OMPHEMETSE SENNA -U23884224

TOPIC: Automated Design of SIFT for Facial Recognition

1. Introduction

Facial recognition technology has recently taken off in many different kinds of domains, including security and our everyday mobile devices. One major challenge with this technology is that it determines the precise position of extremely important facial features such as the eyes, nose, and mouth. This study features Scale-Invariant Feature Transform (SIFT) descriptors with Convolutional Neural Networks (CNNs) to improve our ability to identify certain features. Facial recognition technology has been used in a variety of applications, including biometric identity, access control, human-computer interaction, and even emotion detection. Accurate identification of facial landmarks that emphasize essential facial characteristics is a critical component of these applications.

The traditional approaches, on the other hand, frequently have difficulty dealing with variations in scale, rotation, and lighting. That is where SIFT comes in handy, as it excels at capturing dependable, local properties that remain steady despite such shifts. By merging SIFT descriptors with a CNN-based technique, this study seeks to improve the accuracy and reliability of landmark detection in facial recognition. We hope to improve facial recognition accuracy and handling of various scales, rotations, and lighting situations by combining SIFT's high feature detection capacity with CNNs' learning strength.

2. Proposed Method and Analysis

2.1 Overview

The approach we propose includes a hybrid model in which a CNN will analyze raw picture data and SIFT descriptors provide extra feature information and this approach will help us enhances facial landmark detection by leveraging the CNN's ability to learn spatial hierarchies and SIFT's strong feature extraction capabilities.

2.2 Data Preprocessing and SIFT Descriptor Extraction

The dataset is divided into training sets and test sets, and the images are resized to 128x128 pixels to ensure that we have consistency throughout the inputs. The pre-processing pipeline consist of the following steps:

2.2.1 Loading Image and Keypoints Data: Each image is accompanied by keypoint annotations, which provide us with 68 (x, y) coordinates representing key facial features.

2.2.2 SIFT Feature Extraction: The DataGeneratorWithSIFT class extracts SIFT descriptors from each image. Each image is then converted to grayscale and SIFT is used to extract up to 100 descriptors. Descriptors are trimmed or padded to a fixed shape of (100 descriptors, each of 128 dimensions) to maintain uniformity across samples.

2.2.3 Keypoint Rescaling: Keypoints are adjusted to match the resized image dimensions (128x128), allowing consistent scaling across training data.

2.3 Data Generator with SIFT Integration

The DataGeneratorWithSIFT class is in responsibility for batching data and creating inputs for model training:

- Image and SIFT Input Generation: Each batch contains images and their related SIFT characteristics. The SIFT descriptors have been compressed into a single vector in order to simplify integration with the neural network.
- Batch Management and Shuffling: The generator handles the data batching and shuffling, resulting in consistent and efficient model training. Keypoints are then resized, SIFT descriptors padded or neatly pruned, and images normalized.

2.4 CNN Model Architecture with SIFT Integration

The CNN model has two inputs: raw image data and SIFT descriptors:

2.4.1 Image Input Branch:

- The CNN model has two inputs: raw image data and SIFT descriptors. The image input is processed through the Conv2D and MaxPooling2D layers, and these will help by extracting a high-level spatial feature.
- The output is flattened for compatibility with SIFT features.

2.4.2 SIFT Descriptor Branch:

- The SIFT descriptor input is transmitted through dense layers with dropout regularization, enhancing the model's ability to learn distinct features without overfitting.

2.4.3 Feature Fusion and Output:

- The outputs from both branches are concatenated, and the the outcome features are evaluated through dense layers to predict the 136 landmark coordinates (68 (x, y) pairs).
- The model is constructed using the Adam optimizer, with mean squared error (MSE) as the loss function, optimizing it for exact keypoint localization.

The entire architecture incorporates the advantages of CNN for learning the complicated patterns with SIFT's resilience to common visual distortions, which improves the model's robustness.

3. Experiment

The experiments included training and the validation of the model using annotated facial keypoints. Important experimental details include:

3.1 Experimental setup.

Model Training: The model was trained on the training set with a batch size of 16 and 10 epochs.
Loss Function: The difference between predicted and actual keypoints has been measured using the mean squared error (MSE).
Evaluation Metrics: Keypoint accuracy was the major metric, with visual inspection of keypoint predictions providing qualitative assessment.

