

Person detection in LWIR imagery using image retrieval

Thomas Müller and Daniel Manger

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

ABSTRACT

This paper addresses the detection and localization of persons in LWIR imagery which is useful especially in visual surveillance tasks such as intruder detection in military camps or for gaining situational awareness. A robust image retrieval function is used after a previous hot spot detection and localization step in LWIR using a suitable, extensive image data base that covers a variety of different shapes and appearances of persons in LWIR. The basic idea behind this approach is, in contrast to the visual optical band (VIS), that persons in thermal infrared exhibit somehow similar, weakly individualized signatures which can be matched to a sufficient degree to images in the data base and, thus, can be distinguished from background structures and other objects. Dedicated pre and post processing routines optimize the results and compensate for a possibly occurring lack of image features needed by the image retrieval function. The achieved results document the practical benefit and the robustness of the presented approach.

Keywords: Thermal infrared, MWIR and LWIR, human detection, person recognition, intruder detection, visual surveillance, template matching, cross correlation.

1. INTRODUCTION

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 applications. In this paper, we focus on outdoor applications where persons have to be detected in the environment using a robot equipped with cameras or using stationary cameras. The applications range from intruder detection in military camps and ground security of civil complexes¹ to victim detection after catastrophes². We further focus on thermal infrared imaging since the thermal infrared band (MWIR or LWIR) exhibits some important advantages in these applications. For example, it allows a detection of persons in complete darkness passively (i.e. without the need of active light sources) which is very important in the military domain. Additionally, persons can often be detected more easily or faster in this frequency band than in the VIS or NIR band even in daylight, see figure 1 for illustration. Furthermore, victim detection after building collapses can often only be done using the thermal infrared band because a dust layer covers the scene under which persons are hidden completely when observed with other frequency bands. A further advantage is that LWIR can see through dust and fog in the air by far better than VIS.

When inspecting the LWIR image sequences in the mentioned outdoor applications we have found out (in contrast to indoor scenes, for example) that there are only sparsely distributed hot spots in the environment in a lot of situations (stemming from open windows, open doors, motor blocks of cars, ...). So, in order to solve the person detection and localization problem (which is unsolved in general) we focus on scenarios with a more or less common amount and size of hot areas in the environment (cf. example in the right part of figure 1).

The basic idea behind the presented approach is, that in contrast to the visual optical band (VIS) persons in thermal infrared exhibit somehow similar, weakly individualized signatures which can be matched by the use of an image retrieval technique to images in a data base to a sufficient degree and, thus, can be distinguished from background structures and other objects. Differences in shape (of walking persons for example) or variations in clothes' temperature isolation have to be represented sufficiently by the underlying data base.

Further author information:

Thomas Müller: E-mail: Thomas.Mueller@iosb.fraunhofer.de, Telephone: +49(0)721/6091 458.

Daniel Manger: E-mail: Daniel.Manger@iosb.fraunhofer.de, Telephone: +49(0)721/6091 353.



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

This image retrieval technique has to be combined with a function that reduces the image data to regions of interest that have to be checked. So, the presented approach in this paper is built of two processing steps. In the first step hot spots are detected and localized in the image with a fast scale invariant method to detect and localize persons (and other hot areas) at all different sizes in the image. Afterwards, person's hot spots are distinguished from background structures, clutter and other objects with the image retrieval procedure as a second processing step in order to get an overall system providing fast person detection, localization and recognition in thermal infrared. The first step should detect a high rate of true positives (i.e. should reach a high rate of detected visible persons) where the thereby associated tendency of a higher rate of detected background structures should be no problem due to the second processing step.

Of course, the second step can also be solved with a more traditional approach which we examine in a further paper at this SPIE 2013 conference³. In that paper the second step is seen as a classification problem with two classes ('person' and 'everything else'). It is solved with state of the art classification methods calculating a variety of different image features and using modern classifiers like support vector machine (SVM), Boosting, or Random Forest. A discussion of related work in the discussed field can be found there³.

In this paper, the image retrieval is examined in parallel to that approach in order to optimize the obtainable results in the mentioned challenging application range under its inherent hard real-world conditions. To be more precise, in this paper the image retrieval function is investigated if it can be used as a robust classifier of hot spots to distinguish persons from background structures and other objects in a nearest neighbor classification scheme. Two image retrieval algorithms are investigated: a powerful, generic, rotation invariant and fast state of the art image retrieval system⁴⁻⁷ based on SIFT-features⁸ and a purpose-built, but slower image retrieval function based on improved template matching in order to optimize the obtained results.

