RGB-D-T based Face Recognition

Nikisins, Olegs; Nasrollahi, Kamal; Greitans, Modris; Moeslund, Thomas B.

Published in:
Proceedings 22nd International Conference on Pattern Recognition, ICPR 2014

DOI (link to publication from Publisher):
10.1109/ICPR.2014.302

Publication date:
2014

Document Version
Early version, also known as pre-print

RGB-D-T based Face Recognition

Olegs Nikisins[1], Kamal Nasrollahi[2], Modris Greitans[1] and Thomas B. Moeslund[2]
1) Institute of Electronics and Computer Science, Dzerbenes 14, LV-1006, Riga, Latvia
2) Visual Analysis of People Laboratory, Aalborg University, Sofiendalsvej 11, 9200 Aalborg, Denmark

***Abstract*—Facial images are of critical importance in many real-world applications from gaming to surveillance. The current literature on facial image analysis, from face detection to face and facial expression recognition, are mainly performed in either RGB, Depth (D), or both of these modalities. But, such analyzes have rarely included Thermal (T) modality. This paper paves the way for performing such facial analyzes using synchronized RGB-D-T facial images by introducing a database of 51 persons including facial images of different rotations, illuminations, and expressions. Furthermore, a face recognition algorithm has been developed to use these images. The experimental results show that face recognition using such three modalities provides better results compared to face recognition in any of such modalities in most of the cases.**

I. Introduction

Facial biometrics has a number of motivating points: it is easily collectible, non-contact, universal, non-intrusive and can be used without cooperation of the user [1] making it practical in applications such as gaming, access control, and surveillance. On the other hand face recognition systems have to deal with challenges like noisy data, intra-class variations and spoof attacks [2]. Noisy data can be caused by environmental / illumination conditions, which are not acceptable for normal operation of the particular image capturing device. Head poses, expressions of the face, illumination conditions and occlusions can result in significant intra-class variations [3]. A recent research trend for the reduction of the above mentioned downgrading factors is multi-modal biometrics [4]. An intuitive explanation of the improved performance in this case is that particular kind of noise has unequal impact on different modalities [5].

In order to keep both the ease of usage of the facial biometrics and mutual complementarity of different modalities we have developed a multi-modal face recognition algorithm, which is based on the combination of *RGB, depth and thermal* data. There are some significant attempts to combine different modalities for face recognition, however in most cases it is limited to 2-modal face recognition. Various combination of the modalities for face recognition can be found in literature: RGB-D [1], [6], D-T [7], RGB-T [8], [9]. To the best of our knowledge the only similar combination RGB-D-T is introduced in [10]. However the database utilized in the experimental part of the paper [10] contains only *non-synchronized* frontal view images, which were acquired under controlled lighting conditions.

According to the taxonomy in [11] information fusion can be done at feature level, score level, and rank / decision level. Recent research in the field shows that feature level fusion outperforms other approaches due to preservation of raw information about the class label [2], [5], [12]. Therefore, the proposed multi-modal face recognition algorithm is based on feature level fusion. Some examples of *bi-modal* face recognition algorithms with feature level fusion are introduced in [6], [9]. However most publications in the field incorporate score level fusion strategy [1], [6], [8], [10].

For evaluation purposes we have also developed a multi-modal face database. To the best of our knowledge, this is the first database including the *synchronized* RGB, Depth and Thermal face images. The introduced database has a number of noteworthy advantages. The proposed range of modalities provides a comprehensive experimental platform for various strategies of biometric data fusion. Moreover the database covers some of the most challenging aspects of face recognition [3]: pose, expression and illumination variations, which are organized in *three acquisition sequences*. A special evaluation protocol is also introduced in order to unify the usage of the database and split of the images into Training, Cross-Validation and Test sets. The database is also supplemented with ground-truth data for all modalities providing the coordinates of the face bounding box in the images.

The rest of the paper is organized as follows: a description of the database and evaluation protocol is given in Section II. The details of the proposed multi-modal face recognition algorithm are introduced in Section III. The experimental results are then given in Section IV and finally the paper is concluded in Section V.

II. RGB-D-T Facial Database

A. Hardware configuration

The Microsoft® Kinect for Windows has been used to capture the RGB and depth images. The Near Mode was enabled in the acquisition process enabling the device to see objects as close as 40 centimeters in front of the sensor without losing accuracy. The thermal camera AXIS Q1922 was used for capturing the thermal face images and it was mounted right under the Kinect RGB camera with the distance of 8.5 cm between the lens centers (Figure 1). The resolutions of the RGB, depth, and thermal images are 640×480, 640×480, and 384×288 pixels, respectively.

The capturing setup is schematically represented in Figure 1. The Kinect and thermal cameras were located at a distance of 1 meter from the face at the hight of 1.5 meters from the floor. The same distance was selected for light sources 1, 2, 4, 5 with a hight of 1.6 meters. The lamp 3 was placed behind the capturing devices at a distance of 1.3 meters from the face and at a hight of 1.85 meters. An average height at which the face was located is 1.3 meters. The tripods of the lamps were marked with numbered labels (at a hight of 1.35 meters), which were used to control the rotations of the head in the corresponding capturing scenario. The frontal tripod had two additional labels at the hight of 0.4 and 1.85 meters.

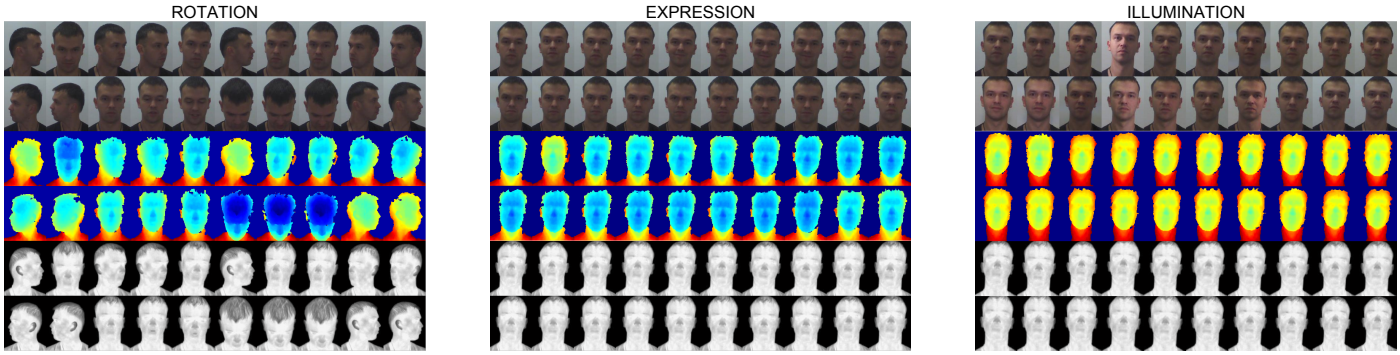

Fig. 2. Examples of RGB (top), corresponding synchronized depth (middle) and thermal (bottom) *training* images for all sequences.

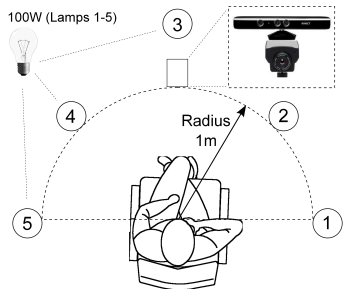

Fig. 1. A schematic representation of the capturing setup.

B. Capturing scenarios

The images of each person in the database are organized in three sets corresponding to *rotation*, *expression* and *illumination* scenarios. The total number of persons in the database is 51. Each capturing sequence (*rotation*, *expression*, *illumination*) has 300 images per person: 100 RGB, 100 Depth and 100 Thermal *synchronized* images. The total number of images per person is 900 resulting in 45900 images in the database.

In the *rotation* acquisition scenario people were asked to keep the neutral expression and to turn the head delaying a look at the numbered labels, which were sticked to the tripods of the lamps (Figure 1). The total of seven markers were used: one for frontal view, four for side turns of the head, one for turning the head up and one for turning down. The acquisition was performed indoors with fluorescent ceiling lighting.

In the *expression* sequence individuals were asked to keep the frontal position of the head and to change the expression of the face according to the list of emotional moods (Ekman universal expressions): neutral, happy, sad, angry, and surprised. The illumination conditions were unchanged.

The *illumination* sequence is intended to study the impact of variable lighting on the recognition performance in the absence of other complicating factors. People were asked to keep the neutral expression of the face and frontal position of the head. The lamps 1 to 5 (Figure 1) were sequentially turned on and off by the operator. Only one lamp was on at a time. The fluorescent ceiling lighting was also switched on. Sample images for all sequences from RGB, Depth and Thermal modalities are displayed in Figure 2.

C. Evaluation protocol

For development purposes the database is supplemented with a Matlab indexing function, which must be used to split the data into Training, Validation and Testing sets. This function is introduced in order to unify both the development and testing of the face recognition algorithms among the researchers who use the database.

The number of *Testing* images in the indexing function per person is *always* 50 for each modality in each capturing scenario. To enable a fair comparison of different algorithms on the introduced database, we ask the researchers to report the results on the Test data (strictly defined by indexing function), which should not be involved in the training process in any manner. The Equal Error Rate (EER) is selected as the performance criteria.

The user of the indexing function can also select up to 50 *Training* images per person for each modality in each capturing scenario if *Validation set is not needed*. If both *Training* and *Validation* sets are needed then the maximal number of images in each set is 25 (the number of Training and Validation images is *always* the same).

The introduced approach has a number of advantages. First, the indexes generated by the function are always strictly defined making the development process reproducible. Second, the testing data is completely separated from the training and validation sets. Moreover, the test set *always* covers a wide spectrum of variability in face appearance, which is present in each particular capturing scenario.

D. Ground-truth data

Each facial image in the database is supplemented with ground-truth data, which includes bounding box parameters of the face. The ground-truth data is generated automatically. One of the most popular face detectors is Viola-Jones algorithm [13]. However in our case this method produced a high number of errors due to significant off-plane rotations of a face in the input images. For this reason we have developed previously unpublished algorithm for the detection of region of interest (ROI), which is based on the depth images. It provides reasonable stability of the ROI on a frame to frame level in the input sequences of the images and does not generate false detections in our data, which are the primary requirements for the ground-truth information. The first step of the ROI

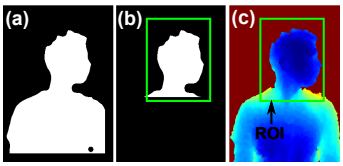

Fig. 3. Generalized steps of the ROI detection algorithm.

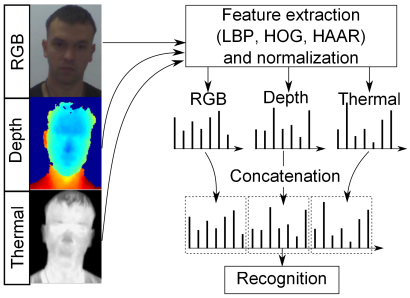

Fig. 4. The block-diagram of the proposed multi-modal face recognition algorithm with feature level fusion.

detection algorithm is *smoothing and binarization* (Figure 3 (a)) of the input depth image (Figure 3 (c)). Next, the pixels corresponding to the body region are set to zero (the part exceeding the predetermined width limit), Figure 3 (b). In the last step the *rectangular region growing algorithm* is applied on the input image (Figure 3 (b)). The algorithm iteratively increments the size of the rectangular region, which is centered in the position of the mean of the binary image. The cost function is the ratio of *sum of pixel values* to the *number of pixels* in the current ROI. The execution is terminated when the cost value is below the specified limit. The detection result is displayed as a green rectangle in Figure 3 (b) - Figure 3 (c).

In order to map the ground-truth data obtained for depth the images to other modalities, a registration of the inputs is needed. In particular the D-to-RGB and D-to-T registration is important. The transformation of the depth images is limited to rotation, scaling and translation. One hundred corresponding points were manually marked for each modality, so as to estimate the D-to-RGB and D-to-T transformation matrices with RANSAC method. Once the transformation matrices are obtained, the binary images (e.g.: Figure 3 (b)) are *first registered* to the corresponding modality. Next, the rectangular region growing algorithm is applied. The detection now represents the ground-truth data for corresponding modality.

III. Multimodal face recognition algorithm

The standard pipeline of automatic face recognition system is based on three major tasks: detection, alignment and recognition (DAR pipeline) [14]. This section is dedicated to the recognition problem, which in our case is multimodal. The performance of the recognition algorithm is always better in the case of manual alignment of the subject [14], but this scenario is usually unrealistic in real-life applications. Despite the fact that the DAR pipeline is not directly discussed here, the performance reported on the proposed database is related to the *automatic* face recognition problem, since ground-truth data was generated automatically.

The generalized block-diagram of the proposed face recognition approach is displayed in the Figure 4. In order to obtain a clear understanding of "as is" possibilities of each particular modality the *preprocessing* of the cropped facial regions is *excluded*. The only preprocessing steps are conversion of RGB data to gray-scale format and resizing of the input facial regions from all modalities to the same scale.

First, the feature extraction is performed for each modality. The list of selected features is Local Binary Patterns (LBP) [15], Histograms of Oriented Gradients (HOG) [16] and HAAR-like features for face recognition [17]. The LBP is a very popular descriptor in the field of face recognition [15], [18]. The simplicity and high discriminative power of LBP in various computer vision tasks motivated the development of various extensions of the paradigm [19]. The literature on the subject of HOG-based face recognition is not so vast, however some significant research is done in the field [20]. Some authors tried to combine the LBP and HOG into the single descriptor [21]. Recent research in the field of HAAR-like features applicability to the face recognition task demonstrated the effectiveness of the descriptor [17], [22]. The above mentioned feature spaces are selected for the evaluation due to their semantically integral nature and potential applicability to any of the modalities in the introduced database. The advantage of the selected features is computational simplicity and the absence of the learning stage. One can argue that learning based features, such as Learning-based (LE) descriptor [23] and Spatial Face Region Descriptor (SFRD) [24] lead to the state-of-the-art performance. However authors in [25] demonstrate that high dimensionality of the feature space is also or even more critical to high performance than the presence of the learning stage. They show that simple *high dimensional* LBP descriptor can achieve significant improvements over both its low-dimensional version and the state-of-the-art [25]. Since concatenation of feature vectors from various modalities is one of the ways to increase the dimensionality of the feature space, we find it is useful to report the results for non-learning based descriptors.

Next, the concatenation of feature vectors from all modalities into a single descriptive vector is performed. Normalization of the feature vectors before concatenation is a critical issue [26]. The normalization principles are discussed later. According to the taxonomy in [11] the utilized concatenation strategy corresponds to *feature level fusion*.

The final step of the pipeline is recognition. Three recognition strategies are tested here: the most simplistic one based on Nearest Neighbor Classifier (NNC), the Weighted NNC based approach [18], [19] and Linear SVM based principle with "One-vs-All" classification scheme. The algorithmic details are discussed later.

A. Dimensionality and normalization of feature vectors

The detailed process of calculation of LBP, HOG and HAAR features is not discussed here since we keep it exactly the same as in the original papers. The explicit details can be found in [15] for LBP , in [16] for HOG and in [17] for HAAR-like features. Only the aspects essential for understanding of further discussions are introduced.

In order to be concatenated, the feature sets must be *compatible* [26]. The compatibility implies that feature spaces in all modalities are 1) semantically similar, 2) with equal dimensionality and 3) are normalized.

Since the same features are extracted from all modalities semantical similarity is ensured automatically.

The equal dimensionality means:

$$N_{RGB} = N_T = N_D = N_m, \tag{1}$$

where N_{RGB}, N_T, N_D are the number of features for RGB, Thermal and Depth modalities, respectively. N_m is the number of features per modality.

In the case of LBP the dimensionality of the feature space N_m^{LBP} is equal to [18]:

$$N_m^{LBP} = K^2 \cdot 2^P, \tag{2}$$

where K is the regioning factor (the LBP transformed image is divided into $K \times K$ regions), P is the number of sampling points on the radius R. The notation (P, R) is usually used for the description of LBP operator [15], [18].

The dimensionality of the HOG feature space per modality is defined as follows [16]:

$$N_m^{HOG} = \nu \cdot \varsigma^2 \cdot \beta, \tag{3}$$

where ν is the number of *square* overlapping blocks (the overlap of the blocks is fixed at 50% in our case), each block is divided into $\varsigma \times \varsigma$ cells, where ς is the block regioning factor; β is the number of orientation bins in the histogram of the cell. The value ν depends on the sizes of the input image and of the cell:

$$\nu = (W/\eta - 1) \cdot (H/\eta - 1), \tag{4}$$

where W and H are correspondingly width and hight of the input image in pixels, η is the size of the *square* cell in pixels. Here we assume that the remainders of divisions (W/η) and (H/η) are zero.

The dimensionality of the HAAR-like feature space is fixed to the value introduced in the original paper [17]: $N_m^{HAAR} = 115$.

Finally, for ensuring the normalization we suppose that $\boldsymbol{x}_m = (x_{m,1}, x_{m,2}, \ldots, x_{m,N_m})$ is the feature vector of modality m. Then the elements of normalized LBP feature vector $\hat{\boldsymbol{x}}_m^{LBP}$ can be found as follows:

$$\hat{x}_{m,i}^{LBP} = x_{m,i}^{LBP} / \|\boldsymbol{x}_m^{LBP}\|_1, i = 1, \ldots, N_m^{LBP}, \tag{5}$$

where $\|\boldsymbol{x}_m\|_1$ is L1-norm of the vector.

The normalization procedure of the HOG vector is the same as in the original paper [16]. HOG feature vector of the modality m is the concatenation of block feature vectors: $\boldsymbol{x}_m^{HOG} = \left(\boldsymbol{x}_{m,1}^{HOG}, \boldsymbol{x}_{m,2}^{HOG}, \ldots, \boldsymbol{x}_{m,\nu}^{HOG}\right)$. The feature vector of *each block* is normalized individually. The elements of normalized feature vector of the block j can be determined as follows:

$$\hat{x}_{m,j,i}^{HOG} = x_{m,j,i}^{HOG} / \sqrt{\|\boldsymbol{x}_{m,j}^{HOG}\|_2^2 + \epsilon^2}, i = 1, \ldots, \varsigma^2 \cdot \beta, \tag{6}$$

where $\|\boldsymbol{x}_{m,j}^{HOG}\|_2$ is L2-norm of the vector, ϵ is a small regularization constant, $\epsilon = 0.01$ in our case.

The HAAR-like features are sensitive to the dynamic range of pixel values, which might vary significantly for different modalities. The following normalization is introduced in this case:

$$\hat{x}_{m,i}^{HAAR} = x_{m,i}^{HAAR} / \|\boldsymbol{x}_m^{HAAR}\|_1, i = 1, \ldots, N_m^{HAAR}, \tag{7}$$

Once normalization is completed the concatenation of modalities into a single *normalized* multi-modal face feature vector $\boldsymbol{x}$ can be performed (Figure 4):

$$\boldsymbol{x} = (\hat{\boldsymbol{x}}_{RGB}, \hat{\boldsymbol{x}}_D, \hat{\boldsymbol{x}}_T). \tag{8}$$

B. Recognition stage

The first sequence of recognition experiments is performed with NNC. Despite of the fact that this recognition approach is the most simplistic, it is still useful for defining the borderline values of the performance of each particular feature and modality.

The second sequence of experiments is based on Weighted Nearest Neighbor Classifier (WNNC). In this instance the features are weighted according to their discriminative power prior the recognition. Following [18], [19] we have used an iterative feature weighting method called a Mini-Batch Discriminative Feature Weighting (MB-DFW) algorithm. The MB-DFW has a number of significant advantages: only *two* training samples per class are needed, which is important for biometric applications. The algorithm incorporates a mini-batch learning principle accelerating the learning process and it is a critical factor in the case of high-dimensional training data (multi-modal feature vectors). The weights obtained in the learning process are *the same for all classes* meaning that no learning is needed if more persons are added to the database. Moreover the algorithm is extended in two levels of feature-level and block-level weighting. In the feature-level weighting each feature in the descriptive vector has a unique weight, while in the block-level weighting the features within the block have the same weight. According to the results in [18], [19] block-level weighting slightly outperforms the feature-level weighting concept.

However, one drawback can be mentioned: the original MB-DFW is not able to operate with more than two training samples per class. For this reason we have extended the MB-DFW concept. Basically, the extension does not modify the MB-DFW algorithm itself, only the process of learning data forming is updated.

Suppose, that $d_{(i,i),(1,k)}^{intra}$ is the value of Squared Euclidean distance between two *weighted intra-class* (of the same person) feature vectors $\tilde{\boldsymbol{x}}_i^{(1)}$ and $\tilde{\boldsymbol{x}}_i^{(k\neq 1)}$, where (k) is the number of the training sample for person i. If N_c^{Train} is the total number of training samples per class / person, then $k = 2, \ldots, N_c^{Train}$, and $d_{(i,j),(1,k)}^{inter}$ is the value of Squared Euclidean distance between two *weighted inter-class* (of different persons) feature vectors $\tilde{\boldsymbol{x}}_i^{(1)}$ and $\tilde{\boldsymbol{x}}_j^{(k\neq 1)}$ with constraint $i \neq j$.

The cost function to be optimized in the MB-DFW is formed from pairs [18], [19]: $\left(d_{(i,i),(1,k)}^{intra}, d_{(i,j),(1,k)}^{inter}\right)$. The number of pairs in our *extended* version is $(N_c^{Train} - 1) \cdot M$, where M is the number of classes. In the original version $N_c^{Train} = 2$ resulting in the total of M pairs.

TABLE 1. EER VALUES (FOR TEST DATA) IN % FOR VARIOUS MODALITIES, CLASSIFIERS, FEATURES AND NUMBER OF TRAINING SAMPLES PER CLASS

		N_c^{Train} :	No training (NNC)				2				5				10				25			
		Mod:	RGB	D	T	All	RGB	D	T	All	RGB	D	T	All	RGB	D	T	All	RGB	D	T	All
WNNC	Rot.	LBP	31.9	34.6	31.3	31.4	27.4	26.1	30	24.8	23.3	21.5	31.1	18.4	23.1	21.3	30.4	18.5	23.6	21.8	30.5	18.5
		HOG	30.1	35.2	34.4	32.5	24.9	38.3	31.6	30.1	21.3	36.2	26.9	23.4	20.5	35.3	25.7	21.8	21.2	35.5	26.4	22.4
		HAAR	24.2	31.8	32.6	27.5				27.6				28				27.9				27.4
	Expr.	LBP	1.2	1.8	1.5	1.1	0.7	2.3	1.9	0.8	0.7	1.8	1.2	0.7	0.6	2	1.2	0.7	0.7	1.9	1.2	0.7
		HOG	1.6	6.1	2.2	1.7	0.7	6.6	0.9	0.6	0.7	6.5	0.9	0.6	0.8	7.1	0.8	0.6	0.8	6.5	0.8	0.6
		HAAR	2.6	8.8	3.2	2.8				2				2.1				2.1				2.1
	Illum.	LBP	15.2	8.8	9.8	8.4	11.2	5.2	4.7	3.8	12.3	4.9	4.4	3.5	9.8	4.7	3.8	3	10	4.8	3.8	3
		HOG	15.7	16.3	8.7	11.2	8	24.6	4.9	4.4	8.2	15.7	4.9	4.5	6.6	13.8	4.4	3.9	6.5	14.3	4.5	4
		HAAR	21.3	23.3	16.2	16				15.2				15.3				15.3				15.2
SVM	Rot.	LBP					13	14.1	17.1	10.6	1.9	3.9	4.9	1.5	0.5	1.1	0.9	0.2	0.1	0.2	0.3	0
		HOG					13.5	22.5	16.7	13.9	2.5	8.2	4.2	2.3	0.6	3.8	1	0.5	0.1	1.6	0.3	0
	Expr.	LBP					0.1	0.1	0.2	0	0.1	0.1	0.1	0	0.1	0.1	0.1	0	0.1	0	0.1	0
		HOG					0	0.6	0.1	0	0	0.2	0.1	0.1	0	0.2	0.1	0	0	0.1	0.1	0
	Illum.	LBP					1.4	0.9	0.7	0.6	0.4	0.3	0.1	0	0	0	0	0	0	0	0	0
		HOG					2.3	4.8	1	1.3	0.6	2.4	0.1	0.3	0	0.7	0	0	0.2	0.4	0	0.1

TABLE 2. THE IMPORTANCE OF EACH MODALITY IN MULTI-MODAL FACE RECOGNITION

	RGB	D	T
Rotation	**0.37**	**0.37**	0.26
Expression	**0.35**	0.15	**0.50**
Illumination	0.06	**0.25**	**0.69**

The last sequence of recognition experiments is based on the Linear SVM classifier with "One-vs-All" technique. This methodology is well known, thus the details are not provided. However, it is worth to mention, that it has a serious drawback: learning is needed every time a new person is added to the database.

IV. EXPERIMENTAL RESULTS

Evaluation of the proposed multi-modal face recognition algorithm is performed on the introduced database. The algorithm has a lot of variables to be optimized. First, the parameters of LBP and HOG features are estimated. The parameters of HAAR-like features are the same as in original paper [17]. In order to do so the grid search is utilized. At the same time the Equation (1) should be satisfied, in other words the parameters affecting the dimensionality of the feature space must hold the *same* values for all modalities. The evaluation criteria which should be minimized is the sum of EER values for all modalities. The EER is calculated for training data, which is selected according to the database evaluation protocol (10 images per person for each modality). Only Expressions and Illumination sequences are utilized in the parameter selection procedure. In all experiments facial regions were resized to $(W = 100, H = 130)$ for LBP and HAAR features and to $(W = 96, H = 128)$ for HOG features. The selected values for the parameters of the features are:

- LBP: $P = 8, K = 5,\ R^{RGB} = 5,\ R^D = 2,\ R^T = 6$;
- HOG: $\eta = 16, \beta = 18, \varsigma = 2$.

The corresponding dimensionalities of the feature spaces: $N_m^{LBP} = 6400$, $N_m^{HOG} = 2520$, and $N_m^{HAAR} = 115$.

Once the best parameters of the features are selected the NNC-based recognition is applied to testing data. The experiment is performed for all types of features and for all capturing scenarios. The resulting EER for each modality separately (RGB, D - depth and T - thermal) and for concatenated modalities (All) appear in Table 1 in the "No training (NNC)" column. The abbreviations "Illum.", "Expr." and "Rot." in Table 1 stand for capturing scenarios Illumination, Expressions and Rotations, respectively.

Next, a WNNC-based face recognition concept is tested. The MB-DFW algorithm is applied in the *block-level* in order to adjust the feature weights according to their discriminative importance. Block-level weighting principle implies that *features extracted from local image neighborhoods have the same weight*. In the case of LBP the facial image is divided into $K \times K$ *regions*, where each region is considered to be a local image neighborhood. Thus all features extracted from particular region will have the same weight, and the total number of *unique* weights per modality is $K \cdot K = 25$. The total number of unique weights in the concatenated feature vector is $M \cdot K \cdot K = 75$, where $M = 3$ is the number of modalities in our setup. In the case of HOG features the smallest unit, which can be considered as local image neighborhood is *cell*. Thus the total number of unique weights per modality is $\nu \cdot \varsigma^2 = 140$. In the case of HAAR-like features the regioning of the input facial image is not incorporated in any form, therefore MB-DFW in the block-level can be applied for concatenated modalities only and the number of unique weights is $M = 3$. Only the training data is utilized in the learning process, where the number of training samples per class is selected from the list $N_c^{Train} = (2, 5, 10, 25)$. The resulting EERs obtained for test data and various scenarios are reported in Table 1.

The last sequence of experiments is performed for SVM based face recognition. The classifier in this case is a linear SVM, which operates in "One-vs-All" mode. Thus the total of C classifiers should be trained, where $C = 51$ is the number of classes in the database. The recognition is then performed by predicting using each binary classifier, and choosing the prediction with the highest confidence score. The results are reported in Table 1 in the SVM-related rows. The linear SVM optimization does not work with HAAR-like features, due to low dimensionality of the feature space.

Towards a clear understanding of the importance of each modality in the multi-modal face recognition process we also calculate the *average modality weight* based on the results obtained with MB-DFW algorithm. LBP features in most cases outperform HOG and HAAR-like descriptors, thus the experiment is limited to LBP features only. In particular each modality has $K \cdot K = 25$ unique weights. Here the *importance* of the modality is considered to be an average of these weights. The results are reported in Table 2. Some of the values in Table 2 are highlighted in bold to demonstrate the *two* most important modalities for the particular capturing scenario.

V. Conclusion

To the best of our knowledge this is the first work observing the task of multi-modal face recognition for *synchronized* RGB-D-T modalities. Since the problem is novel the paper introduce both a new multi-modal face database with specific evaluation protocol and the facial recognition algorithm itself. The database covers some of the most challenging face recognition scenarios: rotation of the head, expression and illumination variations, which are organized in three acquisition sequences. The database, ground-truth information and evaluation protocol will be publicly available for the research community upon the acceptance of the paper.

The face recognition algorithm is based on feature-level fusion concept. The experimental results cover various combinations of classifiers (NNC, WNNC + MB-DFW, Linear SVM) and feature spaces (LBP, HOG, HAAR-like). It is worth mentioning that preprocessing of the input face images was deliberately excluded from the algorithmic pipeline in order to get a clear insight of "as is" possibilities of each particular modality. From experimental results (Tables 1 and 2) a few important conclusions can be made. First, based on the complexity for the recognition the capturing scenarios can be prioritized as follows: rotations (difficult), illumination (less difficult), expressions (the most simple one). Second, the importance of each modality in the recognition process depends on the capturing scenario. However, thermal data constantly holds high impact in the recognition regardless of the scenario. From the list of observed features LBP in most cases provides the best recognition results. However, the dimensionality of the LBP feature vectors is the highest, which possibly leads to high performance [25]. From a classification point of view SVM outperforms the combination of WNNC + MB-DFW. However SVM has a serious drawback, training of the classifier is needed every time a new person is added to the database. Thus, WNNC is preferable if simplified management of the database is needed.

References

[1] A. S. Mian, M. Bennamoun, and R. A. Owens, "An efficient multimodal 2d - 3d hybrid approach to automatic face recognition," *IEEE PAMI*, vol. 29, no. 11, pp. 1927–1943, 2007.

[2] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE PAMI*, vol. 99, no. PrePrints, p. 1, 2013.

[3] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: the problem of compensating for changes in illumination direction," *IEEE PAMI*, vol. 19, pp. 721–732, 1997.

[4] A. Mgelmose, A. Claps, C. Bahnsen, T. Moeslund, and S. Escalera, "Tri-modal person re-identification with rgb, depth and thermal features," *IEEE CVPR Workshop on Perception Beyond the Visible Spectrum*, June 2013.

[5] A. J. Ma, P. C. Yuen, and J.-H. Lai, "Linear dependency modeling for classifier fusion and feature combination," *IEEE PAMI*, vol. 35, no. 5, pp. 1135–1148, 2013.

[6] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition," *Comput. Vis. Image Underst.*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[7] I. A. Kakadiaris, G. Passalis, T. Theoharis, G. Toderici, I. Konstantinidis, and M. N. Murtuza, "Multimodal face recognition: Combination of geometry with physiological information," in *CVPR (2)*. IEEE Computer Society, 2005, pp. 1022–1029.

[8] Y. Zheng and A. Elmaghraby, "A brief survey on multispectral face recognition and multimodal score fusion," in *ISSPIT*, A. Elmaghraby and D. N. Serpanos, Eds. IEEE, 2011, pp. 543–550.

[9] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis, "Face recognition by fusing thermal infrared and visible imagery," *Image Vision Comput.*, vol. 24, no. 7, pp. 727–742, 2006.

[10] K. W. Bowyer, K. I. Chang, P. J. Flynn, and X. Chen, "Face recognition using 2-d, 3-d, and infrared: Is multimodal better than multisample?" *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2000–2012, 2006.

[11] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, pp. 2115–2125, 2003.

[12] P. V. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *ICCV*. IEEE, 2009, pp. 221–228.

[13] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[14] J. Ruiz-del Solar, R. Verschae, and M. Correa, "Recognition of faces in unconstrained environments: A comparative study," *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 1:1–1:19, Jan. 2009.

[15] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *ECCV (1)*, ser. Lecture Notes in Computer Science, T. Pajdla and J. Matas, Eds., vol. 3021. Springer, 2004, pp. 469–481.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR (1)*. IEEE Computer Society, 2005, pp. 886–893.

[17] K. Nasrollahi and T. B. Moeslund, "Are haar-like rectangular features for biometric recognition reducible?" in *CIARP (2)*, ser. Lecture Notes in Computer Science, J. Ruiz-Shulcloper and G. S. di Baja, Eds., vol. 8259. Springer, 2013, pp. 334–341.

[18] O. Nikisins and M. Greitans, "A mini-batch discriminative feature weighting algorithm for lbp - based face recognition," *IEEE IST 2012 - 2012 IEEE International Conference on Imaging Systems and Techniques, Proceedings*, pp. 170–175, 2012.

[19] O. Nikisins, "Weighted multi-scale local binary pattern histograms for face recognition," *AMCM 2013 - 2013 International Conference on Applied Mathematics and Computational Methods, Proceedings*, pp. 76 – 81, 2013.

[20] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recogn. Lett.*, vol. 32, no. 12, pp. 1598–1603, Sep. 2011.

[21] H. Guo, W. R. Schwartz, and L. S. Davis, "Face verification using large feature sets and one shot similarity," in *IJCB*, A. K. Jain, A. Ross, S. Prabhakar, and J. Kim, Eds. IEEE, 2011, pp. 1–8.

[22] K. Nasrollahi and T. B. Moeslund, "Haar-like features for robust real-time face recognition," *IEEE ICIP 2013*, pp. 3073–3077, 2013.

[23] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *CVPR*. IEEE, 2010, pp. 2707–2714.

[24] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *CVPR*. IEEE, 2013, pp. 3554–3561.

[25] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*. IEEE, 2013, pp. 3025–3032.

[26] A. Rattani, D. R. Kisku, M. Bicego, and M. Tistarelli, "Feature level fusion of face and fingerprint biometrics," *CoRR*, vol. abs/1002.2523, 2010.