A Thermal Infrared Face Database With Facial Landmarks and Emotion Labels

Marcin Kopaczka, Raphael Kolk, Justus Schock, Felix Burkhard, and Dorit Merhof

***Abstract*—Thermal infrared imaging is an emerging modality that has gained increasing interest in recent years, mostly due to technical advances resulting in the availability of affordable microbolometer-based IR imaging sensors. However, while sensors are widely available, algorithms for thermal image processing still lack robustness and accuracy when compared to their RGB counterparts. Current methods developed for RGB data make use of machine learning algorithms that require large amounts of labeled images which are currently not available for the thermal domain. In our work, we address the question whether providing a large number of labeled images would allow the application of current image processing methods on the example of solving challenging face analysis tasks. We introduce a high-resolution thermal facial image database with extensive manual annotations and explore how it can be used to adapt methods from the visual domain for infrared images. Additionally, we extend existing approaches for infrared landmark detection with a head pose estimation for improved robustness and analyze the performance of a deep learning method on this task. An evaluation of algorithm performance shows that learning algorithms either outperform available solutions or allow completely new applications that could previously not be addressed. As a conclusion, we prove that investing the effort into acquiring appropriate training data and adapting competitive algorithms is not only a viable approach in analyzing thermal infrared images but can also allow outperforming specific task-designed solutions. The database is freely available for academic use at https://github.com/marcinkopaczka/thermalfaceproject.**

***Index Terms*—thermal infrared, machine learning, face detection, face tracking, facial expression recognition**

I. Introduction

One of the key research areas in computer vision addressed by a vast number of publications is the processing and understanding of images containing human faces. The most often addressed tasks include face detection, facial landmark localization, face recognition and facial expression analysis. Other, more specialized tasks such as affective computing, the extraction of vital signs from videos or analysis of social interaction usually include one or several of the aforementioned tasks in their processing pipeline.

Currently, most face processing is performed either in regular 2D recordings (RGB videos) or with methods that take advantage of an additional depth channel (RGB + D) as provided by devices such as Microsoft's Kinect camera to gain 3D information. For many tasks involving human bodies and faces, RGB + D has replaced complex 3D imaging techniques such as stereographic cameras or marker-based approaches due to its wide availability and ease of use [1] [2]. Most algorithms presented for facial image processing focus therefore on RGB or RGB+D applications. While the capabilities of these imaging techniques are well understood, there is a number of other approaches for image acquisition that often come with unique advantages. One of these methods is thermal or long wave infrared (LWIR) imaging, an emerging modality that has gained growing attention over the last years. It has several benefits compared to regular imaging technologies operating in the visual spectrum: Since LWIR sensors rely on the heat radiation emitted by the objects themselves, they do not require natural or artificial light sources and are therefore invariant to illumination changes, remaining fully operational even in complete darkness. Furthermore, vital signs such as respiratory rate [3] or heart rate (HR) can be extracted from thermal recordings of humans, and recent studies show that physiopsychological and other clinically relevant effects are visible in the IR domain as well [4] [5].

Despite these advantages, two factors have been limiting the widespread use of thermal infrared imaging for commercial and medical use: equipment price and algorithm performance. LWIR cameras operate in wavelengths much longer than the visual spectrum (7 - 14 μ m vs. 380-700 nm) and therefore require special optics and sensor materials. This downside has been addressed by recent advances in sensor technology, most importantly with the introduction of affordable microbolometer array sensors which have already become the most widely used thermal IR detector type [6]. The second shortcoming is the difference in image appearance caused by the different physical phenomena behind visual and thermal imaging: While regular imaging is based on the reflectance and translucence of objects for electromagnetic waves in the visual spectrum, thermal imaging records heat radiation emitted by the objects themselves. These two are not correlated, it is generally not possible to compute the thermal appearance of an object from its color and vice versa. Since regular RGB or monochrome cameras are much more widely used, most algorithms presented are developed for the visual domain and usually cannot be transferred directly to the thermal domain. For example, most texture and color information is completely lost in the thermal spectrum. This is the reason why a number of specialized algorithms for facial image processing has been introduced in the thermal domain. Most of these algorithms take advantage of the fact that foreground-background contrast in thermal images between a human subject and the scenery is usually very high, therefore computationally simple algorithms for thresholding can be used with greater success than in regular photographs [7] [8] [9]. However, these algorithms often have strong inherent requirements towards the subject's

Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany

pose, for example requiring a strict frontal view or a certain body pose to work correctly. These requirements strongly limit the robustness and versatility of such algorithmic approaches. Other methods such as [10] rely on co-acquisition of visible and thermal data with jointly calibrated cameras for the two domains. Landmark detection in the visible domain can be performed more robustly than in infrared data, therefore landmarks are localized in the visual image and their coordinates are subsequently mapped onto the infrared image. While this approach allows to skip the detection in thermal data, it requires a set of of calibrated cameras for both domains.

In the visible domain, this issue has been addressed by moving from algorithmic approaches with fixed rule sets to data-driven machine learning methods. These approaches use a learning algorithm paired with appropriate data to accomplish complex tasks, therefore their robustness is not only determined by the underlying learning algorithm, but to a large extent by the available training data. The successful application of learning-based methods to a large number of challenging image processing tasks in recent years shows that current algorithms are powerful enough to tackle complex problems when provided with sufficient training data. The problem of requiring large amounts of labeled images is even more severe when modern deep learning algorithms need to be trained. Current deep learning architectures clearly outperform both algorithmic approaches as well as established machine learning methods in various image processing tasks, however they require much larger image databases for training. The sizes of notable databases for object identification and/or segmentation such as the PASCAL VOC dataset [11] with about 10.000 manually labeled images, the Microsoft COCO database with 330.000 [12] images and most notably the ImageNet database [13] with 3.2 millions of images reflect the numbers required for training a deep neural net for complex image processing tasks. For face-related tasks such as face detection and facial landmark tracking, current databases in the visual domain are the HELEN database [14] with 2330 images or the LFPW dataset [15] with 1432 images. These databases are designed for the task of face analysis and therefore significantly smaller than general-purpose databases such as ImageNet or COCO, however they additionally provide full manual annotations for a number of facial landmarks.

In our work, we analyze the hypothesis that a large number of tasks for facial image processing in thermal infrared images that are currently solved using specialized rule-based methods or not solved at all can be addressed with modern learning-based approaches when high-quality data is provided. To this end, we design and introduce a fully annotated thermal face database that is inspired by current state-of-the-art databases in the visual domain. To the best of our knowledge, our database is the first dataset to include the same high-quality manual annotations that are used in visual face image databases and can at the same time offer a comparable number of images. We evaluate our hypothesis by using our database for training of a number of machine learning algorithms for different facial image processing tasks. The versatility of our approach is evaluated by using a number of different algorithms that require large amounts of labeled data. In previous publications, we already used earlier versions of the database to perform face detection [16] and facial landmark localization [17]. In this work, which is an extended version of [18], we present the final database in detail and demonstrate its application for facial expression recognition. As a novelty and extension of all previous work, we also re-visit the task of facial landmark detection as this is a key application of our database since it is the first to provide full landmark annotations for thermal images. We introduce a head pose estimation method to increase landmark detection performance on non-frontal faces. Additionally, we evaluate how a recently presented facial landmark detection using deep convolutional networks can be applied to thermal data by using our database. Using facial landmark detection for precise and robust localization of defined points in the image is a highly relevant method to improve the versatility of image analysis algorithms for thermal video data. Using detected landmarks for the localization of relevant facial areas allows applying algorithms that were initially designed and evaluated on non-moving subjects to persons that move their heads without additional constraints. Methods such as metabolism changes that were presented in [19] or Eulerian analysis for heart rate detection in IR data [20] that require non-moving subjects can be fed with data that has been stabilized and frontalized with a landmark detection, therefore requiring no changes in the original algorithm. We have successfully demonstrated in [21] how the robustness of an algorithm for breathing anomaly detection can be improved when using facial landmark localization and believe that further analysis algorithms can also benefit from precise landmark detection.

