Hot Spot Detection and Classification in LWIR Videos for Person Recognition

Michael Teutsch, Thomas Müller

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)
Fraunhoferstr. 1, 76131 Karlsruhe, Germany

ABSTRACT

Person recognition is a key issue in visual surveillance. It is needed in many security applications such as intruder detection in military camps but also for gaining situational awareness in a variety of different safety applications. A solution for LWIR videos coming from a moving camera is presented that is based on hot spot classification to distinguish persons from background clutter and other objects. We especially consider objects in higher distance with small appearance in the image. Hot spots are detected and tracked along the videos. Various image features are extracted from the spots and different classifiers such as SVM or AdaBoost are evaluated and extended to utilize the temporal information. We demonstrate that taking advantage of this temporal context can improve the classification performance.

Keywords: Thermal infrared, IR, MWIR and LWIR, warm area localization, human detection, person classification, visual surveillance, intruder recognition, temporal context.

1. INTRODUCTION

In this paper, we focus on outdoor applications where persons are to be detected in the environment using a moving robot equipped with cameras. The applications range from intruder detection in military camps, border patrol, and ground security of civil complexes[1] to victim detection after catastrophes[2]. In such scenarios, the thermal infrared band (MWIR or LWIR*) offers some important advantages. The possibility to detect persons in complete darkness passively (i.e. without the need of active light sources) is very important in the military domain, for example. Even in daylight, persons can often be detected easier in this frequency band than in the visual-optical (VIS) or NIR band, see Fig. 1 for illustration. Another big advantage is that LWIR cameras can see through dust and fog in the air much better than VIS sensors. Furthermore, victim detection after the collapse of a building is more effective when using LWIR as dust might lay on the ground and persons could be covered completely for other imaging sensors.

When inspecting our LWIR videos in the mentioned outdoor applications we found out that in contrast to indoor scenes there are only sparsely distributed hot spots in the environment. Typical non-human hot spots are coming from open windows, open doors, or motorized objects. So, in order to solve the person detection and localization problem, which is unsolved in general, we focus on scenarios with limited number and size of hot spots in the environment. Examples for such hot spots can be seen in the right part of Fig. 1.

With that focus, a solution for LWIR imagery is presented that is based on a processing chain consisting of two steps. In the first step, hot spots are detected and localized efficiently in the image with a scale invariant approach. Afterwards, a hot spot classification is performed in the second step to distinguish persons from background clutter and other objects. A high rate of true positives (detected visible persons) is to be guaranteed in the first step with the tendency of rather being prone to false positives (detected background structures) than false negatives (missed persons). False positives can be rejected afterwards by the classification module but missed persons are definitely lost. For the second step, the detected hot spots can optionally be tracked in order to provide temporal information making the classification more stable and reliable (see Section 4).

Further author information: Email: {michael.teutsch, thomas.mueller}@iosb.fraunhofer.de

*These two thermal bands are very similar to each other and only differ in dedicated situations that are not relevant here. Therefore, it does not matter if MWIR or LWIR is used. In practice, LWIR is preferred due to the lower sensor costs.

Figure 1. Same scene in VIS (left) and LWIR band (right). The person can be discovered easier in LWIR than in VIS.

Since our image data is coming from a LWIR camera mounted on a moving robot, we do not consider background subtraction for moving object detection. We want to detect moving and stationary persons in close distance as well as in higher distance with small appearance in the image, where popular approaches using local image features such as SIFT[3] do not work reliably. Different classification algorithms are compared and temporal context is utilized to achieve optimal results. Various kinds of image features are extracted from the spots and evaluated for their separability potential. Different methods for features space reduction are implemented and tested. Finally, classifiers such as Support Vector Machine (SVM) or AdaBoost are compared to each other and extended to benefit from the temporal information of the tracked spots.

The challenges are covering close and high object distances reliably and being robust against variable background, weak signal-to-noise ratio (SNR), weak contrast and sensor-specific noise of thermal infrared sensors as well as occurring motion blur due to camera or object motion. Furthermore, the algorithms have to run in real-time, so we do not consider approaches using combined sliding window detection and object recognition such as Histograms of Oriented Gradients (HOG).[4]

1.1 Related Work

For the presentation of related work we chose papers and articles where thermal infrared (MWIR, LWIR) images were used. Especially in driver assistance applications, it is popular to use Near Infrared (NIR)[5–7] or Far Infrared (FIR)[8,9] cameras. Furthermore, some authors try to fuse image data coming from visual optical (VIS) and infrared cameras.[10,11] Those topics will not be discussed here.