The challenge is to robustly cover low and high object distances, to be robust against variable background, weak signal-to-noise ratio (SNR), weak contrast and sensor-specific noise of MWIR/LWIR as well as occurring motion blurring due to camera and/or person motion.

This paper is organized as follows: In section 2, the detection of hot spots will be described and evaluated. Afterwards, the examined image retrieval system based on SIFT-features is introduced and investigated in section 3. Section 4 presents the proposed purpose-built image retrieval based on template matching techniques and presents a novel template distance measurement. In Section 5, the obtained experimental results are presented and discussed. Finally, the conclusions and an outlook to potential future work are given in section 6.

2. HOT SPOT DETECTION

When thinking about and looking for an adequate hot spot detection algorithm, we first came upon the MSER (Maximally Stable Extremal Regions) approach by Matas et al.⁹ because of our experiences with that algorithm in the near past in other applications^{10,11}. So we made our first tests with it and found out that this algorithm

	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

Table 1. Detections and misses of the hot spot detection using MSER (ground truth, generated by human inspection).

fulfills the requirements mentioned in section 1. Then, 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 a hot spot detection for itself without further integrated knowledge about the appearance of persons 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 also with other hot spot detection algorithms because they cannot be avoided due to their nature. And, of course, this topic is not a principal problem due to construction of the proposed processing chain in which the second step filters such unwished effects out. Since, therefore, this aspect effects frequencies but not the approach itself we decided to shift possible examinations of hot spot detection alternatives to a future work package at this point and to start with MSER in order to concentrate on the main aspects intended with this paper which is proving the principal practical benefit of the proposed processing chain and of the classification step in the discussed application range. Future work can afterwards optimize the frequency details.

In the implemented hot spot detector the MSER results are used to calculate bounding boxes around the maximal image regions. The MSER results for minimal 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 (we use 6 pixels for the border size).

In our experiments we used two outdoor image sequences: **sequence 1** with 4580 LWIR images in length 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 collection¹²⁻¹⁴: 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 person misses (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 the words about unwished effects in the above text) are hot spots of persons with a large amount of additional background structure around due to contrast reasons. Such hot spots are useless in the context of the proposed processing chain since they are not represented in the data base. Partly visible persons (when persons appear at the image border for example) are not considered here (they could just be handled in the same way as the fully visible persons, if desired). Figure 2 depicts the detected persons of **sequence 1** and **sequence 2** for illustration of the image data character and the difference between the two sequences.

As the numbers in the table show, there are enough person detections (in the average over time) for practical

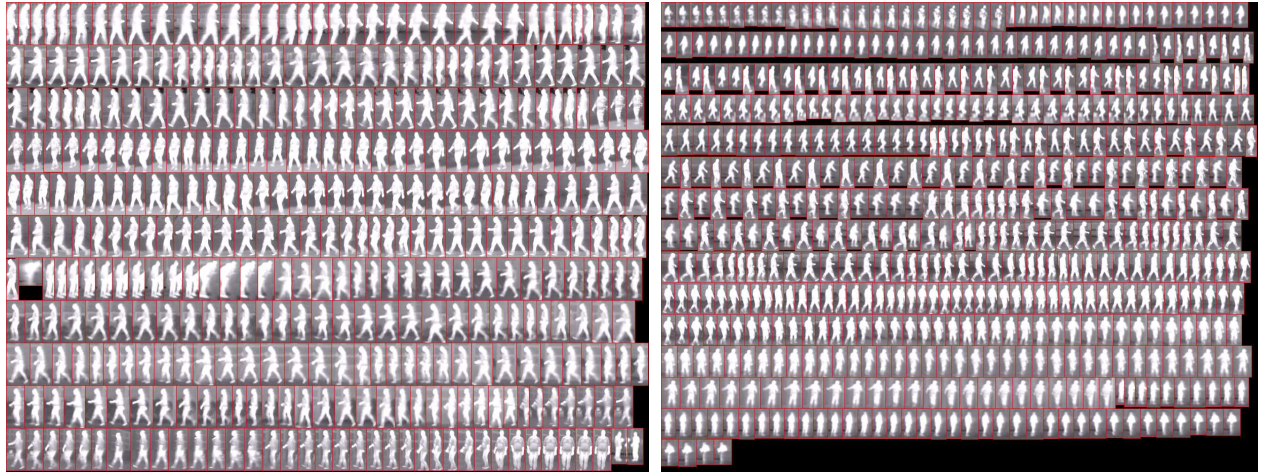


Figure 2. Detected persons of **sequence 1** (left) and **sequence 2** (right).

operation. Because, to our experience, it does not matter in practice if a person is not detected in every single image and it is sufficient to produce (robust) alarms with moderate latency.