Fig. 1: Sample images from the database with the 68-point landmarks shown as overlay.

The paper is structured as follows: In section II we will give a detailed description of our database. Our methods for head pose estimation and neural network-based facial landmark detection are presented in Sections III and IV respectively. In V, we show how the database can be used to train algorithms for facial expression analysis. The presented methods are evaluated in Sec. VI. A discussion of all results and can be found in section VII, followed by a conclusion in section VIII.

II. Thermal Face Database

In this section, we will give an overview of existing databases and give a detailed description of our own contribution.

A. Existing Thermal Face Databases

Several thermal face databases with emphasis on different aspects exist. The most established are the IRIS database [22] with 4228 co-acquired thermal and visual images and the USTC-NVIE database [23] with videos and 236 frames that are manually annotated for facial expression recognition. The spatial resolution of both databases is 320 x 240 pixels. Both databases are freely available online. A third formerly widely used database with of similar size and quality, the Equinox database, is no longer available. Several other databases with higher resolution (up to 640 x 480) have been described in more recent literature [24] and very recently [25], however none of them is publicly available. We downloaded the two freely available datasets and reviewed literature that presented other databases in order to analyze the differences to our database:

- Our database offers the highest spatial resolution of 1024 x 768 pixels.
- At the same time, we have a thermal resolution of 0.03 K. The only authors giving information on their thermal resolution are [25] who state that their thermal resolution is less than 0.05 K. A qualitative comparison between our images and the remaining databases suggests that [25] is indeed the database with most comparable thermal quality while the remaining datasets have a drastically lower thermal resolution and display more noise.
- Our database is the first that comes with full manual annotations for 68 facial landmark points, a prerequisite for complex tracking and localization tasks.
- Additionally to a subset for emotion analysis, we offer database subsets for the detection of facial action units (AU) and a subset with free expressions.

However, our datasets do not contain certain specific data offered by some other databases, with two aspects being the most important. First, we did not include any persons wearing glasses in our database. This was intended as glasses are thermally opaque, making the analysis of action units involving eyes or landmark annotations for the eyes and eyebrows either more difficult or impossible. Furthermore, our database does not contain data from the visible spectrum. Other datasets, such as the IRIS dataset, contain images in the visible spectrum with different lighting setups to analyze light effects or allow developing multimodal methods involving both thermal and visual data. Not including visible data in our database was a design decision that allowed focusing on high-quality thermal data for infrared algorithm development.

B. Recording setup

All images for our database were recorded using an Infratec HD820 high resolution thermal infrared camera with a 1024 x 768 pixel-sized microbolometer sensor with a thermal resolution of 0.03K at 30°C and equipped with a 30 mm f/1.0 prime lens. Subjects were filmed while sitting at a distance of 0.9m to the camera, resulting in a spatial resolution of the face of approximately 0.5 mm per pixel. A thermally neutral backdrop was used for the recordings to minimize background variation. The recordings were acquired as full resolution videos with a frame rate of 30 frames per second. To build the database, each video was screened manually and selected frames were exported for annotation. As a result, the final database contains 2935 images of 90 subjects in total, however not all subjects were filmed in all sequences. The database does not contain any RGB data. The recordings were split into different sequences, each designed for a different task:

- **Sequence A** contains a defined head movement pattern,

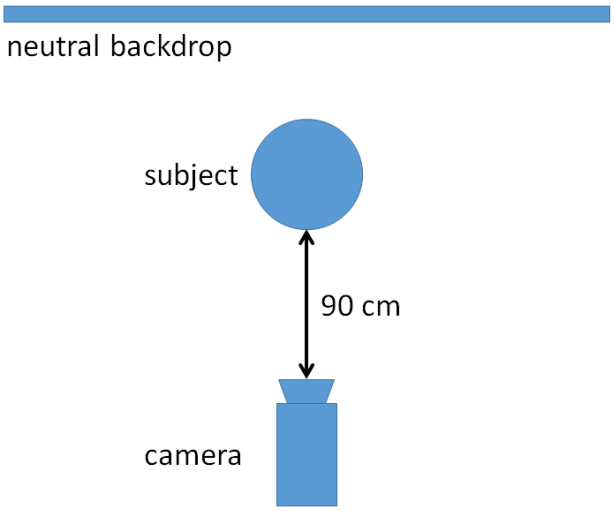

Fig. 2: Database image acquisition setup. Sitting participants were filmed against a thermally neutral backdrop.

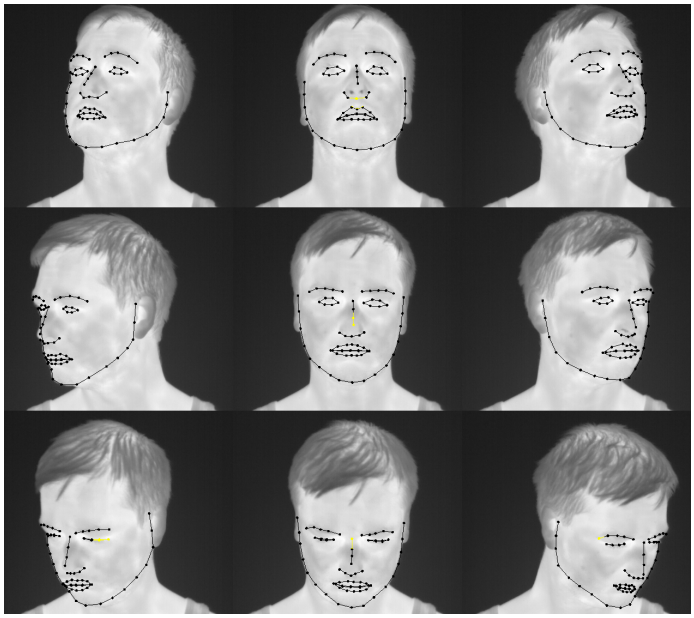

Fig. 3: The nine different head poses from sequence A.

where each participant was instructed to follow a defined S-shaped trajectory (Fig. 3). From this recording, frames at 9 distinct positions (upper left, upper frontal, upper right, frontal right, full frontal, frontal left, lower left, lower frontal, lower right) have been extracted and annotated. This allows for a large number of images with strongly varying head poses that are required for robust face detection and landmark localization.

- **Sequence B** is a set of images showing basic facial action units (AUs) according to the facial action coding system (FACS) introduced by Ekman et al. [26]. Action units are fundamental, elementary facial movements that usually do not appear separately, but in conjunction with other AUs to form complex facial expressions. Due to the large number of action units, only a subset of action units was recorded, namely AU 1+2 (inner and outer brow raiser), AU 4 (brow lowerer), AU 6 + 7 (cheek raiser and lid tightener), AU 9 + 10 (nose wrinkler and upper lip raiser), AU 24 (lip pressor), AU 27 (mouth stretch) and AU 43 (sniff). For all participants with a recording in sequence B, one frame per AU has been annotated. Figure 4 shows the AUs recorded for the database.

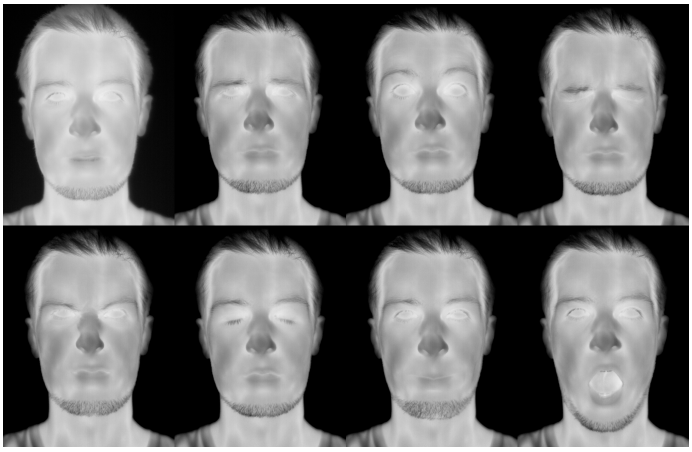

Fig. 4: Elementar action units. Top row from left to right: Neutral, AU 4, AU 1 + 2, AU 6 + 7. Bottom row from left to right: AU 9 + 10, AU 43, AU 24, AU 27.