Dai et al.[12] propose a layerered representation by separation of foreground and background using a generalized Expectation Maximization (EM) algorithm. Pedestrians are classified by a Support Vector Machine (SVM) using shape features and localized by Principal Component Analysis (PCA) using appearance features. Finally, a graph-matching based algorithm is presented for pedestrian tracking. Zhang et al.[13] exploit algorithms for VIS image data and try to adapt and apply them to IR data. Two descriptors (edgelets and Histograms of Oriented Gradients (HOG)) and two classifiers (cascaded SVMs and AdaBoost) are evaluated and provide good results. Jüngling and Arens[14] use SURF features to detect and classify human body parts. Persons are detected using an Implicit Shape Model (ISM). Xia et al.[15] introduce a SUSAN keypoint sliding window searching strategy detecting regions of interest (ROIs). Each region is analyzed using multi-block Local Binary Patterns (LBP) to describe pedestrians and a cascade boosted classifier to detect them. Li et al.[16] propose using a sliding window approach with a combination of HOG features and geometric characteristics as features, and a SVM as classifier. Chen et al.[17] implemented a multi-level spatial-temporal median filter to extract the background frame in scenarios with stationary cameras. Background clutter is suppressed using Principal Component Analysis (PCA). A spatially related fuzzy adaptive resonance theory (ART) neural network is applied to identify ROIs and within each region, another fuzzy ART neural network is used to detect moving persons. Finally, Sun et al.[18] propose an approach similar to Viola-Jones[19] object detection with sliding window, Haar-features, and

AdaBoost classification. Most authors[12–17] are using the publicly available OTCBVS benchmark datasets 01[20] or 03[21] for evaluation. Thus, besides our own datasets, we consider them for our evaluation, too.

The remainder of the paper is organized as follows: In Section 2, the detection of hot spots will be described. Classification in single images and image sequences will be the topic in Section 3 and 4, respectively. Experimental results will be presented and discussed directly in the specific section. Finally, the conclusions and an outlook to potential future work are given in Section 5.

2. HOT SPOT DETECTION

After testing some approaches for hot spot detection based on fixed and adaptive thresholding, we finally decided for the Maximally Stable Extremal Regions (MSER) algorithm as proposed by Matas et al.[22] Along the tested approaches, it shows the best potential to fulfill the requirements mentioned in Section 1, and we already used it successfully in other applications[23,24] in the near past. When comparing this algorithm with a potential (but slightly worse) alternative we discovered that there are some LWIR inherent effects that cannot be solved by only hot spot detection without further considered knowledge about human appearance due to the character of LWIR imagery and the inherent contrast effects and variabilities when working with real world scenes. In other words, these LWIR inherent effects are expected to show up also with other hot spot detection algorithms because they cannot be avoided due to their nature. However, this is not a principal problem due to the construction of the proposed processing chain in which the second step aims to handle those undesired effects.

In the implemented hot spot detector the MSER results are used to calculate ROIs (bounding boxes) for the bright image regions. The MSER results for dark image regions are discarded since most LWIR cameras use bright values for warm image regions. Finally, the calculated bounding boxes are expanded by some border in order to properly capture also the transition from the hot spot to the darker background. Before the ROIs are fed into the classification step they are scaled to a fixed size. We use 3 pixels for the border size and 16×32 pixels for the scaled ROIs.

In our experiments we used two outdoor image sequences: `sequence 1` with 4580 LWIR single images and `sequence 2` with 2162 images showing a similar environment with different persons in different situations. Furthermore, we processed the OSU thermal pedestrian database - dataset 01 of the OTCBVS benchmark dataset collection[20,21,25]: sequences `otcbvs_osu 1` to `otcbvs_osu 10` with 18 to 73 images per sequence, 284 images in total. The images of sequence `otcbvs_osu 3` were inverted before the hot spot detection because hot areas are depicted with dark colors here. Table 1 summarizes the number of correct detections/localizations of persons, the number of missed person (i.e. visible persons in the image where no hot spot was generated, i.e. undiscovered persons), the number of bad hot spots 'persons + background' as well as the number of detections of background structures. Bad hot spots in the sense of the proposed processing chain ('persons + background', see above text about undesired effects) are hot spots of persons with a large amount of additional background stucture due to contrast reasons. Such hot spots are useless in the context of the proposed processing chain since they are not learned in the classification. Fig. 2 depicts the detected persons of `sequence 1` and `sequence 2` for illustration.

As seen in Table 1, there are sufficiently enough person detections for our application. Up to now, it does not matter, if a person is not detected in every single image. Since our LWIR sensors record 25 images per second and it is acceptable to produce an alarm with a delay of some hundred milliseconds, the person detection rate is sufficient for practical operation purposes. However, tracking will be added in future work. The processing time of MSER is below 8 ms per image.