3. IMAGE RETRIEVAL BASED ON SIFT-FEATURES

3.1 Image retrieval system

The aim of content-based image retrieval systems is to compare images with respect to their content. To this end, local image regions are compared using local features. Local features like the popular Scale-Invariant Feature Transform (SIFT)⁸ are used in many different topics of computer vision. They typically detect repeatable salient regions in an image and subsequently encode their local image appearance in a descriptor. Given the two sets of descriptors of two images, similar regions in both images can be searched by determining descriptors which are similar in descriptor space, which is for SIFT usually 128 dimensional. Typically, distances are calculated by L2 norm and a threshold is applied on the distance or on the ratio of closest to second closest distance. The similarity of two images is then often calculated as the number of matching features.

For matching sets of descriptors, various heuristic algorithms have been proposed which can lead to an impressive speedup while sacrificing not too much of the descriptors discriminance^{15,16}. Nevertheless, in large-scale CBIR systems with thousands or millions of images, a pair-wise image comparison of the query image with every image of the database becomes infeasible. Besides, the memory consumption of the image features and their processing during one query prohibit a direct matching of descriptors sets. To solve this, the bag-of-words (BOW) representation has been proposed¹⁷, which quantizes the features by assigning every feature to one element of a set of feature representatives called visual words. Thus, the image matching can be performed with text retrieval methods analyzing the common visual words of images. The set of visual words termed codebook or visual vocabulary is commonly obtained by clustering an independent set of features. Using large codebooks, the representation of an image becomes a very sparse vector indicating the occurring visual words. This sparsity can be exploited by inverted files which store for every visual word a list of references to the images containing at least one feature corresponding to that visual word.

While enabling the construction of fast and efficient systems, the quantization of features also comes with the drawback of loss of information which leads to a reduced accuracy of the overall system. We use two popular extensions to circumvent the loss of accuracy namely Hamming Embedding (HE)¹⁸ and Weak Geometry Consistency (WGC)¹⁸. Both techniques have shown a significant improvement of performance in large scale image retrieval. In previous experiments, we could confirm this for the performance in tattoo image retrieval⁴ with a database of up to 330,000 images.

As rare visual words are assumed to be more discriminative, the similarity of two images given the two BOW vectors is commonly calculated using the tf-idf scheme¹⁷. It weights the BOW vectors according to both the

best hit	#rating values	minimum	maximum	mean value	standard deviation
persons	570	2.63	22.61	7.99	3.38
background	162	0.64	15.93	4.96	2.29

10 best hits	#rating values	minimum	maximum	mean value	standard deviation
persons	5700	0.43	22.61	3.44	2.75
background	1620	0.45	15.93	2.34	1.75

Table 2. Evaluation of the rating values of the image retrieval system (persons of **sequence 1** in data base, person and background spots of **sequence 2** used as query images) when considering the best hit (upper table part) or the 10 best hits (lower table part).

person query			background query		
person found	background found	nothing found	background found	person found	nothing found
226 (39.6%)	33 (5.8%)	311 (54.6%)	2 (1.2%)	0 (0%)	160 (98.8%)

Table 3. Result of the second experiment (persons and background spots of **sequence 1** in data base, person and background images of **sequence 2** used as query images).

local frequency (within the image) and the global frequency (within the entire database). In all experiments in this paper, we use the similarity function of Schmid¹⁹ which is the cosine angle between the weighted BOW vectors which equals the L2 normalized dot product of the vectors. See Jegou et al.¹⁸ for details.

To further increase the performance, we make use of a subsequent re-ranking step, which performs a matching based on the original features. The images are re-ranked according to the number of matches with the query image.

