Machine Vision Group

UNIVERSITY OF OULU

3D Facial Expression Recognition Based on Multi-View and Prior Knowledge Fusion

Nhat Vo, Khanh Tran, Guoying Zhao

IEEE 21st International Workshop on Multimedia Signal Processing

Agenda

- Introduction
- Proposed method
- Experiment
- Discussion

Introduction

UNIVERSITY OF OULU

- **Benefits of 3D Facial Expression Recognition (3D-FER):**
  - Inherent characteristics of 3D face scans that make it robust to lighting and pose variation.
  - 3D geometry information may include important features for FER

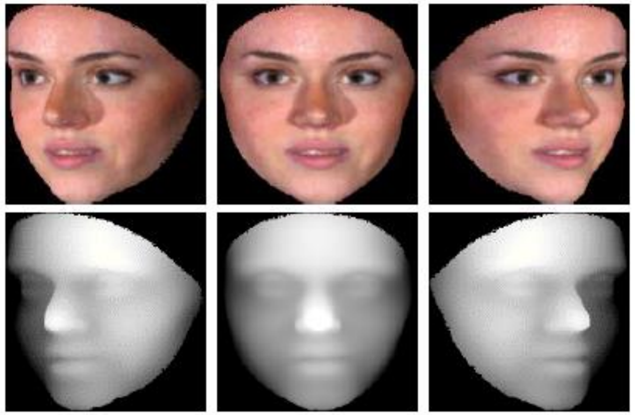
Introduction

- **Three main approaches of 3D-FER**
  - The first group extracts 3D features at landmark or patch locations.
    - **Drawback** : depends heavily on the accuracy of 3D facial landmark detection.
  - The second approach employs the morphable models to get the one-to-one point correspondence among face scans
    - **Drawback** : This approach require an accurate method of dense correspondence among face models.
  - The third one utilizes the 2D representation of the 3D face scans.
    - Reuse traditional solutions in 2D FER for 3D FER and produce better results.
    - **Drawback**: do not fully exploit the information of 3D model

Observations 1

UNIVERSITY OF OULU

- **Benefit of multi-view for the FER**
  - Provide more clues to recognize the low-intensity emotions



- Low-intensity Happy expression looks quite similar to the neutral or surprise expression
- The emotion is expressed more clearly in the side view.

- On the frontal view, Happy and Fear express the same movement on the face.
- On side-view, the fear and happy expressions are quite different

Observations 2

UNIVERSITY OF OULU

- **Benefit of facial prior**
  - Not all information on the face is useful for emotion recognition
    - E.g., face shape, gender, age, etc.
  - The facial attribute maps may contain some areas unrelated to the facial expression.
  - According to Wegrzyn et al. [1], people were mostly relying on the eye and mouth regions when successfully recognizing an emotion.

[1] M. Wegrzyn, et al., “Mapping the emotional face. How individual face parts contribute to successful emotion recognition,” PloS one, 12(5), pp. e0177239, 2017.

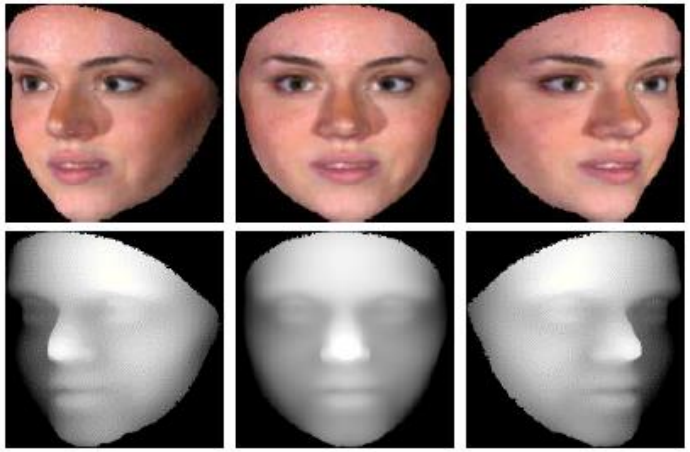
Introduction

Our contribution

- Propose a multi-view CNN architecture for 3D FER
  - Jointly learn the 2D RGB texture and depth images
  - Utilize different views of a 3D face scan
- Incorporate beneficial facial prior knowledge to guide the learning process.
  - Teach the network to predict emotion-related facial areas
  - Learn to extract facial-related features

