A Fully Annotated Thermal Face Database and its Application for Thermal Facial Expression Recognition

Marcin Kopaczka
Institute of Imaging & Computer Vision
RWTH Aachen University
Aachen, Germany
marcin.kopaczka@lfb.rwth-aachen.de

Raphael Kolk
Institute of Imaging & Computer Vision
RWTH Aachen University
Aachen, Germany
raphael.kolk@lfb.rwth-aachen.de

Dorit Merhof
Institute of Imaging & Computer Vision
RWTH Aachen University
Aachen, Germany
dorit.merhof@lfb.rwth-aachen.de

***Abstract*—A large number of algorithms for processing faces in regular photographs and videos has been published in recent years, making this field one of the most active research areas in computer vision. Most current algorithms require a sufficiently large, manually annotated database for training. While several large databases for the visible spectrum are available, no sufficiently large and fully annotated database for the emerging thermal infrared modality has been published so far. Instead, algorithms in the thermal spectrum usually rely on specific assumptions regarding image content, making them less robust than their data-driven counterparts that are based on machine learning methods. We address this shortcoming by introducing a novel high-resolution thermal infrared face database with extensive manual annotations. We describe the database in detail and show that it can be used for advanced image processing tasks by training algorithms for facial expression recognition using the database. The full database itself, all annotations and the complete source code are freely available from the authors for research purposes at lfb.rwth-aachen.de. The code and annotations will be made commonly available under LGPL license, the image data will be available for download upon agreeing to the terms and conditions for image data given on the website.**

***Index Terms*—Thermal Infrared, Database, 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 require one or several of the aforementioned tasks that have to be performed.

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. Most algorithms presented for facial image processing focus therefore on RGB or RGB+D applications. Examples include emotion recognition [1] and face tracking [2] using the Kinect camera. 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. 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 physiophsychological effects are visible in the IR domain as well [4]. Despite these advantages, two factors have been limiting the widespread use

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

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 μm 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. 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 signals 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. Color information, for example, 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.

Many current image processing algorithms are based on machine learning methods and therefore require annotated training data. Classifiers such as neural networks that have been trained on visual images cannot be applied directly to LWIR data due to the large difference in object appearance. So far, learning-based methods have been applied to thermal images only in a few specialized cases, mostly due to the lack of sufficiently large and deliberately annotated image databases that are required to train the algorithms.

We address the second shortcoming by introducing a novel, high-resolution thermal image database with extensive manual annotations (Fig. 1). we show that this database can be used to train machine learning algorithms that require well-annotated image data for training. In this work, we describe the database in detail, followed by showing how it can be used for thermal infrared facial expression recognition. In previous publications, we were already able to show how earlier versions of the database were used for face detection [5] and facial landmark detection [6]. Both methods have been combined and used to detect breathing anomalies in [7]. In this work, we will describe the exact protocol used to acquire the final database and show its performance for the task of facial expression recognition, being a further example of a common image processing task. The database, all annotations and the code required to perform the experiments described here are freely available under BSD license for research purposes upon request, allowing using it for own research and reproducing all results shown here.

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

While a vast number of databases designed for various tasks exists for the visual spectrum, only a few relevant thermal face databases have been presented so far. In the past, the most prominent databases used for facial image processing on the thermal infrared domain were the the Equinox and IRIS databases. However, both resources are no longer available. The only database currently available upon request is the USTC-NVIE thermal image database [8], released in 2010. The database is multimodal, containing both visible and thermal videos that have been acquired simultaneously. It contains videos of 215 participants, 236 still frames of 84 of these participants were manually annotated for facial expression recognition. The spatial resolution of the infrared videos is 320 x 240 pixels. The database contains data sets with both spontaneous and posed emotions as well as images with and without glasses and under different lighting conditions. In contrast to putting emphasis on acquiring multimodal data, we focused on high-resolution thermal recordings and precise manual annotations. Therefore, our database provides:

- High resolution data at 1024 x 768 pixels, much higher than currently available databases that usually work with 320 x 240 pixel data.
- a wide range of head poses instead of the usually fully frontal recordings provided elsewhere
- manually placed and validated landmarks for 68 facial points while other databases provide either no annotations at all or rudimentary annotations for the positions of mouth, eyes and nose.
- a high variation of the expressions shown, starting with basic morphological changes indicated by single AUs (basic facial action units according to the facial action coding system (FACS) introduced by Ekman et al. [9]), followed by fundamental emotions up to arbitrary expressions. To the best of our knowledge, our database is the only set available with AU data for thermal infrared recordings.