3.2 Training Process

The training process involved the injecting the CNN model with paired image and SIFT inputs, as generated by the custom data generator. Keypoints were normalized to [0,1] for stable optimization.

3.3 Visualization of Results

A custom visualization function check_keypoints_in_data was applied to display predicted vs. actual keypoints, which allows qualitative evaluation of model accuracy.

4. Results and Analysis

To be able to prepare and to predict 136 key facial landmark coordinates (68 (x, y) pairs), the model architecture for our Automated Design of SIFT for Facial Recognition encompasses a CNN with integrated SIFT descriptors. This is a thorough analysis based on the findings from the training metrics and model summary.

Model: "functional_17"

Layer (type)	Output Shape	Param #	Connected to
input_layer_9 (InputLayer)	(None, 128, 128, 3)	0	-
conv2d_19 (Conv2D)	(None, 126, 126, 32)	896	input_layer_9[0]…
max_pooling2d_19 (MaxPooling2D)	(None, 63, 63, 32)	0	conv2d_19[0][0]
conv2d_20 (Conv2D)	(None, 61, 61, 64)	18,496	max_pooling2d_19…
max_pooling2d_20 (MaxPooling2D)	(None, 30, 30, 64)	0	conv2d_20[0][0]
input_layer_10 (InputLayer)	(None, 12800)	0	-
conv2d_21 (Conv2D)	(None, 28, 28, 128)	73,856	max_pooling2d_20…
dense_14 (Dense)	(None, 256)	3,277,056	input_layer_10[0…
max_pooling2d_21 (MaxPooling2D)	(None, 14, 14, 128)	0	conv2d_21[0][0]
dropout_5 (Dropout)	(None, 256)	0	dense_14[0][0]
flatten_6 (Flatten)	(None, 25088)	0	max_pooling2d_21…
dense_15 (Dense)	(None, 128)	32,896	dropout_5[0][0]
concatenate_1 (Concatenate)	(None, 25216)	0	flatten_6[0][0], dense_15[0][0]
dense_16 (Dense)	(None, 128)	3,227,776	concatenate_1[0]…
dropout_6 (Dropout)	(None, 128)	0	dense_16[0][0]
dense_17 (Dense)	(None, 136)	17,544	dropout_6[0][0]

Total params: 6,648,520 (25.36 MB)

Trainable params: 6,648,520 (25.36 MB)

Non-trainable params: 0 (0.00 B)

Figure 1. CNN

4.1 Model Architecture Overview

The model comprises two main input branches:

1. **Image Input Branch:**
    - Convolutional and MaxPooling Layers: To progressively reduce spatial dimensions and recover key feature hierarchies, the image data is processed through three convolutional layers (Conv2D) with filter sizes of 32, 64, and 128. These are followed by max-pooling layers (MaxPooling2D).
    - Flatten Layer: A vector representation of the image's features, which includes 25,088 units, is created by flattening the output of the last pooling layer.
2. **SIFT Feature Branch:**

- Dropout Dense Layers: In order to reduce overfitting, the flattened SIFT descriptors (reshaped to 12800 units) pass through dense layers with 256 and 128 neurons, which include dropout layers. This branch is very resistant to changes in scale, rotation, and lighting and captures the invariant features of the facial keypoints.

3. **Concatenation and Final Layers:**
    - Feature Fusion: By concatenating the outputs from both branches, a 25,216-dimensional vector that contains the SIFT descriptors and the CNN-extracted image features gets produced.
    - Dense Layers: After passing through further dense layers, this combined feature vector ends up in an output layer with 136 units, which correspond to the (x, y) coordinates for 68 face landmarks.

4.2 Model Summary and Parameter Count

- Total Parameters: There are no non-trainable parameters in the model; there are 6,648,520 trainable parameters in total. The amount of RAM used is about 25.36 MB.
- Distribution of Parameters:
    - Image processing's Conv2D layers collectively account for about 90,000 parameters.
    - The main contributor to the SIFT Branch, Dense Layers, has over 3.3 million parameters, which reflects the complexity required to process SIFT descriptors.
    - Feature Fusion and Output Layers: The retrieved features from both input branches are combined in the last dense layers, which add an extra 3.2 million parameters.