- Basic emotions are shown in **Sequence C**. Basic or universal emotions according to Ekman [27] are happiness, sadness, surprise, fear, disgust, anger and contempt. The emotions are posed, the database contains no recordings of actual emotions induced by video clips or other means. Three frames of the emotions neutral, happy, sad and surprised have been selected and annotated for each participant. The remaining four emotions are included in the database, however with no annotations.
- Finally, in **Sequence D**, all participants were asked to perform arbitrary head movements and facial expressions. Between 3 and 5 frames from each participant's sequence was selected and annotated. These recordings were used to add realistic facial expression and pose variance to the database in contrast to the posed expressions and poses acquired in the other sequences, as recent research indicates that adding unposed "'in-the-wild"'-images with large variation increases the robustness of face analysiss algorithms. Fig. 6 shows examples from this sequence.

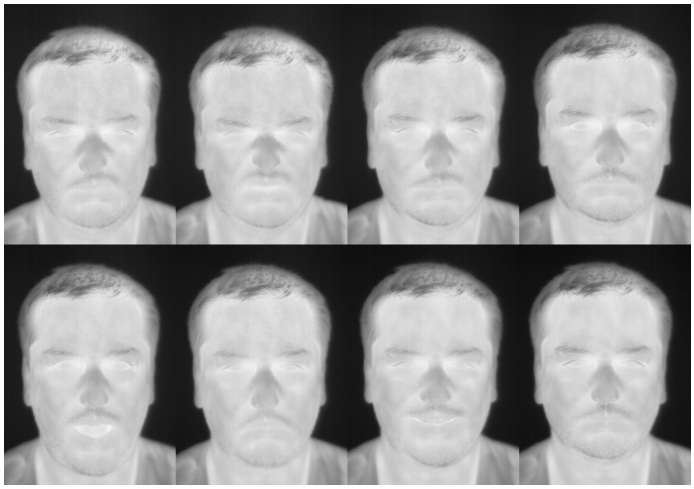

Fig. 5: Basic posed emotions. Top row from left to right: contempt, disgust, anger and fear. Bottom row from left to right: surprise, sadness, happiness and a neutral face.

C. Manual Annotations

All 2935 selected frames were manually annotated with the 68-point landmark set also used for databases such as Helen [14] and LFPW [15]. This extensive set of annotations using a widely established scheme allows using the database for a substantial number of algorithms, allowing assessment of their performance on thermal infrared data. Figure 1 shows examples of annotated frames while Fig. 7 shows the exact localization of the 68 landmark positions in the face. Both the landmarks as well as the connectivity information are stored, allowing selection of landmarks of specific facial areas such as eyes or mouth separately. After landmarking all images, the dataset has been checked for annotation consistency, ensuring that landmark positions correspond to the same facial features in all database images.

Fig. 6: Samples from the free movement sequence.

D. Comparison to Existing Visual and Thermal Databases

Compared to the existing thermal databases, we offer the best resolution and, for the first time, extensive manual land-

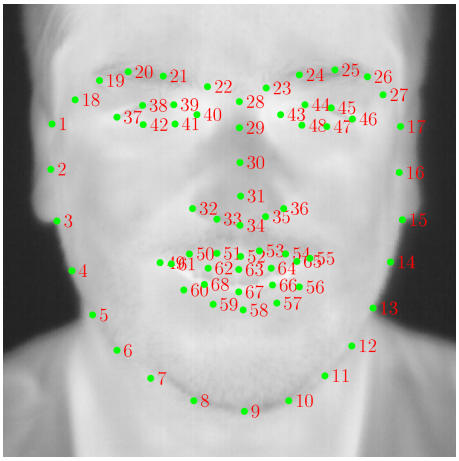

Fig. 7: The 68-point annotation scheme with each point's coordinates in the face.

mark annotations in a standarized and widely used format. Additionally, our different subsets allow using the database for a number of novel tasks that could not be addressed with existing datasets. While we were focusing on the landmark quality and diversity of facial expressions, our database contains no images in the visible domain and therefore no different illuminations as offered by some other dual-modality databases. Our database contains no images of persons with glasses; all participants were asked to remove glasses if they were wearing any. This decision was made to offer unobstructed view at the eyes since they are an importand facial area when identifying emotions. We use a thermally neutral backdrop as in all known thermal face databases. We performed first explorative experiments on recordings with non-neutral backgrounds that show that algorithms trained on our database may be applied to non-neutral backgrounds, however there is no systematic evaluation how different backgrounds may affect the results. Currently this is not a shortcoming since most published work on thermal face processing uses data with neutral backgrounds. Hoerver, current databases and algorithms using them will require a re-evaluation with novel data if LWIR face analysis becomes more widely used in scenarios with unconstrained background conditions.

Our database was inspired by current annotated databases in the visual domain with the goal of providing data with comparable quality and quantity in the thermal spectrum. Compared to state-of-the-art visual databases such as LFPW and Helen, we offer a comparable number of images (Helen contains 2330 manually annotated images, LFPW 1432 of them). However, the two visual datasets use real-world data (so-called 'in-the-wild' images from random photographs taken under unconstrained conditions) with a larger variety in head pose, facial expression and illumination. Acquiring this type of data with a thermal camera is currently not viable since thermal cameras have technical limitations such as their extremely narrow depth of field. Therefore, the database reflects realistic current uses of thermal cameras and the applications given below show that a number of tasks can be accomplished by training algorithms using our data. Nevertheless, should thermal cameras become more widely used in unconstrained environments, the need for an updated database that would be able to address the requirements of these new scenarios might arise.

E. Applications of the Database

Next to using it for the tasks described in this paper, we were able to apply our database to several image processing tasks. In [16], we analyzed the performance of several face detection algorithms by training and evaluating them on the full database. As a result, we were able to show that current face detection methods that perform successfully on visual data, such as the HOG-SVM detector [28], the deformable parts model [29] and pixel intensity comparisons [30] work well on infrared data and, when trained with a sufficiently large dataset, outperform established methods for face detection in thermal infrared images. In another work [17], we were able to apply our database to perform facial landmark localization on thermal infrared images and use the localization algorithms to perform robust landmark tracking in videos. To this end, we trained a feature-based active appearance model [31] that uses the database to learn a statistical face model and subsequently uses gradient-based optimization to fit the model to unseen faces.

F. Obtaining the Database

The database and supplementary Python code for basic analysis are freely available at https://github.com/marcinkopaczka/thermalfaceproject.

III. Head Pose Estimation

The method for facial landmark localization that we trained with our database and presented in [17] is based on feature-based active appearance models (AAMs). While these models allow very precise landmark detection, they have an intrinsic limitation that lowers their robustness. At the start of their optimization, AAMs need to be initialized with an initial shape. Often, this shape is defined by obtaining a facial bounding box that is determined either manually, from a ground truth reference or from a face detection system as described in Sec. [16]. Subsequently, the AAM's mean shape is transformed globally to the scale, rotation and position of the bounding box to give an initial estimate of the facial landmarks. Without any additional a-priori information on the final landmark positions, this is the statistically best initialization since the mean shape represents the mean of the training landmark distribution and therefore the highest probability for all landmark positions. Usually, the mean shape of a sufficiently large face database is a frontal view of the face. Therefore, the initialization is a good starting point for close-to-frontal faces, but may cause the model to diverge or settle for a wrong local minimum when the target face shows strong out-of-plane rotation. We analyzed the results of the AAM on our database and discovered that indeed most of the images with high fitting errors were those displaying non-frontal faces.

