Fast eyes detection in thermal images

Mateusz Knapik¹ · Bogusław Cyganek¹

Received: 27 August 2019 / Revised: 16 June 2020 / Accepted: 21 July 2020 /

Published online: 23 September 2020

© The Author(s) 2020

Abstract

In recent years many methods have been proposed for eye detection. In some cases however, such as driver drowsiness detection, lighting conditions are so challenging that only the thermal imaging is a robust alternative to the visible light sensors. However, thermal images suffer from poor contrast and high noise, which arise due to the physical properties of the long waves processing. In this paper we propose an efficient method for eyes detection based on thermal image processing which can be successfully used in challenging environments. Image pre-processing with novel virtual high dynamic range procedure is proposed, which greatly enhances thermal image contrast and allows for more reliable computation of sparse image descriptors. The bag-of-visual-words approach with clustering was selected for final detections. We compare our method with the YOLOv3 deep learning model. Our method attains high accuracy and fast response in real conditions without computational complexity and requirement of a big dataset associated with the deep neural networks. For quantitative analysis a series of thermal video sequences were recorded in which eye locations were manually annotated. Created dataset was made publicly available on our website.

Keywords Eye detection · Thermal imaging · Virtual high dynamic range · Drowsiness control sparse descriptor

1 Introduction

Although an idea of autonomous cars is taking on popularity, road security still poses a major issue. Systems for self-driving gain new abilities each year. This involves fusing data from multiple sources as well as forces researchers to come up with new and innovative frameworks for data sharing and analysis among connected vehicles, like for example approach presented by Alam et al. in [3]. On the other hand, a key subject in this ecosystem constitute millions of drivers whose state and behavior play an essential role. Drivers

✉ Bogusław Cyganek
cyganek@agh.edu.pl

Mateusz Knapik
mknapik@agh.edu.pl

¹ AGH University of Science and Technology, Al. Adama Mickiewicza 30, 30-059, Krakow, Poland

are still required to be fully aware of their environment. Therefore, drivers' awareness control/monitoring systems can help in this respect [8, 9]. However, to succeed, such systems must be reliable, versatile, and be able to reliably operate in difficult conditions, such as day and night illumination, shocks, temperature, noise, etc., as well as to deal with drivers' variability in appearance and behavior. On the other hand, such assisting system must provide reliable and continuous real-time information to the higher layers of information processing. Lastly, it cannot be a source of distraction itself. We argue, that thermal imaging offers many benefits.

Systems for detecting fatigue driving found much attention in recent times, both from science and the automotive industry. Nevertheless, real-time monitoring of driver's tiredness level, especially using facial-features is a complex task that involves many smaller steps, such as fast data acquisition, image processing and behavior modeling. Moreover, the front-end of almost all such systems requires some form of face localization module. However, we propose a different solution. Due to recent advances in manufacturing of the low-cost long-wave infrared (LWIR) spectrum sensors, thermal imaging became a viable option to consider in consumer-centric applications. However, processing of such images differs in many aspects from the visible spectrum. To be able to successfully conclude about driver's vigilance level, or to monitor her/his behavior, proper image analysis methods are necessary.

In this paper a novel method for eye localization in thermal images is proposed. The main novelty of our approach comes from direct detection approach and prior thermal signal enhancement using the virtual high dynamic range imaging (VHDR) module. Only after that, computation of the scale-invariant feature transform (SIFT) descriptors [24] on a dense grid becomes reliable and efficient in the long-wave infrared images. This step is then followed by the clustering process aimed at selection of distinctive prototype patterns. For final eye localization our method employs dictionary matching, distance computation and verification of the geometric conditions. Such connection of steps and methods allows for very reliable eye detection in thermal images without the need for complex and computationally expensive convolutional neural networks (CNN) training, as shown by experiments conducted with the hand annotated ground truth data. These were also made available from the Internet what constitutes the next contribution of this paper. Moreover, although designed for automotive environment, the proposed method of eyes localization in thermal images is universal and can be used in other domains, such as medical imaging, surveillance, virtual reality, entertainment, just to name a few.

The rest of this paper is organized as follows: Section 2 describes the existing state-of-the-art works that are related to the proposed system. Section 3 is an overview of the system architecture and theory behind methods used throughout this paper. Section 4 presents the results of the conducted experiments to ascertain scientific validity of presented system. Finally, Section 5 presents the conclusions.

2 Related works

In this section an overview of the influencing works related to the localization of facial areas in infrared spectrum is presented and discussed.

2.1 Contrast enhancement

Many methods have been proposed for image contrast enhancement [9]. A recent example is the adaptive contrast enhancement method that preserves the image brightness, proposed

by Khan et al. [26]. By transforming image to the wavelet domain they leverage the multi-resolution analysis. Moreover, Particle Swarm Optimization (PSO) is utilized to select optimal values for final gamma correction. Proposed method maximizes the performance of PSO by means of novel fitness function. However, for contrast enhancement in images with limited dynamic range, the methods based on the virtual high dynamic range imaging were shown to provide superior results [11, 16]. In the method proposed by Cyganek & Woźniak, a set of tone curves is applied to a single input image to create additional (virtual) exposures. The following fusion of such virtual images, combined with range conversion and high-pass filtering, allow for computation of high dynamic range image based solely on a single input with a low dynamic range (LDR) image. In this paper we extend this idea and propose VHDR as a preprocessing step for the thermal images (result of such operation can be seen in Fig. 1), greatly enhancing further object detection with the sparse features descriptors and classification modules, as well as the deep detection neural networks, as will be shown.

2.2 Features acquisition and object detection

Feature computation and selection are primary steps in any classification task [5, 9, 40]. For years the expert designed features played a predominating role, giving the field in recent years to the learned features by the deep neural architectures. The subject is very broad and here we only scratch the surface by providing some examples [1, 2, 21, 22, 25, 34, 35, 42].

