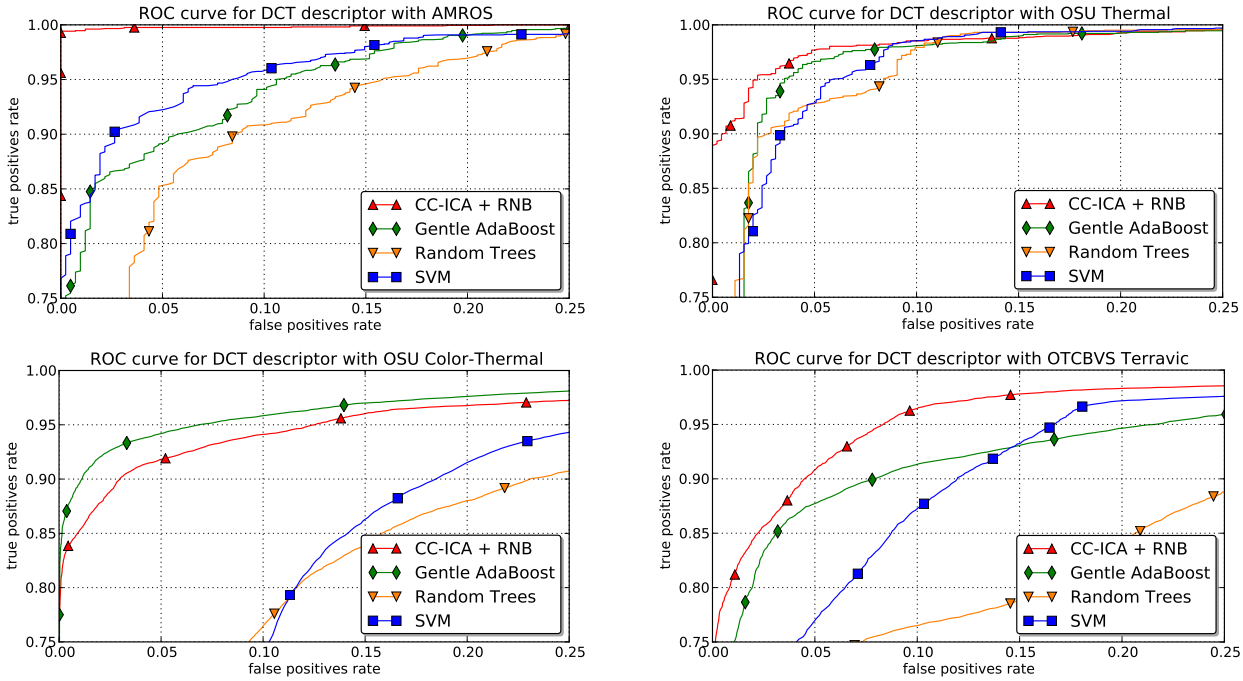

Figure 4. Evaluation results: best ROC curves for DCT descriptor for each dataset.

Table 4. Overall results (GT = Ground Truth, TP = True Positives, FP = False Positives, FN = False Negatives).

test dataset	GT	MSER detection			MSER + RNB		
		TP	FP	FN	TP	FP	FN
AMROS	780 (100%)	767 98.33%	10,349 1,327%	13 1.67%	740 94.87%	64 8.21%	40 5.13%
OTCBVS OSU Thermal	695 (100%)	610 87.77%	514 73.96%	85 12.23%	608 87.48%	94 13.53%	87 12.52%
OTCBVS Terravic Motion IR (without indoor scene / forest scene)	1,515 (100%)	1,265 83.50%	42,685 2,817%	250 16.50%	1,250 82.51%	35 2.31%	265 17.49%

6. Conclusions and Future Work

We presented an approach for detecting persons in real-time in LWIR images acquired by a moving camera. We focused on low resolution person appearances in outdoor surveillance scenarios where it is difficult to recognize persons with visual-optical cameras due to masking background texture or lack of illumination. Our holistic approach consists of MSER hot spot detection and subsequent hot spot classification using a DCT-based descriptor and a modified Random Naïve Bayes (RNB) classifier. Since we use MSER detection instead of background subtraction and features which are tolerant of motion blur, our approach is applicable with a moving camera. Promising results were achieved for three out of four datasets. However, scenes with many hot spots and around 100 background MSERs per image can cause too many false positives. Some ideas for future work are using cascaded classifiers, learning models in different scale, fusion with other methods, or the introduction of tracking.

