Two-Level Attention-based Fusion Learning for RGB-D Face Recognition

Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, and Ali Etemad

Department of Electrical and Computer Engineering & Ingenuity Labs, Queen's University, Kingston, Ontario, Canada

Ambient Intelligence and Interactive Machines (Aiim) Lab

Abstract

With recent advances in RGB-D sensing technologies as well as improvements in machine learning and fusion techniques, RGB-D facial recognition has become an active area of research. A novel attention aware method is proposed to fuse two image modalities, RGB and depth, for enhanced RGB-D facial recognition. The proposed method first extracts features from both modalities using a convolutional feature extractor. These features are then fused using a two layer attention mechanism. The first layer focuses on the fused feature maps generated by the feature extractor, exploiting the relationship between feature maps using LSTM recurrent learning. The second layer focuses on the spatial features of those maps using convolution. Comparative evaluations demonstrate that the proposed method outperforms other state-of-the-art approaches, including both traditional and deep neural network-based methods, on the challenging CurtinFaces and IIIT-D RGB-D benchmark databases, achieving classification accuracies over 98.2% and 99.3% respectively. The proposed attention mechanism is also compared with other attention mechanisms, demonstrating more accurate results.

Contributions

- We introduce a novel multimodal fusion mechanism for RGB-D FR using attention to selectively learn useful information from both RGB and depth modalities
- We perform ablation experiments on a number of variations of attention feature and spatial mechanisms, and demonstrate the performance improvement in attention-based fusion of the two modalities
- Our proposed method outperforms several other methods, setting new state-of-the-art values on two public RGB-D face datasets

Datasets and Preprocessing

We use two public RGB-D datasets for our experiments

a) IIIT-D RGB-D dataset
b) CurtinFaces dataset

Depth image preprocessing is shown in Figure 1; the orange lines represent the clipping planes, to clip out data points that are too close to the camera or very far from it, keeping only face data. The datasets are passed through a dlib CNN-face detector to crop the faces from the entire image.

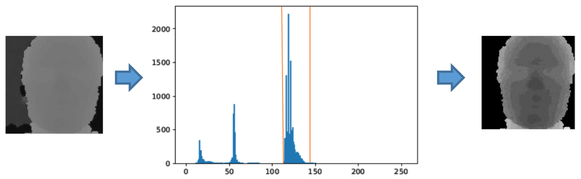

Fig 1. Depth image preprocessing

After the pre-processing the images are passed though a series of geometric transformation for augmentation as listed below.

Rotation	-30° to 30°	Flip	Vertical
Sheer	-16° to 16°	Scale	50% to 150%

Table 1. Image augmentation parameters

Architecture

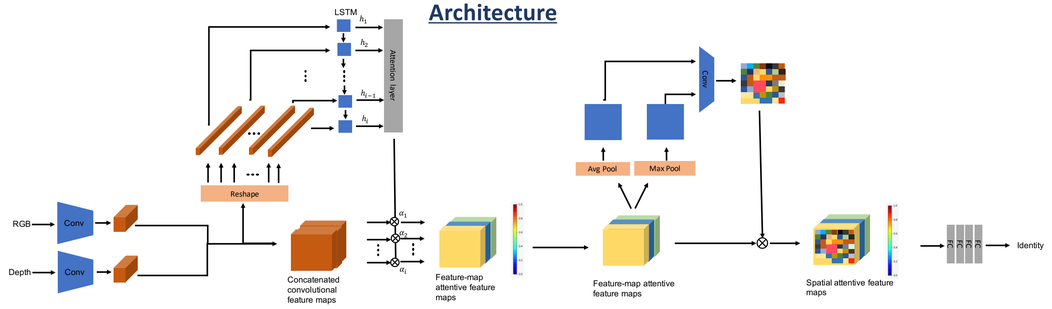

Fig 2. Architecture of the proposed two –level attention network

Our proposed two-level attention mechanism sits on top of convolution feature detectors following the architecture of the VGG network. Both depth and RGB images are passed through the convolutional part of the network to obtain respective feature embeddings which are then fed to the attention module.

- **Feature-Map Attention:** The aim of this mechanism is to help train the network to focus on those feature maps generated in the embedding that have a greater contribution towards our classification task. Here, an LSTM layer first acts as the conditional encoder for each feature map to then calculate the attention weights using the subsequent dense layer.
- **Spatial Attention:** After refining the features through feature-map attention, the network next focuses attention on the spatial axis of the embedding. This module helps the network to identify the most salient features in the embedding and to focus its attention on those features. Here, we use convolution layer to generate an attention map over the spatial dimension of the features.