To address this shortcoming, we propose a head pose estimation method that allows improved initialization of the

AAM. The method is based on a random forest regression [32] similar to the approaches presented in [33] with a random forest classifier and [1], where a random forest regressor is used to estimate head pose in depth images. In contrast to these and other similar approaches, we do not predict the head pose as rotation angles, but use the knowledge that the head pose estimation results will be forwarded to an AAM and predict AAM parameters directly. This is achieved by training an AAM on the database or a cross-validation subset and using this AAM to track a set of videos in which persons rotate their head slowly. The AAM parameters for each frame are stored together with the cropped face. Subsequently, we compute fixed-length HOG feature vectors of all tracked face images and forward these vectors as feature vectors into a random forest regression algorithm [34]. The labels of the vectors are the two AAM parameters that have been stored during tracking. Once the forest has been trained, it allows predicting the first two AAM parameters of an input face. Since the first parameters correspond to the axes with highest variance as computed by the PCA of the AAM, the resulting head pose is much closer to the final pose when the head pose is not frontal. We forward the estimated parameters to the AAM and compute a shape based on the prediction. This shape then replaces the mean shape for optimizer initialization. Figure 8 shows the training and application of the random forest regression.

IV. Facial Landmark Detection and Tracking with Deep Alignment Networks

Deep learning - the use of multilayer neural networks for machine learning tasks - is the currently dominating research area in image processing. The most commonly used networks in computer vision are various types of convolutional neural networks, a sub-type of neural networks that uses locally connected convolutional layers in addition to (or as replacement for) classical fully connected layers. Convolutional networks have achieved outstanding performance in tasks such as image classification, where residual networks [35] or the recently proposed Mask R-CNN [36] are the current benchmark for multi-label classification. Fully convolutional networks [37] such as the U-Net [38] are the current state-of-the-art in image segmentation. Deep learning techniques have also been applied to the task of facial landmark detection, with notable approaches being [39] and [40], where a deep networks are trained to detect a set of facial landmarks together with additional attributes such as head pose, gender and basic facial expressions.

To analyze the suitability of our database for deep learning tasks, we have chosen to train and evaluate the recently proposed deep alignment network (DAN) [41] algorithm with our database. The DAN is designed to detect a set of landmarks - in their original publications the authors use the same 68 landmark points that we use for our database - therefore allowing direct performance comparison with the AAM-based landmark detection described above.

The deep alignment network is a multistage approach in which several stages refine landmark positions predicted by the previous stage. The initial stage receives an input image and as output it yields a transformed version of the image warped into a normalized canonical form together with a first estimate of the facial landmarks as a heatmap and 256 features from the last, fully connected layer of the fist stage. When compared to our above described approach, the first DAN stage can be roughly compared to the head pose estimation step where initial landmark estimations are computed. After the image has been transformed by the first stage, all subsequent stages have an identical layout. Consequently, their input and output is a canonical image, a landmark heatmap and the feature map. These subsequent layers have a role similar to our AAM implementation - they improve landmark positions once initial positions are estimated. Following the results reported by the creators of DAN, we implemented only one single second stage after the initial stage as further stages drastically increase the required training time while are reported to yield no improvement in detection precision. Unlike an AAM result, the network returns the landmark positions and not a full face model. However, since the goal of both algorithms is the landmark detection and the texture model built by the AAM has the sole purpose of improving fitting accuracy and is not used for any further computations, the results of both algorithms are directly comparable and both algorithms can be evaluated on the same data and using an identical error metric.

V. Facial Expression Recognition

Facial expression recognition is a common image processing task and an active research area. A recent overview over different databases and approaches, especially for non-RGB data including multimodal and thermal methods, can be found in [42]. To evaluate how our database can be used for this task, we have analyzed how a set of methods that are already established for facial expression recognition in RGB data can be applied to our images. To this end, we used the annotated emotion images from sequence C to train a facial expression classifier. Different combinations of feature descriptors and machine learning methods that have been proven to work for similar tasks in regular photographs were evaluated. As features, we used the following:

- The coordinates of the manually annotated landmarks. This is a purely geometric feature containing no pixel intensity or neighborhood information.
- The pixel intensities of the faces without any feature extraction applied.
- Histograms of oriented gradients (HOG) [28], Local binary patterns (LBP) [43] and dense scale-invariant features (SIFT) [44] extracted from the faces.

Note that the first item in the list is the only feature that requires explicit landmark coordinates while the remaining features are computed from the pixel values from the bounding boxes of the faces, therefore allowing skipping the landmark localization for expression analysis. The extracted features were then fed into the following classifiers (See [45] for an explanation of the used algorithms):

- Linear SVM, as preliminary experiments have shown that this type of SVM has superior performance for our

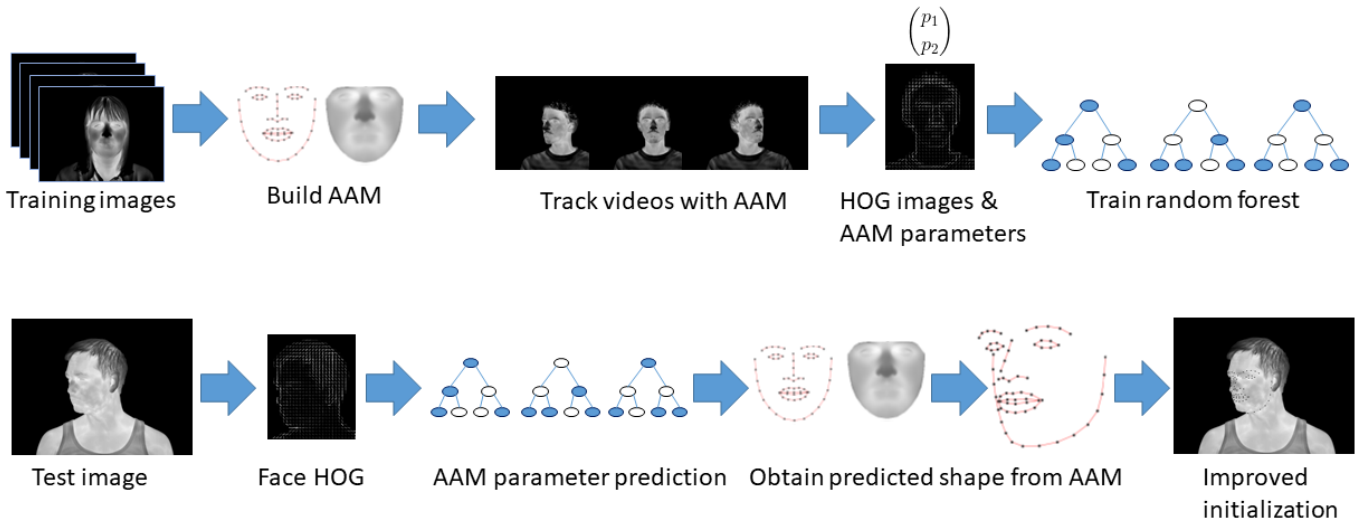

Fig. 8: Schematic of training and application of our head pose estimation. Top row: a set of training images is used to create an AAM. The AAM is subsequently used to track a set of training videos, where the final shape and AAM parameters of each result are stored. To train the random forest, HOG features of every face in the tracked video frames are computed and forwarded to the random forest as features, while the first two shape parameters of the AAM result are used as labels. The trained forest can subsequently be used to predict the shape of a face in a new image. Bottom row: applying the trained random forest for improved initialization. Given an input image, HOG features for the subimage defined by the facial bounding box are extracted. The HOG vector is fed into the previously trained random forest to obtain predictions for the first two AAM parameters. The predictions are forwarded to the AAM which uses these to generate a shape instance. Finally, the shape instance is warped to the facial bounding box and AAM fitting can be initialized with the prediction.

problem than its polynomial and RBF-based variants. Standardizing the features before SVM classification has shown to yield better results.
- A kNN classifier, for which preliminary tests have shown that k=1 and feature standardization give best results.
- A Binary Decision Tree. The tree's split criterion was chosen by the training function.
- Linead Discriminant Analysis (LDA). For LDA computation, the pseudoinverse was chosen over the inverse matrix since not all features had nonzero variance, thereby making direct inverse computation impossible.
- The Naive Bayes Classifier. Since this method is not able to work on invariant features we implemented an additional step that detects and removes invariant features from the feature vectors.
- A Random Forest Classifier. For this method, a forest size of 40 trees has been chosen as initial experiments had shown that using more than 40 trees does not result in performance gain.