The two main advantages of this image retrieval system are that it can work with large data bases and that it is able to robustly find similar images in the data base for a given query image. Because the image retrieval bases on SIFT-features, the hot spots used as query and data base images are scaled in the following experiments to assure a minimum of 200 pixel in height for the images.

3.2 Application directly to LWIR hot spots

In the first experiment it is examined if the rating values associated with the output results of the image retrieval have the potential for distinguishing person hot spots from other ones. This is done as follows. The 380 person detections of **sequence 1** are used as data base. And all 570 person detections as well as the 162 background detections of **sequence 2** are used as query images. Finally, minimum, maximum as well as arithmetic mean and standard deviation are calculated for both query classes assuming a normal distribution. This is done for the (single) best hits in the data base as well as for the 10 best hits, see table 2. Big rating values denote high similarities between query image and found data base image in this software. When looking at the data in the table it can be stated that the ratings of persons and background structures differ statistically in a significant manner. From this, the principal applicability of the image retrieval as classifier can be derived. But the ratings for persons and background structures intersect too much so that the rating value by itself should not be used as criterion to decide if a hot spot is a person or something else. The results degrade significantly when considering the 10 best hits, so just the best hit is used in the following evaluations.

In the second experiment the image retrieval is used as a nearest neighbor classifier. This is done by using the 380 person hot spots and the 1131 background spots of **sequence 1** as data base. The person and background spots of **sequence 2** are used as query data and the number of true and false positives and negatives is counted based on the ground truth that was determined by hand previously for each hot spot. The result is shown in table 3. The results show that the proposed procedure works principally fine, since person and background queries are answered significantly more often with a candidate of the correct class than with a candidate of the other (wrong) class. But there are many cases where the image retrieval software cannot produce a result. The reason is that there are a lot of images for which not enough SIFT-features can be calculated. In order to improve

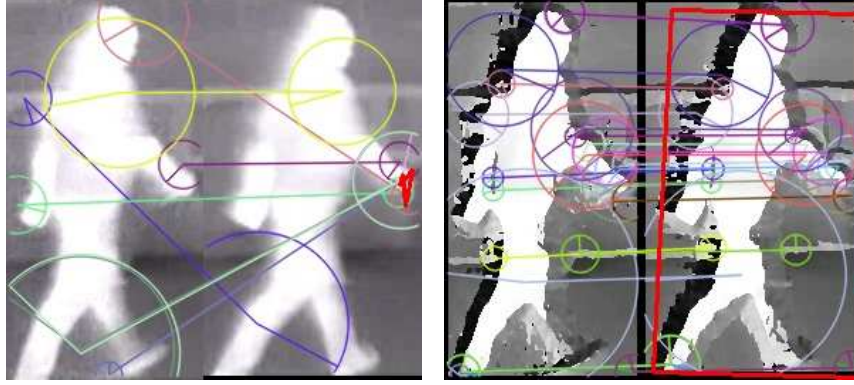


Figure 3. Left: LWIR spots of two different images of **sequence 1** with SIFT-features depicted by overlaid circles and indications of the assignments in between depicted by overlaid lines. Right: the same two spots enriched with gradient direction information leading to more SIFT-features that can be assigned much better. The overlaid rectangle visualizes the homography that can be calculated between the two spots based on the SIFT-features. Such a homography cannot be calculated for the original LWIR spots on the left.

person query		background query		total
person found	background found	background found	person found	correct hits
289 (50.7%)	281 (49.3%)	151 (93.2%)	11 (6.8%)	440 (60.1%)

Table 4. Result of the third experiment (persons and background spots of **sequence 1** in data base, person and background images of **sequence 2** used as query images) when using preprocessed LWIR spots.

that, the LWIR spots are preprocessed in the following subsection before they are used in the image retrieval system in order to increase their potential to produce SIFT-features.

3.3 Application to preprocessed LWIR hot spots

In order to improve the results obtained so far and to enrich the LWIR spots with more useful SIFT-features multiple variants were tested based on the first image derivative: image representing the absolute value of the gradient, the gradient direction or the maximum of the image gradient in gradient direction (NAG²⁰), each combined with or without a threshold for the absolute value of the gradient, while experimenting with black or white background when using the threshold or using the original LWIR image as background. The best results were obtained when using the threshold for the absolute value of the image gradient and overlaying the LWIR image with the gradient direction information coded with values from 0 to 230. Figure 3 shows an example for illustration including the calculated SIFT-features. Table 4 summarizes the results when repeating the second experiment on the preprocessed query and data base spots (third experiment). The results show that this image retrieval system cannot be used in this way as classifier against previous expectations. So, the best results were reached with the original LWIR spots despite the fail of result calculation in some situations. With that result, the potential of the SIFT-based image retrieval system seems to be exhausted. Therefore, a further image retrieval function is constructed and examined in the following section.