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 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 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 relevant frames according to the requirements described in the sequence descriptions below were exported for annotation. As a result, the final database contains 2500 images of 90 subjects in total, however not all subjects were filmed in all sequences. A detailed overview of the images and participants is included in the database as Excel spreadsheet. In contrast to the USTC-NVIE database, the database does not contain a regular RGB video of each session. The recordings were split into different sequences, each designed for a different task:

- **Sequence A** contains a defined head movement pattern, where each participant was instructed to follow a defined S-shaped trajectory (Fig. 2). 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 and at the same time defined head poses.

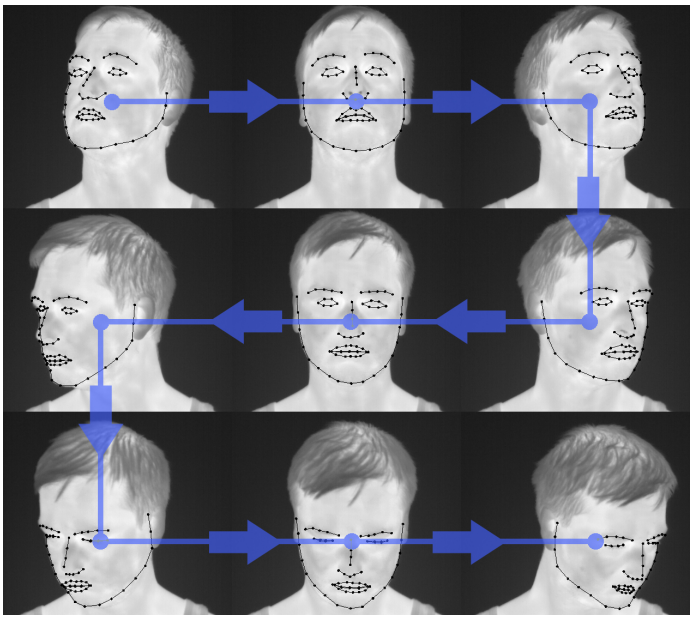

Fig. 2. The nine different head poses from sequence A.

- **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. [9]. 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) (Fig.3). For all participants with a recording in sequence B, one frame per AU has been annotated.
- Basic emotions are shown in **Sequence C**. Basic or universal emotions according to Ekman's work are happiness, sadness, surprise, fear, disgust, anger and contempt. Each participant was instructed to display the requested emotions, 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 per person. 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. Fig. 5 shows examples from this sequence.

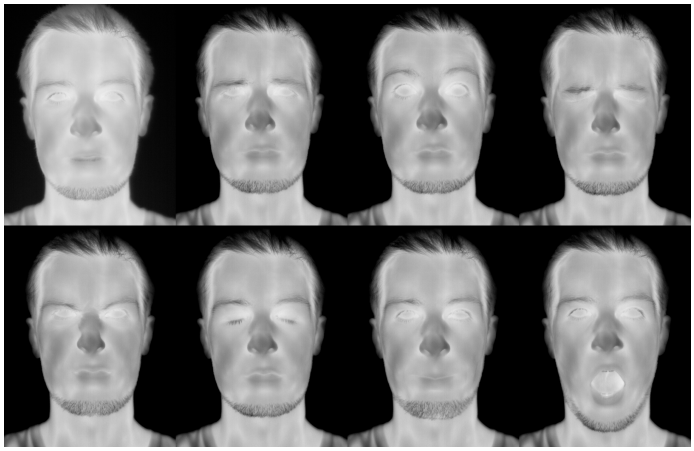

Fig. 3. Elementary action units. Top row from left to right: Neutral, AU 4, AU 1 + 2, AU 6 + 7. Bottom row from left to right: AU 9 + 10, AU 43, AU 24, AU 27.

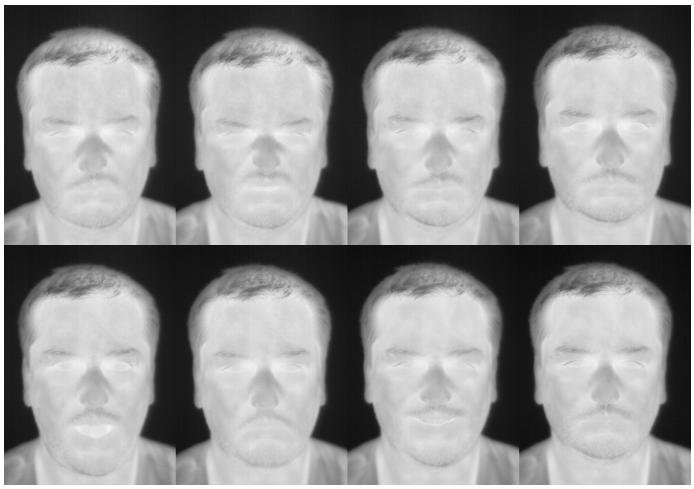

Fig. 4. Basic posed emotions. Top row from left to right: Neutral, happiness, sadness, surprise. Bottom row from left to right: Fear, anger, disgust, contempt.