VI. Experiments and Results

Here, we describe how the algorithms for each task were trained end evaluated. Results of the facial expression analysis evaluation were obtained using leave one-subject-out-cross-validation. In this validation type, we removed all images of a given subject from the database, trained the algorithms on all remaining subjects and tested their performance on the subject previously removed from the database. While requiring a large number of evaluation runs, this method was chosen as it gives the best impression of the overall algorithm and database performance due to the maximal possible overlap of training data with the full database while still allowing evaluation on unseen subjects. We did not perform the same type of cross-validation for the facial landmark detection as the DAN algorithm requires substantially longer for training its neural network than the other used algorithms; training a DAN instance requires about 72 hours on a GeForce 980Ti GPU. Therefore we randomly picked 270 images from 8 subjects from the database as test set and trained the landmark detection algorithms on the remaining images to perform evaluation within reasonable time.

A. Facial landmark detection

We perform the landmark detection evaluation in two steps. Fist, we analyze the performance of several combinations of feature descriptors and fitting algorithms. Once the optimally performing algorithm combination has been found, we evaluate its performance in comparison to the performance of the AAM with random forest initialization and the DAN. We already presented an evaluation of feature descriptors and algorithms in [17], however our database has significantly grown in size since the original publication (695 vs. 2935 images). Furthermore, a new AAM fitting method [46], the Wiberg inverse compositional (WIC) method has been introduced which has shown to be mathematically equivalent to the currently most precise fitting method, the simultaneous inverse compositional (SIC) fitting while being computationally more efficient. We therefore performed the algorithm evaluation again on the full database. Additionally, we evaluated the performance of the

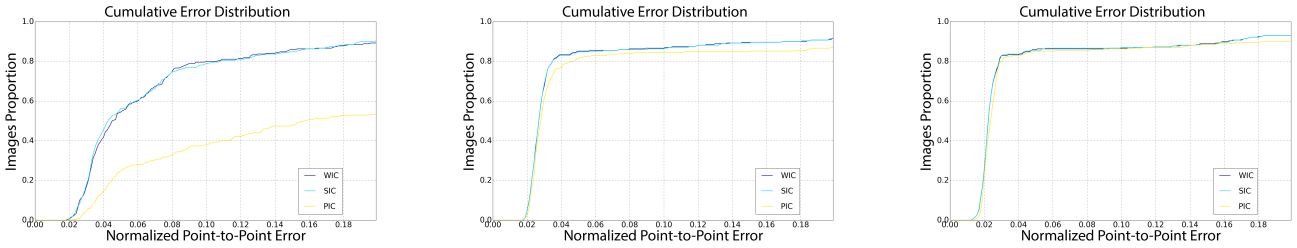

Fig. 9: Cumulative distribution of the fitting performance of the evaluated AAM algorithms. From left to right: intensity based AAM, DSIFT-AAM, HOG-AAM

project-out inverse compositional algorithm (PIC) [47], which is a commonly used fitting algorithm for AAMs in the visual domain, performing significantly faster than SIC and WIC at the cost of slightly reduced precision.

For performance evaluation, The AAM's mean shape was scaled and translated to the face position in the image and subsequently used to initialize landmark positions. AAM fitting was performed for all combinations of the evaluated three algorithms (PIC, SIC, WIC) and feature descriptors (intensity, HOG and DSIFT). To obtain optimal results, the AAM was set to retain 98% of its original variance. The model diagonal was set to 150 pixels. To quantify the fitting error, the widely used normalized point-to-point error introduced by [48] was used, which minimizes the effect of face size and head pose on the final result, and allows efficient comparison of errors across different databases and imaging settings. The error metric E_i computed for each image I_i is the root mean squared distance in pixels between each horizontal and vertical fitted landmark position $\mathrm{x}_{n,f}$, $\mathrm{y}_{n,f}$ after AAM fitting and its corresponding ground truth landmark $\mathrm{x}_{n,g}$, $\mathrm{y}_{n,g}$, accumulated across all N landmarks in the image and normalized by the mean of face width w_i and height h_i:

$$\mathrm{E}_i = \mathrm{N}_i \sqrt{\frac{\sum_{n=1}^{N} \left[(\mathrm{x}_{n,f} - \mathrm{x}_{n,g})^2 + (\mathrm{y}_{n,f} - \mathrm{y}_{n,g})^2 \right]}{2N}} \quad (1)$$

with

$$\mathrm{N}_i = \left(\frac{1}{\frac{1}{2}(w_i + h_i)} \right) \quad (2)$$

The normalized errors computed for the test set are shown in Fig.9. It can be seen that the feature-based approaches clearly outperform the classical intensity-based method when AAMs are used for facial landmark detection in the thermal infrared, with HOG slightly outperforming DSIFT (Note that in our previous evaluation DSIFT was slightly outperforming HOG when the same algorithms were used on the smaller dataset; we therefore assume that DSIFT performs better on smaller datasets while HOG benefits from the increased number of images). Of the three evaluated algorithms, WIC and SIC show very similar performance with the fast PIC method yielding a slightly lower precision for feature-based AAM and a substantially lower precision for the intensity-based approach. Since WIC and SIC have been proven to be mathematically equivalent, we assume that details in the implementation of the algorithms of the used Menpo framework result in the observed small differences in fitting performance. Subsequently, we analyzed how the performance of the best-performing AAM method - HOG features with WIC as fitting algorithm - is affected by adding the random forest initialization described in Sec. III. Additionally, we trained the DAN and evaluated its performance against the plain AAM and the AAM with random forest initialization. The results, which are shown in Fig. 11, indicate that the random forest initialization allows improving the fitting performance of the AAM. The initialization does not improve general fitting precision as it does not shift the overall error curve, indicating that it has no effect on images that can already be fitted well with the default AAM method. However, the error curve shows a strong precision increase for images that had a large fitting error with the traditional AAM method. This observation is backed further by Fig. 10. While the random forest has virtually no impact on the landmark positions of frontal images, its application results in a strong improvement of AAM fitting precision on faces displaying out-of-plane rotation.

The CNN-based DAN method outperforms both AAM-based approaches in terms of fitting precision. At the same time, the DAN performs fitting drastically faster than the AAM (Fig. I). Two main reasons contribute to the high speedup: The DAN uses a feedforward neural network architecture that performs landmark prediction in a single pass, while the AAM's fitting algorithms use an iterative optimization that is much more time-comsuming at runtime. Furthermore, the DAN is implemented with GPU support, allowing it to perform a number of numerical operations at a much higher speed.

Fitter	Feature	Fitting time (s)	fps
WIC	intensity	0,32	3,13
SIC	intensity	0,40	2,50
PIC	intensity	0,16	6,25
WIC	DSIFT	5,66	0,18
SIC	DSIFT	6,85	0,15
PIC	DSIFT	1,05	0,95
WIC	HOG	7,31	0,14
SIC	HOG	8,65	0,12
PIC	HOG	2,79	0,36
DAN		0,03	33.3

TABLE I: Time required by the evaluated algorithms to perform landmark detection on an input face image.