3. SINGLE IMAGE CLASSIFICATION

The evaluation in Table 1 shows that there is a high amount of false positives in between the detected hot spots. Hence, a classification module is introduced to separate true and false positives. This module consists of three submodules: *feature extraction*, *feature reduction*, and *classification*. For each module we tested various standard approaches with different properties. The classifiers in the classification submodule are learned using a set of training samples and evaluated using a set of test samples which is disjoint from the training set. Here is an overview of the implemented and evaluated approaches:

Table 1. Detections and misses of the hot spot detection using MSER.

	detected persons	missed persons	persons + background	background detections
sequence 1	380 (61.5 %)	77 (12.5 %)	160 (26.0 %)	1131
sequence 2	570 (83.6 %)	21 (3.1 %)	91 (13.3 %)	162
otcbvs_osu 1	72 (83.7 %)	14 (16.3 %)	0 (0.0 %)	0
otcbvs_osu 2	84 (97.6 %)	0 (0.0 %)	2 (2.3 %)	32
otcbvs_osu 3	49 (59.8 %)	32 (39.0 %)	1 (1.2 %)	40
otcbvs_osu 4	94 (87.8 %)	8 (7.5 %)	5 (4.7 %)	33
otcbvs_osu 5	73 (86.9 %)	1 (1.2 %)	10 (11.9 %)	58
otcbvs_osu 6	76 (95.0 %)	0 (0.0 %)	4 (5.0 %)	5
otcbvs_osu 7	37 (50.0 %)	36 (48.6 %)	1 (1.4 %)	0
otcbvs_osu 8	79 (92.9 %)	6 (7.1 %)	0 (0.0 %)	1
otcbvs_osu 9	80 (85.1 %)	13 (13.8 %)	1 (1.1 %)	2
otcbvs_osu 10	45 (54.9 %)	32 (39.0 %)	5 (6.1 %)	6
all otcbvs_osu	689 (80.1 %)	142 (16.5 %)	29 (3.4 %)	177

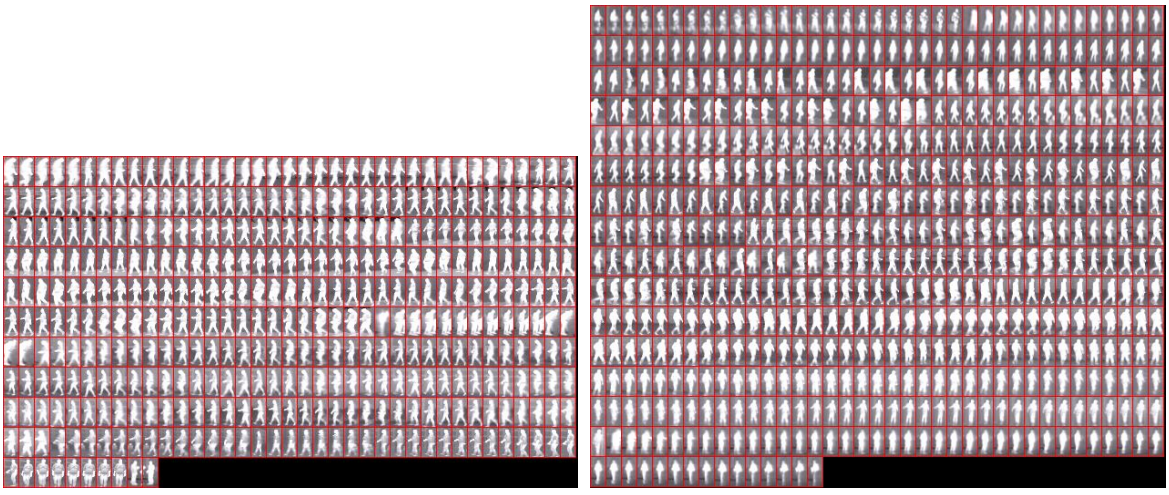

Figure 2. Detected persons of sequence 1 (left) and sequence 2 (right).