Proposed Method

- **Multi-view CNN for 3D facial expression recognition**
  - We employ one frontal view and two side views for projecting the facial expression of 3D face model.
  - We select depth map images and RGB texture images synthesized from the 3D mesh and related texture information
    - The problem of training time and storage memory
    - Oyedotun [2] presents high accuracy by training a model on only depth and RGB texture images



[2] Oyedotun et al; “Facial Expression Recognition via Joint Deep Learning of RGB-Depth Map Latent Representations”; ICCVW 2017

Proposed Method

Multi-view CNN for 3D facial expression recognition

- We design a three stream CNN architecture for learning jointly from the depth maps and RGB texture images of three facial views
- Feature fusion: we use two levels of feature fusion:
    - 1^{st} : Concatenate the feature from fc7 layer
    - 2^{nd} : Utilize fc8 layer

$$L_p = -\log\left(\frac{e^{f_i}}{\sum_j e^{f_j}}\right),$$

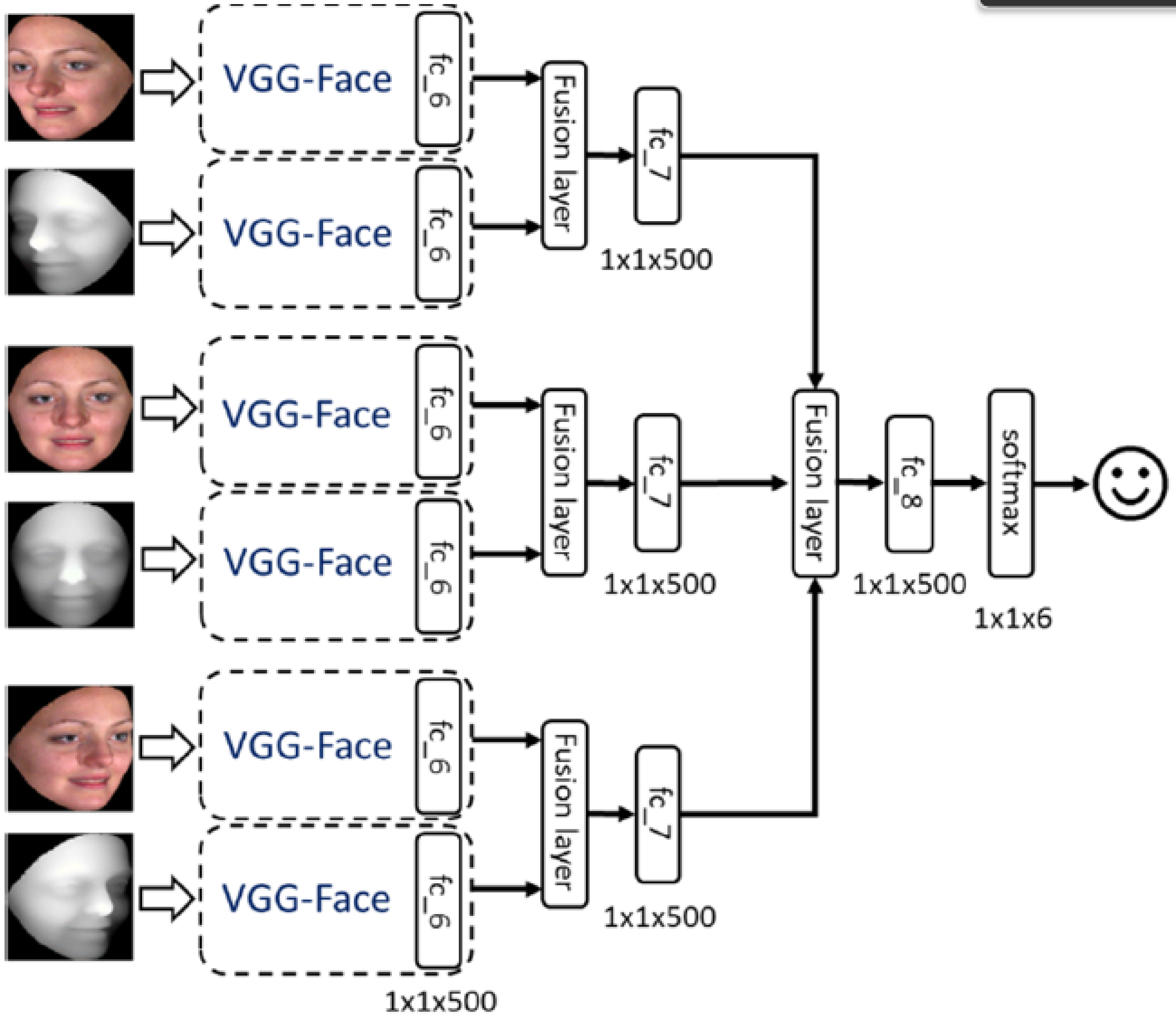
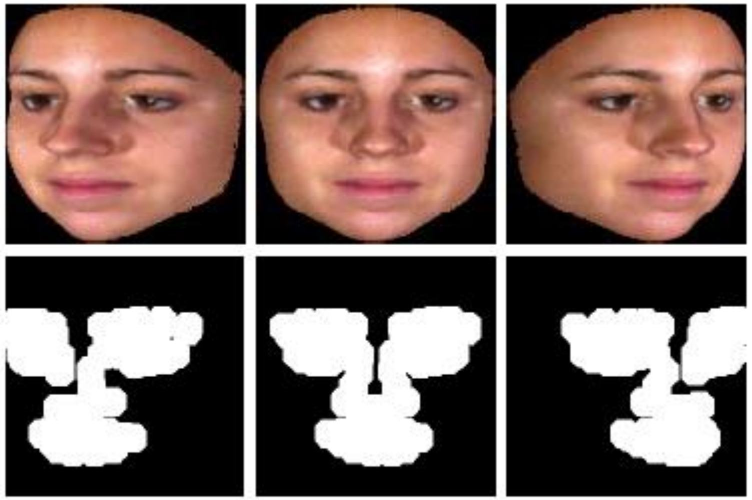