4. IMAGE RETRIEVAL BASED ON TEMPLATE MATCHING

With a purpose-made image retrieval we want to examine now how far the previous results can be improved by substituting the SIFT-based distance measurement by a measurement which evaluates the shape of the hot spot detections.

In a first step every spot is scaled to a fixed, small image height (we use 30 pixels) in order to throw away the redundant shape details in large spots and to reduce the image content to its essential shape. Then, for every

image d in the (scaled) data base the template difference

$$r = r(q, d) := \min_{x,y} m(x, y)$$

between (scaled) query image q and (scaled) data base image d is calculated over all possible translations (x, y) between q and d for which there is enough overlap of at least 25 % between q and d . (x, y) is varied in order to compensate for spatial variations inside the hot spots. m denotes the chosen measurement function for the template distance. The result r is a rating for the similarity of q and d . In contrast to the SIFT-based image retrieval system small values denote good similarities here. The final image retrieval result for a query image q is the data base image d with minimal rating r .

m can be chosen as one of the following template distances m_k , for example as a standard measurement known from the literature^{21–23} like the cross correlation

$$m_1(x, y) := -\frac{1}{|F|} \sum_{(i,j) \in F} d(x+i-a, y+j) \cdot q(i, j) \quad (1)$$

or the normalized cross correlation (NCC)

$$m_2(x, y) := -\frac{1}{|F|} \frac{\sum_{(i,j) \in F} d(x+i-a, y+j) \cdot q(i, j)}{\sum_{(i,j) \in F} d(x+i-a, y+j)^2}, \quad (2)$$

where F is the set of coordinates so that q and d in the sum are both defined. The constant a is just used to place the center of d over the middle of q when q and d have different image widths (see T. Müller and M. Müller²² for details). Due to scaling at the beginning the image heights of q and d are always the same leading to an offset of always 0 between q and d in y -direction in contrast to the x -direction.

T. Müller and M. Müller²² suggest to use the sum of absolute differences

$$m_3(x, y) := \frac{1}{|F|} \sum_{(i,j) \in F} |d(x+i-a, y+j) - q(i, j)| \quad (3)$$

which has been found out in their paper to be better in such a kind of applications.

In this paper here, we propose the following novel measurement m_4 . First, let the *local grayvalue dynamic* of an image $h(x, y)$ be represented by the intervall

$$Z_h(x, y) := [z_{h,\min}(x, y), z_{h,\max}(x, y)] \quad (4)$$

with

$$z_{h,\min}(x, y) := \min \{h(x + \Delta x, y + \Delta y) \mid (\Delta x, \Delta y) \in \{(-1, 0), (0, 0), (1, 0), (0, -1), (0, 1)\}\}, \quad (5)$$

$$z_{h,\max}(x, y) := \max \{h(x + \Delta x, y + \Delta y) \mid (\Delta x, \Delta y) \in \{(-1, 0), (0, 0), (1, 0), (0, -1), (0, 1)\}\} \quad (6)$$

and $p(I, J) \in [0, 100]$ the percentage of the overlap of two intersecting intervalls $I = [i_1, i_2]$, $J = [j_1, j_2] \subset \mathbb{R}$:

$$p(I, J) := 100 \cdot \frac{\min\{i_2, j_2\} - \max\{i_1, j_1\}}{\max\{i_2, j_2\} - \min\{i_1, j_1\}} \quad (7)$$

for $\max\{i_2, j_2\} - \min\{i_1, j_1\} \neq 0$ and $p(I, J) := 100$ otherwise. Furthermore, let g be a scaled and shifted version of the gaussian function, i.e.

$$g(x) := \frac{\alpha}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} + \beta \quad (8)$$

with $\mu := 0$ and $\sigma := 50$ and some fixed constants α and β so that $g(0) = 1$ or $g(0) = 0.8$ and $g(100) = 0$.