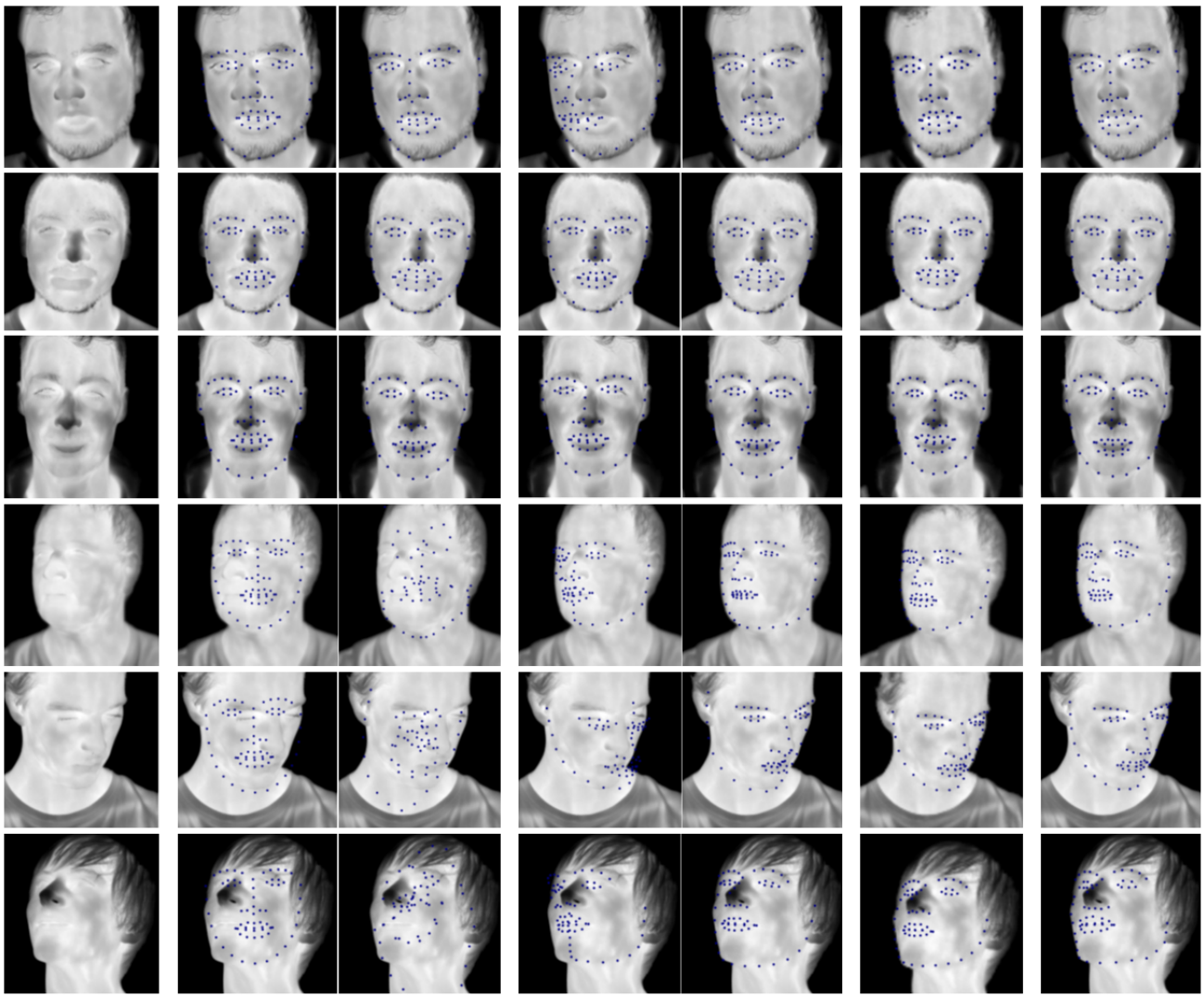

Fig. 10: Fitting results of the three evaluated approaches. From left to right: Original image crop, AAM initalization with mean shape, result of the AAM initialized with mean shape, AAM initialization with random forest prediction, result of the random forest initialized AAM, DAN result, manual ground truth landmarks.

.

B. Facial Expression Recognition Performance

Experiments conducted to determine optimal image size have shown that quadratic images of the faces scaled to a length of $l = 144$ pixels for both sides yield best results. Values below 144 result in lower classification rates, while higher image resolutions did not increase classification performance. Therefore, all results refer to images scaled to $l = 144$. Fig. 13 shows an overview of all tested feature-classifier combinations. It can be seen that the chosen SVM configuration is the best performing method while the basic kNN and decision tree classifiers perform clearly weaker on the expression recognition task. Using feature descriptors for constructing the feature vectors has been shown to deliver results superior to feature vectors created by using raw landmark coordinate or pixel intensity data.

A detailed confusion matrix of the four emotions for the best performing combination - the linear SVM using dense SIFT features - is shown in Fig.12. Happiness is the most clearly detectable expression with only minimal misdetections. On the other hand, sadness is often misclassified as a neutral expression and vice versa. Still, the classification rates are far above chance level in all cases, clearly indicating that a facial expression classifier can be trained using the database.

Our database contains manual landmark annotations for four of the eight recorded emotions while the images for the remaining four emotions are stored with an emotion label but without landmark information. In a study to analyze the combined performance of our algorithms, we performed an automated classification of all eight emotions using the linear SVM with dense SIFT features as above. This algorithm requires a bounding box of the face. To obtain precisely aligned bounding boxes, we used a two-step approach on all images by combining the best performing algorithms of the previous evaluations: First, initial bounding boxes were detected using the DPM algorithm. Subsequently, facial landmark detection

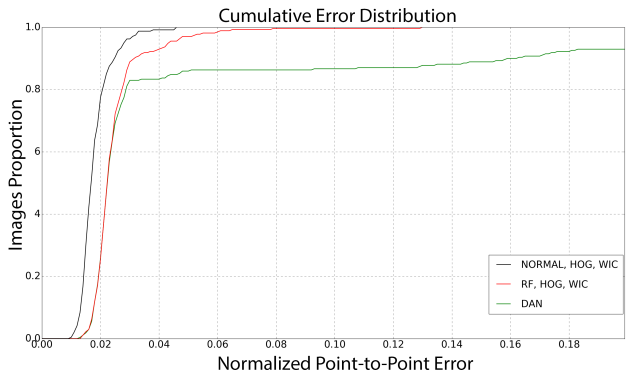

Fig. 11: Cumulative error plot of WIC-fitted HOG-AAM, WIC-fitted HOG-AAM with our suggested random forest-based shape prediction and DAN.

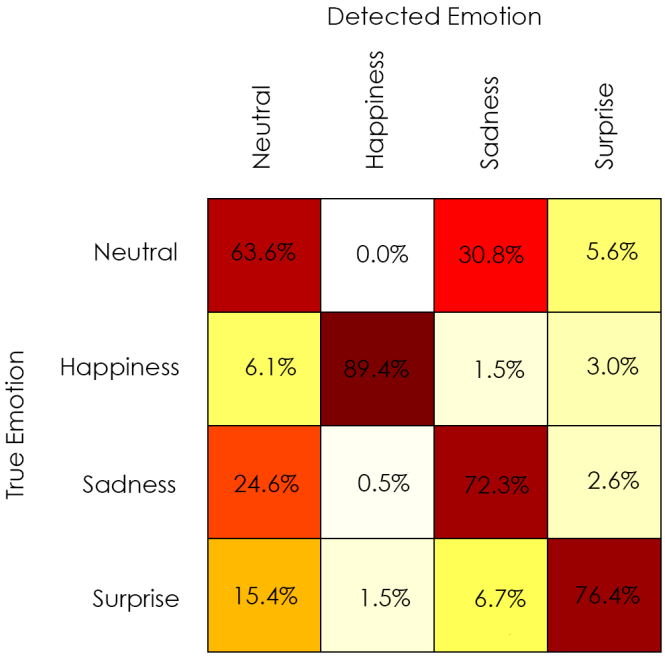

Fig. 12: Confusion Matrix for DSIFT + linear SVM

.