After the attention module, the refined features are then fed to classifier which finally generates the identity label for the particular image pair.

Results

Year	Author	Feat. extractor	Classifier	Accuracy
2013	Goswami et al.	RISE	Random Forest	91.6%
2014	Goswami et al.	RISE	Random Forest	95.3%
2018	Zhang et al.	9 Layers CNN + Inception	FC/Softmax	98.6%
2016	Chowdhury et al.	Autoencoder	FC/Softmax	98.7%
2020	**Proposed**	**VGG + Two-level Attention**	**FC/Softmax**	**99.4%**

Table 2. Performance comparison on the IIIT-D dataset

Year	Author	Feat. extractor	Classifier	Pose Acc.	Ill. Acc.	Avg. Acc.
2013	Li et al.	Discriminat Color Space Trans.	SRC	96.4%	98.2%	97.3%
2016	Li et al.	LBP + Haar + Gabor	SRC			91.3%
2016	Hayat et al.	Covariance Matrix Rep.	SVM			96.4%
2020	**Proposed**	**VGG + Two-level Attention**	**FC/Softmax**	**97.5%**	**98.9%**	**98.2%**

Table 3. Performance comparison on the CurtinFaces dataset

For IIIT-D dataset, our proposed method increases the identification rank-1 accuracy to 99.4%. The proposed method outperformed the state-of-the-art results by Chowdhury et al which uses depth rich features acquired from an autoencoder, achieving a classification accuracy of 98.7%, and Zhang et al, which uses complimentary feature learning to achieve 98.6% accuracy as shown in Table 2.

For CurtinFaces dataset, our proposed model outperforms the state-of-the-art results in both the test sets, with 97.53% accuracy in the pose-expression variation test set, and with 98.88% accuracy in the illumination-expression variation test set as mentioned in Table 3.

tSNE visualizations

To show the effectiveness of our attention mechanism, we visualize tSNE for the feature embeddings for RGB modality, depth modality and our proposed method.

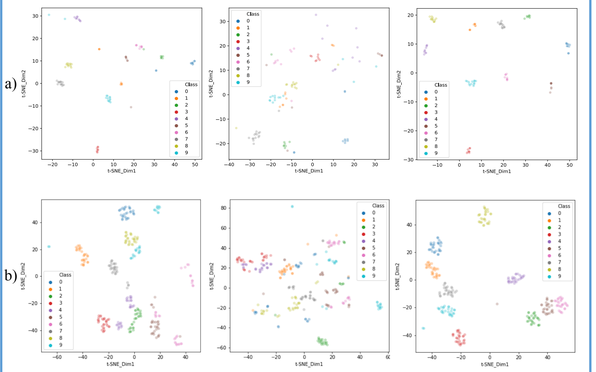

Fig 3. t-SNE visualization of our proposed method and other solutions. Every row in the figure represents a dataset, i.e. a) IIIT-D RGB-D and b) CurtinFaces. Every column corresponds to a network and data modality, i.e.: (I) VGG with only RGB input, (II) VGG with only Depth input and (III) attention-based fusion with RGB-D.

Ablation study

We conduct extensive ablation trials, revealing the relative contributions of the individual components of our method towards the final performance. The results are Shown in Table 4.

	IIIT-D		CurtinFaces	
Model	Accuracy	Pose acc.	Ill. acc.	Avg. acc.
VGG-Face (RGB)	94.1%	92.5%	93.2%	92.8%
VGG-Face (Depth)	68.5%	60.2%	63.2%	61.7%
VGG-Face (RGB + Depth)	95.4%	92.6%	94.2%	93.4%
With Feature-Map Attention	96.4%	97.5%	98.3%	97.8%
With Spatial Attention	96.2%	96.7%	97.3%	97.0%
Proposed	99.4%	98.1%	99.1%	98.6%

Table 4. Ablation study on the two datasets

Conclusion and Future Work

We presented an attention-based network to effectively fuse the RGB and depth modalities of RGB-D images for face recognition. Through our evaluations we validate that our attention aware fusion offers more accurate rank-1 recognition results than the state-of-the-art methods on the IIIT-D RGB-D dataset at 99.4% and on the CurtinFaces dataset at 98.6%.

In future work, we will explore the performance of our proposed solution using different sets of modalities such as thermal, speech, bio-signals. Moreover, we will explore Depth image generation from RGB images to aid the face recognition mechanism, as depth has been shown here to significantly affect the recognition accuracy.