Successful object detection became possible after invention by D. Lowe of the sparse SIFT detector/descriptor framework [24]. Soon, many next sparse descriptors, as well as methods of their applications, were proposed which we briefly outline here. For example Fulkerson et al. proposed to incorporate SIFT features, computed on a dense grid, for localization and categorization of objects in images [14]. Their presented methods focus on reducing size of very large features dictionaries as well as on improvements in computation speed of the SIFT descriptors. Localization of objects is done on a pixel level by usage of local features histograms.

On the other hand, visual image categorization by the method called bag-of-words has gained much popularity [20]. It relies on clustering of the sparse SIFT features to compose 'image words' which are then used for image matching and classification. However, despite its popularity in processing of airy images, the method only recently was applied to other spectra [27]. This is due to poor contrast and noise of thermal images which impedes proper computation of image derivatives, underlying further computations. We directly address this issue in the proposed method, as will be discussed.

Fig. 1 Thermal image (left) and its virtual high dynamic range version (right)

Bag-of-words approach is also used in other computer vision tasks. Method proposed by Zeng et al. in [41] extends this approach for contour matching. Shape contour is decomposed into fragments to form curvature descriptors, that are then matched with SVM classifier.

In more recent papers, deep learning approach is used for facial features recognition. Kwasniewska et al. [19] proposed a system based on the modified Inception v2 [38] neural network that was able to detect and track selected facial-feature in low resolution thermal images with accuracy close to 90%. Performance of their approach was close to real-time, although this was obtained on a high-end workstation equipped with powerful GPU.

Chu et al. in [6] proposed modification to general object detector Faster Region-based Convolutional Neural Networks (Faster R-CNN) for detection of small and occluded objects. Region proposal network was optimized by multi-layer convolution feature fusion (MCFF), so that high-level and low-level features are fused to address scale problem. Addition of online hard example mining (OHEM) improves convergence speed of the model training by focusing on effective samples.

In [36], Sonkusare presented cascaded-framework, combining their own deep learning landmark estimation method with OpenPose system [4] that employs a robust multi-view bootstrapping architecture. Accuracy of their proposed system was reported to achieve as high as 92% at the cost of using two convolutional neural networks. Authors also highlighted the low performance of standard RGB methods when applied to thermal imaging.

This trend continues as enhanced versions of state-of-the-art techniques are presented. In [43] Zhang et al. proposed Mask-Refined R-CNN, which builds on state-of-the-art general object detection model and enhances its segmentation performance. The FPN structure is added to the mask head and the higher-level features are propagated to lower layers. In result, mask prediction is more sensitive to the instance details. On the other hand, this still requires powerful platform, preferably with GPU, to run on.

2.3 Detection in infrared spectrum

In the paper by Wang et al. [39] an automatic system for eyes localization for human subjects, both with glasses and without glasses, is proposed. Vertical and horizontal projections in conjunction with the SVM classifier are used to find characteristic valleys, and then to detect and localize eyeglasses. In a case of subject without them, the image is divided into 15 subregions and algorithm comprising Haar-like features with Adaboost and eight classifiers are used for eye localization based on a voting strategy.

A system for face tracking in thermal images based on detection of eye corners was proposed by Strąkowska et al. [37]. In their approach, a modified Otsu's binarization algorithm, combined with morphological operations, is used to localize human faces. Then H-maxima transformation is used to determine regions of interest (ROI), from which final corners of the eyes are determined by geometrical constraints.

On the other hand, Marzec et al. described a system for fast eye localization from thermal images using neural networks [28]. Their proposed two-stage method starts with automatic finding of a face area by searching parts of the image with highest average intensity. Then identification of potential ROIs is done by thresholding. Second stage is a specialized decision block, based on neural network, that finally classifies pre-designated areas.

2.4 Fatigue monitoring

For years computer visions systems have been designed for drivers' fatigue monitoring in order to enhance road security. These heavily rely on detection of faces, various parts of

the faces, as well as behavior patterns. For example, in the system proposed by Cyganek and Gruszczyński [10], images from near infrared camera are fed into the cascade of classifiers. They used specialized iris-pupil model for the initial detection of regions of interest. Potential candidates are sorted by intensity ratio between iris and pupil area. Then a tensor classifier, based on the higher order singular value decomposition of tensors, is trained with the geometrically deformed versions of real eye prototypes. An accuracy of over 96% was reported.

Khan et al. presented system for automatic driver fatigue recognition based on analysis of facial expressions [17]. Proposed algorithm splits image processing into three stages. In the first one, image is transformed by applying a discrete wavelet transform. Simultaneously, original image is downsampled into smaller blocks and classified as either informative or non-informative. Then features with high variance are selected using discrete cosine transform. Final classification of facial expression is done by chain of binary SVM classifiers. Authors reported that accuracy rate of proposed system in expression recognition task outperforms other similar techniques.

Another technique for automatic facial expression recognition was proposed by Munir et al. in [30]. Combination of Fast Fourier Transform and Contrast Limited Adaptive Histogram Equalization (FFT + CLAHE) was used to compensate poor illumination in source images. Classification was done on merged binary pattern code (MBPC) descriptors with dimensionality reduction by means of PCA.

In our previous paper [18], we presented a system for driver's fatigue detection based on yawn detection in thermal images. Although it also has the face detection and localization module, it is fundamentally different from the approach presented in this paper. Namely, we start with image binarization using the Otsu's method to obtain the binary mask of the hottest regions of the image. In next step, pixel summation of the binary mask is done and face boundaries are computed using curve fitting method. Then non-overlapping regions with highest temperature are investigated to find the best candidates for the corners of the eyes.