- **Feature extraction:**
  1. **HOG:** Histograms of Oriented Gradients (HOG) as proposed by Dalal and Triggs[4] are calculated by analyzing the edge structure of an object. The ROI is subdivided in blocks of 16×16 pixels with cells of 8×8 pixels. In each cell gradient orientations are collected in histograms of 9 bins each and normalized block-wise using L2-hys-norm. The histograms are concatenated. To apply the algorithm implementation in the OpenCV library,[26] we upscaled the ROIs to 48×96 pixels.
  2. **DCT:** The ROI is subdivided in blocks and each block is processed with the Discrete Cosinus Transform (DCT). The DCT coefficients for each block are concatenated as proposed by Ekenel et al.[27] We considered block sizes of 8×8 and 16×16 pixels.
  3. **COOC:** The number of co-occurrences (COOC) of similar pixel gray-values in fixed offsets are stored in a matrix[28] and evaluated using Haralick features[28] such as contrast, correlation, entropy, and many more.
  4. **LBP:** Local Binary Patterns (LBP) are a unique description of a pixel's neighborhood. We consider rotation-invariant uniform LBP[29] and local gray-value variance VAR[29] calculated either in the whole ROI (global) or in blocks and cells (local) similar to HOG with blockwise L2-norm.
  5. **Moments:** This is a feature mix containing improved Hu moments,[30] central moments, and Haralick features[28] calculated on the original ROI image, the gradient image, and the LBP image.
- **Feature reduction:**
  1. **None:** It is quite common to not use any feature reduction at all. The benefit of feature reduction is often pretty small. This means the classification rates do not become significantly better and the time saved during classification due to smaller feature vector size is added by the necessary transform.
  2. **PCA:** The Principal Component Analysis (PCA) is used for data-driven feature reduction without considering the class labels of the training set but the variance of all samples. It is assumed that best class separability is given in the direction where the sample distribution has its highest variance.
  3. **LDA:** The Linear Discriminant Analysis (LDA) is used for data-driven feature reduction considering the class labels of the training set. While the inner-class variance is to be minimized, the between-class variance is to be maximized.
- **Classification:**
  1. **SVM:** Support Vector Machines (SVM) are widely used in machine learning. We use the implementation provided in OpenCV[26] with a radial basis function (RBF) kernel as we achieved better results compared to a linear kernel. Furthermore, we apply 3-fold cross validation during the learning process for better generalization.
  2. **AdaBoost:** Boosting is a popular classification meta-algorithm combining many weak classifiers to a strong one. In most cases, linear classifiers or decision trees are chosen as weak classifiers. We apply Real AdaBoost[31] as implemented in OpenCV.[26]
  3. **Random Trees:** Random Trees (RT)[26] is another meta-algorithm combining weak decision trees by a voting scheme. The class is chosen with the highest number of votes. It is called *random* since the feature subset and the subset of training samples are chosen randomly for each weak classifier. It has originally been introduced as *Random Forest* by Breiman.[32]

More detailed information about the briefly described approaches can be found in the cited papers. In the evaluation we will focus on SVM and AdaBoost as they appeared to be more stable and robust than RT in our experiments. We also tried k-Nearest Neighbor (k-NN) and Bayes classifiers but their results have been even less stable in the experiments than with the classifiers mentioned before, although they outperformed the mentioned ones in a few test cases. One more reason to consider only SVM and AdaBoost for the evaluation is that we can plot Receiver Operating Characteristic (ROC) curves for them. ROC curves show the performance of a classifier by plotting the true positive rate against the false positive rate for a variable decision function value threshold.

Table 2. Area Under Curve (AUC) for Experiment 1.

feature reduct.	classifier	test	feature extraction							
			HOG	DCT_8	DCT_{16}	COOC	LBP_G	LBP_L	LBP_{G+L}	Moments
NONE	SVM	$1 \rightarrow 2$	0.759	0.9886	0.9902	0.9640	0.9549	0.8734	0.8771	0.9943
		$2 \rightarrow 1$	0.626	0.9871	0.9941	0.9394	0.8939	0.8241	0.8413	0.9846
	AdaBoost	$1 \rightarrow 2$	0.768	0.9870	0.9739	0.9897	0.9359	0.8725	0.9332	0.9947
		$2 \rightarrow 1$	0.598	0.9864	0.9716	0.9579	0.9012	0.7795	0.8495	0.9876
PCA	SVM	$1 \rightarrow 2$	0.737	0.9829	0.9885	0.9678	0.8968	0.8165	0.8740	0.9862
		$2 \rightarrow 1$	0.731	0.9273	0.9202	0.9179	0.8699	0.8298	0.8229	0.9800
	AdaBoost	$1 \rightarrow 2$	0.665	0.9711	0.9575	0.9819	0.8842	0.8109	0.8144	0.9864
		$2 \rightarrow 1$	0.706	0.9150	0.9050	0.9460	0.8742	0.7048	0.7804	0.9290
LDA	SVM	$1 \rightarrow 2$	0.715	0.9181	0.9477	0.8610	0.9393	0.9171	0.9201	0.9675
		$2 \rightarrow 1$	0.534	0.9085	0.9180	0.8757	0.8787	0.8615	0.9394	0.9347
	AdaBoost	$1 \rightarrow 2$	0.686	0.7638	0.8788	0.9139	0.9404	0.8893	0.8965	0.9057
		$2 \rightarrow 1$	0.558	0.7475	0.8251	0.8364	0.8519	0.7952	0.8131	0.8387

The closer the curve approximates to the point $(1, 0)$, which is 100 % true positives and 0 % false positives, the better is the classifier. Since we have many results to present, they are organized in a table with the Area Under Curve (AUC) as performance measure. A perfect classifier has an AUC of 1.0, a poor classifier 0.5 (close to guessing) or less.