With the use of the distance function

$$n(I, J) := \begin{cases} 1, & \text{if } I \cap J = \emptyset \\ g(p(I, J)), & \text{if } I \cap J \neq \emptyset \end{cases} \quad (9)$$

for two intervals I and J we finally define

$$m_4(x, y) := \frac{100}{|F|} \sum_{(i,j) \in F} n(Z_d(x+i-a, y+j), Z_q(i, j)) \quad (10)$$

$m_4(x, y)$ is a value between 0 and 100 and can be interpreted as a percent value measuring how similar two images d and q are with respect to offset (x, y) in their common intersection area. if $d = q$ holds in this area, $m_4(x, y)$ is zero. In comparison to m_3 in formula 3, in m_4 not only the grayvalues of two images are compared but the local grayvalue dynamics. If both dynamics are low in homogeneous image regions and the grayvalues are equal, the contribution to m_4 is low. If both dynamics are big but similar the contribution to m_4 is low, too. The more the grayvalues and/or the dynamics differ (for example when comparing a homogeneous region to an image edge region or dot feature), the greater is the contribution to the distance measurement m_4 . The gaussian component g in the formula controls how important and less important contributions are weighted relative to each other. And the choice of $g(0)$ controls the weighting between intersecting and non intersecting grayvalue dynamics. In some experiments, we concretely compared the results obtained with $g(0) = 1$ and $g(0) = 0.8$ in order to find out, if the concrete choice of $g(0)$ makes a significant difference. Since we found out, that it makes a difference and $g(0) = 0.8$ led to better results in our experiments we worked with that value. In further experiments we substituted the function $g(x)$ in the above formulas with $g^v(x)$, $v \in \mathbb{N}$ in order to evaluate the dependencies of the obtained results from the weighting function. The results are presented in the following section among the other made experiments.

5. EXPERIMENTAL RESULTS

In the following experiments the image retrieval variants based on template matching which were presented in the previous section are evaluated. This is done in a way similar to the evaluations of the SIFT-based image retrieval. Person and background spots are used as data base while person and background structures from different image sequences are used as query images in order to measure the rates of correct and incorrect assignments when using the image retrieval as a nearest neighbor classifier. Table 5 documents the performed experiments and table 6 summarizes the obtained results. Experiments 1 and 2 show that m_1 and m_2 are obviously completely inappropriate in this kind of application. Always the same few background images were found in the data base leading to very low success rates. In contrast to that, experiment 3 turns out m_3 as a good template distance which confirms the experiences made in the paper of T. Müller and M. Müller²² that m_3 is a far better choice in such kinds of applications than the common distance measurements m_1 and m_2 .

As experiment 4 shows, the results with m_3 can be improved significantly by using the proposed novel distance measurement m_4 . By increasing the exponent v from 1 up to 4 (experiments 5 and 6), the results can even further be improved. This means that it is advantageous to lower the influence of smaller contributions in the sum of m_4 in relation to the bigger contributions so that the bigger ones (i.e. big differences between the compared templates) are emphasized more. The results (nearly) remain the same when v is further increased to 6 or 8 as experiments 7 and 8 show.

In experiment 9 we tried to match the spots of the `otcbvs_osu`-sequences against a database consisting of the spots from `sequence 1` and `sequence 2`. Despite the success reached in the experiments before with m_4 , the quite low success rates in this experiment show that it is difficult to match very different image bases against each other, though. In other words, the training of the person detector (i.e. the generation of the data base for later queries) should be done under somehow similar conditions as the practical application later. But, of course, this statement is well known in this field and cannot be changed anyway - at most this problem can be weakened.

The experiments 10 and 11 show the results for different v when going from the `sequence 1` and `sequence 2` scenario completely into the `otcbvs_osu` discourse. The data base is built with the first three sequences

Experiment 1	$m := m_1$
Experiment 2	$m := m_2$
Experiment 3	$m := m_3$
Experiment 4	$m := m_4, v := 1$
Experiment 5	$m := m_4, v := 2$
Experiment 6	$m := m_4, v := 4$
Experiment 7	$m := m_4, v := 6$
Experiment 8	$m := m_4, v := 8$
Experiment 9	$m := m_4, v := 6$, using the person and background spots of sequence 1 and sequence 2 as database and otcbvs_osu 1 to otcbvs_osu 10 as query images
Experiment 10	$m := m_4, v := 4$, using the person and background spots of otcbvs_osu 1 to otcbvs_osu 3 as database and otcbvs_osu 4 to otcbvs_osu 10 as query images
Experiment 11	$m := m_4, v := 8$, using the person and background spots of otcbvs_osu 1 to otcbvs_osu 3 as database and otcbvs_osu 4 to otcbvs_osu 10 as query images
Experiment 12	$m := m_4, v := 8$, using the person and background spots of otcbvs_osu 8 to otcbvs_osu 10 as database and otcbvs_osu 1 to otcbvs_osu 7 as query images