To the best of our knowledge, presented in this paper method is the first one that attains a reliable eye detection in thermal that uses the virtual high dynamic range technique, to enhance performance of the dense grid of scale-invariant feature descriptors, combined with the bag-of-visual-words approach. Although it relies on sparse features, it has many benefits also when compared to the novel deep neural network based systems, as will be presented.

3 Architecture of the proposed system

In this work we concentrate exclusively on eye detection in thermal video. Our system is envisioned as fast and efficient building block that can be further utilized by drivers' assisting system. An overall architecture of such integration is presented in Fig. 2. It consists of multiple paths of processing data coming from different sensors and operating on various physical effects or radiation spectra. However, compared with the visible and near-infrared (NIR) signals, this longwave-infrared spectrum allows for reliable and least distractive operation in a driver's vicinity [9, 10].

With the advent of machine learning and deep neural networks, computer vision based on RGB images improved substantially, both in academic research, as well as in commercial products. However, challenges like computational burden, a necessity of the large scale annotated training sets as well as transfer of knowledge to images from distant domains and spectra, makes us believe that classic methods of image analysis and classifications

Fig. 2 Overall architecture of the driver drowsiness control system in a car with a highlighted path of eye detection in thermal images presented in this paper

are still vital and important to explore, especially for usage in embedded and automotive systems. It should be clearly mentioned here that for this problem we have also evaluated the deep learning with the convolutional neural networks. In our preliminary works, we tested regression model with the Inception v2 [38] back-end, pretrained on the ImageNet [12], but the results were unsatisfactory. This was caused by the fundamental differences between the RGB and thermal images, combined with a relatively low number of training examples. Moreover, using deep learning techniques required much greater processing power than in the proposed system.

Eye detection module presented in this paper utilizes dense scale-invariant feature transform descriptors to generate sparse representation of thermal image that is later used for localization. Extracted descriptors are compared with features from visual words dictionary to obtain distances. The low confidence areas are removed by morphological erosion and only best features are passed for further processing. Eye candidate regions are formed by clustering selected descriptors and later checked by series of geometric constraints to draw a final result. However, as already mentioned, to ensure proper operation of the SIFT features in thermal spectrum, the input images need to be preprocessed with proposed virtual high dynamic range module (VHDR module). Overall architecture of proposed eye detection system described in this paper is presented in Fig. 3 and utilized methods are presented in detail in further sections.

3.1 Image enhancement with the VHDR

Thermal imaging is characterized by low contrast and low dynamic range. This is a serious impediment which affects subsequent processing and detection modules in image analysis pipeline. On the other hand, recently developed methods of high dynamic range (HDR) image generation, also called Virtual HDR (VHDR) since they are obtained from a single view rather than a series of various exposures, were shown to enhance image details in different conditions [11, 16]. However, this effect was tested exclusively on airy images. Based

Fig. 3 General overview of the proposed method

on our experiments we observed that VHDR has positive effect also in LWIR spectrum. Therefore, in this paper we propose a novel thermal image preprocessing method with the VHDR module. Thanks to this, further image transformations using sparse image descriptors (SIFT descriptor in our case) leads to much better results, as will be discussed. In this section an overview of the VHDR method proposed by Cyganek & Wozniak is presented [11].

In their approach, the input low dynamic range (LDR) image is processed by a number of tone adjustment curves, which provide corresponding number of tone converted images. Their purpose is to nonlinearly enhance different intensity ranges of the input image, for better details exposition, as well as to enhance contrast. Those generated virtual exposures are then fused back to form single VHDR image.

To obtain a series of virtual exposures, three variants of the logistic function are used, that is convex, concave and S-shaped, as shown in Fig. 4. These can be computed based on the following formulas [11]

$$s_1(x, k, x_0) = \frac{P_{max}}{1 + e^{-Dk \frac{(x-x_0)}{R}}} \quad (1)$$

$$s_2(x, k, x_0) = \left(\frac{|x_0 - P_{mid}|}{P_{mid}} + 1 \right) s_1(x, k, x_0) + s_1(0, k, x_0) \quad (2)$$

where x_0 and k are parameters denoting the middle intensity value and scale, respectively, D stands for a domain range constant, and $P_{max} = 255$ denotes a maximal pixel range of the input LDR image, whereas $P_{mid} = 127.5$ denotes a middle value for 8-bit images. The parameter D controls the range of values of s_1 , for which the logistic function is far from saturation, i.e. it is effectively different from 0 and 1. This parameter was set to 2 times 6 around x_0 , which makes $D = 12$.

Selecting different values of x_0 and k in (2) allows for generation of different versions of the sigmoid function. This way, the concave, convex and S like shape, are obtained. Each of them serves enhancement of different aspects (features) of the input image. This way generated virtual expositions need to be combined together into one output VHDR image. For this purpose each pixel i_h of the output image is obtained as a weighted sum of all V pixels i_v , as follows

$$i_h = \frac{\sum_{v=1}^V w_v i_v}{\sum_{v=1}^V w_v} \quad (3)$$

$$w_v = e^{-\lambda \left(\frac{i_v - P_{mid}}{P_{mid}} \right)^2} \quad (4)$$

Fig. 4 The sigmoid function for virtual exposures generation with $x_0 = 127.5$ and different values of k . Axis: the input brightness vs. the output brightness

where V stands for a number of the generated virtual expositions and λ is a parameter controlling slope of the weighting function (set between 2.5–3.5). In our experiments a total of six different exposures was used which were generated from the following settings: $x_0 = 0$, $k \in \{0.3, 0.2\}$, and $x_0 = 127.5$, $k \in \{0.4, 0.5\}$, as well as $x_0 = 255$, $k \in \{0.4, 0.3\}$.