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5

[2] L. Breiman. Random forests. *Machine Learning*, 45(1), 2001. 3, 4

[3] M. Bressan and J. Vitria. Improving Naive Bayes using Class-Conditional ICA. In *Springer LNCS*, volume 2527, 2002. 4, 5

[4] H. Cao, K. Yamaguchi, T. Naito, and Y. Ninomiya. Pedestrian Recognition Using Second-Order HOG Feature. In *ACCV*, 2009. 1

[5] B. Chen, W. Wang, and Q. Qin. Robust multi-stage approach for the detection of moving target from infrared imagery. *SPIE Optical Engineering*, 51(6), June 2012. 1, 2, 3

[6] C. Dai, Y. Zheng, and X. Li. Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery. In *CVPR Workshops*, 2005. 1, 2

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 1, 3, 6

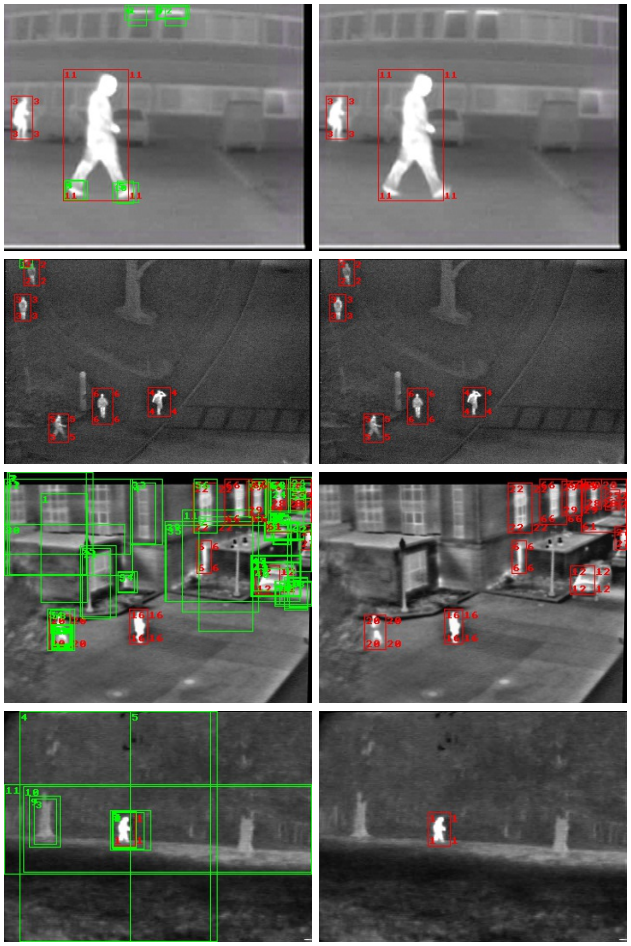

Figure 5. Results for AMROS (row 1), OSU Thermal (row 2), OSU Color-Thermal (row 3), and Terravic Motion IR (row 4). Detected persons are highlighted in red and background in green.

[8] J. W. Davis and M. A. Keck. A two-stage approach to person detection in thermal imagery. In *IEEE Workshop on Application of Computer Vision (WACV/MOTION)*, 2005. 1, 2

[9] J. W. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106, 2007. 1, 2

[10] P. Dollar, Z. Tu, P. Perona, and S. Belongie. Integral Channel Features. In *BMVC*, 2009. 2, 3, 6

[11] P. Domingos and M. Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29, 1997. 3

[12] H. K. Ekenel and R. Stiefelhagen. Local appearance based face recognition using discrete cosine transform. In *EUSIPCO*, 2005. 3, 6