Three different experiments have been performed to find out about the stability and robustness of the features and the classifiers:

1. Manually labeled ground truth data for training and manually labeled ground truth data for testing.
2. Manually labeled ground truth data for training and MSER detection ROIs for testing.
3. MSER detection ROIs for training and MSER detection ROIs for testing.

3.1 Experiment 1: Ground Truth against Ground Truth

Our own two datasets `sequence 1` and `sequence 2` have been labeled manually for persons and background as ground truth. Manual labeling of false positives is possible as we search for hot spots which appear brightly in the image. This should be a better ground truth than randomly chosen false positives as it is common in visual-optical (VIS) images and videos. In `sequence 1` 1152 persons and 1664 background objects have been labeled and in `sequence 2` 930 persons and 2923 background objects. We evaluated with cross validation by training with `sequence 1` and testing with `sequence 2` first, and then training with `sequence 2` and testing with `sequence 1`. Table 2 shows the results. Besides the different features, feature reduction methods, and classifiers, the cross validation is entered in column *test* where $1 \rightarrow 2$ stands for training with `sequence 1` and testing with `sequence 2`. The classifier performance is evaluated with the Area Under Curve (AUC) coming from the ROC curves. DCT_8 and DCT_{16} describe DCT features with block size 8×8 and 16×16 pixels, respectively. LBP_G are global LBP, LBP_L are local LBP, and LBP_{G+L} are both features combined by concatenation.

The table shows that the features choice is strongly influencing the results. HOG are not performing well since there is nearly no visible texture except of the person contours in our LWIR data. Furthermore, the upscaling to 48×96 pixels is very similar to strong image smoothing. Dalal and Triggs[4] point out, that smoothing is significantly decreasing the classification performance. Thus, HOG are not considered anymore for further experiments in this paper which does not mean that they are unsuitable in general for IR image processing as seen in other papers.[13,16] On the other hand, DCT_{16} and Moments are performing best which could be the result of varying image sharpness (blur) in our sequences as seen in Fig. 2. LBP_G gives the most stable AUC values along all feature reduction methods and classifiers. Both feature reduction algorithms do not improve the performance. LDA is not considered anymore for further experiments as it performs worst. SVM and AdaBoost provide similar results with a slight advantage for SVM.

Table 3. Area Under Curve (AUC) for Experiment 2.

feature reduction	classifier	test	feature extraction						
			DCT_8	DCT_{16}	COOC	LBP_G	LBP_L	LBP_{G+L}	Moments
NONE	SVM	$1 \rightarrow 2$	0.9632	0.9542	0.9929	0.9774	0.9336	0.9341	0.9704
		$2 \rightarrow 1$	0.9729	0.9365	0.9399	0.8756	0.7639	0.7903	0.9556
		$1 \rightarrow$ osu	0.5874	0.6113	0.5592	0.5047	0.3443	0.3455	0.6795
		$2 \rightarrow$ osu	0.7144	0.6248	0.5218	0.6505	0.8821	0.8563	0.6423
	AdaBoost	$1 \rightarrow 2$	0.9860	0.9550	0.9984	0.9764	0.8620	0.9487	0.9589
		$2 \rightarrow 1$	0.9331	0.9196	0.9886	0.8464	0.7187	0.8430	0.8833
		$1 \rightarrow$ osu	0.6396	0.6097	0.5891	0.5445	0.3616	0.5108	0.6817
		$2 \rightarrow$ osu	0.6643	0.6996	0.6259	0.6503	0.8658	0.7521	0.6435
PCA	SVM	$1 \rightarrow 2$	0.9325	0.9421	0.9627	0.9074	0.8503	0.8846	0.9121
		$2 \rightarrow 1$	0.8381	0.8482	0.9340	0.8695	0.7855	0.7507	0.9282
		$1 \rightarrow$ osu	0.5345	0.6640	0.3622	0.6710	0.3230	0.3982	0.6878
		$2 \rightarrow$ osu	0.7830	0.8669	0.4961	0.5586	0.5737	0.8685	0.7961
	AdaBoost	$1 \rightarrow 2$	0.8754	0.9835	0.9911	0.8928	0.8754	0.8547	0.9198
		$2 \rightarrow 1$	0.8572	0.8379	0.9195	0.8321	0.6643	0.7229	0.8046
		$1 \rightarrow$ osu	0.5861	0.6884	0.4713	0.4955	0.4373	0.4751	0.6267
		$2 \rightarrow$ osu	0.7907	0.6436	0.5682	0.5161	0.6188	0.8142	0.7579

3.2 Experiment 2: Ground Truth against MSER

Besides the aim of achieving good classification results, we also want to find out how biased our dataset is. Therefore, we evaluate classifiers trained on `sequence 1` ground truth against `sequence 2` and `otcbvs_osu` MSER results and classifiers trained on `sequence 2` ground truth against `sequence 1` and `otcbvs_osu` MSER results. Since the `otcbvs_osu` dataset has been successfully processed by several authors,[16,17] we do not use it for training, but only for testing. If we get good results on this dataset, too, this indicates good generalization abilities. We do not consider partial MSER detections in this experiment, yet, but only fully detected persons and background.