B. Manual Annotations

All 2935 selected frames were manually annotated with the 68-point landmark set also used for databases such as Helen [10] and LFPW [11]. 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. 6 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

Fig. 5. Samples from the free movement sequence.

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.

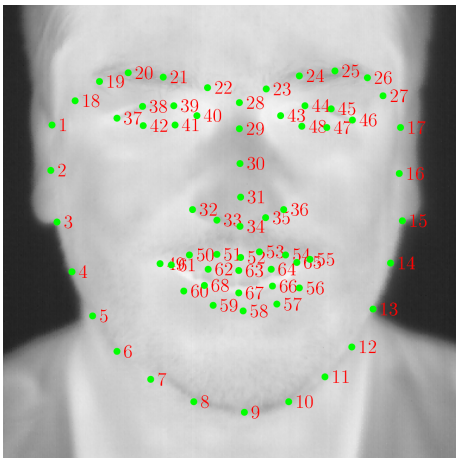

Fig. 6. The 68-point annotation scheme with each point's coordinates in the face.

III. Facial Expression Recognition

Facial expression recognition has been an active research area for the past years. A recent overview over different databases and approaches, especially for non-RGB data, can be found in [12]. 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 evaluate the capabilities of our database, 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 tested. 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. A similar approach was used in [13] for feature description in visual images.
- The pixel intensities of the faces without any feature extraction applied.
- Histograms of oriented gradients (HOG) [14], Local binary patterns (LBP) [15] and dense scale-invariant features (SIFT) [16] extracted from the faces.

The features were then fed into the following classifiers:

- Linear SVM [17], as preliminary experiments have shown that this type of SVM has superior performance for our problem than its polynomial and radial basis function-based variants. Standardizing the features before SVM classification has shown to yield better results.
- A k-nearest-neighbours (kNN) classifier, for which preliminary tests have shown that k=1 and feature standardization give best results.
- A binary decision tree (BDT). The tree's split criterion was chosen by the training function.
- Linear 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 (NB). 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 [18] (RF). 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.

A. 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 $l = 144$. Results 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. Fig. 7 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 landmark coordinate or pixel intensity data only.

A detailed confusion matrix of the four emotions for the best performing combination - the linear SVM using dense SIFT features - is shown in Fig.8. Happiness is the most

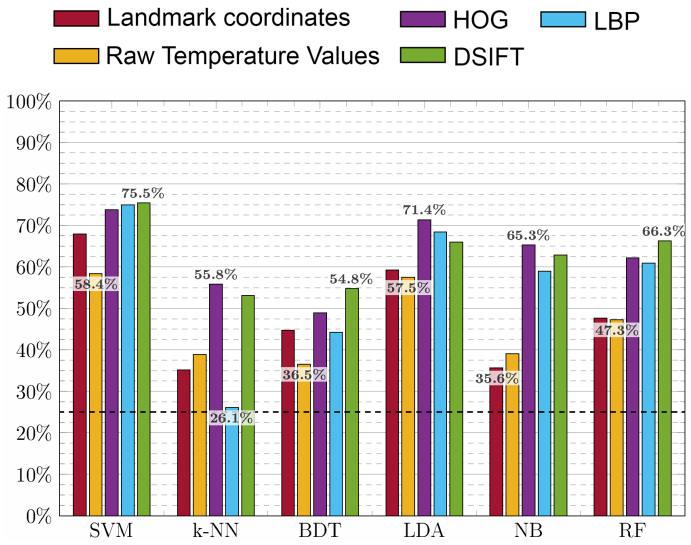

Fig. 7. True positive rates of the tested feature-classifier combinations. The dashed line indicates the success rate of a random guess (25%).

clearly detectable expression with only minimal misdetections. On the other hand, sadness is often misclassified as a neutral expression and vice versa.

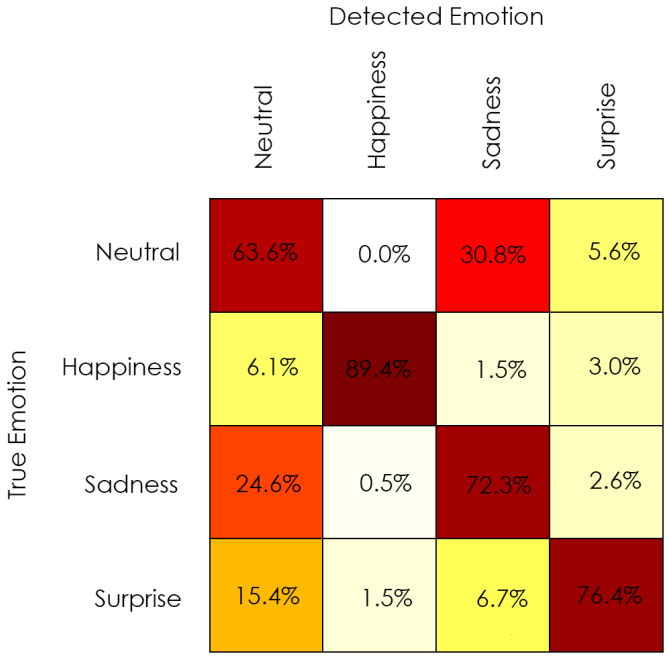

Fig. 8. Confusion Matrix for DSIFT + linear SVM.

Since the chosen feature-classifier-combination does not require full manual annotations, we were able to perform a full eight-class classification on all facial expression images including the non-annotated set, the results of which are shown in Fig. 9. As reference, each image has been shown to three humans and the sum of all results is shown in Fig. 10. In total, our method achieves superhuman classification accuracy.

A closer investigation of the results shows that humans pick the neutral class with a significantly higher preference, an indicator that humans tend to classify a face as neutral when in doubt while misclassifications are more evenly spread among classes when automatic classification is used.

	Neutral	Happiness	Sadness	Surprise	Fear	Anger	Disgust	Contempt
Neutral	47.7%	0.0%	14.4%	2.1%	13.3%	9.2%	5.1%	8.2%
Happiness	1.5%	81.8%	1.5%	0.5%	1.5%	2.5%	10.1%	0.5%
Sadness	19.5%	0.5%	45.6%	1.5%	7.7%	9.2%	6.2%	9.7%
Surprise	5.6%	1.5%	2.1%	55.4%	20.5%	1.5%	5.6%	7.7%
Fear	17.2%	0.0%	12.2%	16.1%	27.8%	2.2%	10.0%	14.4%
Anger	5.6%	2.5%	9.1%	1.5%	6.1%	49.0%	14.6%	11.6%
Disgust	5.8%	1.6%	9.0%	3.2%	9.0%	20.1%	36.0%	15.3%
Contempt	11.9%	0.0%	13.2%	6.9%	10.1%	13.8%	20.8%	23.3%

Fig. 9. Confusion Matrix for DSIFT + linear SVM for 8 facial expressions.

	Neutral	Happiness	Sadness	Surprise	Fear	Anger	Disgust	Contempt
Neutral	59.1%	4.8%	10.5%	5.3%	6.0%	5.0%	3.8%	5.5%
Happiness	1.7%	94.5%	0.7%	0.5%	0.2%	0.5%	1.2%	0.7%
Sadness	36.3%	3.3%	36.5%	2.9%	2.6%	3.8%	5.5%	9.1%
Surprise	12.5%	10.1%	1.9%	54.3%	11.5%	1.4%	2.4%	5.8%
Fear	33.2%	6.6%	5.3%	26.1%	10.8%	7.7%	2.9%	7.4%
Anger	29.4%	5.7%	10.0%	3.8%	6.7%	30.3%	8.6%	5.5%
Disgust	11.9%	9.9%	13.6%	6.7%	6.7%	11.6%	25.7%	14.1%
Contempt	23.8%	2.9%	18.8%	7.5%	5.8%	9.0%	15.4%	16.8%

Fig. 10. Confusion Matrix for 8 facial expressions classified by humans.

IV. Conclusion and Future Work

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 thoroughly described the database's image acquisition procedure and its contents. Afterwards, the database was used for studies in thermal infrared facial expression recognition. We were able to show that the task can be solved robustly by using learning-based approaches trained using our database and that the database allows training of classifiers that, on average, perform better than humans on the given task.

In the near future, we will use our database to establish a full image processing pipeline for facial image processing and evaluate its capabilities by using it for medical purposes such as pain and stress detection.

References

[1] M. Szwoch and P. Pieniazek, "Facial emotion recognition using depth data," in *2015 8th International Conference on Human System Interaction (HSI)*, June 2015, pp. 271–277.

[2] N. Smolyanskiy, C. Huitema, L. Liang, and S. E. Anderson, "Real-time 3d face tracking based on active appearance model constrained by depth data," *Image and Vision Computing*, vol. 32, no. 11, pp. 860 – 869, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885614001310

[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] 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.

[6] 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.

[7] M. Kopaczka, O. Oezkan, 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 Workshop (ICIAP)*, 2017.

[8] 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.

[9] 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.

[10] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 679–692.

[11] 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.

[12] 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.

[13] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges," in *Proceedings of the British Machine Vision Conference*, 2013.

[14] 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.

[15] 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.

[16] 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.

[17] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324