3.2 Computation of the sparse image descriptors

Scale-Invariant Feature Transform proposed by Lowe in [24] is one of the most well-known algorithms for detection and descriptor generation of local features in images. Due to its high invariance to uniform scaling, orientation and illumination changes it is used in many applications including image-stitching, video tracking, match moving and, as in case of our method, object detection and recognition. It consists of two separate blocks, namely keypoint detector and keypoint descriptor, that can be used independently. Following similar methodology to [14], we decided to skip key point detection part of the original method. Instead, SIFT descriptors computed on a dense grid are used to create sparse representation of the entire image.

SIFT descriptor is a vector that represents orientation histograms computed in neighborhood of a key-point. It is created by dividing window of a fixed size (often 16 by 16 pixels) into 16 sub-blocks. Then orientation histogram with 8 bins is computed for every sub-block. In total, SIFT descriptor has of 128 values ranging from 0 to 255. Further details on its computation can be found in the paper by D. Lowe [24], as well as on the VLFeat website¹, which software was also used in our experiments.

3.3 Detection with dense grid of features

3.3.1 Dictionary of visual words

Proposed method requires construction of a dictionary of visual words. Dense grid of feature descriptors is extracted from images of human eyes taken by long-range infrared spectrum camera. Samples from this database can be seen in Fig. 5. Both the sampling distance d_{desc} and descriptor sub-block size s_{block} have been chosen experimentally. Best results were obtained with descriptors sized to cover about 90% of the region of interest. This is due to the fact that smaller descriptors are very generic and not distinctive enough to differentiate between them. On the other hand, positional accuracy is harmed when descriptors are too big. Computed feature descriptors could be directly used to form a dictionary of features. However, having large number of examples could lead to over-fitting as well as directly increase computational complexity. Hence, a codebook reduction method has to be applied. For this purpose, our method utilizes the *k-means* clustering algorithm to reduce the number of prototype feature descriptors. The *k-means* clustering starts by generating k clusters with arbitrarily chosen center points (depending on a particular implementation) and assigning each sample vector into cluster of the nearest mean point. Then centroids of every cluster are recomputed and assignments are recomputed. This process repeats until the stop conditions are met [7, 9].

Final dictionary of visual words is constructed using the clusters' centroids. It is important to note, that varying dictionary size not only influences detection performance, but it is

¹<http://www.vlfeat.org>

Fig. 5 Exemplary eye images from the thermal database, not processed (left), and VHDR processed (right) which show more details

a primary factor determining computational performance. Thus, its value should be chosen empirically, as will be shown later.

3.3.2 Distance map

In our approach, eyes are localization based on the sparse image representation. Dense grid of feature descriptors is computed for every image using the same grid dimensions as for the dictionary. We can further optimize this process by leveraging the thermal characteristics of the image. In our method we use Otsu's thresholding method [31] to mask-out unnecessary background and compute feature descriptors only in relevant parts of the image, reducing computational burden on all stages of the processing pipeline as well as increasing robustness of our system at the same time.

Then, for every feature in a grid, we compute its Euclidean distance from the nearest word in the visual words dictionary. This way the distance map D_f is obtained, as follows:

$$D_f(i, j) = \min [\|f(i, j) - x_0\|, \|f(i, j) - x_1\|, \dots, \|f(i, j) - x_{k-1}\|] \quad (5)$$

where $f(i, j)$ is feature descriptor obtained from grid position (i, j) , $x_0, x_1, \dots, x_{k-1}$ represents words in dictionary and k denotes the size of the dictionary.

Then, to refine the distance map, we use the morphological erosion with the square 3x3 structural element (Fig. 6). This step lowers noise level in the distance map, thus local minimas are easier to find. It is worth noting, that we evaluated different pooling methods for this purpose, like averaging or lowest value point neighbourhood, but results were unsatisfactory.

3.3.3 Distance map thresholding

As further processing steps require sorting of the distance map, we can further optimize this process and reduce false candidates by considering only values under some specific threshold. We can obtain this value by computing distances between training samples and final

Fig. 6 Visualization of distance map (darker colors mean lower distance, such points are potential regions of interest) before (images 6a and 6c) and after morphological erosion (images 6b and 6d)

visual words dictionary. In our experiments, we tested two different thresholding methods. Namely, we evaluated two types of outlier detection metrics, average intensity plus two standard deviations described in [29] (referred to as a *Sigma* threshold), as well as the Median of Absolute Differences, presented in [23], that is claimed to be less prone to outliers in training data (referred to as a *MAD* threshold).

However, based on our observation, although distance map thresholding can be beneficial from optimization perspective it doesn't have measurable effect on detection performance. This is because initial image thresholding and morphological filtering greatly reduces impact of false regions of interest.

3.3.4 Clustering and geometric constrains

Pixels from the distance map (now associated with their position (x_i, y_i) and henceforth referred to as points) are then sorted by their intensities in ascending order. Lower values represent better candidates. However, in order to further reduce sorting and clustering time we only store L_f points. In other words, only points with lowest values will be passed to further steps. This value was selected empirically, as it depends on the descriptor size and image resolution.

Further on, the method called Salient Point Clustering, presented in [9, 375] and implemented as shown in Algorithm 1, is used to group points into clusters. The clustering method starts with arbitrarily selecting one point within set of unassigned points to create first cluster. Then, Euclidian distances are calculated between that one and all other points. Every point that lies within arbitrarily chosen maximum distance d_c , is added to the cluster and the process is repeated.

This clustering method is very simple and efficient, but could lead to suboptimal results. However, in our case, its robustness is improved by aforementioned thresholding and further processing in form of geometric constrains. To form input for geometric verification and to locate the clusters' centers in the image, shape of each cluster is approximated by an ellipse.