[13] T. Elguebaly and N. Bouguila. A Nonparametric Bayesian Approach for Enhanced Pedestrian Detection and Foreground Segmentation. In *CVPR Workshops*, 2011. 1, 2, 3

[14] L. Fan and K. L. Poh. A Comparative Study of PCA, ICA and Class-Conditional ICA for Naïve Bayes Classifier. In *Springer LNCS*, volume 4507, 2007. 5

[15] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. *Journ. of Computer and System Sciences*, 55, 1997. 3

[16] A. Gaszczak, T. P. Breckon, and J. Han. Real-time people and vehicle detection from UAV imagery. In *Proc. of SPIE*, volume 7878, 2011. 1

[17] J. Ge, Y. Luo, and G. Tei. Real-Time Pedestrian Detection and Tracking at Nighttime for Driver-Assistance Systems. *IEEE Transactions on ITS*, 10(2), June 2009. 2

[18] M. Godec, C. Leistner, A. Saffari, and H. Bischof. On-line Random Naive Bayes for Tracking. In *ICPR*, 2010. 4

[19] C. K. Heng, S. Yokomitsu, Y. Matsumoto, and H. Tamura. Shrink Boost for Selecting Multi-LBP Histogram Features in Object Detection. In *CVPR*, 2012. 1, 2, 3, 6

[20] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, NY, 2001. 4

[21] K. Jüngling and M. Arens. Feature based person detection beyond the visible spectrum. In *CVPR Workshops*, 2009. 2

[22] A. Leykin, Y. Ran, and R. Hammoud. Thermal-Visible Video Fusion for Moving Target Tracking and Pedestrian Classification. In *CVPR*, 2007. 1, 2

[23] W. Li, D. Zheng, T. Zhao, and M. Yang. An Effective Approach to Pedestrian Detection in Thermal Imagery. In *Proc. Intern. Conf. on Natural Computation (ICNC)*, 2012. 2

[24] J. Matas, O. Chum, M. Urban, and T. Pajdla. RobustWide Baseline Stereo from Maximally Stable Extremal Regions. In *BMVC*, 2002. 1, 2

[25] R. Miezianko. IEEE OTCBVS WS series bench. In *Terravic Research Infrared Database*, 2006. 2

[26] R. Miezianko and D. Pokrajac. People Detection in Low Resolution Infrared Videos. In *CVPR Workshops*, 2008. 2

[27] S. Munder and D. M. Gavrila. An Experimental Study on Pedestrian Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11), Nov. 2006. 3

[28] D. Nister and H. Stewenius. Linear Time Maximally Stable Extremal Regions. In *Springer LNCS*, volume 5303, 2008. 2

[29] A. Prinzie and D. V. den Poel. Random Multiclass Classification: Generalizing Random Forests to Random MNL and Random NB. In *Springer LNCS*, volume 4653, 2007. 3, 4

[30] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers Inc., 1992. 3

[31] M. Teutsch and T. Müller. Hot Spot Detection and Classification in LWIR Videos for Person Recognition. In *Proceedings of SPIE*, volume 8744, 2013. 2, 3, 6

[32] K. M. Ting and Z. Zheng. A Study of AdaBoost with Naive Bayesian Classifiers: Weakness and Improvement. *Computational Intelligence*, 19(2), 2003. 4

[33] J. Wang, G. Bebis, and R. Miller. Robust Video-Based Surveillance by Integrating Target Detection with Tracking. In *CVPR Workshops*, 2006. 1, 2

[34] T. Watanabe, S. Ito, and K. Yokoi. Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection. In *Springer LNCS*, volume 5414, 2009. 1

[35] H. Zhang, B. Zhao, L. Tang, and J. Li. Variational-Based Contour Tracking in Infrared Imagery. In *Intern. Congress on Image and Signal Processing (CISP)*, 2009. 1, 2, 3

[36] L. Zhang, B. Wu, and R. Nevatia. Pedestrian Detection in Infrared Images based on Local Shape Features. In *CVPR Workshops*, 2007. 2