The results are visualized in Table 3. Since the AUC values during the evaluation of our sequences are consistently good, the manually labeled ground truth and the MSER detection results seem to be pretty similar. With Co-occurrence and Moments very good results are achieved. Unfortunately, the evaluation of the `otcbvs_osu` dataset causes a strong performance decrease. Our datasets are different from the `otcbvs_osu` sequences and this results in a performance worse than guessing in some cases. However, Moments provide the best trade-off between the datasets together with DCT. It seems that `sequence 2` is a more representative training set than `sequence 1` due to the consistently better AUC values for the evaluation with `otcbvs_osu`. Again, SVM and AdaBoost perform similar.

3.3 Experiment 3: MSER against MSER

In the final experiment of this section, MSER detection are to be used for training and evaluated with MSER detections as well. In this experiment we also add partial MSER detections of persons to the training and test sets. The motivation is to achieve better classifier generalization. We do not consider DCT_8 and LBP_{G+L} here since they did not perform significantly different than DCT_{16}, LBP_G, or LBP_L.

The AUC values in Table 4 are in most cases better for the evaluation of the `otcbvs_osu` dataset. This shows that with the new training data set consisting of MSER hot spots of background, persons, and partially detected persons is less biased than the manually labeled ground truth data. In some cases, PCA performs better than no feature reduction. This might be due to the curse of dimensionality which is already appearing there are too less training samples in a too high-dimensional feature space. Feature reduction with PCA can decrease the negative effects coming from the curse of dimensionality. The best performance is again achieved when using Moments as features. This brings us to the conclusion that future work should focus on this kind of features when good generalization abilities are to be guaranteed.

Table 4. Area Under Curve (AUC) for Experiment 3.

feature reduction	classifier	test	feature extraction				
			DCT_{16}	COOC	LBP_G	LBP_L	Moments
NONE	SVM	$1 \rightarrow 2$	0.9380	0.9768	0.8880	0.8914	0.9949
		$2 \rightarrow 1$	0.8983	0.9131	0.8529	0.9002	0.9331
		$1 \rightarrow$ osu	0.4275	0.6190	0.7444	0.4204	0.6926
		$2 \rightarrow$ osu	0.7636	0.5022	0.7420	0.8153	0.5686
	AdaBoost	$1 \rightarrow 2$	0.9429	0.9821	0.8826	0.9167	0.9942
		$2 \rightarrow 1$	0.8357	0.8206	0.8579	0.8902	0.8747
		$1 \rightarrow$ osu	0.3315	0.6104	0.6283	0.3916	0.6655
		$2 \rightarrow$ osu	0.7605	0.6224	0.7196	0.8382	0.7908
PCA	SVM	$1 \rightarrow 2$	0.9280	0.9785	0.8366	0.8240	0.9810
		$2 \rightarrow 1$	0.9215	0.9131	0.8761	0.8684	0.9448
		$1 \rightarrow$ osu	0.3748	0.5357	0.5931	0.3451	0.6509
		$2 \rightarrow$ osu	0.7073	0.5853	0.8176	0.7884	0.7389
	AdaBoost	$1 \rightarrow 2$	0.8190	0.9830	0.8591	0.7890	0.9899
		$2 \rightarrow 1$	0.8494	0.9090	0.8227	0.8211	0.8970
		$1 \rightarrow$ osu	0.3759	0.6910	0.4665	0.4366	0.7068
		$2 \rightarrow$ osu	0.8385	0.6973	0.5492	0.6747	0.7810

Table 5. Area Under Curve (AUC) for classification with 11 timesteps of temporal context (history).

feature reduction	classifier	test	feature extraction						
			DCT_8	DCT_{16}	COOC	LBP_G	LBP_L	LBP_{G+L}	Moments
NONE	SVM	$1 \rightarrow 2$	0.9929	0.9948	0.9777	0.9738	0.9066	0.9051	0.9983
		$2 \rightarrow 1$	0.9909	0.9955	0.9473	0.9085	0.8371	0.8532	0.9922
	AdaBoost	$1 \rightarrow 2$	0.9924	0.9807	0.9959	0.9566	0.9056	0.9483	0.9973
		$2 \rightarrow 1$	0.9924	0.9798	0.9671	0.9287	0.7937	0.8653	0.9899