For each cluster, we start by assigning a weight w_i for every point in a cluster based on its intensity. To simplify notation, we denote $c(x_i, y_i)$ as i th point in a cluster c .

$$w(i) = c_{max} - c(x_i, y_i) \quad (6)$$

where c_{max} represents maximal intensity in this cluster. Therefore, we assign bigger weights to points with smaller intensity, as they constitute better matches with the visual words in the dictionary. Next, the location of center point, denoted as c_p , is computed using the weighted average of all points:

$$W = \sum_i^N w_i \quad (7)$$

$$c_p = \frac{1}{W} \sum c(x_i, y_i) \cdot w_i = \left(\frac{1}{W} \sum_i^N c^{x_i} \cdot w_i, \frac{1}{W} \sum_i^N c^{y_i} \cdot w_i \right) \quad (8)$$

where W is the sum of all weights, N is a number of points in cluster c and x_i and y_i denote the position of an i th point. Because we expect regions of interest to form elliptical shapes in the distance map, we use the inertia tensor method described in [9, 62–65] to approximate every cluster by an ellipse. We first define function m_{ab} , called the *moment of inertia* in a direction ab .

$$m_{ab} = \sum_{i=1}^N \sum_{j=1}^N (x_i - p_c^x)^a (y_i - p_c^y)^b \quad (9)$$

Then we can compute the central point based on three moments:

$$x_c = \frac{m_{10}}{m_{00}}, y_c = \frac{m_{01}}{m_{00}}, (for \text{ all } m_{00} \neq 0) \quad (10)$$

Based on the above, the inertia tensor can be expressed in terms of the central moments, as follows

$$T = \begin{bmatrix} m_{20} & -m_{11} \\ -m_{11} & m_{02} \end{bmatrix} \quad (11)$$

Lastly, we use the inertia tensor to calculate the length L and width K of the elliptical blob, as follows:

$$L = \sqrt{\lambda_1} \quad (12)$$

$$K = \sqrt{\lambda_2}, \lambda_1 \geq \lambda_2 \quad (13)$$

where λ_1 and λ_2 are eigenvalues of the inertia tensor T .

Clusters are then sorted in ascending order using the average intensity of all points within it, effectively promoting clusters that have descriptors with smallest distance from the dictionary.

Algorithm 1 Salient point clustering.

Input: Set of unclustered points p_{in} . Maximum distance d_c .

Output: Set of points grouped into clusters.

```

1: while  $p_{in}$  has unassigned points do
2:   get first unassigned point from  $p_{in}$  as  $p_0$ 
3:   assign  $p_0$  as a first point of a new cluster  $c_i$ 
4:   for all unassigned points in  $p_{in}$  do
5:     if distance between  $p_0$  and selected point  $p < d_c$  then
6:       mark  $p$  as assigned to cluster  $c_i$ 
  
```

Last processing stage consists of verification of the geometrical constraints. Due to a direct localization scheme, previous stages have not made any assumptions about clusters' absolute or relative position. Differently from methods that rely on visibility of both eyes, this feature allows for detection of a single eye region. However, we can still leverage geometrical relations when both eyes are visible in the frame. In our method, we match best clusters into pairs based on their shape, size and relative position. This operates in accordance with Algorithm 2.

Algorithm 2 Geometrical constraints verification.

Input: Set of clusters. Constants g_{ecc} , g_{min} , g_{max} , e_{min} , e_{max} , θ_{max} .

Output: Detected eye regions

```

1: for all clusters  $c$  do
2:   if  $c$  eccentricity  $\geq g_{ecc}$  then
3:     continue
4:   if  $c$  width  $\geq g_{max}$  or length  $\geq g_{max}$  then
5:     continue
6:   if  $c$  width  $\leq g_{min}$  or length  $\leq g_{min}$  then
7:     continue
8:   add  $c$  to correct clusters list
9: if correct clusters length == 0 then
10:  return empty list
11: else if correct clusters length == 1 then
12:  return one proposed region
13: else
14:  get first two clusters from correct clusters list as  $c_a$  and  $c_b$ 
15:  set  $p_a$ ,  $p_b$  to be the center point of respective cluster
16:  if  $\text{atan2}(p_a, p_b) > \theta_{max}$  deg then
17:    return empty list
18:   $D$  denotes the Euclidian distance between points
19:  if  $D(p_a, p_b) < e_{min}$  or  $D(p_a, p_b) > e_{max}$  then
20:    return empty list
21:  return clusters  $c_a$  and  $c_b$  as proposed regions
  
```

Fig. 7 Examples of correct detections. Predicted eyes positions are marked with green crosshair

Figure 7 presents sample detections obtained from our method. Examples of single eye and incorrect detections are presented in Fig. 8.

4 Experimental results

Presented system was implemented in Python, leveraging using the well-known libraries like numpy, scipy and scikit-learn. As mentioned before, the VLfeat library was used for computation of the SIFT descriptors. Experiments presented in this section were run on a laptop computer equipped with 16GB of RAM, and 4-core processor i7-6700HQ with the 2.6GHz clock, operating under the 64-bit version of Windows 10[®]. Processing time of around 105 ms to 190 ms per frame was achieved, depending on the size of the dictionary and the source image. However, we foresee that further rewriting to C++ can greatly improve the performance. Additionally, many stages of the proposed algorithms can be computed in parallel. Thus, future implementation with help of the OpenMP[®] or parallel computing platforms like CUDA[®] or OpenCL[®], can bring even higher speedup.

Fig. 8 Examples of single detections (left) and incorrect detections (right). Predicted eyes positions are marked with green crosshair

For the quantitative evaluation, the two thermal images datasets have been used. The first one, henceforth referred to as YawnDB) is our own thermal database.² It contains over 3500 images of human faces obtained from four subjects doing normal activities like talking, looking around, driving a car and yawning. Database was created using the FLIR® A35 camera. Images were manually labeled to provide ground-truth frames which are used for quantitative evaluation. The second used is the UL-FMTV database with thermal images [15]. It was created with the high resolution thermal camera Phoenix Indigo IR. From this database we extracted faces of 128 individuals and annotated each image with eyes positions by hand.

The experiments were conducted to answer the following questions:

- How many sample images are necessary to achieve a reliable detection?
- What is the accuracy of the proposed eye detector?
- Can the proposed image pre-processing improve the detection robustness?
- How our method compares with the deep learning object detectors?

²<http://home.agh.edu.pl/~mknapiak/>

Table 1 Parameters used in our experiments

Parameter	Value	Used in	Comment
d_{desc}	3 pixels	Dense grid sampling	Distance in vertical and horizontal direction between feature descriptors.
w_h, w_w	3 by 3	Morphological erosion	Size of structural element.
d_c	10 pixels	Algorithm 1	Maximal distance in pixels for neighboring points.
g_{ecc}	0.85	Algorithm 2	Maximal eccentricity of the cluster.
g_{min}, g_{max}	2×2 pixels, 50×50 pixels	Algorithm 2	Cluster length and width constraints.
e_{min}, e_{max}	0.8, 3	Algorithm 2	Ratio between distance between two clusters and their average diameter.
θ_{max}	60 deg	Algorithm 2	Maximal angle between two clusters.

- Can our method be used as a source of annotated data for semi-supervised deep learning?

Each test was run 10 times and average results are presented. For each test round the dataset was randomly split in half to form the training and validation subsets. The quantitative results were measured in terms of the true-positive (TP), false-positive (FP) and false-negative (FN) parameters. From these, the so called F value was computed. True-positive is reported for each correctly recognized eye position, that is, if a region found by an algorithm is within 10 pixels from the ground truth, which is equivalent to 3% of the input image resolution (in case of resolution 320×256). On the other hand, for every incorrectly detected eye, a false-positive counter is incremented. If algorithm fails to find an eye, then it is accounted as a false-negative. Parameters used in the conducted experiments are presented in Table 1. Full results of the experiments are presented in Table 2. Examples of correct and incorrect detections are shown in Figs. 7 and 8.

We first started with UL-FMTV dataset. This database was made in controlled environment and with high resolution thermal camera, so it is a good baseline for other experiments. For this purpose, we split the images into two equal groups, training data and test data, each with pictures of 62 individuals (one image per person, different people in each group).

In the first step, we performed tests using low dynamic range images with dictionaries ranging from 10 to 350 words, essentially doubling dictionary size with every test. Detector achieved average accuracy of 0.888 (F-Value). As can be seen, there is no significant improvement when dictionary size is increased, (see Fig. 9). This is caused by inherently low contrast of thermal images. Because of this, discriminative properties of descriptors are low and introduction of additional words to the dictionary leads to decreased performance.

We then input the same images to our algorithm, but pre-processed with the VHDR method, described in Section 3.1. As can be expected, an increased contrast and enhanced details in the VHDR images contributed to improvement in detection performance. On average, accuracy was higher by 1 percentage point as compared to the non-processed thermal images, averaging at 0.901 (F-Value). Moreover, if we look at a VHDR image, it is visible that increased contrast reveals finer details that allow for better usage of the SIFT descriptors.

In the era of Convolutional Neural Networks it is important to ask if methods rooted in classical computer vision are still relevant. We think that these are still worth researching,

Table 2 F1 value measured on different models. Best result for each test is underlined. Overall best result in bold

Model	Training samples	Dictionary size	TP	FP	FN	Precision	Recall	F1
Ours normal	62	10	99	4	21	0,961	0,825	0,888
		20	102	5	17	0,953	0,857	,903
		50	100	6	18	0,943	0,847	0,893
		100	99	5	20	0,952	0,832	0,888
		150	97	5	22	0,951	0,815	0,878
		300	97	5	22	0,951	0,815	0,878
	31	10	95	6	23	0,941	0,805	0,868
		20	101	5	18	0,953	0,849	,898
		50	100	5	19	0,952	0,840	0,893
		100	87	9	28	0,906	0,757	0,825
		150	85	10	29	0,895	0,746	0,813
		300	83	10	31	0,892	0,728	0,802
	16	10	98	6	20	0,942	0,831	,883
		20	97	5	22	0,951	0,815	0,878
		50	86	10	28	0,896	0,754	0,819
		100	87	14	23	0,861	0,791	0,825
		150	88	13	23	0,871	0,793	0,830
		300	90	12	22	0,882	0,804	0,841
Ours VHDR	62	10	102	6	16	0,944	0,864	0,903
		20	105	5	14	0,955	0,882	,917
		50	102	7	15	0,936	0,872	0,903
		100	102	5	17	0,953	0,857	0,903
		150	100	6	18	0,943	0,847	0,893
		300	99	6	19	0,943	0,839	0,888
	31	10	99	9	16	0,917	0,861	0,888
		20	104	4	16	0,963	0,867	,912
		50	102	5	17	0,953	0,857	0,903
		100	100	5	19	0,952	0,840	0,893
		150	103	5	16	0,954	0,866	0,907
		300	104	4	16	0,963	0,867	0,912
	16	10	99	5	20	0,952	0,832	0,888
		20	92	9	23	0,911	0,800	0,852
		50	94	10	20	0,904	0,825	0,862
		100	96	8	20	0,923	0,828	0,873
		150	102	7	15	0,936	0,872	,903
		300	96	9	19	0,914	0,835	0,873
YOLOv3 Normal	62	n/a	108	5	16	0,956	0,871	,911
	31		38	0	86	1,000	0,306	0,469
	16		39	0	85	1,000	0,315	0,479
YOLOv3 VHDR	62		121	5	3	0,960	0,976	0,968
	31		89	1	35	0,989	0,718	0,832
	16		108	5	16	0,956	0,871	0,911