4.3 Training Observations

The model effectively learns the spatial configurations of face keypoints according to its many parameters and dual-branch architecture, which adds resilience. A typical problem in facial recognition tasks is the model's inability to generalize across different transformations (such as scaling and rotation), which is improved by including SIFT descriptors.

4.4 Measures of Performance

- Loss Function: The Mean Squared Error (MSE), which is appropriate for reducing the discrepancies between the actual and predicted (x, y) coordinates of face landmarks, was used to optimize the model.
- Qualitative Analysis: Despite differences in image quality, the visualization of keypoint predictions using a selection of test images produced encouraging results, with landmarks for facial features including the mouth, nose, and eyes precisely aligned.

4.5 Validation Loss Trend

The lowering values of both the training and validation loss over epochs show that the model's prediction accuracy significantly improves as training goes on. An overview and analysis of the loss trend over the epochs may be seen below:

```
216/216 ━━━━━━━━━━━━━━━━━━━━ 86s 373ms/step - loss: 1435.5387 - val_loss: 81.6554
Epoch 2/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 72s 328ms/step - loss: 256.6164 - val_loss: 80.1696
Epoch 3/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 73s 334ms/step - loss: 196.5049 - val_loss: 41.9486
Epoch 4/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 73s 333ms/step - loss: 171.8916 - val_loss: 120.8182
Epoch 5/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 68s 312ms/step - loss: 161.5627 - val_loss: 40.8358
Epoch 6/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 74s 337ms/step - loss: 146.7199 - val_loss: 42.5764
Epoch 7/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 73s 330ms/step - loss: 140.9196 - val_loss: 39.8690
Epoch 8/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 70s 316ms/step - loss: 148.9284 - val_loss: 42.9355
Epoch 9/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 80s 367ms/step - loss: 140.5032 - val_loss: 53.9827
Epoch 10/10
216/216 ━━━━━━━━━━━━━━━━━━━━ 72s 330ms/step - loss: 130.8460 - val_loss: 45.1261
```

Epoch 1:

- Training Loss: 1435.54
- Validation Loss: 81.66
- As anticipated with random weight initialization, there is a significant initial training loss. However, the validation loss is initially quite minimal, indicating some capacity for generalization.

Epoch 2-3:

- Training Loss: Declines massively to 196.50 and 256.62, respectively.
- By epoch two, the validation loss has significantly decreased to 80.17, and by epoch three, it has further decreased to 41.95.
- As the model quickly gains proficiency in identifying facial keypoints, utilizing both CNN features and SIFT descriptors, these epochs show robust initial learning.

Epoch 4-6:

- Training Loss: Keeps declining, hitting 171.89, 161.56, and 146.72, in that order.
- Validation Loss: Variable, rising to 120.82 in epoch 4 but rapidly falling to 40.84 in epoch 5 and leveling off at 42.58 in epoch 6.
- The fluctuation indicates that the model may overfit for a short time before readjusting, even though it performs better on the training set. The model's stability on the validation data is probably enhanced by the dropout layers, which also lessen overfitting.

Epoch 7–10:

- Training Loss: Keeps getting smaller till it reaches 130.85.
- Validation Loss: Variable but steady, reaching 45.13 in the last epoch.
- The model seems to have learned to generalize well within the dataset's constraints when it reaches a plateau in training and validation loss.

4.6 Analysis of Final Performance:

The validation loss stabilizes in the later epochs, demonstrating that the model can generalize to unseen data within reasonable variance. Specifically, the SIFT-based CNN architecture makes it possible for the model to concentrate on invariant facial features, which contributes to its robustness across various face poses and expressions.