with the HOG-WIC AAM was performed and initialized from the detected bounding box. The algorithms were re-evaluated using the detected landmarks and the bounding boxes of the landmark positions. To assess human performance on the same task and to evaluate the validity of the final bounding boxes, the images were cropped to their bounding boxes and shown to three humans. The humans had to assess the emotion and to report if any bounding box had been incorrectly aligned, which was not the case. Subsequently, we performed the same algorithm evaluation as for the analysis of four emotions (Fig. 15). Again, the linear SVM is the best performing classifier, however the HOG feature descriptor outperforms it in the case of eight depicted emotions. A detailed confusion matrix for the HOG-SVM algorithm is shown in Fig. 14. The average detection rate of the algorithm (46.7%) outperforms

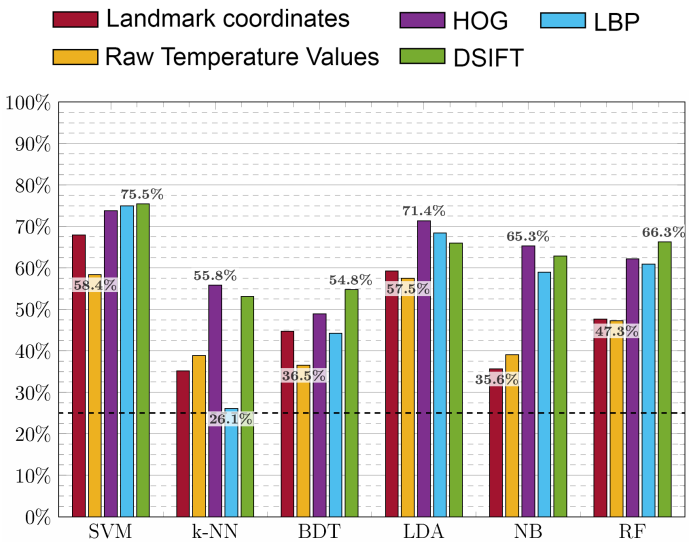

Fig. 13: True positive rates of the tested feature-classifier combinations. The dashed line indicates the success chance of a random guess (25%).

human humans who assessed 42.0 of the images correctly. At the same time, the figure shows that humans assign the neutral expression to a high number of images, while the HOG-SVM has no such preference and misclassifications are spread more evenly. A possible reason might be that humans tend to classify a face as neutral when in doubt.

VII. Discussion

The key results of our studies are two-fold: First, we were able to show that different computer vision tasks involving processing of facial images in thermal infrared recordings can be solved using our newly introduced database. We allow not only the evaluation of face detection and facial expression analysis tasks that have been addressed before, but also allow performing facial landmark detection using AAMs and DANs, two methods that require sufficiently large image databases for training. Training of these advanced methods is not possible with existing thermal databases since existing datasets offer only strongly limited resolutions and, most importantly, no manual landmark annotations.

Second, in all cases where approaches based on image features and machine learning were used, these methods clearly outperform algorithm-based approaches in terms of robustness. In facial landmark detection using AAMs, feature-based models outperform intensity-based approaches. When applied to facial expression analysis, the database allows training of state-of-the-art facial expression detection algorithms that make use of machine learning and feature extraction such as HOG-SVM. The successful training of a current deep convolutional network with our database and its application to the challenging task of facial landmark detection proves that our database is also suitable for being used with modern deep learning methods.

The database has been evaluated on several algorithms that have been developed for RGB data. Therefore we can now

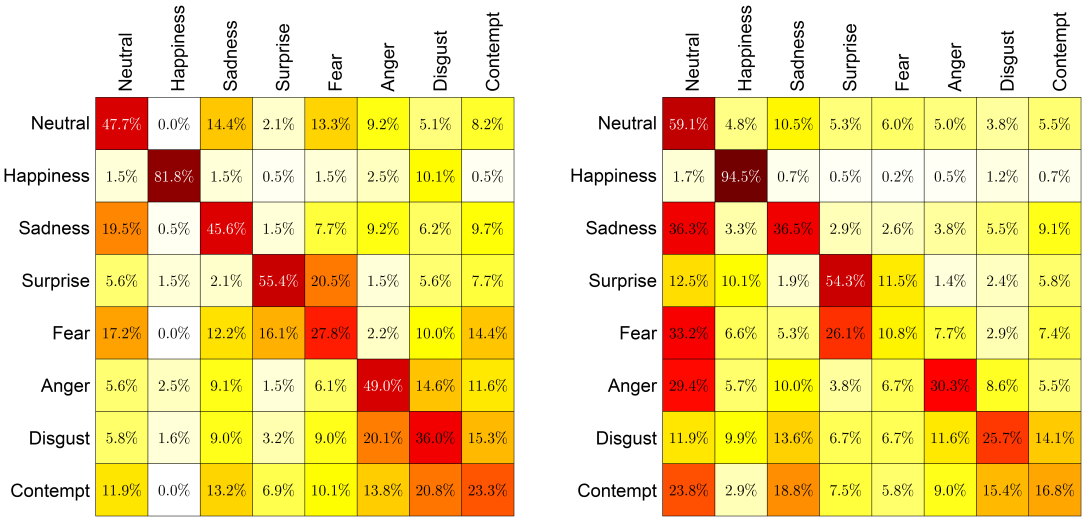

Fig. 14: Performance of our algorithm (left) and humans (right) on the facial expression recognition task. Average human classification rate is 42.0%, HOG-SVM achieves 46.7%.

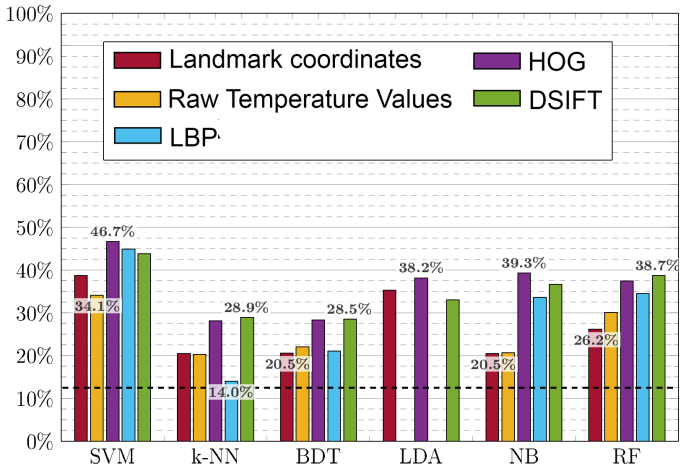

Fig. 15: True positive rates of the tested feature-classifier combinations for eight emotions using automatically detected landmarks and bounding boxes. The dashed line indicates the success chance of a random guess (12.5%). The LDA implementation for LBP and raw temperature failed to train and crashed, therefore no values are reported.

for the first time compare the performance of facial landmark detection algorithms in thermal and visual data. Here, we refer to the results provided by the authors of the original work on established RGB datasets. The authors of the feature-based active appearance models [31] report that their algorithm achieved a normalized point-to-point error of less than 0.02 for 80% of the images in the LFPW test set and that all test images had a fitting error below 0.05. The authors of the deep alignment network reported that around 92% of the images of the menpo test set (a novel dataset introduced for the CVPR 2017 facial landmark detection challenge [49]) had an error of 0.02 or less and their failure rate was 1.74%. Our results displayed in Fig. 11 show that the fitting accuracy is qualitatively similar, but still slightly lower in thermal images. While the failure rate is comparable when using the DAN or a random-forest-initialized AAM, the non-initialized AAM described in [31] has strong difficulties in finding the landmarks in strongly rotated faces. 78% of our images have a fitting error when fitted with the DAN, however it can be seen that the AAM has a slightly lower maximal precision. Possible reasons for this difference can only be assumed. Next to the possiblity that facial landmark detection is a generally a more difficult task in thermal images than in RGB data due to the lower contrast of the data - an assumption backed by the observation that only very current algorithms are able to perform the detection at all while the established intensity-based AAM yielded significantly lower results as shown in Fig. 9 - there might be differences in the datasets. Our database has a slightly lower resolution than its RGB counterparts; while the faces were re-scaled to lower resolutions for fitting for both RGB and thermal data there might still be landmark-relevant information in the RGB data that is not contained in our recordings due to the resolution difference. Also, landmarking the images in the thermal domain was a difficult task for the human annotators as well due to the low contrast and texture information in the thermal data. It is possible that the landmarks in the visual datasets are more consistently placed, therefore allowing improved training and detection performance of the algorithms. A way of confirming this assumption would be the comparison of landmark consistency across several human annotators for both thermal and visible data in the future.