Fig. 9 Detection accuracy on UL-FMTV with respect to the dictionary size

especially for modalities other than the visual spectrum images, where we have potential problems like knowledge transfer to distant domains, very small number of training samples or runtime platforms with limited computational power. For comparison we selected YOLOv3 which is one of the best known single shot general object detectors based on deep learning [33]. It is implemented in Keras framework with TensorFlow backend [32] and pretrained on ImageNet [12]. We wanted to compare its performance on our dataset, both without and with VHDR preprocessing, as well as compare its speed when run in the CPU-only mode, i.e. conditions common in embedded systems. We started with the same images as in the baseline and trained every model on 62 examples, followed in the next run with 31 and, lastly, with 16 training samples. Then evaluation was done on 62 unique images for each training run. Results are presented in Fig. 10. Our method achieves high accuracy even

Fig. 10 Detection accuracy on UL-FMTV with respect to the number of training samples

with very low number of training samples. On the other hand, deep learning model needs at least 62 examples to match method proposed in this paper. Moreover it is worth noting that pre-processing the training images with VHDR greatly improves performance of the YOLOv3-based detector, increasing F1 score by 5 percentage points, thus proving, that this technique is somehow general and useful for object detection in thermal images also with the deep neural networks.

Because of insufficient size of the thermal training data, we envision the semi-supervised training of deep learning models as one more application for our method. Namely, in the scenarios where dataset is not big enough for convolutional neural networks, simpler models, like the one proposed in this article, could be used to automatic data annotation, that is later used for deep neural networks training. This idea is left for further research.

Regarding the detection speed, on our testing platform we achieved 1.4 FPS with YOLOv3 (when run solely on CPU to simulate platforms without discrete GPU, like embedded systems). In comparison, our model achieved 4 FPS on the same data. It is worth noticing that our implementation does not take advantage of the possible optimizations - for example computation of the feature descriptors can be constrained to only masked regions, so higher speeds can be achieved. Method optimization is also left for future research.

In the case of the YawnDB dataset we skipped tests with the deep learning models due to a low number of unique individuals in that dataset. It was done to not distort results by overfitting during training. However, performance of the proposed model was very similar to that of UL-FMTV, therefore we decided not to include it in favor of simplifying figures and tables.

Finally, we would like to report that for data classification also the k-NN and the SVM methods were tested, but the proposed bag-of-words approach gave the best results and therefore only this approach is discussed in our paper.

5 Conclusion

In this paper a method is proposed for efficient eye localization in thermal images for the purpose of driver drowsiness monitoring. The main novelty is the proposed preprocessing step which converts low dynamic range thermal image into its high dynamic range version. As shown and experimentally verified, it allows for much higher detection accuracy due to the significant detail enhancement. Such pre-conditioned signal allows for efficient usage of the SIFT descriptors sampled on a dense grid. These are then used to create bag-of-words, that are employed for eye regions localization. As shown, the proposed system achieves high accuracy and fast response in real conditions. The method has been compared with the YOLOv3 deep neural network based object detector. We showed that also in this case VHDR preprocessing increases detection accuracy for the available low exemplar thermal training datasets. Nevertheless, it can be expected that substantially larger number of training data can further improve performance of deep detectors. In this context it will be also interesting to evaluate usage of the proposed model as an automatic annotation tool for the deep learning based general object detectors. This is left for future research, however.

Our method can be easily reimplemented in C++ to allow for full real-time operation. Being based on fast and computationally light SIFT descriptors it can be used in low-power or embedded systems, where implementing deep neural networks is frequently difficult or infeasible due to computational complexity or costs. Last but not least, although the method was designed as a front-end of drivers' monitoring systems operating with thermal images, it can be easily incorporated in other domains, such as medicine, surveillance or gaming.

In future work, we plan to further investigate potential cooperation of the method presented in this paper with modern deep neural detectors. As already mentioned, this method can provide initial training exemplars for the network training or in the inference mode it can operate as one of the member classifiers in an ensemble. It would be also interesting to evaluate performance of other feature descriptors, e.g co-occurrence texture and spatial structure from Scoot facial sketch metric presented in [13], that could serve as a replacement of the SIFT descriptor. Other research direction would be to use the hierarchical dictionaries proposed by Fulkerson in [14] to further improve robustness and detection speed.

Acknowledgments This work was supported by the Polish National Science Center NCN under the grant no. 2014/15/B/ST6/00609.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abualigah L, Shehab M, Alshinwan M, Alabool H (2019) Salp swarm algorithm: a comprehensive survey. *Neural Comput Applic*, pp 1–21
2. Abualigah LM, Khader AT, Hanandeh ES (2018) A hybrid strategy for krill herd algorithm with harmony search algorithm to improve the data clustering. *Intel Decision Technol* 12(1):3–14
3. Alam F, Mehmood R, Katib I, Altowaijri SM, Albeshri A (2019) Taawun: a decision fusion and feature specific road detection approach for connected autonomous vehicles. *Mobile Networks and Applications*, pp 1–17
4. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields 2017. In: *IEEE conference on computer vision and pattern recognition (CVPR)*. <https://doi.org/10.1109/cvpr.2017.143>
5. Chellappa R, Turaga P (2014) *Feature Selection*. Springer, Boston, pp 291–294. https://doi.org/10.1007/978-0-387-31439-6_299
6. Chu J, Guo Z, Leng L (2018) Object detection based on multi-layer convolution feature fusion and online hard example mining. *IEEE Access* 6:19959–19967
7. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: *Workshop on statistical learning in computer vision, ECCV*, vol 1, Prague, pp 1–2
8. Cyganek B (2007) Road signs recognition by the scale-space template matching in the log-polar domain. In: Martí J, Benedí JM, Mendonça AM, Serrat J. (eds) *Pattern recognition and image analysis*. Springer, Berlin, pp 330–337
9. Cyganek B (2013) *Object detection and recognition in digital images: theory and practice*. Wiley
10. Cyganek B, Gruszczyński S (2014) Hybrid computer vision system for drivers' eye recognition and fatigue monitoring. *Neurocomputing* 126:78–94
11. Cyganek B, Woźniak M (2018) Virtual high dynamic range imaging for underwater drone navigation. In: *Proceedings of the 6th IIAE international conference on industrial application engineering 2018*
12. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: *CVPR 09*
13. Fan DP, Zhang S, Wu YH, Liu Y, Cheng MM, Ren B, Rosin PL, Ji R (2019) Scoot: A perceptual metric for facial sketches
14. Fulkerson B, Vedaldi A, Soatto S (2008) Localizing objects with smart dictionaries. In: Forsyth D, Torr P, Zisserman A (eds) *Computer vision – ECCV 2008*. Springer, Berlin, pp 179–192
15. Ghiass R (2014) *Face recognition using infrared vision*. Ph.D thesis. <https://doi.org/10.13140/2.1.4447.3122>
16. Ke W, Wang T, Chiu C (2009) Hardware efficient virtual high dynamic range reproduction. *16th IEEE International Conference on Image Processing (ICIP'09)* pp 2665–2668