Table 5. Performed experiments with the image retrieval function based on template matching. See text for variants and parameters of that approach. In the experiments 1 to 8 persons and background spots of **sequence 1** were used as data base and person and background images of **sequence 2** were used as query images.

	person query		background query		total correct hits
	person found	background found	background found	person found	
Experiment 1	2 (0.4 %)	568 (99.6 %)	161 (99.4 %)	1 (0.6 %)	163 (22.3 %)
Experiment 2	0 (0.0 %)	570 (100.0 %)	160 (98.8 %)	2 (1.2 %)	160 (21.9 %)
Experiment 3	402 (70.5 %)	168 (29.5 %)	161 (99.4 %)	1 (0.6 %)	563 (76.9 %)
Experiment 4	504 (88.4 %)	66 (11.6 %)	159 (98.1 %)	3 (1.9 %)	663 (90.6 %)
Experiment 5	528 (92.6 %)	42 (7.4 %)	160 (98.8 %)	2 (1.2 %)	688 (94.0 %)
Experiment 6	541 (94.9 %)	29 (5.1 %)	158 (97.5 %)	4 (2.5 %)	699 (95.5 %)
Experiment 7	541 (94.9 %)	29 (5.1 %)	158 (97.5 %)	4 (2.5 %)	699 (95.5 %)
Experiment 8	543 (95.3 %)	27 (4.7 %)	158 (97.5 %)	4 (2.5 %)	701 (95.8 %)
Experiment 9	174 (25.3 %)	515 (74.7 %)	152 (85.9 %)	25 (14.1 %)	326 (37.6 %)
Experiment 10	480 (99.2 %)	4 (0.8 %)	88 (83.8 %)	17 (16.2 %)	568 (96.4 %)
Experiment 11	476 (98.3 %)	8 (1.7 %)	92 (87.6 %)	13 (12.4 %)	568 (96.4 %)
Experiment 12	478 (98.6 %)	7 (1.4 %)	47 (28.0 %)	121 (72.0 %)	525 (80.3 %)

Table 6. Result of the experiments documented in table 5 using the proposed image retrieval that is based on template matching.

otcbvs_osu 1 to otcvbs_osu 3 and the query images stem from the other seven otcvbs_osu sequences. As can be seen in the obtained results shown in table 6 this works fine again. The results are even better than with using sequence 1 and sequence 2 as data base and query.

Finally, in experiment 12 the three last sequences of the otcvbs_osu data set are used for the data base instead of the first three ones as in experiment 11. The results for person queries are quite the same as in experiment 11 while for the background queries worse results are obtained.

6. CONCLUSIONS AND FUTURE WORK

In this contribution it could be shown that a person detection, localization, and recognition system for thermal imagery in a wide application range can be realized by the proposed two-step processing scheme. After a hot spot detection is done in the first step to find potential thermal signatures of persons at arbitrary scale in the image, an appropriate image retrieval function can be used in a second step as classifier to decide which detected spots are persons and which are not. The second step also compensates for weaknesses of the first step that occur naturally and inherently in thermal imagery.

A state of the art image retrieval system based on SIFT-features that works robust in the visual-optical band has shown weaknesses in the thermal band due to a possible lack of SIFT-features in such images which could not be compensated with the methods examined in this paper. Therefore, a novel purpose-built (but by far slower) image retrieval function based on template matching was constructed in order to compare image content by shape. It could be proved that this method works well and fills out the role as classifier in a nearest neighbor classification scheme. Two image data sets were evaluated for which classification success rates of 95.8% and 96.4% could be reached.

Despite an adequate and significant down scaling of the processed images, the template matching based image retrieval does not work in real-time for large data bases. The time complexity is $O(n)$ with n denoting the number of images in the data base. So, future work should deal with speeding this up somehow - directly by optimizing the developed algorithms, by looking for algorithmic alternatives or by reducing large data bases with some intelligent algorithm.