Fig. Multiview CNN architecture for 3D facial expression recognition.

Proposed Method



Learning with attention using facial prior knowledge

- Inspired from the previous research, not all information on face are useful
- Incorporate the facial prior information to the training process in the manner multi-task learning.
    - Feature extraction on each view are connected to a FCN.

$$L_l^k = -\beta \sum_{j \in Y_+} \log\left(\sigma(a_j)\right) - (1-\beta) \sum_{j \in Y_-} \log(1-\sigma(a_j)),$$

$$L = L_p + \frac{1}{N} \sum_{k=1}^{N} L_l^k,$$

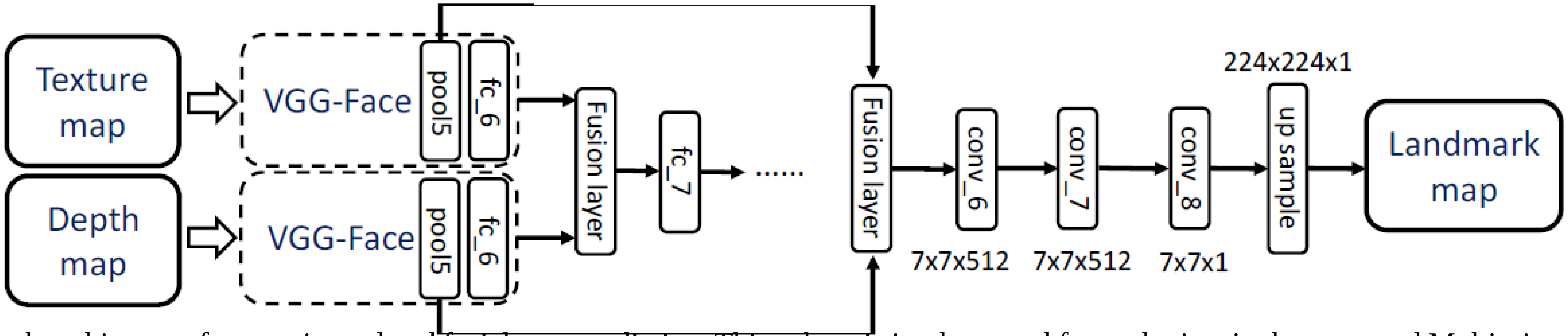


Fig. The network architecture for emotion-related facial area prediction. This subnet is implemented for each view in the proposed Multi-view CNN architecture.