17. Khan SA, Hussain S, Xiaoming S, Yang S (2018) An effective framework for driver fatigue recognition based on intelligent facial expressions analysis. *IEEE Access* 6:67459–67468
18. Knapik M, Cyganek B (2019) Driver's fatigue recognition based on yawn detection in thermal images. *Neurocomputing* 338:274–292
19. Kwaśniewska A, Rumiński J, Rad P (2017) Deep features class activation map for thermal face detection and tracking. In: 2017 10th international conference on human system interactions (HSI), pp 41–47. <https://doi.org/10.1109/HSI.2017.8004993>
20. Lazebnik S, Schmid C, Ponce J (2006)
21. Leng L, Li M, Kim C, Bi X (2017) Dual-source discrimination power analysis for multi-instance contactless palmpoint recognition. *Multimedia Tools and Applications* 76(1):333–354
22. Leng L, Zhang J, Khan MK, Chen X, Alghathbar K (2011) Dynamic weighted discrimination power analysis : A novel approach for face and palmpoint recognition in dct domain
23. Leys C, Ley C, Klein O, Bernard P, Licata L (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* 49(4):764–766
24. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 60(2):91–110
25. Leng L, Zhang J, Jing X, Khan MK, Alghathbar K (2010) Dynamic weighted discrimination power analysis in dct domain for face and palmpoint recognition. In: 2010 International conference on information and communication technology convergence (ICTC), pp 467–471
26. Mahmood A, Khan SA, Hussain S, Almaghayreh EM (2019) An adaptive image contrast enhancement technique for low-contrast images. *IEEE Access* 7:161584–161593
27. Malpani SS, Narasimhadhan AC, Thermal AV (2016) Vision human classification and localization using bag of visual word. In: 2016 IEEE Region 10 conference (TENCON), pp 3135–3139
28. Marzec M, Lamża A, Wróbel Z, Dziech A (2016) Fast eye localization from thermal images using neural networks. *Multimedia Tools and Applications*
29. Miller J (1991) Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology Section A* 43(4):907–912
30. Munir A, Hussain A, Khan SA, Nadeem M, Arshid S (2018) Illumination invariant facial expression recognition using selected merged binary patterns for real world images. *Optik* 158:1016–1025. <https://doi.org/10.1016/j.ijleo.2018.01.003>, <http://www.sciencedirect.com/science/article/pii/S0030402618300032>
31. Otsu N (1979) A threshold selection method from gray level histograms. *IEEE Trans Systems, Man and Cybernetics* 9:62–66. Minimize inter class variance
32. qqwweee (2018) Keras yolov3. <https://github.com/qqwweee/keras-yolo3>
33. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement
34. Shehab M, Abualigah L, Al Hamad H, Alabool H, Alshinwan M, Khasawneh AM (2019) Moth-flame optimization algorithm: variants and applications. *Neural Comput Applic*, pp 1–26
35. Shehab M, Alshawabkah H, Abualigah L, Nagham AM (2020) Enhanced a hybrid moth-flame optimization algorithm using new selection schemes. *Engineering with Computers*, pp 1–26
36. Sonkusare S, Ahmedt-Aristizabal D, Aburn MJ, Nguyen VT, Pang T, Frydman S, Denman S, Fookes C, Breakspear M, Guo CC (2019) Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. *Scientific Reports* 9(1):4729
37. Strąkowska M, Strąkowski R (2015) Automatic eye corners detection and tracking algorithm in sequence of thermal medical images. *Measure Autom Monitor* 61(6):199–202
38. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision 2016. In: IEEE conference on computer vision and pattern recognition (CVPR). <https://doi.org/10.1109/cvpr.2016.308>
39. Wang S, Liu Z, Shen P, Ji Q (2013) Eye localization from thermal infrared images. *Pattern Recogn* 46(10):2613–2621
40. Wang S, Tang J, Liu H (2017) Feature Selection. Springer, Boston, pp 503–511. https://doi.org/10.1007/978-1-4899-7687-1_101
41. Zeng J, Liu M, Fu X, Gu R, Leng L (2019) Curvature bag of words model for shape recognition. *IEEE Access* 7:57163–57171
42. Zhang L, Mistry K, Neoh SC, Lim CP (2016) Intelligent facial emotion recognition using moth-firefly optimization. *Knowledge-Based Systems* 111:248–267. <https://doi.org/10.1016/j.knosys.2016.08.018>, <http://www.sciencedirect.com/science/article/pii/S0950705116302799>
43. Zhang Y, Chu J, Leng L, Miao J (2020) Mask-refined r-cnn: a network for refining object details in instance segmentation. *Sensors* 20(4):1010