Future work should also deal with increasing the detection rates of person hot spots. Also, for example, by decreasing the rate of detections where persons are mixed up with a large amount of background structure due to contrast effects.

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] M. Teutsch, T. Müller: "Hot Spot Detection and Classification in LWIR Videos for Person Recognition". SPIE Defense, Security, and Sensing, Baltimore Convention Center Baltimore, Maryland, USA, 29th April - 3rd May 2013. In Proceedings of the SPIE: Sensor Data and Information Exploitation: Automatic Target Recognition XXIII, 2013.
- [4] D. Manger: "Large-Scale Tattoo Image Retrieval". In 2012 Canadian Conference on Computer and Robot Vision CRV, 2012.
- [5] D. Manger: "Tattoo Image Retrieval for Forensics". In 6th European Academy of Forensic Science Conference EAFS, 2012.
- [6] D. Manger: "Content-based Tattoo Image Retrieval". In 21st International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society ANZFSS, 2012.
- [7] D. Manger: "Demo: A Tattoo Image Retrieval System". IEEE International Workshop on Information Forensics and Security WIFS, 2012.
- [8] D. G. Lowe: "Distinctive Image Features from Scale-Invariant Keypoints". In International Journal of Computer Vision IJCV, 60(2), Springer, 2004, pp. 91–110.

- [9] J. Matas, O. Chum, M. Urban, and T. Pajdla: “*Robust wide baseline stereo from maximally stable extremal regions*”. In Proceedings of the British Machine Vision Conference (BMVC), Sept. 2002.
- [10] M. Teutsch and W. Krüger: “*Classification of small Boats in Infrared Images for maritime Surveillance*”. In Proceedings of the 2nd NURC WaterSide Security Conference (WSS), Marina di Carrara, Italy, Nov. 2010.
- [11] M. Teutsch and W. Krüger: “*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), Chicago, IL, USA, July 2011.
- [12] “*OTCBVS Benchmark Dataset Collection*”. Internet page. Online at 3th April 2013: <http://www.cse.ohio-state.edu/otcbvs-bench/> .
- [13] J. W. Davis and M. A. Keck: “*A Two-Stage Approach to Person Detection in Thermal Imagery*”. In Proceedings of the Seventh IEEE Workshop on Application of Computer Vision (WACV/MOTION), January 2005, pp. 364–369.
- [14] J. W. Davis and V. Sharma: “*Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery*”. In Computer Vision and Image Understanding Vol. 106, No. 2-3, May 2007, pp. 162–182.
- [15] M. Muja, D. G. Lowe: “*Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration*”. In International Conference on Computer Vision Theory and Applications VISSAPP, 2009.
- [16] D. Nister, H. Stewenius: “*Scalable Recognition with a Vocabulary Tree*”. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, Vol. 2, 2006, pp. 2161–2168.
- [17] J. Sivic, A. Zisserman: “*Video Google: A Text Retrieval Approach to Object Matching in Videos*”. In Ninth IEEE International Conference on Computer Vision Proceedings, 2003, pp. 1470–1477.
- [18] H. Jegou, M. Douze, and C. Schmid: “*Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search*”. In European Conference on Computer Vision ECCV, 2008, pp. 304–317.
- [19] C. Schmid: “*Improving Bag-Of-Features for Large Scale Image Search*”. In International Journal of Computer Vision IJCV, 2011, pp. 316–336.
- [20] V. Gengenbach: “*Einsatz von Rückkopplungen in der Bildauswertung bei einem Hand-Auge-System zur automatischen Demontage*”. PhD Thesis, University of Karlsruhe, Juli 1994. Released in: Dissertationen zur künstlichen Intelligenz (DISKI), Vol. 72, infix publishing company, Sankt Augustin, 1994.
- [21] R. Brunelli: “*Template Matching Techniques in Computer Vision: Theory and Practice*”. Wiley, ISBN 978-0-470-51706-2, 2009.
- [22] T. Müller and M. Müller: “*CART V: Recent Advancements in Computer-Aided Camouflage Assessment*”. SPIE Defense, Security & Sensing Symposium, Orlando, Florida, USA, 25th - 29th April 2011. In Proceedings of the SPIE: Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXII, 2011.
- [23] Introduction to Template Matching at Wikipedia: http://en.wikipedia.org/wiki/Template_matching .