We assume that many of the results of our research can be transferred to other fields of computer vision, where scientists are facing the decision between acquiring and labeling large datasets for available machine learning methods or developing custom-made algorithms for the problem at hand. For our

evaluated tasks, the answer can be clearly given, as data-driven machine learning methods outperformed manual and statistical algorithms in all cases. The comparison of established AAMs and novel DANs, where the more recent method drastically outperforms the former approach, indicates that more advanced algorithms perform better on the same data when the data itself meets all required criteria regarding its amount and quality of annotations. Since machine learning is a currently highly active research area, we assume that algorithms of the near future will make the gap between learning-based approaches and non-learning methods even wider. Therefore, when acquiring and labeling sufficient data for a given problem is technically possible and machine learning algorithms for similar tasks exist, then our results indicate that adapting the existing algorithms to the new data has a high chance of not only being successful, but also of outperforming expert-designed algorithms for the given problem. Such custom-tailored methods are possible solutions if gathering training data is not possible or not viable, or if the computing power of the machines applied for the image processing tasks is highly limited since machine-learning based algorithms, and this applies even more to deep learning methods, are very demanding on computer hardware. So if the algorithms need to be executed on machines with low computational performance, such as embedded or highly power-efficient processors, then the resources available might only allow algorithm-based solutions.

VIII. Conclusion

In our work, we have introduced a new, fully annotated high resolution thermal face image database for different computer vision tasks and evaluated how different algorithms perform on commonly appearing problems when trained using the database. We have througly described the database's image acquisition procedure and its contents. Afterwards, the database was used to evaluate its suitability for facial landmark detection and facial expression recognition. We were able to show that both tasks can be solved robustly by using learning-based approaches that are trained using our database. Evaluation has shown that the learning-based approaches, several of which have not been used for these tasks in the thermal infrared domain before, clearly outperform previously presented methods. In conclusion we were able to show that using a sufficiently large and well annotated database can be used to train different learning-based algorithms wich should be preferred over algorithm-based approaches due to their increased performance.

References

[1] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 617–624.

[2] F. A. Kondori, S. Yousefi, H. Li, S. Sonning, and S. Sonning, "3d head pose estimation using the kinect," in *Wireless communications and signal processing (WCSP), 2011 international conference on*. IEEE, 2011, pp. 1–4.

[3] F. Al-Khalidi, R. Saatchi, H. Elphick, and D. Burke, "An evaluation of thermal imaging based respiration rate monitoring in children," *American Journal of Engineering and Applied Sciences*, vol. 4, no. 4, pp. 586–597, 2011.

[4] S. Ioannou, V. Gallese, and A. Merla, "Thermal infrared imaging in psychophysiology: potentialities and limits," *Psychophysiology*, vol. 51, no. 10, pp. 951–963, 2014.

[5] B. Lahiri, S. Bagavathiappan, T. Jayakumar, and J. Philip, "Medical applications of infrared thermography: a review," *Infrared Physics & Technology*, vol. 55, no. 4, pp. 221–235, 2012.

[6] A. Rogalski and K. Chrzanowski, "Infrared devices and techniques (revision)," *Metrology and Measurement Systems*, vol. 21, no. 4, pp. 565–618, 2014.

[7] W. K. Wong, J. H. Hui, J. B. M. Desa, N. I. N. B. Ishak, A. B. Sulaiman, and Y. B. M. Nor, "Face detection in thermal imaging using head curve geometry," in *Image and Signal Processing (CISP), 2012 5th International Congress on*. IEEE, 2012, pp. 881–884.

[8] M. Chakraborty, S. K. Raman, S. Mukhopadhyay, S. Patsa, N. Anjum, and J. G. Ray, "High precision automated face localization in thermal images: oral cancer dataset as test case," *Proc. SPIE*, vol. 10133, pp. 1 013 326–1 013 326–7, 2017. [Online]. Available: http://dx.doi.org/10.1117/12.2254236

[9] M. Marzec, R. Koprowski, and Z. Wróbel, "Methods of face localization in thermograms," *Biocybernetics and Biomedical Engineering*, vol. 35, no. 2, pp. 138–146, 2015.

[10] J.-G. Wang and E. Sung, "Facial feature extraction in an infrared image by proxy with a visible face image," *IEEE Transactions on Instrumentation and Measurement*, vol. 56, no. 5, pp. 2057–2066, 2007.

[11] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[14] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European conference on computer vision*. Springer, 2012, pp. 679–692.

[15] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[16] M. Kopaczka, J. Nestler, and D. Merhof, "Face detection in thermal infrared images: A comparison of algorithm- and machine-learning-based approaches," in *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2017.

[17] M. Kopaczka, K. Acar, and D. Merhof, "Robust facial landmark detection and face tracking in thermal infrared images using active appearance models," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, Rome, Italy, February 2016, pp. 150–158.

[18] M. Kopaczka, R. Kolk, and D. Merhof, "A fully annotated thermal face database and its application for thermal facial expression recognition," in *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2018, pp. 1956 – 1961.

[19] G. Koukiou and V. Anastassopoulos, "Drunk person identification using local difference patterns," in *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*. IEEE, 2016, pp. 401–405.

[20] S. Bennett, T. N. El Harake, R. Goubran, and F. Knoefel, "Adaptive eulerian video processing of thermal video: An experimental analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 10, pp. 2516–2524, 2017.

[21] M. Kopaczka, Ö. Özkan, and D. Merhof, "Face tracking and respiratory signal analysis for the detection of sleep apnea in thermal infrared videos with head movement," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 163–170.

[22] B. Abidi. IRIS Thermal/Visible face Database. Accessed (2018, July 23th). [Online]. Available: http://vcipl-okstate.org/pbvs/bench/index.html

[23] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *Multimedia, IEEE Transactions on*, vol. 12, no. 7, pp. 682–691, Nov 2010.

[24] R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague, "Infrared face recognition: A comprehensive review of methodologies and databases," *Pattern Recognition*, vol. 47, no. 9, pp. 2807–2824, 2014.

[25] M. Kowalski and A. Grudzień, "High resolution thermal face dataset for face and expression recognition," *Metrology and Measurement Systems*, vol. 2, 06 2018.
[26] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
[27] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.
[29] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
[30] N. Markuš, M. Frljak, I. S. Pandžić, J. Ahlberg, and R. Forchheimer, "Object detection with pixel intensity comparisons organized in decision trees," *arXiv preprint arXiv:1305.4537*, 2013.
[31] E. Antonakos, J. Alabort-i Medina, G. Tzimiropoulos, and S. P. Zafeiriou, "Feature-based lucas–kanade and active appearance models," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2617–2632, 2015.
[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324
[33] M.-J. Kang, H.-Y. Lee, and J.-W. Kang, "Head pose estimation using random forest and texture analysis," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–4.
[34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.
[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[39] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.
[40] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
[41] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, vol. 3, no. 5, 2017, p. 6.
[42] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.
[43] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
[44] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.
[45] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, January 2006. [Online]. Available: https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/
[46] J. Alabort-i Medina and S. Zafeiriou, "A unified framework for compositional fitting of active appearance models," *International Journal of Computer Vision*, vol. 121, no. 1, pp. 26–64, 2017.
[47] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.
[48] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.
[49] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution."