RGB-D-T based Face Recognition

Olegs Nikisins1, Kamal Nasrollahi2, Modris Greitans1 and Thomas B. Moeslund2

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 $640 \times 480$, $640 \times 480$, and $384 \times 288$ 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.
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
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 Neighboor 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:
\[ N_{RGB} = N_T = N_D = N_m, \]
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 \( N_{m,RGB} \) is equal to [18]:
\[ N_{m,RGB} = K^2 \cdot 2^P, \]
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]:
\[ N_{m,HOG} = \nu \cdot \varsigma^2 \cdot \beta, \]
where \( \nu \) 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 \( \varsigma \) is the block regioning factor; \( \beta \) is the number of orientation bins in the histogram of the cell. The value \( \nu \) depends on the sizes of the input image and of the cell:
\[ \nu = (W/\eta - 1) \cdot (H/\eta - 1), \]
where \( W \) and \( H \) are correspondingly width and height of the image in pixels, \( \eta \) is the size of the square cell in pixels. Here we assume that the remainders of divisions \((W/\eta)\) and \((H/\eta)\) 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 \( x_m = (x_{m,1}, x_{m,2}, \ldots, x_{m,N_m}) \) is the feature vector of modality \( m \). Then the elements of normalized LBP feature vector \( \hat{x}_{m,1} \) can be found as follows:
\[ \hat{x}_{m,1} = x_{m,1} / \| x_{m,1} \|_2, \]
where \( \| x_{m,1} \|_2 \) is L2-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:
\[ x_{m,HOG} = (x_{m,1,1}, x_{m,1,2}, \ldots, x_{m,\varsigma^2,HOG}). \]
The feature vector of each block is normalized individually. The elements of normalized feature vector of the block \( j \) can be determined as follows:
\[ \hat{x}_{m,j,i} = \frac{x_{m,j,i}}{\| x_{m,j,i} \|_2^2 + \epsilon^2}, i = 1, \ldots, \varsigma^2 \cdot \beta, \]
where \( \| x_{m,j,i} \|_2^2 \) is L2-norm of the vector, \( \epsilon \) 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:
\[ \hat{x}_{m,i} = \frac{x_{m,i}}{\| x_{m,i} \|_1}, i = 1, \ldots, N_{m,HAAR} \]

Once normalization is completed the concatenation of modalities into a single normalized multi-modal face feature vector \( x \) can be performed (Figure 4):
\[ x = (\hat{x}_{RGB}, \hat{x}_D, \hat{x}_T). \]

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

The cost function to be optimized in the MB-DFW is formed from pairs [18], [19]:
\[ \left[ \left( d_{(i,i),(1,k)}^{intra}, d_{(i,j),(1,k)}^{inter} \right) \right]. \]

The number of pairs in our extended version is \( (N_c^{\text{Train}} - 1) \cdot M \), where \( M \) is the number of classes. In the original version \( N_c^{\text{Train}} = 2 \) resulting in the total of \( M \) pairs.
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

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.

#### TABLE 1. EER values (for test data) in % for various modalities, classifiers, features and number of training samples per class

<table>
<thead>
<tr>
<th>Mod.</th>
<th>RGB</th>
<th>D</th>
<th>T</th>
<th>All</th>
<th>RGB</th>
<th>D</th>
<th>T</th>
<th>All</th>
<th>RGB</th>
<th>D</th>
<th>T</th>
<th>All</th>
<th>RGB</th>
<th>D</th>
<th>T</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP</td>
<td>31.9</td>
<td>34.6</td>
<td>31.3</td>
<td>31.2</td>
<td>27.4</td>
<td>28.1</td>
<td>30.0</td>
<td>24.8</td>
<td>33.3</td>
<td>31.1</td>
<td>31.2</td>
<td>27.9</td>
<td>23.3</td>
<td>21.5</td>
<td>31.8</td>
<td>24.4</td>
</tr>
<tr>
<td>HOG</td>
<td>30.1</td>
<td>35.2</td>
<td>34.4</td>
<td>32.5</td>
<td>24.9</td>
<td>38.3</td>
<td>31.6</td>
<td>30.1</td>
<td>28.5</td>
<td>35.3</td>
<td>25.7</td>
<td>21.8</td>
<td>21.2</td>
<td>35.5</td>
<td>26.4</td>
<td>22.4</td>
</tr>
<tr>
<td>HAAR</td>
<td>24.2</td>
<td>31.8</td>
<td>32.6</td>
<td>27.5</td>
<td>24.3</td>
<td>31.8</td>
<td>32.6</td>
<td>27.5</td>
<td>24.3</td>
<td>31.8</td>
<td>32.6</td>
<td>27.5</td>
<td>24.3</td>
<td>31.8</td>
<td>32.6</td>
<td>27.5</td>
</tr>
</tbody>
</table>

#### TABLE 2. The importance of each modality in multi-modal face recognition

<table>
<thead>
<tr>
<th>Mod.</th>
<th>RGB</th>
<th>D</th>
<th>T</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP</td>
<td>13</td>
<td>14.1</td>
<td>17.1</td>
<td>10.6</td>
</tr>
<tr>
<td>HOG</td>
<td>13.5</td>
<td>22.5</td>
<td>16.7</td>
<td>13.9</td>
</tr>
<tr>
<td>HAAR</td>
<td>13.5</td>
<td>22.5</td>
<td>16.7</td>
<td>13.9</td>
</tr>
</tbody>
</table>

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 $K - K = 25$ 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