Table 6. Area Under Curve (AUC) for classification with 21 timesteps of temporal context (history).

feature reduction	classifier	test	feature extraction						
			DCT_8	DCT_{16}	COOC	LBP_G	LBP_L	LBP_{G+L}	Moments
NONE	SVM	$1 \rightarrow 2$	0.9943	0.9953	0.9854	0.9789	0.9192	0.9178	0.9986
		$2 \rightarrow 1$	0.9917	0.9957	0.9501	0.9130	0.8396	0.8558	0.9936
	AdaBoost	$1 \rightarrow 2$	0.9943	0.9840	0.9968	0.9627	0.9147	0.9536	0.9976
		$2 \rightarrow 1$	0.9940	0.9826	0.9707	0.9380	0.7956	0.8707	0.9913

4. IMAGE SEQUENCE CLASSIFICATION

In this section, temporal information is introduced. This is possible for the ground truth data only since we do not track the hot spots detected by MSER, yet. Classification with temporal context is implemented quite simple in this paper: each classifier has a history of n timesteps. As soon as there are more than one classifications for the same tracked object available, the final class is determined by voting. Actually, this is not affecting the ROC curves, yet, since they are generated using the decision function values of SVM and AdaBoost. To influence the ROC curves, too, we calculate the mean decision function value in each timestep along the history for each classifier.

Table 5 and Table 6 show the results for a history of 11 and 21 timesteps, respectively. For nearly each case a significant performance improvement is achieved, and a longer history consistently causes better performance as seen when comparing Table 5 and 6. Highest correct classification rates are reached for DCT and Moments. We consider these results as very promising. The maximum in the ROC curve is at 99.9 % true positives with 0.095 %

false positives for Moments features. Future work will include more sophisticated fusion of temporal information such as concatenation of feature vectors coming from different timesteps or the introduction of spatio-temporal features like LBP-TOP[33] or Spatio-Temporal HOG (STHOG).[34]

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a processing chain for person detection in thermal infrared (LWIR) image sequences consisting of MSER hot spot detection and subsequent person classification. The proposed processing chain appears to be a promising approach for person detection, localization, and recognition. Our experiments show that it works well in real world outdoor scenes providing reasonable detection rates with MSER and very good classification performance for different features and classifiers along the detected hot spots. Generalization abilities have been analyzed using the publicly available OTCBVS benchmark dataset for testing classifiers trained on our datasets.

As described in Section 2, alternatives to or modifications of the MSER algorithm should be researched in order to optimize the obtained person detection rates. Especially systematic effects disturbing the detection rates such as merged background and person or missed persons in general are to be studied and improved here. Additionally, the calculation of the bounding boxes could be done hierarchically to produce a set of bounding boxes for each maximal image region instead of just calculating one bounding box considering also different intensity levels of the hot areas. Future work in classification will cover further analysis of features and how to utilize temporal information.

REFERENCES

[1] "AMROS." Internet page. Online at 3th April 2013: http://www.iosb.fraunhofer.de/servlet/is/4593/.

[2] "SENEKA - Sensor network with mobile robots for disaster management." Internet page. Online at 3th April 2013: http://www.iosb.fraunhofer.de/servlet/is/34099/.

[3] Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints," *Intern. Journal of Computer Vision* **60**(2), 91–110 (2004).

[4] Dalal, N. and Triggs, B., "Histograms of Oriented Gradients for Human Detection," in [*Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*], 886–893 (June 2005).

[5] Dong, J., Ge, J., and Luo, Y., "Nighttime Pedestrian Detection with Near Infrared using Cascaded Classifiers," in [*Proceedings of the IEEE International Conference on Image Processing (ICIP)*], **6**, 185–188 (Sept. 2007).

[6] Lee, Y.-S., Chan, Y.-M., Fu, L.-C., Hsiao, P.-Y., Chuangs, L.-A., Chen, Y.-H., and Luo, M.-F., "Nighttime Pedestrian Detection by Selecting Strong Near-Infrared Parts and Enhanced Spatially Local Model," in [*Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*], 1765–1770 (Sept. 2012).

[7] Soga, M., Hiratsuka, S., Fukamachi, H., and Ninomiya, Y., "Pedestrian Detection for a Near Infrared Imaging System," in [*Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems (ITSC)*], 1167–1172 (Oct. 2008).

[8] Miron, A., Besbes, B., Rogozan, A., Ainouz, S., and Bensrhair, A., "Intensity Self Similarity Features for Pedestrian Detection in Far-Infrared Images," in [*Proceedings of the IEEE 2012 Intelligent Vehicles Symposium (IV)*], 1120–1125 (June 2012).

[9] Olmeda, D., Armingol, J. M., and de la Escalera, A., "Discrete Features for Rapid Pedestrian Detection in Infrared Images," in [*Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*], 3067–3072 (Oct. 2012).

[10] Chen, Y. and Han, C., "Night-time Pedestrian Detection by Visual-Infrared Video Fusion," in [*Proceedings of the 7th World Congress on Intelligent Control and Automation (WCICA)*], 5079–5084 (June 2008).

[11] Krotosky, S. J. and Trivedi, M. M., "Person Surveillance Using Visual and Infrared Imagery," *IEEE Transactions on Circuits and Systems for Video Technology* **18**, 1096–1105 (Aug. 2008).

[12] Dai, C., Zheng, Y., and Li, X., "Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery," in [*Proceedings of the 2005 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2005).

[13] Zhang, L., Wu, B., and Nevatia, R., "Pedestrian Detection in Infrared Images based on Local Shape Features," in [*Proceedings of the 2007 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2007).

[14] Jüngling, K. and Arens, M., "Feature based person detection beyond the visible spectrum," in [*Proceedings of the 2009 IEEE CVPR Workshop on Computer Vision Beyond The Visible Spectrum (OTCBVS)*], 30–37 (June 2009).

[15] Xia, D., Sun, H., and Shen, Z., "Real-time Infrared Pedestrian Detection Based on Multi-block LBP," in [*Proceedings of the 2010 International Conference on Computer Application and System Modeling (ICCASM)*], **12**, 139–142 (Oct. 2010).

[16] Li, W., Zheng, D., Zhao, T., and Yang, M., "An Effective Approach to Pedestrian Detection in Thermal Imagery," in [*Proceedings of the 2012 8th International Conference on Natural Computation (ICNC)*], 325–329 (May 2012).

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

[18] Sun, H., Wang, C., and Wang, B., "Night vision pedestrian detection using a forward-looking infrared camera," in [*Proceedings of the 2011 International Workshop on Multi-Platform/Multi-Sensor Remote Sensing and Mapping (M2RSM)*], (Jan. 2011).

[19] Viola, P. and Jones, M., "Robust Real-time Object Detection," *International Journal of Computer Vision* (2001).

[20] Davis, J. W. and Keck, M. A., "A two-stage approach to person detection in thermal imagery," in [*Proceedings of the Seventh IEEE Workshop on Application of Computer Vision (WACV/MOTION)*], 364–369 (Jan. 2005).

[21] Davis, J. W. and Sharma, V., "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding* **106**, 162–182 (May 2007).

[22] Matas, J., Chum, O., Urban, M., and Pajdla, T., "Robust wide baseline stereo from maximally stable extremal regions," in [*Proceedings of the British Machine Vision Conference (BMVC)*], (Sept. 2002).

[23] Teutsch, M. and Krüger, W., "Classification of small Boats in Infrared Images for maritime Surveillance," in [*Proceedings of the 2nd NURC WaterSide Security Conference (WSS)*], (Nov. 2010).

[24] Teutsch, M. and Krüger, W., "Fusion of Region and Point-Feature Detections for Measurement Reconstruction in Multi-Target Kalman Tracking," in [*Proceedings of the International Conference on Information Fusion (FUSION)*], (July 2011).

[25] "OTCBVS Benchmark Dataset Collection." Internet page. Online at 3th April 2013: http://www.cse.ohio-state.edu/otcbvs-bench/.

[26] Bradski, G., "The OpenCV Library," *Dr. Dobb's Journal of Software Tools* (2000).

[27] Ekenel, H. K. and Stiefelhagen, R., "Local appearance based face recognition using discrete cosine transform," in [*Proceedings of the 13th European Signal Processing Conference (EUSIPCO)*], (2005).

[28] Haralick, R. M., Shanmugam, K., and Dinstein, I., "Textural Features for Image Classification," *IEEE Transactions on Systems, Man and Cybernetics* **3**, 610–621 (Nov. 1973).

[29] Ojala, T., Pietikäinen, M., and Mäenpää, T., "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 971–987 (July 2002).

[30] Li, Y., "Reforming the theory of invariant moments for pattern recognition," *Pattern Recognition* **25**, 723–730 (July 1992).

[31] Friedman, J., Hastie, T., and Tibshirani, R., "Additive logistic regression: a statistical view of boosting," *Annals of Statistics* **28**, 2000 (1998).

[32] Breiman, L., "Random forests," *Machine Learning* **45**, 5–32 (Oct. 2001).

[33] Zhao, G. and Pietikäinen, M., "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**, 915–928 (June 2007).

[34] Reddy, K. K., Cuntoor, N., Perera, A., and Hoogs, A., "Human Action Recognition in Large-Scale Datasets Using Histogram of Spatiotemporal Gradients," in [*Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*], (Sept. 2012).