Experiments

Dataset:

- BU3DFE: 100 subjects with six types of expression and 4 levels of expression intensity
  - Subset I: includes expressions with two higher levels of expression intensity
  - Subset II: consist of all four levels of expression intensity
- Bosphorus: 65 subjects perform the six prototypical expressions, one sample for each expression

Experimental protocol:

- BU3DFE Subset I: 40 subjects to the validation and 60 subjects to the training and testing (54-versus-6-subject-partition experiments)
- BU3DFE Subset II and Bosphorus : 10-fold cross-validation training
- Maximum of 1000 training epochs
- Adam optimizer

Experiments

UNIVERSITY OF OULU

Comparison with the state-of-the-art

BU3DFE Subset I

Methods	Feature	Accuracy
Li et al. [32]	normals, curv./hist.	82.01
Zhen et al. [33]	coordinates, normals, shape index	84.50
Yang et al. [34]	depth, normals, curv./scattering	84.80
Li et al. [35]	meshHOG/SIFT meshHOS/HSOG	86.32
Li et al. [20]	depth, normal, curv., RGB, maps, deep feature	86.86
Oyedotun et al. [21]	depth, RGB, deep feature	89.31
Multi-view CNN	multiview, depth, RGB, deep feature	89.68
Multi-view CNN (with prior)	**multiview, depth, RGB, deep feature**	**91.39**

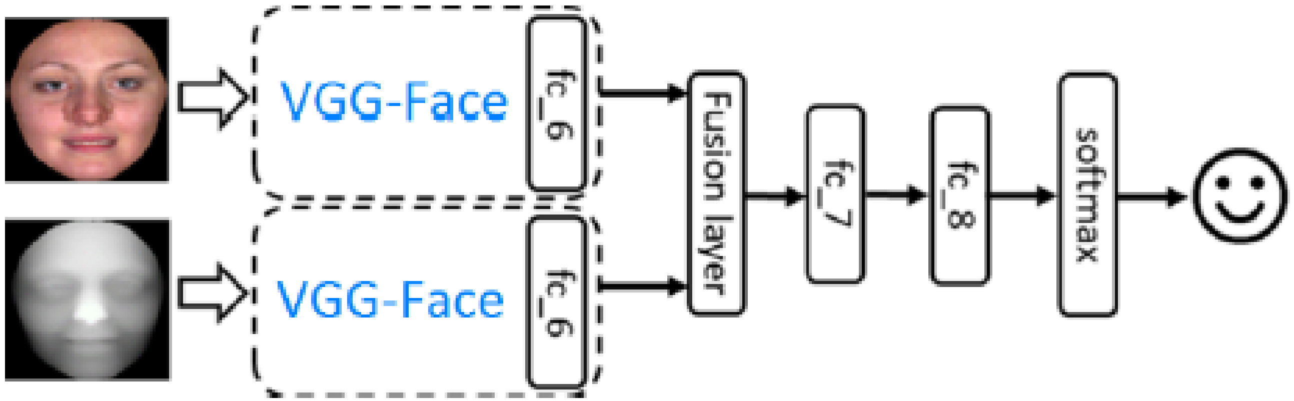
BU3DFE Subset II and Bosporus

Methods	BU-3DFE Subset II	Bosphorus
Li et al. [35]	80.42	79.72
Yang et al. [34]	80.46	77.50
Li et al. [20]	81.33	80.00
Multi-view CNN	83.54	81.94
Multi-view CNN (with prior)	**84.30**	**82.40**

Experiments

Ablation studies

- Multi-view CNN vs Single-view CNN
- Hierarchical fusion vs single feature fusion



Methods	BU-3DFE Subset I	BU-3DFE Subset II	Bosphorus
Single-view CNN	87.91	80.99	80.78
Multi-view CNN (without hierarchical fusion)	88.43	82.92	81.48
Multi-view CNN	**89.68**	**83.54**	**81.94**

Discussion

- Proposing a novel CNN model for 3D Facial Expression Recognition
- Our method presents promising results compared with existing methods
- Plan for improvement and exploration
  - Study the importance of each view
  - Extending to 4D data

THANKS FOR ATTENTION
Q/A