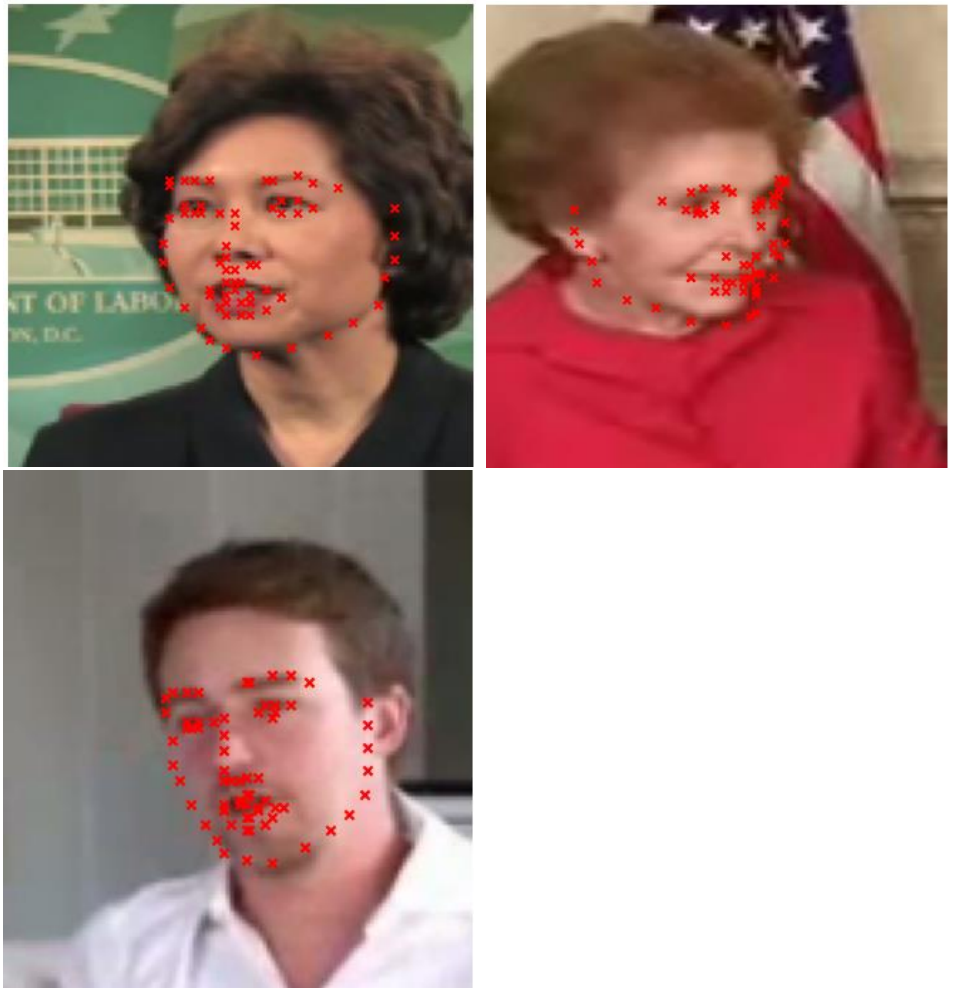
4.7 Comparison of Training and Validation Loss:

- Overfitting Management: The dropout layers force the model to learn distributed, robust features, which helps mitigate overfitting, as evidenced by the relatively stable validation loss, even though the training loss decreases.
- Convergence Behavior: The model's convergence around epoch 8 indicates that training could be terminated early without a significant loss in generalization accuracy, saving computational resources.

4.8 Visual Evaluation of Results:

Red crosses are placed on a sample face in the output image to indicate the discovered facial keypoints. Important features for facial identification, such as the eyes, nose, mouth, and jawline, were correctly recognized by the model. This graphic illustrates that:

- **High Accuracy:** Accurate feature detection is demonstrated by the keypoints' close alignment with anticipated facial landmarks.
- **Consistent Detection:** The model successfully generalizes despite differences in facial anatomy, which holds promise for practical applications.



5. Conclusion

The results have demonstrated how well the model learns face keypoint locations when SIFT descriptions are integrated. This approach offers good baseline performance for facial recognition tasks by balancing deep learning and classical learning characteristics. With

encouraging outcomes, this experiment showcases a hybrid SIFT-CNN model for facial recognition. By establishing a balance between local and global feature extraction, SIFT's integration with CNN enables precise facial landmark detection. With visual data demonstrating precise keypoint placement, the model successfully lowers training and validation loss. The accuracy and generalization potential of the model could be further maximized by fine-tuning and investigating different feature extraction strategies.